frontiers | Research Topics

Machine learning and immersive technologies for user-centered digital healthcare innovation

Edited by
Federico Colecchia, Eleonora Ceccaldi, Daniele Giunchi,
Fang Wang and Rui Qin

Published in
Frontiers in Virtual Reality
Frontiers in Big Data
Frontiers in Artificial Intelligence

ISSN 1664-8714
ISBN 978-2-8325-6431-8
DOI 10.3389/978-2-8325-6431-8

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Machine learning and immersive technologies for user-centered digital healthcare innovation

Topic editors

Federico Colecchia — Brunel University London, United Kingdom
Eleonora Ceccaldi — University of Genoa, Italy
Daniele Giunchi — University College London, United Kingdom
Fang Wang — Brunel University London, United Kingdom
Rui Qin — Manchester Metropolitan University, United Kingdom

Citation

Colecchia, F., Ceccaldi, E., Giunchi, D., Wang, F., Qin, R., eds. (2025). *Machine learning and immersive technologies for user-centered digital healthcare innovation*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-6431-8

Table of contents

frontiers | Frontiers in Big Data

TYPE Editorial
PUBLISHED 14 February 2025
DOI 10.3389/fdata.2025.1567941

OPEN ACCESS

EDITED AND REVIEWED BY
Thomas Hartung,
Johns Hopkins University, United States

*CORRESPONDENCE
Federico Colecchia
federico.colecchia@brunel.ac.uk

RECEIVED 28 January 2025
ACCEPTED 03 February 2025
PUBLISHED 14 February 2025

CITATION
Colecchia F, Giunchi D, Qin R, Ceccaldi E and Wang F (2025) Editorial: Machine learning and immersive technologies for user-centered digital healthcare innovation.
Front. Big Data 8:1567941.
doi: 10.3389/fdata.2025.1567941

Editorial: Machine learning and immersive technologies for user-centered digital healthcare innovation

Federico Colecchia[1]*, Daniele Giunchi[2], Rui Qin[3], Eleonora Ceccaldi[4] and Fang Wang[5]

[1]Brunel Design School, Brunel University of London, Uxbridge, United Kingdom, [2]Department of Computer Science, University College London, London, United Kingdom, [3]School of Computing and Mathematical Sciences, University of Leicester, Leicester, East Midlands, United Kingdom, [4]CasaPaganini - InfoMus, DIBRIS, University of Genoa, Genoa, Liguria, Italy, [5]Department of Computer Science, Brunel University of London, Uxbridge, United Kingdom

KEYWORDS

digital innovation for health and wellbeing, interdisciplinary collaboration, artificial intelligence, immersive technologies, machine learning, virtual reality, augmented reality, mixed reality

Editorial on the Research Topic
Machine learning and immersive technologies for user-centered digital healthcare innovation

User-centered design for digital healthcare innovation

Modern digital technologies such as machine learning and immersive technologies, including virtual reality and augmented reality, hold potential for enabling disruptive innovations to promote individuals' health and wellbeing. However, the adoption of such technologies, including the use of data-driven tools to support healthcare professionals' decision-making and applications relying on consumer electronic devices for the benefit of individuals, is often hindered by issues that do not necessarily arise from technological limitations but rather are user-centered in nature. Whether new technologies become successfully embedded within individuals' daily routines and professionals' workflows often depends on the way in which ethical issues directly impacting on user trust have been addressed at the design concept generation, development, and deployment stages.

There is increasing recognition of a need to facilitate further convergence between the development of emerging technologies for promoting individuals' health and wellbeing and user-centered design research, with a view to achieving positive impact on individuals, care professionals, and healthcare systems. In addressing current development trends relating to user-centered digital innovation for health and wellbeing based on machine learning and immersive technologies, this Research Topic across Frontiers in Artificial Intelligence, Frontiers in Virtual Reality, and Frontiers in Big Data has attracted 13 contributions including original research articles, reviews, perspectives, as well as theoretical and methodological contributions, thereby providing a snapshot of recent and ongoing research and development.

Machine learning and immersive technologies: a case study

Whereas the Research Topic title explicitly refers to healthcare innovation, the broader scope has turned this article collection into an opportunity for reflection on a range of topics of relevance to both health and wellbeing. Topics have included the use of technologies for enhancing the provision of medical education and training, for improving workforce wellbeing, and for augmenting art therapy programmes with a view to increasing therapeutic compliance. Methods employed include Agile data science techniques, "CRoss Industry Standard Process for Data Mining" (CRISP-SM), "Preferred Reporting Items for Systematic reviews and Meta-Analyses" (PRISMA), thematic literature review, mini review, primary research (specifically, collection of data from university nursing students and surgeons), "Simulation Effectiveness Tool – Modified" (SET-M), established data science techniques, and human factors engineering methods. Key research themes that have emerged from the Research Topic are discussed below, followed by a reflection on priorities for further research and development.

Interdisciplinarity, ethics, and stakeholder engagement

The articles have highlighted a need for knowledge and expertise from across academic disciplines and professional practice to converge and underpin the development of digital innovations promoting individuals' health and wellbeing. Relevant disciplines and domains include computer science, user-centered design, human-computer interaction, engineering, human factors engineering, and the social sciences. Technology end user and stakeholder values, expectations, and requirements need to be addressed if new methods enabled by modern digital technologies are to be sustainably employed (Kolyshkina and Simoff; Goisauf and Cano Abadía; Khatun et al.; Buche et al.). Interestingly, the articles have generally suggested the importance of embedding end user and stakeholder perspectives within technology development workflows, although only a minority of the studies have explicitly articulated the need for extensive involvement of human-centered design researchers and practitioners for scaffolding and facilitating iterative development and evaluation (Khatun et al.). Unsurprisingly, the need to address ethical concerns appears intertwined with the recognized desirability of research objectives and methods to deliver deeper integration across discipline boundaries. This is illustrated by research advocating the adoption of intersectional social sciences perspectives within AI development for cancer diagnostics (Goisauf and Cano Abadía) and by studies focusing on the representativity of machine learning training data with a view to reducing health inequalities affecting specific ethnic groups in relation to the provision of diagnostic services (Khatun et al.).

Production-ready systems

The design and development of production-ready AI-based systems designed for flexibility and maintainability over time have received significant attention in recent years, particularly in the information systems, human-computer interaction, and engineering design literature. This is illustrated by Research Topic articles focusing on the definition of architectural requirements for healthcare cost estimation systems relying on dedicated predictive numerical models (Jackson et al.), and by studies delivering prototype models to enable healthcare professionals to query different Electronic Medical Record systems using intuitive interfaces based on natural language (Marshan et al.). Such efforts have achieved a balance between adapting research pipelines for production environments and identifying optimized architectural specifications from an information systems perspective.

Medical education and training

The emphasis in recent academic and professional discourse on opportunities afforded by immersive technologies, including virtual reality, augmented reality, and mixed reality, for achieving more efficient and inclusive delivery of medical educational and training programmes is reflected in this article collection. An interesting review study has focused on a comparison between technology-augmented methods and established approaches (Xu et al.). Reported benefits include enhanced student and trainee motivation, satisfaction, and learning outcomes, although the possible occurrence of undesired consequences of the use of immersive technologies, including cybersickness following prolonged exposure, has been noted. Interestingly, one study has focused on real-time detection of cybersickness with a view to reducing detrimental effects on user experience (Yalcin et al.). Proposed innovations based on mixed reality to streamline urology anatomy training and to facilitate pre-operative urology planning have attracted positive feedback from both university nursing students and surgeons, which encourages further research toward more extensive clinical validation (Sánchez-Margallo et al.). An interesting study has focused on an integration between generative AI and immersive technologies for designing augmented reality filters, with a view to improving medical students' perceptions of self-efficacy in recognizing selected disease manifestations (Stuart et al.).

Individuals' wellbeing

The potential of modern digital technologies for improving individuals' wellbeing has been the subject of recent research, which is reflected in this article collection. The breadth of contributions received illustrates the potential of artificial intelligence and immersive technologies for improving individuals' wellbeing, particularly in clinical and workplace settings. A theoretical model has been presented, explaining the psychological benefits of virtual immersion for oncology patients with emphasis on

distraction for alleviating anxiety and pain (Buche et al.). Opportunities have been identified for artistic expression within virtual reality environments to increase therapeutic compliance and to improve wellbeing outcomes for individuals in relation to psychotherapy and neurorehabilitation (Hadjipanayi et al.). A study has identified features of immersive technologies that hold potential for improving motor rehabilitation compliance and efficacy with stroke patients when used in combination with traditional approaches (Bui et al.). Such features include those enabling real-time movement tracking and the provision of reinforced feedback in line with established neurorehabilitation principles. A review of modern digital technologies for estimating individuals' wellbeing in workplace settings has generated useful recommendations on how real-time posture detection techniques can best be combined with the adoption of established human factors engineering best practices (Ataguba and Orji). This has enabled the identification of optimized algorithms to be employed in conjunction with physiological sensing methods toward the design of healthier workplaces.

Promoting a design-driven user-centered research and development agenda

Overall, the articles published under this Research Topic have highlighted the importance of conducting interdisciplinary research when tackling challenges at the intersection of technology development with human-centered design and human factors engineering. Integrative capabilities across academic disciplines and research methodologies—with emphasis on modern design research and design professional practice—have been identified as an important enabler of challenge-driven research and responsible innovation. Such insights are relevant to the United Nations Sustainable Development Goal number 3 ("Good health and wellbeing") and more broadly (Colecchia et al., 2024). A balanced distribution has been achieved in this collection of articles between applications of immersive technologies (Buche et al.; Xu et al.; Yalcin et al.; Sánchez-Margallo et al.; Stuart et al.; Hadjipanayi et al.; Bui et al.) and applications of machine learning and artificial intelligence (Kolyshkina and Simoff; Goisauf and Cano Abadía; Khatun et al.; Jackson et al.; Marshan et al.; Yalcin et al.). The authors speculate that future research is likely to reflect a convergence of immersive technologies and artificial intelligence in relation to the promotion of individuals' health and wellbeing. If that is the case, it is anticipated that the emphasis will be on human-centered design, participatory design, and methods addressing ethical issues of privacy, transparency, equitability, and fairness. One potential area of convergence relates to the development of personalized immersive experiences designed for inclusivity. It is expected that reliance on human-centered and participatory design methods will prove useful for scaffolding iterative design with significant involvement of technology end users and stakeholders. This is illustrated by the article discussing the use of artistic experiences within virtual reality environments to increase therapeutic compliance and to improve wellbeing outcomes (Hadjipanayi et al.). The articles have also highlighted several limitations with the technology state of the art, which calls for additional emphasis on interdisciplinary research, human-centered design, and inclusive design research moving forward. Such limitations include the following: digital access barriers and reduced digital literacy across user groups; the generally reduced availability of dedicated features for visually-impaired individuals—and for individuals with specific characteristics more broadly—compared with mainstream users; undesired effects from the use of head-mounted immersive displays—including cybersickness; the presence of different skill sets within interdisciplinary AI development teams, potentially reducing the benefits of Agile development. Moreover, although the articles published under this Research Topic have not focused on this aspect, future development will also need to address elements of clinical validation of digital technologies in the context of the relevant regulatory frameworks.

The authors argue that higher education institutions should take the lead in promoting long-term sustainable innovation in collaboration with research, healthcare, and commercial organizations. Non-academic organizations are sometimes better positioned for liaising with technology end users and practitioners, whose involvement in participatory design activities should be promoted wherever possible—ideally starting with the generation of early-stage design concepts. On the other hand, higher education institutions are ideally positioned to contribute expert knowledge and should lead on challenge-driven interdisciplinary research and on the promotion of postgraduate collaborative student projects. The emphasis should be on facilitating transfer of knowledge about modern human-centered design methods to non-academic organizations, with a view to creating favorable conditions for the achievement of broader and sustainable positive impact of research on individuals, society, and the economy.

Author contributions

FC: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Writing – original draft, Writing – review & editing. DG: Data curation, Formal analysis, Writing – review & editing. RQ: Data curation, Writing – review & editing. EC: Writing – review & editing. FW: Writing – review & editing.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Colecchia, F., Ceschin, F., and Harrison, D. (2024). Interdisciplinary integrative capabilities as a catalyst of responsible technology-enabled innovation: a higher education case study of Design MSc dissertation projects. *Int J Technol Des Educ*. doi: 10.1007/s10798-024-09901-w

frontiers in Big Data

PERSPECTIVE
published: 08 January 2019
doi: 10.3389/fdata.2018.00007

The Agile Deployment of Machine Learning Models in Healthcare

Stuart Jackson, Maha Yaqub and Cheng-Xi Li*

Analytics Center of Excellence, IBM Watson Health, Ann Arbor, MI, United States

The continuous delivery of applied machine learning models in healthcare is often hampered by the existence of isolated product deployments with poorly developed architectures and limited or non-existent maintenance plans. For example, actuarial models in healthcare are often trained in total separation from the client-facing software that implements the models in real-world settings. In practice, such systems prove difficult to maintain, to calibrate on new populations, and to re-engineer to include newer design features and capabilities. Here, we briefly describe our product team's ongoing efforts at translating an existing research pipeline into an integrated, production-ready system for healthcare cost estimation, using an agile methodology. In doing so, we illustrate several nearly universal implementation challenges for machine learning models in healthcare, and provide concrete recommendations on how to proactively address these issues.

Keywords: agile, analytics engineering, continuous delivery, health informatics, machine learning

OPEN ACCESS

Edited by:
Enrico Capobianco,
University of Miami, United States

Reviewed by:
Martin Romacker,
Roche, Switzerland
Reinhard Schneider,
University of Luxembourg,
Luxembourg

****Correspondence:***
Stuart Jackson
stuart.jackson@ibm.com;
stuart.jackson@nyu.edu

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Big Data

Received: *28 August 2018*
Accepted: *13 December 2018*
Published: *08 January 2019*

Citation:
Jackson S, Yaqub M and Li C-X (2019) The Agile Deployment of Machine Learning Models in Healthcare. Front. Big Data 1:7. doi: 10.3389/fdata.2018.00007

1. INTRODUCTION

Contemporary software engineering is driven by a number of key themes, such as agile development cycles and the continuous delivery of production software (Fowler and Highsmith, 2001; Shore and Warden, 2008). Such approaches allow for neater partition of development work related to current and future software capabilities, and help to streamline maintenance flows, product documentation, and development team communication. Unfortunately, the continuous deployment of predictive analytics is often hampered by poorly thought-out maintenance plans and non-agile methods of deployment, a phenomenon experienced across widespread industries (Demirkan and Dal, 2014), including in health informatics settings (Reeser-Stout, 2018). In this short Perspective, we describe our team's ongoing efforts and the lessons learned so far in the agile deployment of a new predictive analytics model related to healthcare cost estimation. While this model is designed for a specific use case (i.e., predicting cost in the US Medicaid population), our integrated deployment strategy is more general, and could transfer easily to other claims-based models.

We begin below with a brief overview of the typical challenges and maintenance issues experienced when deploying machine learning models in health informatics settings, using actuarial models as an example. We then describe our use of agile methods in a new actuarial product deployment, emphasizing the hybrid nature of agile data science, the important concepts of iteration and experimentation, and the unique challenges faced and solutions developed to fulfill key product requirements. Along the way, we provide a high-level description of the model that was trained and productionized, and an illustration of how internal maintenance and client use can occur side-by-side in the integrated production codebase. Finally, we conclude by providing some

general recommendations for hybrid development teams to consider when tasked with developing and deploying a new healthcare analytics product.

In passing, note that an earlier version of the research model we deployed was developed by a separate team at IBM, and has been described in detail in their separate methods paper (Ramamurthy et al., 2017). As such, we refrain from discussing this model from a deeper research or design perspective. Our aim in this short Perspective is to describe the challenges faced in refining and productionizing one instantiation of a research model, and the considerations required in sustaining continuous delivery and internal maintenance of a new healthcare analytic. In the spirit of continuous delivery, these efforts are necessarily ongoing.

2. ACTUARIAL MODELS IN HEALTHCARE

Accurate healthcare cost estimation is of critical importance to medical organizations, governments, and societies at large, with healthcare expenditures a primary drain on public resources worldwide. The availability of reliable cost estimates for a population can aid insurance plan administrators and other healthcare professionals in effective resource planning, in risk adjustment, and in developing strategies for population health management (Duncan, 2011). A wide variety of predictive algorithms have been developed over the years for estimating healthcare costs from administrative claims data, including numerous proprietary models (Winkelman and Mehmud, 2007). These models are often tailored for very specific patient populations, use cases, or input data needs; yet, a common goal of such models is the prospective identification of future high-cost claimants, often using linear or tree-based regression methods (Meenan et al., 2003; Bertsimas et al., 2008).

While the development of a high-quality risk model is itself a challenge, the successful, long-term deployment of such a model in applied settings is equally challenging, requiring careful consideration of the potential maintenance issues that could arise from changing industry, client, or technical needs. For example:

- Regular (e.g., yearly) updates might be required when new training or scoring data become available
- Irregular updates might be required when industry reference files (e.g., ICD diagnosis codes) change
- Minor model improvements or technical corrections and bug fixes might be required on an *ad-hoc* basis
- Major model improvements or new functionalities might be required subject to evolving client needs
- Changes to deployment hardware or other architectural constraints may need to be accommodated

Following the completion of any such maintenance task, an additional period of code review, model retraining, and product testing might also be necessary. Yet, the typical research model lacks the continuously-integrated organization necessary to efficiently handle such common maintenance issues. Below, we describe how our team adopted an agile framework in deploying a new actuarial model into production, streamlining the model training, and production process to support effective continuous delivery.

3. PRINCIPLES OF AGILE DATA SCIENCE

3.1. Avoiding the "Pull of the Waterfall"

Our product team was tasked with training and deploying a new claims-based risk model, which we approached initially from an agile framework. Agile software development practices are now industry standard, supporting efficient methods of collaboration and effective ways of getting work done (Fowler and Highsmith, 2001; Shore and Warden, 2008). As the field of data science evolves, however, it is increasingly clear that existing agile methods will need to adapt to successfully support this hybrid development domain (Jurney, 2017; Reeser-Stout, 2018). For example, the typical time allowed for a research data science project conflicts sharply with the standard agile development cycle. This can have the effect of forcing otherwise agile predictive analytics projects toward more sequential development cycles–the so-called "pull of the waterfall" (Jurney, 2017). This can be particularly problematic in healthcare research and product work, where additional constraints are often at play (e.g., strict data access rules).

To avoid the sequential handover of work from one group (e.g., analytics) to another (e.g., engineering), we established early on a hybrid development squad (**Figure 1A**), which facilitated the direct interaction between data scientists, software engineers, a quality assurance (QA) engineer, and a product owner. We also adopted the development and deployment terminology common in software engineering. For example, as a given development phase was completed, our product code was scheduled to pass first to a QA testing phase (or "TST"), then onto user-acceptance testing ("UAT"), and only then into production. While such terminology is somewhat foreign to many research data scientists, these methods prove essential in a highly collaborative, production context. In contrast, prior attempts at formalizing methods for predictive analytics development, such as CRISP-DM (Shearer, 2000), are too far removed from contemporary software engineering practices, having little to say about code deployment or collaboration across multi-functional (and often remotely-located) teams. That being said, there is enormous potential for the refinement of improved hybrid agile methodologies that more smoothly integrate with standard research science practices. Below, we describe one such hybrid strategy that we adopted during our product deployment.

3.2. Iteration and Experimentation

Our product team actively experienced the conflict between research and software development worlds, learning the hard way that specifying a final production date ahead of time is often incompatible with doing successful, agile data science (Jurney, 2017). To overcome this conflict, we evolved a process that roughly centered around two key ideas–*iteration* and *experimentation*. In the spirit of agile, our iterative process encompassed groups of tasks completed over individual three-week sprints, often involving extremely dynamic code changes

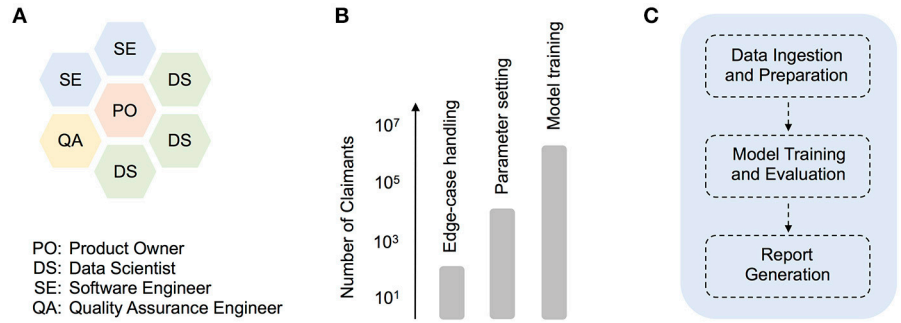

FIGURE 1 | The agile deployment of machine learning models in healthcare. **(A)** Organization of roles in our hybrid data science development squad, **(B)** Synthetic illustration of how different sets of production tasks typically warranted working with claimant dataset sizes of a particular scale, from technical debugging with smaller input datasets, to product model training with large-scale input data, **(C)** Schematic depicting higher-level modularization within the core model pipeline.

that ranged across the entire product pipeline. As a rule-of-thumb, we aimed to not only have the code running end-to-end at key iteration milestones, but more importantly, to better understand the data flow and model behavior at key points in the pipeline. Only then was the model pipeline deemed worthy of intermediate delivery to other users (e.g., QA engineer).

As iterations progressed, however, the importance of parallel experimentation quickly became apparent, both from a technical debugging and model training perspective (**Figure 1B**). For example, when building a client-facing product that deals with medical claims data, comprehensive edge-case handling is a particularly challenging technical issue. In our case, smaller input datasets often included blank or missing claims data for time ranges that the model expected (i.e., empty months), and revealed bugs in parts of the original pipeline dealing with the aggregation of disease and cost information. Given the numerous time-related, cross-dependencies in the prediction pipeline, the development of appropriate code fixes benefited greatly from over-and-back interactions with a QA engineer, and trial code runs with smaller datasets of varying size (e.g., 10^1 or 10^3 claimant records). In general, experiments with smaller dataset sizes were essential throughout development, for performing quick, technical debugging. As the product code was better refined, experiments on larger datasets (e.g., 10^5 claimant records) allowed for deeper code and model understanding (e.g., parameter exploration and tuning). Finally, as a given release date approached, production model fits were carried out on a formally-curated, large training dataset, with up to several million unique claimant records (**Figure 1B**).

3.3. Reinforcing the "Hybrid" Nature of the Work

How successful was this hybrid agile approach from a data science perspective? All of the data scientists agreed that this more foundational, agile approach to development provided clear advantages over isolated development styles, removing crucially the need for analytics developers to deliver code to a separate production team in a sequential fashion. The hybrid approach was self-reinforcing, in the sense that it encouraged all squad members to play multiple, interacting roles throughout development. For example, while data scientists played the major role in finalizing the core prediction model (described in detail later), the QA and software engineers had numerous opportunities to examine and refine this code, providing complementary feedback which improved the overall quality of the data science work. Likewise, while the QA and software engineers were primarily responsible for smooth deployment of the end-to-end pipeline (described in detail later), the data scientists spent substantial time facilitating this process through proper packaging and documentation of code. This facilitated the deployment process, and again, involved constant cross-squad interaction and feedback. The end results were successful, iterative deployments of the full codebase into production.

One broader advantage of this hybrid system is the ability to more easily organize data science projects at the level of multiple squads, thereby maximizing resource use and collaboration potential. For example, the hybrid squad described here (aka *Mercury*), worked independently of several other "planet" squads (e.g., *Jupiter*), although with some higher-level direction from a scrum master working across multiple teams. While this type of organization is already common in traditional software engineering environments (e.g., tribes, chapters, etc.), we believe it requires the creation of truly hybrid squads, involving both data scientists and software engineers, to be successful in an analytic development context.

4. CHALLENGES IN ANALYTIC DEVELOPMENT

4.1. Providing Multiple Model Types in One Platform

We now describe in detail several key requirements of the software development product, and the associated development and coding challenges we faced. In a following subsection, we illustrate how we tackled these problems and gauged the success of the improved processes.

The key requirements of our analytic product related to the functionality to provide access to multiple model results for a single input claims dataset. Specifically, actuarial predictions

TABLE 1 | Parameters of the twelve model variants deployed to fulfill diverse end-user requirements.

No.	Model type	No.	Model type
1	Concurrent, Total Cost	7	Prospective, Total Cost
2	Concurrent, Total Cost ($100k)	8	Prospective, Total Cost ($100k)
3	Concurrent, Total Cost ($250k)	9	Prospective, Total Cost ($250k)
4	Concurrent, Medical Cost Only	10	Prospective, Medical Cost Only
5	Concurrent, Medical Cost Only ($100k)	11	Prospective, Medical Cost Only ($100k)
6	Concurrent, Medical Cost Only ($250k)	12	Prospective, Medical Cost Only ($250k)

"Concurrent" models predict annual healthcare costs for the same 12-month period as the input data; "Prospective" models provide predictions for the subsequent 12-month period (i.e., next year). Note that 'Total Cost' refers to models that predict combined medical and pharmacy costs, and that the dollar values in brackets refer to truncation thresholds applied to outlier claimant data. See text for further details.

for up to twelve model variants were necessary (**Table 1**), with models varying in terms of the time period of prediction (e.g., concurrent year vs. prospective year), the cost components being predicted (e.g., medical costs only vs. medical and pharmacy costs), and the form of thresholding or truncation applied to outliers (e.g., no truncation vs. $100k or $250k truncation). These model variants were selected to support diverse risk-adjustment use cases, from the retrospective measurement of provider performance (e.g., using concurrent year models), to the estimation of a population's future healthcare costs for resource planning purposes (e.g., using prospective year models). While the original research model we inherited provided some powerful functionality in this regard, there were numerous engineering challenges to face from a production and deployment perspective. Crucially, several of the capabilities below had to be productionized to work identically in both training and deployment (i.e., scoring) scenarios, thereby supporting a more easily maintainable codebase and product. For example:

- The product training data cohort needed to be defined, and needed to accommodate all twelve model conditions, in terms of the time range of claims data, pharmacy costs availability, and the exclusion of certain data sources with unreliable or incomplete costs (e.g., capitated health plan data, claimants that are dual-eligible for Medicare, etc.). An inadequetely-defined training cohort would negatively affect the functioning of all of the data ingestion, preparation, and scoring modules (described in detail later), and the lack of a cohort definition module would make all future updates to the product unnecessarily cumbersome.
- The functionality for selecting and running only a subset of models (based on user input) needed to be developed. While input requests to run the analytic were initially sent in one-at-a-time, along the way it was decided that the end user should have the ability to run multiple versions of the model in one request (e.g., all twelve variants or only a subset). For example, dependent on the particular risk-adjustment use case, an end user might request results from only the concurrent year models, or only those with a specific truncation threshold (e.g., $100k).
- The flexibility to change key model parameters automatically and "online" (i.e., during actual training or scoring) was similarly required. In some instances, these changes involved relatively minor parameter updates (e.g., switching from one truncation threshold to another). For others, substantial pipeline rerouting was necessary (e.g., switching from local directory operations during training to directory operations which are dynamically set during deployment).

4.2. Solutions That Satisfy the Key Product Requirements

How did we tackle these challenges, and how did we measure the success of the resulting processes? By implementing a variety of sustainable coding practices, we developed solutions to these issues as follows:

- To ensure integrity of our training cohort, we first developed a formal "cohort definition plan" (similar in spirit to a CONSORT diagram). This plan involved several stages, including the key steps of: (a) selecting a large random sample of patient IDs (e.g., 5 million) covering a time range of interest (e.g., 2013-2017); (b) excluding those subset of IDs linked to capitated health plans or having dual-eligible for Medicare status (as claims costs from these subgroups of patients are often incomplete); (c) extracting the complete enrollment and medical claims data for all remaining valid patient IDs.
- This plan was then implemented in a series of code modules, used to extract data from internal, proprietary Medicaid databases. To verify the success of this implementation, we monitored the smooth running and completion of the data extraction code, and ensured that the resulting cohort data mapped correctly to a formal data dictionary that we had prepared. The data dictionary in particular acts as a fundamental reference file for all users of the production analytic, and was a crucial milestone in our development. With minimal modifications, the overall cohort definition module can be used in future analytic developments (e.g., after new training data is obtained or industry reference files are updated).
- To ensure that the analytic had the functionality to take specific user requests and to update parameters "on the fly," we began by making the decision to keep the code as flexible as possible and not to hardcode parameters for any of the model variants. Our team then developed a series of input-level scripts (primarily in shell scripting languages), as well as later module-specific templates (in Python), that updated dependent on the specific user input. For example, if the user requested results for models with $100k truncation only (models 2, 5, 8, and 11; see **Table 1**), the analytic proceeded to automatically update relevant parts of the code and configuration files for each of these models in turn.
- After finalizing these cohort definition and model specification techniques independently, the overall set of solutions was tested through extensive model running at key iteration milestones (e.g., during quality assurance and user-acceptance

testing). The above functionalities, which served to provide multiple model types in one platform, were successfully deployed in each of our iterative releases.

5. DEPLOYING AN END-TO-END SOLUTION

5.1. The Core Model Pipeline

We deployed an end-to-end healthcare cost estimation solution that can be maintained internally with relative ease (i.e., recalibrated or extended in functionality), and continuously pushed to cloud production environments where client scoring can occur. The product we deployed was refined from a previously developed research pipeline, described in detail elsewhere (Ramamurthy et al., 2017), and contains a family of cost models that we trained on Medicaid claims data. As described earlier, models varied along a number of parameter dimensions, including the time period of prediction (e.g., concurrent year vs. prospective year), the cost components being predicted (e.g., medical costs only vs. medical and pharmacy costs), and the form of thresholding or truncation applied to outliers (e.g., no truncation vs. $100k or $250k truncation). Model inputs included basic demographics (e.g., age and gender), enrollment details (e.g., number of months enrolled), and diagnosis information. In passing, note that while we avoid discussing Medicaid data in detail here, focusing instead on our general agile development framework, the interested reader can find numerous sources discussing specific disease prevalence and hospitalization issues in these claimant populations e.g., Trudnak et al. (2014).

To facilitate maintenance and future development, the pipeline utilizes a core set of code modules, which at a high-level, perform essentially three functions (**Figure 1C**). First, a sequence of *data ingestion and preparation* modules read in the required input files (including enrollment, claims, and auxiliary input files), and perform batch processing on these in order to aggregate raw input data into intermediate database tables. Operations such as dummy-encoding, feature enrichment, and sparse matrix creation are also carried out at this stage, to improve the efficiency of later data handling. In the second high-level phase of processing, data is passed to modules that support *model training and evaluation*. The data subsets required for training or evaluation are isolated at this stage, and in the case of internal training, a multi-stage regression model is trained and saved. Model parameters can be configured in advance using configuration files. Model evaluation or scoring is then performed, either on a separate test dataset (in internal training mode) or directly on client data (in client scoring mode). Finally, in the *report generation* step, patient-level predictions (e.g., costs and risk scores), as well as overall model performance reports, are saved to file.

5.2. The Integrated Product Codebase

Refining a product codebase is a collaborative effort involving multiple people contributing to the same overall product vision. To make this happen, it is important to host the code in proper software deployment platforms, and to use technologies that support efficient collaboration. This helps to ensure that maintenance tasks can be carried out easily without affecting the core pipeline. For example, changes to industry reference files (e.g., ICD diagnosis codes) or other evolving industry requirements (e.g., the use of social determinants information) can be smoothly incorporated into the product data model and modular pipeline. The integrated product codebase and deployment process that we refined allows for easy modification of code components and model recalibration, without significant effects on code integration and product delivery. Below we describe key characteristics of this integrated product pipeline.

The integrated product codebase is defined by three key platform characteristics–*version control*, *containerization*, and *continuous integration*. First, by refining the final codebase in an environment that supports version control (e.g., GitLab; https://about.gitlab.com), we ensured that every team member had access to and could modify the same codebase. This facilitated efficient collaboration on the final product, and limited the need for having standalone versions of the code existing in different places. Second, to control the vast array of packages required in code running, we adopted a containerized approach to code delivery. Even a single faulty or missing package can cause critical breakages in a code pipeline similar to the one we deployed. To avoid this problem, container technologies (e.g., Docker; https://www.docker.com) allow one to host code in a virtual environment that has all the required software packages pre-installed. This approach facilitated deployment on a production server and eliminated the need for team members to individually sift through package installation requirements, saving a considerable amount of time. Finally, the pipeline was integrated by software engineers into final testing and production layers, with the aim of automating the code building and establishing continuous integration of the product. Open-source continuous integration tools (e.g., Jenkins; https://jenkins.io) allowed the team to monitor the code deployment in real-time and quickly identify any defects.

6. CONCLUSION AND KEY RECOMMENDATIONS

We provided here a brief overview of our attempts at refining an agile data science methodology to support a new healthcare analytic deployment, emphasizing the hybrid nature of agile data science and the important roles played by team iteration and model experimentation. There is clearly enormous potential for the development of more formal approaches to agile data science, both in healthcare and elsewhere, which we hope this brief overview has illustrated. As a starting point, we provide the following general recommendations, when faced with the challenges of any new healthcare analytic deployment:

- Track your work: Incorporate a formal agile tracking tool into your work from the outset, and organize each piece of your work into a separate "user story." Tracking systems encourage teams to remain actively engaged and to communicate clearly, behaviors which are particularly important in hybrid teams, where skill sets might overlap less than in traditional software

engineering teams. In addition, agile approaches encourage the use of the "backlog" to keep track of upcoming tasks, as not every feature or user story will be complete for a given product release. We recommend using it. For example, at a preliminary milestone in our development work, non-critical aspects of the output file formatting were not fully finalized; by adding appropriate notes to the backlog, the team was able to more easily monitor the status of this and other remaining tasks across iterations.

- Investigate and implement: For some data science issues (e.g., finalizing production parameter settings), it is important to allow sufficient time for problem understanding. We have found it beneficial in such cases to pair "investigation" and "implementation" user stories. What do we mean by this, and why is it important in a data science context? The purpose of creating an investigation user story is to allow the team sufficient scope and time to research complex model details, and thereby more clearly define the ideal boundaries of the prediction pipeline. The specifics of a computational model are often more nuanced in implementation than traditional software components, and implementation errors often have subtle effects that are difficult to detect. In one example, our squad investigated methods for handling the well-known medical claims "run-out" issue (i.e., the time lag between recent medical services and subsequent claims payment processing, which can be several months in duration). After detailed investigation, we developed a plan in sync with industry standards, accommodating the processing of relevant medical claims paid within a 3-month run-out window after the end of a claim year. By performing and closing out this investigation story, team flow was at least maintained, even if production code was not necessarily updated significantly. We then implemented this plan in a separate user story, and measured the implementation success on small samples of data, by comparing aggregated claims costs from the pipeline to manually aggregated costs. We believe this "investigate then implement" approach to work definition is particularly useful in a hybrid squad context, as it reinforces continual communication and transfer of learning throughout the diversely-skilled squad.
- Release in increments: Develop an incremental product release strategy, and communicate this plan clearly and early to others. This should help in ensuring that realistic deadlines are formed, and that these are driven primarily by the team's estimation of the workload (not by external stakeholder needs). More specifically, the release process should comprise of a strategic set of deadlines which cater appropriately to development team resources and incremental product release goals. For example, in our development work, we first scheduled an early or "Beta" release directed toward an internal client. By doing this, the deployed codebase was put through the standard release testing processes (e.g., quality assurance and user-acceptance testing), without any changes or biases introduced by the data science team for a defined period of time. This allowed time for independent feedback regarding the codebase and for completion of remaining backlog tasks (e.g., improvements to the formatting of output reports). It also allowed for fine-tuning and retraining of the core prediction model on a larger training dataset, thereby improving the overall performance and quality of a later "Production" release.

In conclusion, we believe the hybrid squad model has many benefits over isolated teams when doing data science software development. In addition to improved communication and collaboration, as well as the removal of sequential handover of work, the hybrid squad model provides substantial opportunity for skills transfer and innovation that would otherwise not occur. While it is still early days for the hybrid squad system we have described, the potential has been obvious to everyone involved, including at the team management level. That being said, the recommendations above illustrate some likely areas of difficulty for new hybrid squads, which we suspect will typically arise in setting sensible release strategies and deadlines. Yet, we firmly believe that hybrid analytics teams are the future, and sincerely hope that others can build on the recommendations outlined here.

DATA AVAILABILITY STATEMENT

No datasets were generated for the purpose of writing this manuscript, and all relevant information is contained here. Proprietary code referred to in the manuscript is not publicly available (property of IBM).

AUTHOR CONTRIBUTIONS

All authors contributed to the conception and design of the paper. SJ wrote the first draft, and MY, C-XL, and SJ wrote additional sections. All authors contributed to manuscript revision and approved the final version.

ACKNOWLEDGMENTS

The authors would like to thank their IBM Watson Health colleagues for providing guidance during different phases of this production work, their colleagues in IBM Research for sharing earlier research code, and their manager (Rajashree Joshi) for encouraging the drafting and submission of this manuscript.

REFERENCES

Bertsimas, D., Bjarnadottir, M. V., Kane, M. A., Kryder, J. C., Pandey, R., Vempala, S., et al. (2008). Algorithmic prediction of health-care costs. *Operat. Res.* 56, 1382–1392. doi: 10.1287/opre.1080.0619

Demirkan, H. and Dal, B. (2014). The data economy: why do so many analytics projects fail? *Analytics Magazine* Available online at: [http://analytics-magazine.org/the-data-economy-why-do-so-many-analytics-projects-fail/] (Accessed December 20, 2018).

Duncan, I. (2011). *Healthcare Risk Adjustment and Predictive Modeling*. Winsted, CT:Actex.

Fowler, M. and Highsmith, J. (2001). The agile manifesto. *Softw. Dev.* 9, 28–35. Available online at: http://www.drdobbs.com/open-source/the-agile-manifesto/184414755?queryText=the+agile+manifesto (Accessed December 20, 2018).

Jurney, R. (2017). *Agile Data Science 2.0*. Sebastopol, CA:O'Reilly.

Meenan, R. T., Goodman, M. J., Fishman, P. A., Hornbrook, M. C., O'Keeffe-Rosetti, M. C., et al. (2003). Using risk-adjustment models to identify high-cost risks. *Med. Care* 41, 1301–1312. doi: 10.1097/01.MLR.0000094480.13057.75

Ramamurthy, K. N., Wei, D., Ray, E., Singh, M., Iyengar, V., Katz-Rogozhnikov, D., et al. (2017). "A configurable, big data system for on-demand healthcare cost prediction," in *2017 IEEE International Conference on Big Data* (Boston, MA), 1524–1533.

Reeser-Stout, S. (2018). "A hybrid approach to the use of agile in health IT [conference presentation]," in *Healthcare Information and Management Systems Society (HIMSS), Annual Conference 2018* (Las Vegas, NV).

Shearer, C. (2000). The CRISP-DM model: the new blueprint for data mining. *J. Data Warehousing* 5, 13–22. Available online at: https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf (Accessed December 20, 2018).

Shore, J. and Warden, S. (2008). *The Art of Agile Development*. Sebastopol, CA:O'Reilly.

Trudnak, T., Kelley, D., Zerzan, J., Griffith, K., Jiang, H. J., and Fairbrother, G. L. (2014). Medicaid admissions and readmissions: understanding the prevalence, payment, and most common diagnoses. *Health Affairs* 33, 1337–1344. doi: 10.1377/hlthaff.2013.0632

Winkelman, R. and Mehmud, S. (2007). A comparative analysis of claims-based tools for health risk assessment. *Society of Actuaries Report* Available online at: http://www.soa.org/research-reports/2007/hlth-risk-assement/ (Accessed December 20, 2018).

Conflict of Interest Statement: At the time of original drafting, the authors were all full-time employees of IBM. The authors share this perspective with the sole aim of contributing to open dialogue on the topics described in the manuscript.

frontiers in Big Data

ORIGINAL RESEARCH
published: 26 May 2021
doi: 10.3389/fdata.2021.660206

Interpretability of Machine Learning Solutions in Public Healthcare: The CRISP-ML Approach

Inna Kolyshkina[1]* *and Simeon Simoff*[2,3]

[1] *Analytikk Consulting, Sydney, NSW, Australia,* [2] *School of Computer, Data and Mathematical Sciences, Western Sydney University, Sydney, NSW, Australia,* [3] *MARCS Institute for Brain, Behaviour and Development, Western Sydney University, Sydney, NSW, Australia*

Public healthcare has a history of cautious adoption for artificial intelligence (AI) systems. The rapid growth of data collection and linking capabilities combined with the increasing diversity of the data-driven AI techniques, including machine learning (ML), has brought both ubiquitous opportunities for data analytics projects and increased demands for the regulation and accountability of the outcomes of these projects. As a result, the area of interpretability and explainability of ML is gaining significant research momentum. While there has been some progress in the development of ML methods, the methodological side has shown limited progress. This limits the practicality of using ML in the health domain: the issues with explaining the outcomes of ML algorithms to medical practitioners and policy makers in public health has been a recognized obstacle to the broader adoption of data science approaches in this domain. This study builds on the earlier work which introduced CRISP-ML, a methodology that determines the interpretability level required by stakeholders for a successful real-world solution and then helps in achieving it. CRISP-ML was built on the strengths of CRISP-DM, addressing the gaps in handling interpretability. Its application in the Public Healthcare sector follows its successful deployment in a number of recent real-world projects across several industries and fields, including credit risk, insurance, utilities, and sport. This study elaborates on the CRISP-ML methodology on the determination, measurement, and achievement of the necessary level of interpretability of ML solutions in the Public Healthcare sector. It demonstrates how CRISP-ML addressed the problems with data diversity, the unstructured nature of data, and relatively low linkage between diverse data sets in the healthcare domain. The characteristics of the case study, used in the study, are typical for healthcare data, and CRISP-ML managed to deliver on these issues, ensuring the required level of interpretability of the ML solutions discussed in the project. The approach used ensured that interpretability requirements were met, taking into account public healthcare specifics, regulatory requirements, project stakeholders, project objectives, and data characteristics. The study concludes with the three main directions for the development of the presented cross-industry standard process.

Keywords: machine learning, interpretability, public health, data science methodology, CRISP-ML, necessary level of interpretability, interpretability matrix, cross-industry standard process

OPEN ACCESS

Edited by:
Kok-Leong Ong,
La Trobe University, Australia

Reviewed by:
Md. Anisur Rahman,
Charles Sturt University, Australia
Yafei Han,
Massachusetts Institute of Technology, United States

***Correspondence:**
Inna Kolyshkina
inna@analytikk.com

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Big Data

Received: *29 January 2021*
Accepted: *07 April 2021*
Published: *26 May 2021*

Citation:
Kolyshkina I and Simoff S (2021) Interpretability of Machine Learning Solutions in Public Healthcare: The CRISP-ML Approach. Front. Big Data 4:660206. doi: 10.3389/fdata.2021.660206

1. INTRODUCTION AND BACKGROUND TO THE PROBLEM

Contemporary data collection and linking capabilities, combined with the growing diversity of the data-driven artificial intelligence (AI) techniques, including machine learning (ML) techniques, and the broader deployment of these techniques in data science and analytics, have had a profound impact on decision-making across many areas of human endeavors. In this context, public healthcare sets priority requirements toward the robustness, security (Qayyum et al., 2021), and interpretability (Stiglic et al., 2020) of ML solutions. We use the term *solution* to denote the algorithmic decision-making scenarios involving ML and AI algorithms (Davenport and Kalakota, 2019). While the early AI solutions for healthcare, like expert systems, possessed limited explanatory mechanisms (Darlington, 2011), these mechanisms proved to have an important role in clinical decision-making and, hence, made healthcare practitioners, clinicians, health economists, patients, and other stakeholders aware about the need to have such capabilities.

Healthcare domain imposes a broad spectrum of unique challenges to contemporary ML solutions, placing much higher demands with respect to interpretability, comprehensibility, explainability, fidelity, and performance of ML solutions (Ahmad et al., 2018). Among these properties of ML solutions, interpretability is particularly important for human-centric areas like healthcare, where it is crucial for the end users to not only have access to an accurate model but also to trust the validity and accuracy of the model, as well as understand how the model works, what recommendation has been made by the model, and why. These aspects have been emphasized by a number of recent studies, most notably in Caruana et al. (2015) and Holzinger et al. (2017), and summarized in the study by Ahmad et al. (2018).

Healthcare, similar to government and business digital services, manufacturing with its industrial internet of things and creative industries, experienced the much celebrated manifestations of "big data," "small data," "rich data," and the increased impact of ML solutions operating with these data. Consequently, the interpretability of such solutions and the explainability of the impact of the judgements they assist to make or have made and, where needed, the rationale of recommended actions and behavior are becoming essential requirements of contemporary analytics, especially in society-critical domains of health, medical analysis, automation, defense, security, finance, and planning. This shift has been further accentuated by the growing worldwide commitment of governments, industries, and individual organizations to address their endeavors toward the United Nations Sustainable Development Goals[1] and by the data-dependent scientific and technological challenges faced by the rapid response to the COVID-19 pandemic. The later challenges highlight and reinforce the central role of healthcare, backed by science, technology, lateral thinking, and innovative solutions in societal and economic recovery.

Some state-of-the-art overviews, such as Doshi-Velez and Kim (2017) and Gilpin et al. (2019) related to interpretability, as well as more method-focused papers, like Lipton (2018) and Molnar et al. (2019), tend to use interpretability and explainability interchangeably. They also report that the interpretability of ML solutions and the underlying models is not well-defined. The study related to interpretability is scattered throughout a number of disciplines, such as AI, ML, human-computer interaction (HCI), visualization, cognition, and social sciences (Miller, 2019), to name a few of the areas. In addition, the current research seems to focus on particular categories or techniques instead of addressing the overall concept of interpretability.

Recent systematic review studies, Gilpin et al. (2018) and Mittelstadt et al. (2019), have clarified some differences and relationships between interpretability and explainability in the context of ML and AI. In these domains, interpretability refers to the degree of human interpretability of a given model, including "black box" models (Mittelstadt et al., 2019). Machine interpretability of the outcomes of ML algorithms is treated separately. Explanability refers primarily to the number of ways to communicate an ML solution to others (Hansen and Rieger, 2019), i.e., the "ways of exchanging information about a phenomenon, in this case the functionality of a model or the rationale and criteria for a decision, to different stakeholders." Both properties of ML solutions are central to the broader adoption of such solutions in diverse high-stake healthcare scenarios, e.g., predicting the risk of complications to the health condition of a patient or the impact of treatment change.

While some authors (for instance, Hansen and Rieger, 2019; Mittelstadt et al., 2019; Samek and Müller, 2019) consider interpretability as an important component of explainability of ML solutions in AI, we view interpretability and explainability as complementary to each other, with interpretability being fundamental in ensuring trust in the results, transparency of the approach, confidence in deploying the results, and, where needed, quality of the maintenance of ML solutions. Further, in this study, we used the term interpretability in a broader sense, which subsumes communication and information exchange aspects of explainability.

We considered two connected aspects of the development of the overall concept of interpretability in ML solutions:

1. *methods*, which include the range of interpretable ML algorithms and interpretability solutions for AI/ML algorithms;
2. *methodologies* in data science, which consider explicitly the achievement of the necessary (for the project) interpretability of the ML solutions.

There is a wide collection of interpretable ML methods and methods for the interpretation of ML models. Murdoch et al. (2019) provide a compact and systematic approach toward their categorization and evaluation. Methods are categorized into model-based and *post-hoc* interpretation methods. They are evaluated using predictive accuracy, descriptive accuracy, and relevancy, the PDR framework (Murdoch et al., 2019), where relevancy is evaluated against human audience. The framework also provides common terminology for practitioners. Guidotti et al. (2018) and Carvalho et al. (2019) provide extensive systematic overviews with elaborate frameworks of the state-of-the-art of interpretability methods. Mi et al. (2020) provide

[1] https://www.un.org/sustainabledevelopment/sustainable-development-goals/ and https://sdgs.un.org/goals.

broader taxonomy and comparative experiments, which can help practitioners in selecting suitable models with complementary features for addressing interpretability problems in ML solutions.

Model interpretability and explainability are crucial for clinical and healthcare practice, especially, since not only non-linear models but also inherently more interpretable ones, like decision trees, if large and complex, become difficult to comprehend (Ahmad et al., 2018).

On the other hand, working with data in the healthcare domain is complex at every step, starting from establishing and finding the relevant, typically numerous, diverse, and heterogeneous data sources required to address the research objective; integrating and mapping these data sources; identifying and resolving data quality issues; pre-processing and feature engineering without losing information or distorting it; and finally using the resulting high-dimensional, complex, sometimes unstructured, data to build a high-performing interpretable model. This complexity further supports the argument for the development of ML methodologies which explicitly embed interpretability through the data science project life cycle and ensure the achievement of the level of interpretability of ML solutions that had been agreed for the project. Interpretability of an ML solution can serve a variety of stakeholders involved in data science projects in connection with the implementation of their outcomes.

Interpretability of an ML solution can serve a variety of stakeholders, involved in data science projects and related to the implementation of their outcomes in algorithmic decision making (Berendt and Preibusch, 2017). For instance, the human-centric visual analytics methodology "Extract-Explain-Generate" for interrogating biomedical data (Kennedy et al., 2008) explicitly relates different stakeholders (molecular biologist, clinician, analysts, and managers) with specific areas of knowledge extraction and understanding associated with the management of patients. This study is focused on addressing the methodological challenges and opportunities of broad embedding of interpretability (including the selection of methods of interpretability that are appropriate for a project, given its objectives and constraints).

2. CHALLENGES AND OPPORTUNITIES IN CREATING METHODOLOGIES WHICH CONSISTENTLY EMBED INTERPRETABILITY

In order to progress with the adoption of ML in healthcare, a consistent and comprehensive methodology is needed: first, to minimize the risk of project failures, and second, to establish and ensure the needed level of interpretability of the ML solution while addressing the above-discussed diverse requirements to ML solutions. The rationale supporting these needs is built on a broader set of arguments about:

- the high proportion of data science project failures, including those in healthcare;
- the need to support an agreed level of interpretability and explainability of ML solutions;
- the need for consistent measurement and evaluation of interpretability of ML solutions; and
- the emerging need for standard methodology, which explicitly embeds mechanisms to manage the achievement of the level of interpretability of ML solutions required by stakeholders through the project.

Further, in this section, we use these arguments as dimensions around which we elaborate the challenges and opportunities for the design of cross-industry data science methodology, which is capable of handling interpretability of ML solutions under the complexity of the healthcare domain.

2.1. High Proportion of Data Science Project Failures

Recent reports, which include healthcare-related organizations, estimate that up to 85% of data science/ML/AI projects do not achieve their stated goals. The latest NewVantage Partners Big Data and AI Executive Survey, based on the responses from C-Executives from 85 blue-chip companies of which 22% are from Healthcare and Life Sciences, noted that only 39% of companies are managing data as an asset (NewVantage Partners LLC, 2021). Fujimaki (2020) emphasized that "the economic downturn caused by the COVID-19 pandemic has placed increased pressure on data science and BI teams to deliver more with less. In this type of environment, AI/ML project failure is simply not acceptable." On the other hand, the NewVantage Partners survey (NewVantage Partners LLC, 2021) emphasized that, over the 10 years of conducting these surveys, organizations continue to struggle with their transformation into data-driven organizations, with only 29% achieving transformational business outcomes. Only 24% have created a data-driven organization, a decline from 37.8%, and only 24% have forged a data culture (NewVantage Partners LLC, 2021), a result which, to a certain extent, is counterintuitive to the overall expectation of the impact of AI technologies to decision-making and which projected benefits from the adoption of such technologies.

A number of sources (e.g., vander Meulen and Thomas, 2018; Kaggle, 2020; NewVantage Partners LLC, 2021) established that a key reason for these failures is linked to the lack of proper process and methodology in areas, such as requirement gathering, realistic project timeline establishment, task coordination, communication, and designing a suitable project management framework (see also Goodwin, 2011; Stieglitz, 2012; Espinosa and Armour, 2016). Earlier works have suggested (see, e.g., Saltz, 2015) that improved methodologies are needed as the existing ones do not cover many important aspects and tasks, including those related to interpretability (Mariscal et al., 2010). Further, studies have shown that the biased focus on the tools and systems has limited the ability to gain value from the effort of organizational analytics effort (Ransbotham et al., 2015) and that data science projects need to increase their focus on process and task coordination (Grady et al., 2014; Gao et al., 2015; Espinosa and Armour, 2016). A recent Gartner Consulting report also emphasizes the role of processes and methodology (Chandler and Oestreich, 2015) and practitioners agree with this view (for examples and analyses from diverse practical perspectives see

Goodson, 2016; Arcidiacono, 2017; Roberts, 2017; Violino, 2017; Jain, 2019).

2.2. Support for the Required Level of Interpretability and Explainability of ML Solutions

In parallel with the above-discussed tendencies, there is pressure on the creation of frameworks/methodologies, which can ensure the necessary interpretability for sufficient explainability of the output of the ML solutions. While it has been suggested, in recent years, that it is only a matter of time before ML will be universally used in healthcare, building ML solutions in the health domain proves to be challenging (Ahmad et al., 2018). On the one hand, the demands for explainability, model fidelity, and performance in general in healthcare are much higher than in most other domains (Ahmad et al., 2018). In order to build the trust in ML solutions and incorporate them in routine clinical and healthcare practice, medical professionals need to clearly understand how and why an ML solution-driven decision has been made (Holzinger et al., 2017; Vellido, 2020).

This is further affected by the fact that the ML algorithms that achieve a high level of predictive performance, e.g., boosted trees (Chen and Guestrin, 2016) or deep neural networks (Goodfellow et al., 2016), are quite complex and usually difficult to interpret. In fact, some researchers argue that performance and interpretability of an algorithm are in reverse dependence (Ahmad et al., 2018; Molnar et al., 2019). Additionally, while there are a number of techniques aiming to explain the output of the models that are not directly interpretable, as many authors note (e.g., Holzinger et al., 2017; Gilpin et al., 2019; Rudin, 2019; Gosiewska et al., 2020), current explanatory approaches, while promising, do not seem to be sufficiently mature. Molnar et al. (2019) found that the reliability of some of these methods deteriorates if the number of features is large or if the level of feature interactions is high, which is often the case in health data. Further, Gosiewska and Biecek (2020) showed that current popular methods for explaining the output of ML models, like SHAP (Lundberg and Lee, 2017) and LIME (Ribeiro et al., 2016), produce inconsistent results, while Alvarez-Melis and Jaakkola (2018) found that the currently popular interpretability frameworks, particularly model-agnostic perturbation-based methods, are often not robust to small changes of the input, which clearly is not acceptable in the health domain.

There is a firm recognition of the impact of ML solutions in economics, including health economics, especially in addressing "predictive policy" problems (Athey, 2019). Many authors (e.g., Holzinger et al., 2017; Dawson et al., 2019; Rudin, 2019) note that in the high-stake areas (e.g., medical field, healthcare) solutions, in which the inner workings are not transparent (Weller, 2019), can be unfair, unreliable, inaccurate, and even harmful. Such views are reflected in the legislation on data-driven algorithmic decision-making, which affects citizens across the world. The European Union's General Data Protection Regulation (GDPR) (EU, 2016), which entered into force in May 2018, is an example of such early legislation. In the context of the emerging algorithmic economy, there are also warnings to policymakers to be aware of the potential impact of legislations like GDPR on the development of new AI and ML solutions (Wallace and Castro, 2018).

These developments increased the pressure on creation of frameworks and methodologies, which can ensure sufficient interpretability of ML solutions. In healthcare, such pressure is amplified by the nature of the interactive processes, wherein neither humans nor the algorithms operate with unbiased data (Sun et al., 2020).

Major technology developers, including Google, IBM, and Microsoft, recommend responsible interpretability practices (see, e.g., Google, 2019), including the development of common design principles for human-interpretable machine learning solutions (Lage et al., 2019).

2.3. Consistent Measurement and Evaluation of Interpretability of ML Solutions

While there are a number of suggested approaches to measuring interpretability (Molnar et al., 2019), a consensus on the ways of measuring or evaluating the level of interpretability has not been reached. For example, Gilpin et al. (2019) found that the best type of explanation metrics is not clear. Murdoch et al. (2019) mentioned that, currently, there is confusion about the interpretability notion and a lack of clarity about how the proposed interpretation approaches can be evaluated and compared against each other and how to choose a suitable interpretation method for a given issue and audience. The PDR framework (Murdoch et al., 2019), mentioned earlier, is a step in the direction of developing consistent evaluations. Murdoch et al. (2019) further note that there is limited guidance on how interpretability can actually be used in data science life cycles.

2.4. The Emerging Need for Standard Methodology for Handling Interpretability

Having a good methodology is important for the success of a data science project. To our knowledge, there is no formal standard for methodology in the data science projects (see Saltz and Shamshurin, 2016). Through the years, the CRISP-DM methodology (Shearer, 2000) created in the late 1990s has become a de-facto standard, as evidenced from a range of works (see, e.g., Huang et al., 2014; Niño et al., 2015; Fahmy et al., 2017; Pradeep and Kallimani, 2017; Abasova et al., 2018; Ahmed et al., 2018). An important factor of its success is the fact that it is industry, tool, and application agnostic (Mariscal et al., 2010). However, the research community has emphasized that, since its creation, CRISP-DM had not been updated to reflect the evolution of the data science process needs (Mariscal et al., 2010; Ahmed et al., 2018). While various extensions and refined versions of the methodology, including IBM's Analytics Solutions Unified Method for Data Mining (ASUM-DM) and Microsoft's Team Data Science Process (TDSP), were proposed to compensate the weaknesses of CRISP-DM, at this stage, none of them has become the standard. In the more recent years, variations of CRISP-DM tailored for the healthcare (Catley et al.,

2009) and medical domain, such as CRISP-MED-DM (Niaksu, 2015), have been suggested. The majority of organisations that apply a data analysis methodology prefers extensions of CRISP-DM (Schäfer et al., 2018). Such extensions are fragmented and either propose additional elements into the data analysis process, or focus on organisational aspects without the necessary integration of domain-related factors (Plotnikova, 2018). These might be the reasons for the observed decline of its usage as reported in studies by Piatetsky-Shapiro (2014), Bhardwaj et al. (2015), and Saltz and Shamshurin (2016). Finally, while methodologies from related fields, like the agile approach used in software development, are being considered for use in data science projects, there is no clear clarity on whether they are fully suitable for the purpose, as indicated by Larson and Chang (2016); therefore, we did not include them in the current scope.

This overall lack of consensus has provided an opportunity to reflect on the philosophy of the CRISP-DM methodology and create a comprehensive data science methodology, through which interpretability is embedded consistently into an ML solution. Such methodology faces a list of requirements:

- It has to take into account the different perspectives and aspects of interpretability, including model and process explainability and interpretability;
- It has to consider the desiderata of explainable AI (fidelity, understandability, sufficiency, low construction overhead, and efficiency) as summarized in Hansen and Rieger (2019);
- It needs to support consistent interaction of local and global interpretability of ML solutions with other established key factors in data science projects, including predictive accuracy, bias, noise, sensitivity, faithfulness, and domain specifics;

In addition, healthcare researchers have indicated that the choice of interpretable models depends on the use case (Ahmad et al., 2018).

In order to standardize the expectations for interpretability, some of these requirements have been addressed in the recently proposed CRISP-ML methodology (Kolyshkina and Simoff, 2019). In section 3, we will briefly discuss the major concepts differentiating CRISP-ML methodology. The CRISP-ML approach includes the concepts of *necessary level of interpretability* (NLI) and *interpretability matrix* (IM), described in detail by Kolyshkina and Simoff (2019), and therefore aligns well with the view of health researchers that the choice of interpretable models depends upon the application and use case for which explanations are required (Ahmad et al., 2018). To illustrate that, in section 4, we present a use case in the public health field that illustrates the typical challenges met and the ways CRISP-ML helped to address and resolve them.

3. CRISP-ML METHODOLOGY—TOWARD INTERPRETABILITY-CENTRIC CREATION OF ML SOLUTIONS

The CRISP-ML methodology (Kolyshkina and Simoff, 2019) of building interpretability of an ML solution is based on revision and update of CRISP-DM to address the opportunities discussed in section 2. It follows the CRISP-DM approach in terms of being industry-, tool-, and application-neutral. CRISP-ML accommodates the necessary elements to work with diverse ML techniques and create the right level of interpretability through the whole ML solution creation process. Its seven stages are described in **Figure 1**), which is an updated version of the CRISP-ML methodology diagram in the study by Kolyshkina and Simoff (2019).

Central to CRISP-ML is the concept of necessary level of interpretability of an ML solution. From this view point, CRISP-ML can be differentiated as a methodology of establishing and building the necessary level of interpretability of a business ML solution. In line with Google's guidelines on the responsible AI practices in the interpretability area (Google, 2019) and expanding on the approach proposed by Gleicher (2016), we have specified the concept of minimal necessary level of interpretability of a business ML solution as the combination of the degree of accuracy of the underlying algorithm and the extent of understanding the inputs, inner workings, the outputs, the user interface, and the deployment aspects of the solution, which is required to achieve the project goals. If this level is not achieved, the solution will be inadequate for the purpose. This level needs to be established and documented at the initiation stage of the project as part of requirement collection (see Stage 1 in **Figure 1**).

We then describe an ML solution as sufficiently interpretable or not based on whether or not it achieved the required level of interpretability. Obviously, this level will differ from one project to another depending on the business goals. If individuals are directly and strongly affected by the solution-driven decision, e.g., in medical diagnostics or legal settings, then both the ability to understand and trust the internal logic of the model, as well as the ability of the solution to explain individual predictions, are of highest priority. In other cases, when an ML solution is used in order to inform business decisions about policy, strategy, or interventions aimed to improve the business outcome of interest, then it is necessary to understand and trust the internal logic of the model that is of most value, while individual predictions are not the focus of the stakeholders. For example, in one of our projects, an Australian state organization wished to establish what factors influenced the proportion of children with developmental issues and what interventions can be undertaken in specific areas of the state in order to reduce that proportion. The historical, socioeconomic, and geographic data provided for the project was aggregated at a geographic level of high granularity.

In other cases, e.g., in the case of an online purchase recommender solution, the overall outcome, such as increase in sales volume, may be of higher importance than interpretability of the model. Similar requirements of solution interpretability were in a project where an organization owned assets that were located in remote areas and were often damaged by birds or animals nests. The organization wished to lower their maintenance cost and planning by identifying as soon as possible the assets where such nests were present instead of doing expensive examination of each asset. This was achieved by building a ML solution that classified Google Earth images of the assets into those with and without nests. In this project, it was

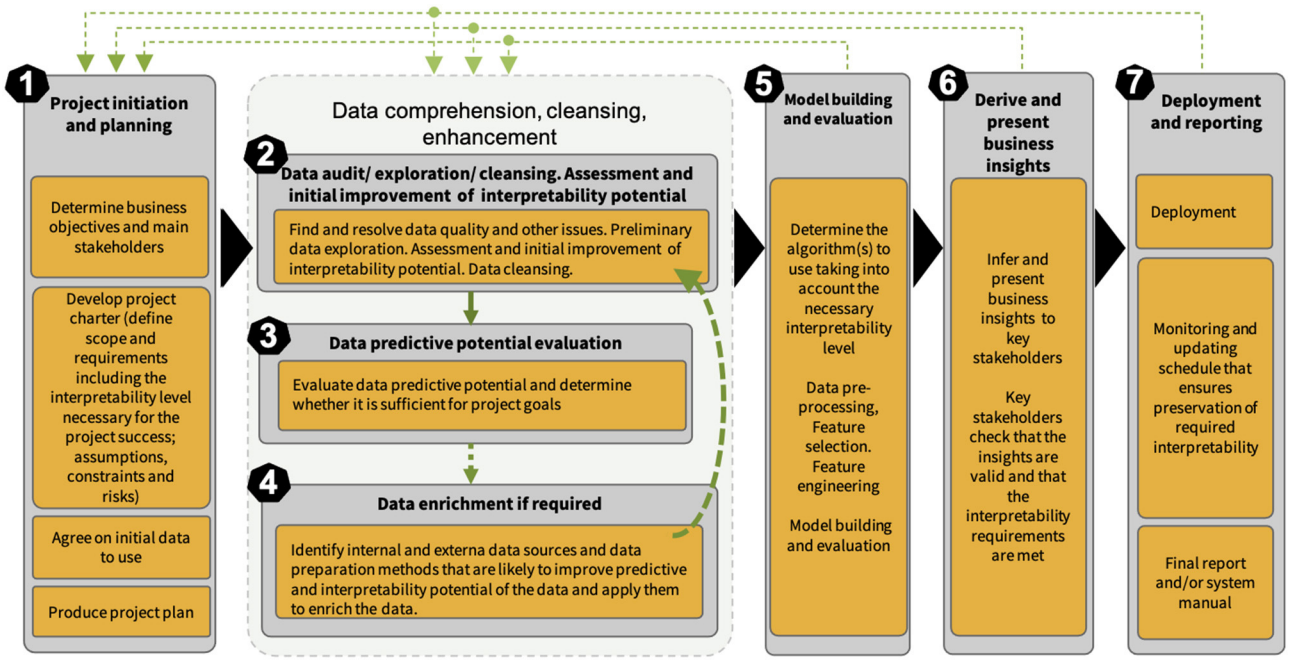

FIGURE 1 | Conceptual framework of CRISP-ML methodology.

important to identify a proportion of assets that were as high as possible with nests on them, while misclassifying an individual asset image was not of great concern.

The recently published CRISP-ML(Q) (Studer et al., 2020) proposes an incremental extension of CRISP-DM with the monitoring and maintenance phases. While the study mentions "model explainability" referring to the technical aspects of the underlying model, it does not consider interpretability and explainability in a systematic way as CRISP-ML (Kolyshkina and Simoff, 2019). Interpretability is now one of the most important and quickly developing universal requirements, not only a "best practice" requirement in some industries. It is also a legal requirement. CRISP-ML (Kolyshkina and Simoff, 2019) ensures that the necessary interpretability level is identified at the requirement collection stage. The methodology then ensures that participants establish the activities for each stakeholder group at each process stage that are required to achieve this level. CRISP-ML (Kolyshkina and Simoff, 2019) includes stages 3 and 4 (data predictive potential assessment and data enrichment in **Figure 1**), which are not present in CRISP-ML(Q) (Studer et al., 2020). As indicated in Kolyshkina and Simoff (2019), skipping these important phases can result in potential scope creep and even business project failure.

In Kolyshkina and Simoff (2019), the individual stages of the CRISP-ML methodology were presented in detail. Each stage was illustrated with examples from cases from a diverse range of domains. There, the emphasis was on the versatility of CRISP-ML as a industry-neutral methodology, including its approach to interpretability. In this study, we focus on a single case study from health-related domain in order to present a comprehensive coverage of each stage and the connections between the stages, and provide examples of how the required level of interpretability of the solution is achieved through carefully crafted involvement of the stakeholders as well as decisions made at each stage. This study does not provide comparative evaluation of CRISP-ML methodology in comparison to CRISP-DM (Shearer, 2000), ASUM-DM (IBM Analytics, 2015), TDSP (Microsoft, 2020), and other methodologies discussed by Kolyshkina and Simoff (2019). The purpose of the study is to demonstrate, in a robust way, the mechanics of explicit management of interpretability in ML through the project structure and life cycle of a data science methodology. Broader comparative evaluation of the methodology is the subject of a separate study.

The structure of the CRISP-ML process methodology has embedded flexibility in it, indicated by the cycles, which link the model-centric stages back to the early data-centric stages, as shown in **Figure 1**. Changes inevitably occur in any project over the course of the project life cycle, and CRISP-ML reflects that. The most typical changes, related to data availability, quality, and analysis findings, occur mostly at stages 2–4, as shown in **Figure 1**. This is illustrated in our case study and was discussed in detail in the study by Kolyshkina and Simoff (2019). Less often changes occur at stages 5–7 in **Figure 1**. From experiential observations, such changes are more likely to occur in longer projects with a volume of work requiring more than 6–8 months for completion. They are usually driven by amendments in project scope and requirements including the necessary level of interpretability (NLI), that are caused by factors external to the analytical part of the project. These factors can be global, such

as environmental, political, or legislative factors; organization-specific (e.g., updates in the organizational IT structure, the way of data storage or changes in the stakeholder team), or they could be related to the progress in ML and ML-related technical areas (e.g., the advent of a new, better performing predictive algorithm).

In this study, we present the stages of CRISP-ML in a rigid manner, around the backbone of the CRISP-ML process, represented by the solid black triangle arrows in **Figure 1** to maintain the emphasis on the mechanisms for handling interpretability in each of these steps, rather than exploring the iterative nature of the approach. For consistency of the demonstration, we draw all detailed examples through the study from the specific public health case study. As a result, we are able to illustrate in more depth how we sustain the level of interpretability through the process structure of the project. The study complements the study by Kolyshkina and Simoff (2019), where, through the examples drawn from a variety of cases, we demonstrated the versatility of CRISP-ML. The methodological treatment of interpretability in evolving scenarios and options is beyond the scope of this study.

4. CASE STUDY ILLUSTRATING THE ACHIEVEMENT OF THE NLI OF MACHINE LEARNING SOLUTION

In this study, we will describe a detailed real-world case study in which, by going through each project stage, we illustrate how CRISP-ML facilitates data science project stakeholders in establishing and achieving the necessary level of interpretability of ML solution.

We would like to emphasize that the specific analytic techniques and tools mentioned in the respective stages of the case study are relevant specifically to this particular study. They illustrate the approach and the content of the interpretability mechanisms of CRISP-ML. However, there are many other available methods and method combinations that can achieve the objectives of this and other projects.

We place a particular focus on the aspects and stages of CRISP-ML from the perspective of demonstrating the flow and impact of interpretability requirements and on how they have been translated into the necessary level of interpretability of the final ML solution. Further, the structure of this section follows the stages of CRISP-ML process structure in **Figure 1**. All sensitive data and information have been masked and altered to protect privacy and confidentiality, without loss of the sensible aspects relevant to this presentation.

4.1. Background. High-Level Project Objectives and Data Description

An Australian State Workers Compensation organization sought to predict, at an early stage of a claim, the likelihood of the claim becoming long-term, i.e., a worker staying on income support for 1 year or more from the date of lodgement. A further requirement was that the prediction model should be easily interpretable by the business.

The data that the analysis was to be based upon were identified by the organizational experts, based on the outcomes for about 20,000 claims incurred in the recent years, and included the following information:

- injured worker attributes, e.g., date of birth, gender, occupation, average weekly earnings, residential address;
- injury attributes, e.g., injury date, the information on the nature, location, mechanism, and agency of injury coded according to the National Type of Occurrence Classification System[2];
- employer attributes (size, industry classification);
- details of all worker's income support or similar payments.

4.2. Building the Project Interpretability Matrix: An Overall Approach

Interpretability matrix is usually built at Stage 1 of the project as part of the requirement collection process. Data science practitioners recognize Stage 1 as crucial for the overall project success (see, e.g., PMI, 2017), as well as from the solution interpretability building perspective (Kolyshkina and Simoff, 2019).

The IM as a structure for capturing and translating interpretability requirements into specific actions and activities is generalized. However, the specific content of its cells depends on the project. Kolyshkina and Simoff (2019) demonstrated the CRISP-ML stages consistently applied to different projects across a number of industries, data sets, and data types.

It covers the activities needed to start up the data science project: (a) the identification of key stakeholders; (b) documenting project objectives and scope; (c) collecting requirements; (d) agreeing upon initial data; (e) preparing a detailed scope statement; and (f) developing project schedule and plan. The deliverable of this stage was a project charter documenting the above activities.

4.2.1. Interpretability-Related Aspects of the Project Charter: Business Objectives, Main Stakeholders, and Interpretability Level

We will describe in more detail the aspects of the project charter that were directly related to this study, specifically the established business objectives, main stakeholders, and the established necessary interpretability requirements.

4.2.1.1. Business objectives and main stakeholders.

The established objectives included:

1. Build an ML system that will explain what factors and to what extent influence the outcome, i.e., claim duration;
2. Allow the organization to derive business insights that will help make data-driven accurate decisions regarding what changes can be done to improve the outcome, i.e., reduce the likelihood of a long claim by a specified percentage;

[2]Type of Occurrence Classification System (3rd Edition, Revision 1), Australian Government—Australian Safety and Compensation Council, Canberra, 2008, https://www.safeworkaustralia.gov.au/doc/type-occurrence-classification-system-toocs-3rd-edition-may-2008).

3. Be accurate, robust, and work with real-world organizational data;
4. Have easy-to-understand outputs that would make sense to the executive team and end users (case managers) and that the end users could trust;
5. Present the output as business rules that are easy to understand for end users and to deploy, monitor, and update in organizational data.
6. Ensure that the overall ML solution is easy to understand and implement by the Information Technology (IT) team of the organization and to monitor/update the Business Intelligence (BI) team of the organization.

The *main stakeholders* were identified as follows: Executive team (E); End Users/Domain Experts, i.e., Case management team (DE); Information Technology team who would implement the solution in the organizational data (IT); Business Intelligence team who would monitor the solution performance and update the underlying model (BI); and Modeling team (M). These abbreviations are used further in the descriptions of the stages of the IM.

4.2.1.2. The established necessary interpretability level.

The necessary interpretability level (Kolyshkina and Simoff, 2019) was established as follows.

- The E, IT, and DE teams needed to have a clear understanding of all internal and external data inputs to be used: their reliability, quality, and whether the internal inputs were representative of the organizational data that the solution would be deployed on.
- The E and DE teams needed to have a clear understanding of the high-level data processing approach (e.g., missing values treatment, aggregation level), as well as high-level modeling approach and its proven validity.
- The outputs needed to be provided in the form of easily understandable business rules. The E and DE teams needed to gain a clear understanding of the rules and to be able to assess their business validity and usefulness from the business point of view.
- The BI team, who would monitor the solution performance and update it as required, need to have a clear understanding of:
  - the data processing stage, as well as the modeling algorithm, its validity, and suitability from the ML point of view;
  - how to assess the solution performance and how the solution needs to be audited, monitored, and updated, as well as how often this should occur.
- The IT team, who would deploy the solution needed to have a clear understanding of the format of the output and confirm that it can be deployed in the organizational data within the existing constraints (e.g., resources, cost) and without disrupting the existing IT systems.

4.2.2. Creating the Project IM: An Overall Approach

The next step is to create and fill out the IM, whose rows show CRISP-ML stages, and columns represent key stakeholders. In each cell of the matrix, we showed what needs to be done by each stakeholder at each project stage to ensure that the required level of solution interpretability is achieved. Matrix cells can be grouped horizontally when there are common requirements for a group of stakeholders. Matrix cells can be grouped vertically when there are common requirements for a specific stakeholder across a number of stages in CRISP-ML. This matrix, once completed, becomes part of the business requirements document. The activities it outlines are integrated into the project plan and are reviewed and updated along with the project plan.

4.2.2.1. Definition of stakeholder involvement extent.

We define the extent of involvement of a stakeholder group needed to achieve the necessary interpretability level in a particular project stage as follows:

- high extent of involvement—the stakeholder group needs to be directly and actively involved in the solution development process to ensure that the NLI is achieved at the stage;
- medium extent of involvement—the stakeholder group needs to receive detailed regular updates on the progress of the stage and get directly involved in the work from time to time to ensure that the NLI is achieved at the stage. For example, this can refer to DE and IT providing information helping to better understand data sources and business processes of the organization.
- low extent of involvement—the stakeholder group is kept informed on the general progress of the stage.

In **Figure 2**, green color background indicates high extent of involvement of a stakeholder group, yellow color shows medium extent of involvement, and the cells with no color in the background show low level of involvement. Depending on the project, the coloring of the cells of the IM will vary. For example, if it had not been necessary to provide knowledge transfer ("Ongoing knowledge and skill development") to the BI team, then their involvement in Stage 2–5 would have been low and the respective cells would have been left with no color in the background.

4.2.2.2. High-level IM diagram.

Figure 2 shows a high-level interpretability matrix for the project.

4.3. Entries to the Project Interpretability Matrix at Each Stage of CRISP-ML

Further, we discuss entries to the project IM at each stage of CRISP-ML.

4.3.1. Stage 1

The content of the interpretability matrix related to the project initiation and planning stage (i.e., the first row of the matrix) has been discussed in detail above and is summarized in **Figure 3**.

4.3.2. Stages 2–4

Stages 2–4 in **Figure 1** are mainly data-related and form the data comprehension, cleansing, and enhancement mega-stage.

	Key stakeholders				
Project stage	**E**	**DE**	**BI**	**IT**	**M**
Stage 1. Project initiation and planning	Requirement gathering; finalising scope; agreeing initial data to use; discussing relevant domain aspects. Preparing project charter. Project sign off				
Stage 2. Data audit. Quality and sanity checks	Kept informed on general progress	Regular consultations; clarification of domain-related aspects. Provide feedback and domain insight	Working with M team. Ongoing knowledge and skill development	Kept informed on general progress	Data audit. Establishing and resolving any data issues. Interpretability and predictive potential of the data assessment and improvement. Underlying model building and evaluation.
Stage 3. Evaluation of data predictive potential					
Stage 4. Data enrichment if needed					
Stage 5. Model building and evaluation					
Stage 6. Derive and present business insights	Evaluate business insights and checking that requirements are met. Review of M team presentation and report	Consulting M team in preparing report and presentation for E	Consult M team in preparing report and presentation for E	Consult about deployment aspects	Derive business insights. Check that the requirements including interpretability level are met. Prepare report and presentation for E team outlining this.
Stage 7. Deploying. Monitoring. Updating	Kept informed on general progress	Provide feedback and domain insight	Reviewing solution manual and other documentation. Developing solution building skills	Solution deployment	Preparing solution manual and other documentation. Consulting in deployment if required. Knowledge sharing with BI team.

FIGURE 2 | High-level interpretability matrix for the project.

Further, we consider the content of the interpretability matrix for each of these stages, they are represented by the second, third, and fourth rows of interpretability matrix.

4.3.2.1. **Stage 2.**

Data audit, exploration, and cleansing played a key role in achieving the interpretability level needed for the project.

Figure 4 demonstrates the content of the interpretability matrix at this stage.

This stage established that the data contained characteristics that significantly complicated the modeling, such as a large degree of random variation, multicollinearity, and a highly categorical nature of many potentially important predictors. These findings helped guide the selection of the modeling and data pre-processing approach.

Random variation. During workshops with E, DE, and other industry experts, it became clear that there were certain "truths" that pervaded the industry, and we used these to engage with subject matter experts (SME) and promote the value of our modeling project. One such "truth" was that claim duration was influenced principally by nature and location of injury, but in combination with the age of the injured worker, and specifically, older workers tended to have longer duration claims. Our analysis demonstrated the enormous amount of random variation that existed in the data. For example, age, body location, and injury type only explained 3–7% of variation in claim duration. There was agreement among the experts that the

	Key stakeholders				
Project stage	**E**	**DE**	**BI**	**IT**	**M**
Stage 1. Project initiation and planning	Stage aim: establish key stakeholders; refine and document detailed project objectives, gather requirements including the necessary level of interpretability; agree upon initial data to use; establish scope; estimate effort, duration and costs; do risk analysis; develop project schedule				
	Work together with the other stakeholders to establish, refine and document detailed project objectives; agree scope Gather requirements including the necessary interpretability level of the solution including the validity of the input data and their preprocessing, format of the output, minimal required accuracy and proportion of explained data, inner workings and implementation aspects as well as who needs to understand what aspects of the solution to what extent to ensure this interpretability level is achieved.				
	Gain a high-level understanding of the approach for the ML solution; the solution building process and potential outcomes. Help modellers to gather requirements and identify suitable data sources.				Gain in-depth understanding of the requirements including the necessary interpretability level. Gain high-level understanding of the data agreed to be used; domain aspects relevant to the project goals achievement and sources for information available on these aspects (e.g., access to experts and documentation)

FIGURE 3 | Interpretability matrix content for Stage 1.

	Key stakeholders				
Project stage	**E**	**DE**	**BI**	**IT**	**M**
Stage 2. Data audit, exploration and cleansing	Stage aim: ensure that data inputs into the solution are of the required quality and validity, make sense to DEs and suits the documented requirements of IT team. Identify and address				
	Kept informed on general progress	Provide domain-related knowledge to help MT to build data understanding and assess data quality and validity	Provide data-related knowledge to help MT to build data understanding and assess data quality and validity. Ongoing knowledge and skill development.	Kept informed on general progress	Develop data understanding and quality/ validity assessment. Identify and address any data issues in consultation with DE and BI teams

FIGURE 4 | Interpretability matrix content for Stage 2.

industry "truths" were insufficient to accurately triage claims and that different approaches were needed.

Our exploratory analysis revealed strong random variation in the data, confirming the prevalent view among the workers' compensation experts that it is the intangible factors, like the injured worker's mindset and relationship with the employer, that play the key role in the speed of recovery and returning to work. The challenge for the modeling, therefore, was to uncover the predictors that represent these intangibles.

Sparseness. Most of the available variables were categorical with large numbers of categories. For example, the variable "Injury Nature" has 143 categories and "Body Location of Injury" has 76 categories. Further, some categories had relatively few observations which made any analysis involving them potentially unreliable and not statistically valid. Such sparseness presented another data challenge.

Multicollinearity. There was a high degree of multicollinearity between numerical variables in the data.

Data pre-processing. First, we reduced the sparseness among categories by combining some categorical levels in consultation with SMEs to ensure that the changes made business sense. Second, we used a combination of correlation analysis, as well as advanced clustering and feature selection approaches, e.g., Random Forests (see, e.g., Shi and Horvath, 2006) and PCAMIX method using iterative relocation algorithm and ascendant hierarchical clustering (Chavent et al., 2012) to reduce multicollinearity and exclude any redundant variables.

4.3.2.2. Stage 3.

Figure 5 shows the content of the interpretability matrix related to the evaluation of the predictive potential of the data (i.e., the third row of the matrix). This stage is often either omitted or not stated explicitly in other processes/frameworks (Kolyshkina and Simoff, 2019); however, it is crucial for the project success because it establishes whether the information in the data is sufficient for achieving the project goals.

To efficiently evaluate what accuracy could be achieved with the initially supplied data, we employed the following different data science methods that have proven their excellence at extracting maximum predictive power from the data: Deep Neural Nets, Random Forests, XGBoost, and Elastic Net. The results were consistent for all the methods used and showed that only a small proportion of the variability of claim duration was explained by the information available in the data. Therefore, the predictive potential of the initially supplied data, containing claim and worker's data history, indicated that the data set is insufficient for the project objectives. Data enrichment was required.

These findings were discussed with DE who then were invited to share their business knowledge about sources that could enrich the initial data predictive power.

4.3.2.3. Stage 4.

Data enrichment. **Figure 6** shows the content of the interpretability matrix related to the data enrichment stage. Based on the DE feedback and results of external research, we enriched the data with additional variables, including:

- lag between injury occurrence and claim lodgement (claim reporting lag);
- information on the treatment received (e.g., type of providers visited, number of visits, provider specialty);
- information on the use of medications and, specifically, on whether a potent opioid was used.

We assessed the predictive value of the enriched data in the same way as before (see section 4.3.2.2), and found that there was a significant increase in the proportion of variability explained by the model. Of particular relevance was the incorporation of the prior claim history of claimants, including previous claim count, type and nature of injury, and any similarity with the current injury.

Further, the data enrichment was a key step in building further trust of the DE team. The fact that the model showed that the

	Key stakeholders				
Project stage	**E**	**DE**	**BI**	**IT**	**M**
Stage 3. Data predictive potential evaluation	Stage aim: assess whether the initially identified data is sufficient for project purposes				
	Kept informed on general progress	Consultation and clarification of any domain-related aspects	Consultation and clarification of any domain-related aspects. Knowledge and skills building	Kept informed on general progress	Assess data predictive potential using different highly predictive ML methods

FIGURE 5 | Interpretability matrix content for Stage 3.

cost of a claim can be significantly dependent on the providers a worker visited built further trust in the solution, because it confirmed the hunch of domain experts that they previously had not had enough evidence to prove.

4.3.3. Stage 5

Figure 7 shows the content of the interpretability matrix for the model building and evaluation stage. To achieve the right interpretability level, it is crucial that modelers choose the right technique that will balance the required outcome interpretability with the required level of accuracy of the model, which is often a challenge (see, e.g., Freitas, 2014), as well as with other requirements/constraints (e.g., the needed functional form of the algorithm). In our case, it was required that the model explained at least 70% of variability.

At this stage, the ML techniques to be used for modeling are selected, taking into account the predictive power of the model, its suitability for the domain and the task, and the NLI. The data is pre-processed, and modeled, and the model performance is evaluated. The solution output was required to be produced in the form of business rules, and therefore, the feature engineering methods and modeling algorithms used included rule-based techniques, e.g., decision trees, and association rules-based methods.

4.3.4. Stage 6

Figure 8 shows how the interpretability matrix reflects the role of interpretability in the formulation of business insights necessary to achieve the project goals and in helping the E and DE to understand the derived business insights and to develop trust in

	Key stakeholders				
Project stage	**E**	**DE**	**BI**	**IT**	**M**
Stage 4. Data enrichment if needed	Stage aim: additional data sources that can potentially improve the predictive potential are identified via research and DE advice, the new data are extracted, audited, cleansed and added to the initial data. Then predictive potential of the enriched data is assessed to check if the necessary level of predictive potential is achieved				
	Kept informed on general progress	Help identify data sources suitable for enrichment. Regular consultations; clarification of domain-related aspects. Provide feedback and domain insight	Help identify data sources suitable for enrichment. Working with M team. Ongoing knowledge and skill development	Kept informed on general progress	Access the new data, extract, audit, cleanse it and added to the initial data. Then assess the predictive potential of the enriched data to check if the necessary level of predictive potential is achieved

FIGURE 6 | Interpretability matrix content for Stage 4.

	Key stakeholders				
Project stage	**E**	**DE**	**BI**	**IT**	**M**
Stage 5. Model building and evaluation	Stage aim: select ML approach to be used in development of the model underlying the solution. Evaluate the resulting model performance taking into account the interpretability requirements as well as statistical performance aspects				
	Kept informed on general progress	Regular consultations; clarification of domain-related aspects. Provide feedback and domain insight	Working with M team. Ongoing knowledge and skill development	Kept informed on general progress	Select the ML techniques to be used for modelling taking into account interpretability requirements. Build the model and evaluate its performance i taking into account the interpretability requirements as well as statistical performance aspects

FIGURE 7 | Interpretability matrix content for Stage 5.

them. Modelers, BI and DEs, prepared a detailed presentation for the E, explaining not only the learnings from the solution but also the high-level model structure and its accuracy.

4.3.5. Stage 7

The final model provided the mechanism for the organization to allocate claims to risk segments based on the information known at early stages. From the technical point of view, the business rules were confirmed by the E, DE, and IT to be easy to deploy as they are readily expressed as SQL code. Based on this success, a modified version of claims triage was deployed into production.

Figure 9 shows the shift of responsibilities for ensuring the achieved interpretability level is maintained during the future use of the solution. At this stage, the deployment was being scheduled, and the monitoring/updating process and schedule was prepared, based on the technical report provided by the M team that included project code, the solution manual, and updating and monitoring recommendations.

5. CONCLUSIONS

This study contributes toward addressing the problem for providing organizations with capabilities to ensure that the ML solutions they develop to improve decision-making are transparent and easy to understand and interpret. If needed, the logic behind the decisions can be explained to any external party. Such capability is essential in many areas, especially in health-related fields. It allows the end users to confidently interpret the ML output use to make successful evidence-based decisions.

In an earlier study (Kolyshkina and Simoff, 2019), we introduced CRISP-ML, a methodology of determining

	Key stakeholders				
Project stage	**E**	**DE**	**BI**	**IT**	**M**
Stage 6. Derive and present business insights	Stage aim: E and DE develop understanding of business insights derived, their validity and importance as well as whether the required interpretability level has been achieved				
	Evaluate business insights and interpretability level achieved	Consulting M team in preparing report and presentation for E	Consulting M team in preparing report and presentation for E	Consult about deployment aspects	Derive business insights and assess interpretability level achieved. Prepare report and presentation for E

FIGURE 8 | Interpretability matrix content for Stage 6.

	Key stakeholders				
Project stage	**E**	**DE**	**BI**	**IT**	**M**
Stage 7. Deploying. Monitoring. Updating	Stage aim: Deployment. Developing solution manual and other related documentation. Creating solution monitoring and updating schedule. Knowledge transfer.				
	Kept informed on general progress	Provide feedback and domain insight	Reviewing solution manual and other documentation. Developing solution monitoring and updating skills	Solution deployment	Preparing solution manual and other documentation. Creating monitoring and updating schedule. Consulting in deployment if required. Knowledge sharing with BI team.

FIGURE 9 | Interpretability matrix content for Stage 7 includes activities ensuring the achieved interpretability level is maintained during the future utilization of the solution.

the interpretability level required for the successful real-world solution and then achieving it *via* integration of the interpretability aspects into its overall framework instead of just the algorithm creation stage. CRISP-ML combines practical, common-sense approach with statistical rigor and enables organizations to establish shared understanding across all key stakeholders about the solution and its use and build trust in the solution outputs across all relevant parts of the organization. In this study, we illustrated CRISP-ML with a detailed case study of building an ML solution in the Public Health sector. An Australian state workplace insurer sought to use their data to establish clear business rules that would identify, at an earlier stage of a claim, those with high probability of becoming serious/long-term. We showed how the necessary level of solution interpretability was determined and achieved. First, we showed how it was established by working with the key stakeholders (Executive team, end users, IT team, etc.). Then, we described how the activities that were required to be included at each stage of building the ML solution to ensure that this level is achieved was determined. Finally, we described how these activities were integrated into each stage.

The study demonstrated how CRISP-ML addressed the problems with data diversity, unstructured nature of the data, and relatively low linkage between diverse data sets in the healthcare domain (Catley et al., 2009; Niaksu, 2015). The characteristics of the case study which we used are typical for healthcare data, and CRISP-ML managed to deliver on these issues, ensuring the required interpretability of the ML solutions in the project.

While we have not completed formal evaluation of CRISP-ML, there are two aspects which indicate that the use of this methodology improves the chances of success of data science projects. On the one hand, CRISP-ML is built on the strengths of CRISP-DM, which made it the preferred and effective methodology (Piatetsky-Shapiro, 2014; Saltz et al., 2017), addressing its identified limitations in previous works (e.g., Mariscal et al., 2010). On the other hand, CRISP-ML has been successfully deployed in a number of recent real-world projects across several industries and fields, including credit risk, insurance, utilities, and sport. It ensured on meeting the interpretability requirements of the organizations, regardless of industry specifics, regulatory requirements, types of stakeholders involved, project objectives, and data characteristics, such as type (structured as well as unstructured), size, or complexity level.

CRISP-ML is a living organism and, as such, it responds to the rapid progress in the development of ML algorithms and the evolution of the legislation for their adoption. Consequently, CRISP-ML development includes three directions: (i) the development of a richer set of quantitative measures of interpretability features for human interpretable machine learning, (ii) the development of the methodology and respective protocols for machine interpretation, and (iii) the development of formal process support. The first one is being extended in a way to provide input to the development and evaluation of common design principles for human interpretable ML solutions in line with that described in the study by Lage et al. (2019). This strategic development adds the necessary agility for the relevance of the presented cross-industry standard process.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication.

FUNDING

Elements of the study on interpretability in ML solutions are partially supported by the Australian Research Council Discovery Project (grant no.: DP180100893).

REFERENCES

Abasova, J., Janosik, J., Simoncicova, V., and Tanuska, P. (2018). "Proposal of effective preprocessing techniques of financial data," in *2018 IEEE 22nd International Conference on Intelligent Engineering Systems (INES)* (IEEE), 293–298. doi: 10.1109/INES.2018.8523922

Ahmad, M. A., Eckert, C., Teredesai, A., and McKelvey, G. (2018). Interpretable machine learning in healthcare. *IEEE Intell. Inform. Bull.* 19, 1–7. Available online at: https://www.comp.hkbu.edu.hk/~cib/2018/Aug/iib_vol19no1.pdf

Ahmed, B., Dannhauser, T., and Philip, N. (2018). "A lean design thinking methodology (LDTM) for machine learning and modern data projects," in *Proceedings of 2018 10th Computer Science and Electronic Engineering (CEEC)* (IEEE), 11–14. doi: 10.1109/CEEC.2018.8674234

Alvarez-Melis, D., and Jaakkola, T. S. (2018). "Towards robust interpretability with self-explaining neural networks," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems, NIPS'18* (Red Hook, NY: Curran Associates Inc.), 7786–7795.

Arcidiacono, G. (2017). Comparative research about high failure rate of it projects and opportunities to improve. *PM World J.* VI, 1–10. Available online at: https://pmworldlibrary.net/wp-content/uploads/2017/02/pmwj55-Feb2017-Arcidiacono-high-failure-rate-of-it-projects-featured-paper.pdf

Athey, S. (2019). "The impact of machine learning on economics," in *The Economics of Artificial Intelligence: An Agenda*, eds A. Agrawal, J. Gans, and A. Goldfarb (Chicago, IL: University of Chicago Press), 507–547.

Berendt, B., and Preibusch, S. (2017). Toward accountable discrimination-aware data mining: the importance of keeping the human in the loop and under the looking glass. *Big Data* 5, 135–152. doi: 10.1089/big.2016.0055

Bhardwaj, A., Bhattacherjee, S., Chavan, A., Deshpande, A., Elmore, A. J., Madden, S., et al. (2015). "DataHub: collaborative data science and & dataset version management at scale," in *Proceedings of the 7th Biennial Conference on Innovative Data Systems Research (CIDR'15), January 4–7, 2015* (Asilomar, CA).

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. (2015). "Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '15* (New York, NY: Association for Computing Machinery), 1721–1730. doi: 10.1145/2783258.2788613

Carvalho, D. V., Pereira, E. M., and Cardoso, J. S. (2019). Machine learning interpretability: a survey on methods and metrics. *Electronics* 8:832. doi: 10.3390/electronics8080832

Catley, C., Smith, K., McGregor, C., and Tracy, M. (2009). "Extending crisp-dm to incorporate temporal data mining of multidimensional medical data streams: a neonatal intensive care unit case study," in *22nd IEEE International Symposium on Computer-Based Medical Systems*, 1–5. doi: 10.1109/CBMS.2009.5255394

Chandler, N., and Oestreich, T. (2015). *Use Analytic Business Processes to Drive Business Performance*. Technical report, Gartner.

Chavent, M., Liquet, B., Kuentz, V., and Saracco, J. (2012). Clustofvar: an R package for the clustering of variables. *J. Statist. Softw.* 50, 1–16. doi: 10.18637/jss.v050.i13

Chen, T., and Guestrin, C. (2016). "Xgboost: a scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16* (New York, NY: Association for Computing Machinery), 785–794. doi: 10.1145/2939672.2939785

Darlington, K. W. (2011). Designing for explanation in health care applications of expert systems. *SAGE Open* 1, 1–9. doi: 10.1177/2158244011408618

Davenport, T., and Kalakota, R. (2019). Digital technology: the potential for artificial intelligence in healthcare. *Future Healthc. J.* 6, 94–98. doi: 10.7861/futurehosp.6-2-94

Dawson, D., Schleiger, E., Horton, J., McLaughlin, J., Robinson, C., Quezada, G., et al. (2019). *Artificial Intelligence: Australia's Ethics Framework*. Technical report, Data61 CSIRO, Australia.

Doshi-Velez, F., and Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv* 1702.08608.

Espinosa, J. A., and Armour, F. (2016). "The big data analytics gold rush: a research framework for coordination and governance," in *Proceedings of the 49th Hawaii International Conference on System Sciences (HICSS)*, 1112–1121. doi: 10.1109/HICSS.2016.141

EU (2016). General data protection regulation (GDPR). *Off. J. Eur. Union* L 119.

Fahmy, A. F., Mohamed, H. K., and Yousef, A. H. (2017). "A data mining experimentation framework to improve six sigma projects," in *2017 13th International Computer Engineering Conference (ICENCO)*, 243–249. doi: 10.1109/ICENCO.2017.8289795

Freitas, A. A. (2014). Comprehensible classification models: a position paper. *SIGKDD Explor. Newslett.* 15, 1–10. doi: 10.1145/2594473.2594475

Fujimaki, R. (2020). *Most Data Science Projects Fail, But Yours Doesn't Have To*. Datanami. Available online at: https://www.datanami.com/2020/10/01/most-data-science-projects-fail-but-yours-doesnt-have-to/

Gao, J., Koronios, A., and Selle, S. (2015). "Towards a process view on critical success factors in big data analytics projects," in *Proceedings of the 21st Americas Conference on Information Systems (AMCIS)*, 1–14.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., and Kagal, L. (2018). "Explaining explanations: an overview of interpretability of machine learning," in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 80–89. doi: 10.1109/DSAA.2018.00018

Gilpin, L. H., Testart, C., Fruchter, N., and Adebayo, J. (2019). Explaining explanations to society. *CoRR* abs/1901.06560.

Gleicher, M. (2016). A framework for considering comprehensibility in modeling. *Big Data* 4, 75–88. doi: 10.1089/big.2016.0007

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

Goodson, M. (2016). *Reasons Why Data Projects Fail*. KDnuggets.

Goodwin, B. (2011). *Poor Communication to Blame for Business Intelligence Failure, Says Gartner*. Computer Weekly.

Google (2019). *Google AI: Responsible AI Practices–Interpretability*. Technical report, Google AI.

Gosiewska, A., and Biecek, P. (2020). Do not trust additive explanations. *arXiv* 1903.11420.

Gosiewska, A., Woznica, K., and Biecek, P. (2020). Interpretable meta-measure for model performance. *arXiv* 2006.02293.

Grady, N. W., Underwood, M., Roy, A., and Chang, W. L. (2014). "Big data: challenges, practices and technologies: NIST big data public working group workshop at IEEE big data 2014," in *Proceedings of IEEE International Conference on Big Data (Big Data 2014)*, 11–15. doi: 10.1109/BigData.2014.7004470

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Comput. Surv.* 51, 93:1–93:42. doi: 10.1145/3236009

Hansen, L. K., and Rieger, L. (2019). "Interpretability in intelligent systems–a new concept?" in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Volume 11700 of LNAI*, eds W. Samek, G. Montvon, A. Vedaldi, L. K. Hansen, and K. R. Müller (Springer Nature), 41–49. doi: 10.1007/978-3-030-28954-6_3

Holzinger, A., Biemann, C., Pattichis, C. S., and Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain? *arXiv* 1712.09923.

Huang, W., McGregor, C., and James, A. (2014). "A comprehensive framework design for continuous quality improvement within the neonatal intensive care unit: integration of the SPOE, CRISP-DM and PaJMa models," in *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, 289–292. doi: 10.1109/BHI.2014.6864360

IBM Analytics (2015). *IBM Analytics Solutions Unified Method (ASUM)*. Available online at: http://i2t.icesi.edu.co/ASUM-DM_External/index.htm

Jain, P. (2019). *Top 5 Reasons for Data Science Project Failure*. Medium: Data Series.

Kaggle (2020). *State of Machine Learning and Data Science 2020 Survey*. Technical report, Kaggle. Available online at: https://www.kaggle.com/c/kaggle-survey-2020

Kennedy, P., Simoff, S. J., Catchpoole, D. R., Skillicorn, D. B., Ubaudi, F., and Al-Oqaily, A. (2008). "Integrative visual data mining of biomedical data: investigating cases in chronic fatigue syndrome and acute lymphoblastic leukaemia," in *Visual Data Mining: Theory, Techniqus and Tools for Visual Analytics, Volume 4404 of LNCS*, eds S. J. Simoff, M. H. Böhlen, and A. Mazeika (Berlin; Heidelberg:Springer), 367–388. doi: 10.1007/978-3-540-71080-6_21

Kolyshkina, I., and Simoff, S. (2019). "Interpretability of machine learning solutions in industrial decision engineering," in *Data Mining*, eds T. D. Le, K. L. Ong, Y. Zhao, W. H. Jin, S. Wong, L. Liu, et al. (Singapore: Springer Singapore), 156–170. doi: 10.1007/978-981-15-1699-3_13

Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S. J., et al. (2019). "Human evaluation of models built for interpretability," in *The Proceedings of the Seventh AAAI Conference on Human Computation and Crowdsourcing (HCOMP-19)*, Vol. 7, 59–67.

Larson, D., and Chang, V. (2016). A review and future direction of agile, business intelligence, analytics and data science. *Int. J. Inform. Manage.* 36, 700–710. doi: 10.1016/j.ijinfomgt.2016.04.013

Lipton, Z. C. (2018). The mythos of model interpretability. *ACM Queue* 16, 30:31–30:57. doi: 10.1145/3236386.3241340

Lundberg, S. M., and Lee, S. I. (2017). "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30*, eds I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, et al. (Curran Associates, Inc.), 4765–4774.

Mariscal, G., Marbán, O., and Fernández, C. (2010). A survey of data mining and knowledge discovery process models and methodologies. *Knowl. Eng. Rev.* 25, 137–166. doi: 10.1017/S0269888910000032

Mi, J. X., Li, A. D., and Zhou, L. F. (2020). Review study of interpretation methods for future interpretable machine learning. *IEEE Access* 8, 191969–191985. doi: 10.1109/ACCESS.2020.3032756

Microsoft (2020). *Team Data Science Process*. Microsoft.

Miller, T. (2019). Explanation in artificial intelligence: insights from the social sciences. *Artif. Intell.* 267, 1–38. doi: 10.1016/j.artint.2018.07.007

Mittelstadt, B., Russell, C., and Wachter, S. (2019). "Explaining explanations in AI," in *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT '19* (New York, NY: Association for Computing Machinery), 279–288. doi: 10.1145/3287560.3287574

Molnar, C., Casalicchio, G., and Bischl, B. (2019). Quantifying interpretability of arbitrary machine learning models through functional decomposition. *arXiv* 1904.03867.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci. U.S.A.* 116, 22071–22080. doi: 10.1073/pnas.1900654116

NewVantage Partners LLC (2021). *Big Data and AI Executive Survey 2021: The Journey to Becoming Data-Driven: A Progress Report on the State of Corporate Data Initiatives*. Technical report. Boston, MA; New York, NY; San Francisco, CA; Raleigh, NC: NewVantagePartners LLC.

Niaksu, O. (2015). Crisp data mining methodology extension for medical domain. *Baltic J. Mod. Comput.* 3, 92–109. Available online at: https://www.bjmc.lu.lv/fileadmin/user_upload/lu_portal/projekti/bjmc/Contents/3_2_2_Niaksu.pdf

Niño, M., Blanco, J. M., and Illarramendi, A. (2015). "Business understanding, challenges and issues of Big Data Analytics for the servitization of a capital equipment manufacturer," in *2015 IEEE International Conference on Big Data, Oct 29–Nov 01, 2015*, eds H. Ho, B. C. Ooi, M. J. Zaki, X. Hu, L. Haas, V. Kumar, et al. (Santa Clara, CA), 1368–1377.

Piatetsky-Shapiro, G. (2014). CRISP-DM, still the top methodology for analytics, data mining, or data science projects. *KDnuggets News* 14.

Plotnikova, V. (2018). "Towards a data mining methodology for the banking domain," in *Proceedings of the Doctoral Consortium Papers Presented at the 30th International Conference on Advanced Information Systems Engineering (CAiSE 2018)*, eds M. Kirikova, A. Lupeikiene, and E. Teniente, 46–54.

PMI (2017). *PMBOK® Guide, 6th Edn*. Project Management Institute.

Pradeep, S., and Kallimani, J. S. (2017). "A survey on various challenges and aspects in handling big data," in *Proceedings of the 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT)*, 1–5. doi: 10.1109/ICEECCOT.2017.8284606

Qayyum, A., Qadir, J., Bilal, M., and Al-Fuqaha, A. (2021). Secure and robust machine learning for healthcare: a survey. *IEEE Rev. Biomed. Eng*. 14, 156–180. doi: 10.1109/RBME.2020.3013489

Ransbotham, S., Kiron, D., and Prentice, P. K. (2015). *Minding the Analytics Gap*. MIT Sloan Management Review.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should I trust you?": explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16)* (New York, NY: ACM), 1135–1144. doi: 10.1145/2939672.2939778

Roberts, J. (2017). *4 Reasons Why Most Data Science Projects Fail*. CIO Dive.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell*. 1, 206–215. doi: 10.1038/s42256-019-0048-x

Saltz, J. S. (2015). "The need for new processes, methodologies and tools to support big data teams and improve big data project effectiveness," in *Proceedings of 2015 IEEE International Conference on Big Data (Big Data)*, 2066–2071. doi: 10.1109/BigData.2015.7363988

Saltz, J. S., and Shamshurin, I. (2016). "Big data team process methodologies: a literature review and the identification of key factors for a project's success," in *Proceedings of 2016 IEEE International Conference on Big Data (Big Data)*, 2872–2879. doi: 10.1109/BigData.2016.7840936

Saltz, J. S., Shamshurin, I., and Crowston, K. (2017). "Comapring data science project management methodologies via a controlled experiment," in *HICSS*. doi: 10.24251/HICSS.2017.120

Samek, W., and Müller, K. R. (2019). "Towards explainable artificial intelligence," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Volume 11700 of LNAI*, eds W. Samek, G. Montvon, A. Vedaldi, L. K. Hansen, and K. R. Müller (Springer Nature), 5–22. doi: 10.1007/978-3-030-28954-6_1

Schäfer, F., Zeiselmair, C., Becker, J., and Otten, H. (2018). "Synthesizing CRISP-DM and quality management: a data mining approach for production processes," in *2018 IEEE International Conference on Technology Management, Operations and Decisions (ICTMOD)*, 190–195. doi: 10.1109/ITMC.2018.8691266

Shearer, C. (2000). The CRISP-DM Model: The new blueprint for data mining. *J. Data Warehousing* 5, 13–22. Available online at: https://mineracaodedados.files.wordpress.com/2012/04/the-crisp-dm-model-the-new-blueprint-for-data-mining-shearer-colin.pdf

Shi, T., and Horvath, S. (2006). Unsupervised learning with random forest predictors. *J. Comput. Graph. Stat*. 15, 118–138. doi: 10.1198/106186006X94072

Stieglitz, C. (2012). "Beginning at the end-requirements gathering lessons from a flowchart junkie," in *PMI® Global Congress 2012–North America, Vancouver, British Columbia, Canada* (Newtown Square, PA: Project Management Institute). Available online at: https://www.pmi.org/learning/library/requirements-gathering-lessons-flowchart-junkie-5981

Stiglic, G., Kocbek, P., Fijacko, N., Zitnik, M., Verbert, K., and Cilar, L. (2020). Interpretability of machine learning-based prediction models in healthcare. *WIREs Data Mining Knowl. Discov*. 10, 1–13. doi: 10.1002/widm.1379

Studer, S., Bui, T. B., Drescher, C., Hanuschkin, A., Winkler, L., Peters, S., et al. (2020). Towards CRISP-ML(Q): a machine learning process model with quality assurance methodology. *arXiv* 2003.05155.

Sun, W., Nasraoui, O., and Shafto, P. (2020). Evolution and impact of bias in human and machine learning algorithm interaction. *PLoS ONE* 15:e0235502. doi: 10.1371/journal.pone.0235502

vander Meulen R, and Thomas, M. (2018). *Gartner Survey Shows Organizations Are Slow to Advance in Data and Analytics*. Gartner Press Release.

Vellido, A. (2020). The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Comput. Appl*. 32, 18069–18083. doi: 10.1007/s00521-019-04051-w

Violino, B. (2017). *7 Sure-Fire Ways to Fail at Data Analytics*. CIO.

Wallace, N., and Castro, D. (2018). *The Impact of the EU's New Data Protection Regulation on AI*. Technical report, Center for Data Innovation.

Weller, A. (2019). "Transparency: motivations and challenges," in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Volume 11700 of LNAI*, eds W. Samek, G. Montvon, A. Vedaldi, L. K. Hansen, and K. R. Müller (Springer Nature), 23–40. doi: 10.1007/978-3-030-28954-6_2

Conflict of Interest: IK was employed by the company Analytikk Consulting Services.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Virtual Reality

SYSTEMATIC REVIEW
published: 06 July 2021
doi: 10.3389/frvir.2021.692103

HMD-Based Virtual and Augmented Reality in Medical Education: A Systematic Review

Xuanhui Xu, Eleni Mangina and Abraham G. Campbell*

School of Computer Science, University College Dublin, Dublin, Ireland

Background: Virtual Reality (VR) and Augmented Reality (AR) technologies provide a novel experiential learning environment that can revolutionize medical education. These technologies have limitless potential as they provide in effect an infinite number of anatomical models to aid in foundational medical education. The 3D teaching models used within these environments are generated from medical data such as magnetic resonance imaging (MRI) or computed tomography (CT), which can be dissected and regenerated without limitations.

Methods: A systematic review was carried out for existing articles until February 11, 2020, in EMBASE, PubMed, Scopus, ProQuest, Cochrane Reviews, CNKI, and OneSearch (University College Dublin Library) using the following search terms: (Virtual Reality OR Augmented Reality OR mixed reality) AND ["head-mounted" OR "face-mounted" OR "helmet-mounted" OR "head-worn" OR oculus OR vive OR HTC OR hololens OR "smart glasses" OR headset AND (training OR teaching OR education)] AND (anatomy OR anatomical OR medicine OR medical OR clinic OR clinical OR surgery OR surgeon OR surgical) AND (trial OR experiment OR study OR randomized OR randomised OR controlled OR control) NOT (rehabilitation OR recovery OR treatment) NOT ("systematic review" OR "review of literature" OR "literature review"). PRISMA guidelines were adhered to in reporting the results. All studies that examined people who are or were medical-related (novel or expert users) were included.

Result: The electronic searches generated a total of 1,241 studies. After removing duplicates, 848 remained. Of those, 801 studies were excluded because the studies did not meet the criteria after reviewing the abstract. The full text of the remaining 47 studies was reviewed. After applying inclusion criteria and exclusion criteria, a total of 17 studies (1,050 participants) were identified for inclusion in the review.

Conclusion: The systematic review provides the current state of the art on head-mounted device applications in medical education. Moreover, the study discusses trends toward the future and directions for further research in head-mounted VR and AR for medical education.

Keywords: virtual reality, augmented reality, head-mounted display, surgery, medicine, systematic review, education

OPEN ACCESS

Edited by:
Philip Breedon, Nottingham Trent University, United Kingdom

Reviewed by:
Deborah Richards, Macquarie University, Australia
Hongchuan Yu, Bournemouth University, United Kingdom

****Correspondence:***
Xuanhui Xu
xuanhui.xu@ucdconnect.ie

Specialty section:
This article was submitted to Virtual Reality in Medicine, a section of the journal Frontiers in Virtual Reality

Received: *07 April 2021*
Accepted: *04 June 2021*
Published: *06 July 2021*

Citation:
Xu X, Mangina E and Campbell AG (2021) HMD-Based Virtual and Augmented Reality in Medical Education: A Systematic Review. Front. Virtual Real. 2:692103. doi: 10.3389/frvir.2021.692103

1 INTRODUCTION

Retaining knowledge in education is challenging. Medical students learn complex structures and anatomy mainly based on teaching material such as books or pictures traditionally, and for some educational institutions with more resources, students may have a chance to dissect actual cadavers. However, the paper-based learning material might cause misunderstandings as it is hard to imagine the 3D relationship between components based on 2D materials. Teaching resources such as real-life cadavers are limited and critically have strict storage restrictions based on health and safety rules.

Therefore, Virtual Reality (VR) and Augmented Reality (AR) interventions based on simulations could offer a possible solution or, at the very least, ease this bottleneck in medical education. They would improve spatial awareness when compared to 2D teaching materials and provide infinite teaching materials that can be the foundation for advancing the accessibility of content for medical education. VR and AR technologies provide a close-to-reality experience for users in industry, education, and gaming. Among all VR/AR formats, the head-mounted display (HMD) provides the most immersive environment, tracking a user's motion and maintains the position of spatial information around them. The 3D teaching models can be generated from medical data like magnetic resonance imaging (MRI) or computed tomography (CT), which allow them to be dissected and regenerated without any limits.

Simulators in general are widely used in medical education and assessment. To date, several systematic reviews have investigated the efficacy of VR simulation training in laparoscopy (Larsen et al., 2012; Alaker et al., 2016). The results showed that VR laparoscopy simulation provided an effective and ethical way to train residents' surgical skills. VR simulation can play an important role in addressing the issue of low training efficiency. However, the main VR simulators used as interventions were LapSim®[1] and Simendo®[2], which did not utilize HMDs. Based on the current authors' knowledge, another systematic review in the usage of HMD-based VR and AR in medical education does not exist. Previous studies have identified some situations where HMDs are suitable for skill acquisition (Jensen and Konradsen, 2018), including cognitive skills related to remembering and understanding spatial information and knowledge. As learning a kinesthetic-based medical skill highly relies on spatial cognition, the immersion provided by an HMD logically then becomes a natural requirement for this review, to explore if VR and AR may potentially benefit medical skill acquisition.

This study focuses on a systematic review to evaluate the effectiveness of applying HMD for VR or AR applications in medical education that can benefit medical training. To this end, the systematic review will answer the following questions that are proposed in the protocol (Section 1.1):

- Compared with the standard teaching method or other types of simulators, what are the comparative effectiveness of HMD VR or AR usage in medical teaching? (Advantages)
- What are the disadvantages of using HMD VR or AR, and which one has lower side effects? (Disadvantages)
- Is there a definitive advantage of HMD VR and AR when used for increasing the efficiency of teaching in medicine? (Proof)
- Do HMD VR and AR have the potential to be support tools for medical education? (Support)

1.1 Protocol

A systematic review was carried out until February 11, 2020. PRISMA guidelines were adhered to in reporting the results of this study (Moher et al., 2009). Methods of the analysis and inclusion criteria were specified in advance and documented in a protocol. The protocol has been registered in PROSPERO, the international prospective register of systematic reviews, where it can be accessed (Registration number: CRD42020165310)[3].

1.2 Search Strategy

The literature search and initial screening were conducted by XX; abstract screening was conducted independently by three authors (XX, EM, and AC), while the disagreement was confirmed by discussion; full article screening and data extraction were conducted by XX. Databases searched were EMBASE, PubMed, Scopus, ProQuest, Cochrane Reviews, CNKI, and OneSearch (University College Dublin Library) on title, abstract, and keywords; searches from Google Scholar are also acceptable. The terms used for searching were as follows: (Virtual Reality OR Augmented Reality OR mixed reality) AND ["head-mounted" OR "face-mounted" OR "helmet-mounted" OR "head-worn" OR oculus OR vive OR HTC OR hololens OR "smart glasses" OR headset AND (training OR teaching OR education)] AND (anatomy OR anatomical OR medicine OR medical OR clinic OR clinical OR surgery OR surgeon OR surgical) AND (trial OR experiment OR study OR randomized OR randomised OR controlled OR control) NOT (rehabilitation OR recovery OR treatment) NOT ("systematic review" OR "review of literature" OR "literature review").

1.3 Selection Criteria

All studies examining the general adult human population or healthy adult humans and people who are or were medical-related (novel or experts) were included. Studies in which individuals were selected with extreme motion sickness, other diagnosed illness, or disability and studies in which individuals were not medical-related are excluded. No year publication limits were set. English and Chinese text publications were included as

[1]https://surgicalscience.com/systems/lapsim/

[2]https://www.simendo.eu/

[3]Xu, X, Abraham, G. C., and Eleni, M. Can the head-mounted device improve medical education quality? Protocol for a systematic review. PROSPERO 2020 CRD42020165310 Available from https://www.crd.york.ac.uk/prospero/display_record.php?ID=CRD42020165310

one author was a native Chinese speaker, which allowed a unique chance to expand the search. The search was last updated on February 11, 2020. The titles and abstracts database searches were screened to identify potentially relevant records for full-text screening. The titles and abstracts of all remaining records were screened for eligibility to identify records for full-text screening. All records identified for full-text screening were screened to identify records for inclusion in the review. All data that were potentially relevant to the review were then extracted from the studies selected for final inclusion and collated in a spreadsheet as follows: details of publication, participant characteristics, sample size, setting, intervention, study design, data type, and result.

A meta-analysis was not undertaken due to the considerable heterogeneity among the studies included in this review. Therefore, a descriptive approach to data synthesis was adapted, whereby summaries of included studies will be presented. Included studies will be presented in line with the outcomes identified from active interventions that involve HMD VR or AR, specifically, changes in surgery or anatomy training or outcomes related to the trainer or trainee's experience (satisfaction and motivation), population characteristics (study design and study outcome measures), and methodological approach (randomized control studies and crossover studies).

Data were extracted using a standardized template to capture information relating to PICO: population (grade and sex), intervention (characteristics of the HMD VR and AR tool), comparator (traditional/other teaching methods), and outcomes (assessment score, time, and subjective feeling).

1.4 Quality Assessment

The study's quality information was collected using the RoB 2.0 tool (Sterne et al., 2019) for assessing the risk of bias. The risk of bias plot was created by using the Robvis tool (McGuinness and Higgins, 2021).

2 RESULTS

2.1 Study Selection

The electronic searches generated a total of 1,241 studies. After removing duplicates, 848 remained. Of these, 801 studies were excluded because the studies did not meet the criteria after reviewing the abstract. The full text of the remaining 47 studies was reviewed. Among those, 30 studies were discarded due to the reasons in the flowchart (**Figure 1**). After applying the inclusion criteria and exclusion criteria, 17 studies were identified for inclusion in the review (**Table 1**). No unpublished relevant studies were obtained. **Figure 2** shows what year the screened and included studies were published, illustrating the increasing interest in the topic over the last number of years.

2.2 Study Characteristics

2.2.1 Methods

Among all studies selected, 15 were randomized controlled trials (RCTs), nine of which were published in English (Coulter et al., 2007; Stepan et al., 2017; Pulijala et al., 2018; Alismail et al., 2019; Logishetty et al., 2019a; Logishetty et al., 2019b; Rojas-Muñoz et al., 2019; Frederiksen et al., 2020; Zackoff et al., 2020) and seven of which were published in Chinese (Meng et al., 2018; Cai et al., 2019; Chen et al., 2019; Jiang et al., 2019; Wang H. et al., 2019; Wang P. et al., 2019). One was a randomized single-blinded crossover study (Harrington et al., 2018) in English, and one was a six-week pre-post comparison study (Logishetty et al., 2020) in English.

2.2.2 Participants

The included studies involved 1,050 participants and 978 of those participated in RCTs (**Figure 3**). The main inclusion criteria entailed medical students (first-year to master students), surgical trainees, and nursing interns.

2.2.3 Intervention

The interventions applied in the studies were VR headsets or AR headsets. A total of 11 studies used VR HMD as interventions, including Oculus Rift (n = 4), HTC VIVE (n = 4), Gear VR (n = 2), and customized HMD (n = 1). The rest of the six studies used AR HMD as interventions, including HoloLens (n = 5) and AiRScouter glasses (n = 1).

2.2.4 Outcomes

Out of 15 RCTs studies, four studies compared the VR/AR HMD method with the traditional teaching method, such as paper materials, 2D videos, and slides (Pulijala et al., 2018; Cai et al., 2019; Jiang et al., 2019; Wang H. et al., 2019). Five studies compared VR/AR HMD with other teaching methods such as desktop and real person guide (Coulter et al., 2007; Logishetty et al., 2019a; Logishetty et al., 2019b; Rojas-Muñoz et al., 2019; Wang H. et al., 2019). Five studies used VR/AR HMD as additional training to support education (Stepan et al., 2017; Meng et al., 2018; Alismail et al., 2019; Chen et al., 2019; Zackoff et al., 2020). One study used HMD to increase the immersive feeling (Frederiksen et al., 2020).

2.2.4.1 Compared with Traditional Method

Four studies compared VR/AR intervention with traditional teaching methods (**Table 2**). Pulijala et al. (2018) designed an RCT (n = 91) to be able to compare immersive VR training with traditional teaching. The study group used Oculus Rift with Leap Motion tracker to interact with the anatomy data and viewed 360° videos of an operating room, while the control used a standard PowerPoint presentation and viewed 2D video of similar content. The knowledge gained was significantly increased in scores for both the VR group (p = 0.024) and the control group (p = 0.025); however, the participants who used the VR headset performed better overall, especially for the early stage (first-year and second-year) residents. This is common in AR/VR training, where it has been found that the nonexperts appear to benefit the most from the experience (Pringle et al. 2018). Another example comes from Cai et al. (2019) who conducted a similar controlled study (n = 50). The study and control groups were given theoretical training using the virtual 3D model generated from CT and MRI scans simultaneously. The intervention was then applied to the study group, where they used the HTC VIVE VR headset to watch real

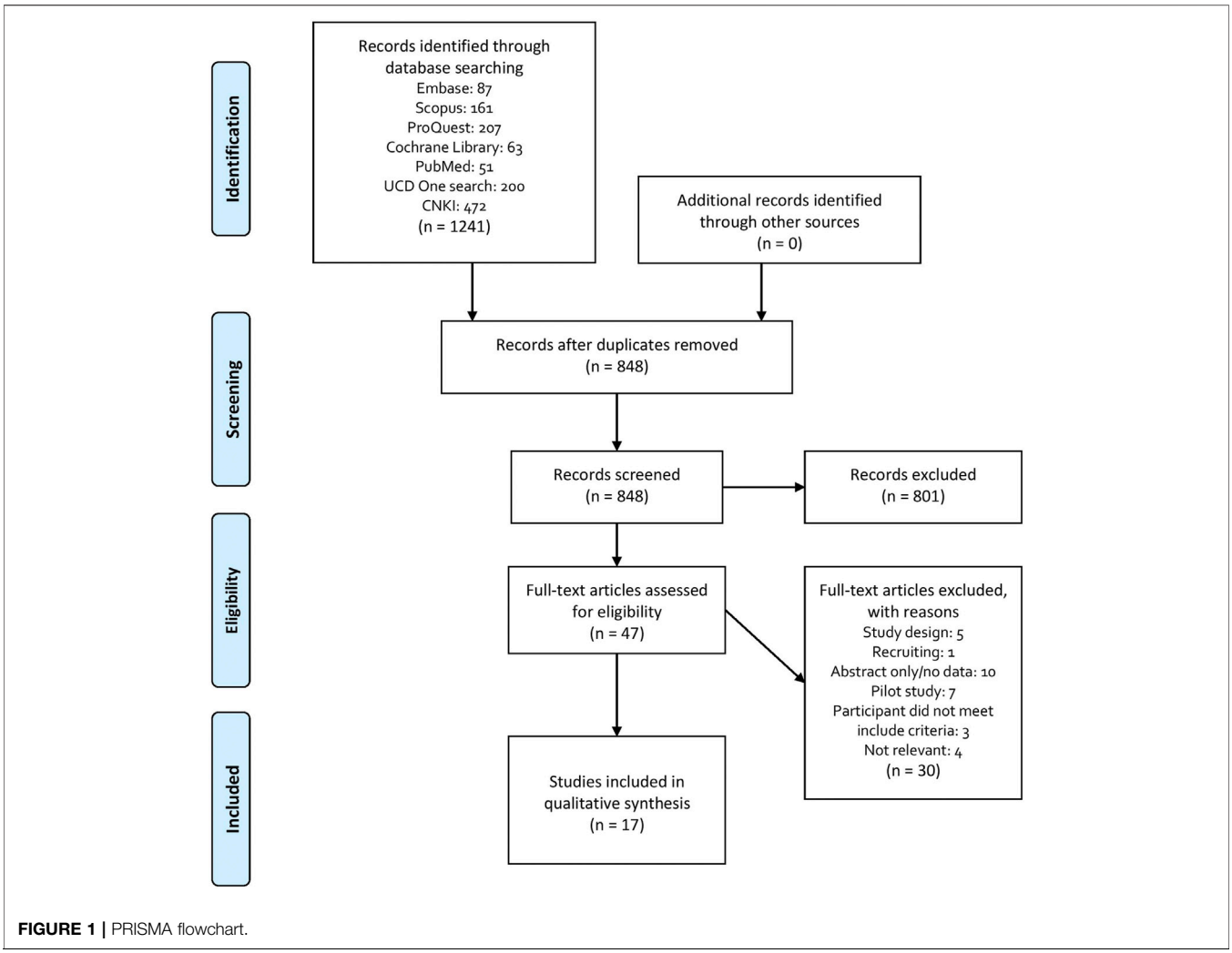

FIGURE 1 | PRISMA flowchart.

operation 360° video to learn the anatomy and operation process, while the control group learned from presentation slides, anatomy pictures, models, and 2D videos explained by the lecturers. The control group also entered the operation room to observe the operation process. The results showed a significant difference in test score between the intervention and the control groups ($p = 0.023$).

Jiang et al. (2019) designed an RCT ($n = 52$) to evaluate the effect of the application of mixed reality technology in teaching spine and spinal cord injury. They developed a mixed reality teaching model with real patient case's MRI 3D reconstruction. The lecturer equipping HoloLens demonstrated the operation process to the study group, and the teaching content was delivered through a monitor. Students in the study group used HoloLens. They learned the process in a simulated environment with all virtual content being synchronized on all headsets, while the control group was taught through the traditional method, including slides and paper teaching materials. The posttest result did not show a significant difference in score between the two groups ($p > 0.05$), but the participants in the study group had a better understanding of the 3D structure ($p < 0.01$). By utilizing the same approach, Wang P. et al. (2019) conducted an RCT experiment ($n = 120$) to explore the effect of this technology in hepatobiliary surgery. Theoretical and surgical operation assessment showed a significant difference in score between the study and control groups, which was different from the previous study ($p < 0.05$). The study group's error rate was significantly lower than the control group ($p < 0.05$).

2.2.4.2 Compared with Nontraditional Method

Five studies compared VR/AR HMD with other teaching methods, such as customized simulators or an expert one-to-one guide (**Table 3**). The earliest randomized controlled study ($n = 25$) explored HMD's effect on medical education learning performance was conducted back in 2007 (Coulter et al., 2007). The study group wore a stereoscopic HMD as a fully immersed system, while the control group used a simulation *via* a PC monitor as a partially immersed system. The result showed significant difference in the pre/posttest in overall ($p < 0.001$), study group ($p < 0.001$), and control group ($p = 0.024$). The study group showed a higher gain than the control group. Logishetty et al. (2019a) conducted an RCT ($n = 24$) to determine that the

TABLE 1 | General data of 17 included studies.

Author	Sample size	Intervention	Study design	Result
Frederiksen et al. (2020)	First-year student n = 31, intervention n = 16, control n = 15	Oculus Rift, 360° video	Randomized control trial	Control: not capable Both groups' total time ($p < 0.001$) and blood loss ($p < 0.001$); cognitive load increase (study group: 43.1%, control group: 23%, $p < 0.001$); motion sickness ($p = 0.62$)
Logishetty et al. (2019a)	Postgraduate student n = 24, intervention n = 12, control n = 12	HTC VIVE	Randomized control trial	Control: conventional preparation materials Procedure-based assessment (PBA) level (study group level 3b, control group level 2a, $p < 0.001$); PBA satisfactory ($p < 0.001$); task-specific checklist ($p < 0.001$); inclination and anteversion error from target ($p < 0.001$); operative time ($p = 0.005$)
Logishetty et al. (2019b)	Final-year medical student n = 24, intervention n = 12, control n = 12	MicronTracker HoloLens	Randomized simulation trial	Control: surgeon guide Study group improvement $p < 0.001$; control group improvement $p < 0.001$; improvement between groups $p = 0.281$
Pulijala et al. (2018)	Master student n = 91, intervention n = 51, control n = 40	Oculus Rift with Leap Motion	Multicentre parallel single-blind randomized controlled trial	Control: traditional teaching method Self-confidence level $p = 0.034$; knowledge improvement (study group performed better, $p > 0.05$)
Stepan et al. (2017)	Medical students n = 66, intervention n = 33, control n = 33	Oculus Rift	Randomized controlled study	Control: no VR as additional method post intervention, or retention quizzes assessments ($p > 0.05$); engaging, enjoyable, and useful (all $p < 0.01$); motivation assessment ($p < 0.01$)
Zackoff et al. (2020)	Third-year medical student n = 168, intervention n = 78, control n = 90	Oculus Rift	Randomized controlled prospective study	Control: no VR as additional method Consideration/interpretation of mental status ($p < 0.01$); assignment of the appropriate respiratory status assessment ($p < 0.01$); recognition of a need for escalation of care ($p = 0.0004$)
Rojas-Muñoz et al. (2019)	Medical students n = 20	HoloLens, Kinect	User study STAR AR HMD vs. conventional telestrator	Control: conventional telestrator Place error (task 1, $p < 0.001$; task 2, $p = 0.01$); time (study group longer, task 1, $p < 0.001$; study group longer, task 2, $p = 0.013$); focus shifts (task 1, $p < 0.001$; task 2, $p = 0.0038$)
Alismail et al. (2019)	Healthcare medical center n = 32, intervention n = 15, control n = 17	AiRScouter WD-200B glasses	Randomized study	Control: no AR as additional method time for ventilation (study group longer, $p = 0.005$); per cent adherence to the intubation checklist ($p < 0.001$)
Coulter et al. (2007)	n = 25, fully immersive n = 13, partially immersive n = 12	Customized HMD	Randomized study	Control: PC Time ($p = 0.004$); both groups' score pre-post ($p < 0.001$, study group: $p < 0.001$, control group: $p = 0.024$)
Jiang et al. (2019)	Sophomore undergraduate n = 52, intervention n = 26, control n = 26	HoloLens	Randomized study	Control: traditional teaching method Three-dimensional construction and enhancement of class atmosphere ($p < 0.01$); class satisfaction and initiative of learning ($p < 0.05$)
Wang H. et al. (2019)	Nursing student n = 125, intervention n = 62, control n = 63	HTC VIVE	Randomized study	Control: intravenous injection simulator Both groups' posttest scores ($p < 0.05$); knowledge improvement (study group performed better, $p > 0.05$); critical thinking and clinical reasoning and clinical learning ($p < 0.05$)
Wang P. et al. (2019)	Surgery postgraduate n = 120, intervention n = 60, control n = 60	HoloLens	Randomized study	Control: traditional teaching method Theoretical examination and surgical operation assessment ($p < 0.05$); error rate ($p < 0.05$); satisfaction ($p < 0.05$)
Cai et al. (2019)	Clinical medicine student n = 50, intervention n = 25, control n = 25	HTC VIVE	Randomized study	Control: traditional teaching method Average score ($p < 0.05$); satisfaction ($p < 0.05$)
Chen et al. (2019)	Neurosurgery interns n = 80, intervention n = 40, control n = 40	HoloLens	Randomized study	Control: no AR as additional method Mastery degree of lateral ventricle anatomy ($p < 0.05$); proficiency in puncture procedure ($p < 0.05$); confidence ($p < 0.05$); first-pass success rate of puncture ($p < 0.05$)
Meng et al. (2019)	Undergraduate intern n = 70, intervention n = 35, control n = 35	Gear VR	Randomized study	Control: no AR as additional method Number of patients ($p > 0.05$); theoretical exam ($p < 0.05$); satisfaction ($p > 0.05$)

(Continued on following page)

TABLE 1 | (*Continued*) General data of 17 included studies.

Author	Sample size	Intervention	Study design	Result
Harrington et al. (2018)	Undergraduate n = 40, group A n = 20, group B n = 20	Gear VR	Randomized single-blinded crossover study	Higher engagement level ($p < 0.0001$); across time periods ($p < 0.0007$); lower TUIT ($p < 0.0001$); across time periods ($p < 0.0005$)
Logishetty et al. (2020)	Orthopedic residents n = 32	HTC VIVE	6-week pre-post comparison with expert performance	Reached expert levels 9 of 10 metrics; procedural errors reduced by 79%; assistive prompts reduced by 70%; procedural duration reduced by 28%

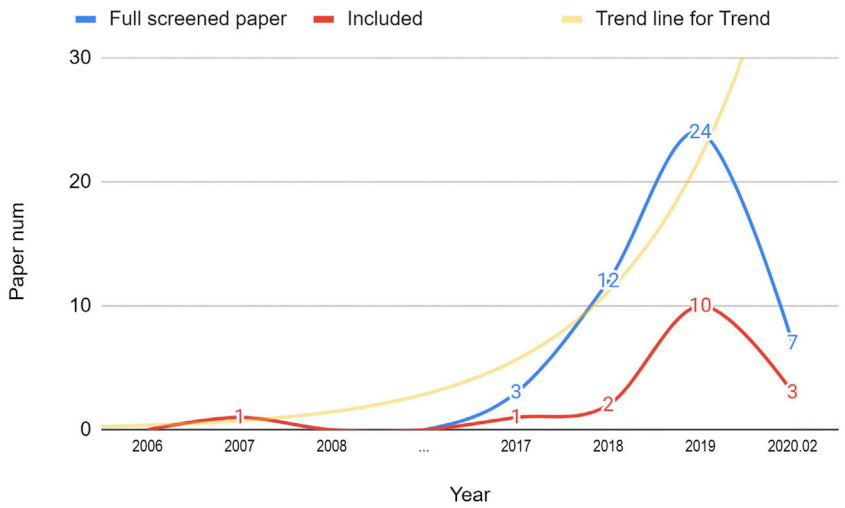

FIGURE 2 | Number of screened and included studies published in the given years.

training effectiveness of using a VR headset was higher than conventional preparation for performing total hip arthroplasty (THA). All participants received standard guidelines and materials to ensure they had similar basic knowledge before the experiment. The study group was enrolled in a six-week VR training program equipped with the HTC VIVE VR headset, while the conventional group received only given preparatory material. The PBA component score and the task-specific checklist score were significantly higher in the VR group than in the control group ($p < 0.001$), which indicated that VR-trained surgeons performed at a higher level than controls. Moreover, the VR group performed faster to complete the procedure ($p = 0.03$) and was more accurate in component orientation (mean error 4° vs. 16°). Another randomized controlled trial was conducted (n = 125) by Wang P. et al. (2019). All participants were trained under the teaching mode of "real person training + model assistance + virtual reality." The study group used the HTC VIVE VR headset as an immersive VR training method, while the control group used an intravenous injection simulator as a nonimmersive VR training method. In the theory test, both study and control groups significantly improved scores ($p < 0.001$). The study group had a higher mean score, but there was no significant difference between groups ($p = 0.136$). However, in the injection test, the study group had significantly higher scores ($p = 0.027$), demonstrating that the immersive VR training method has similar teaching effectiveness to the customized training tool.

Logishetty et al. (2019b) developed an enhanced AR headset capable of tracking bony anatomy in relation to an implant and designed a randomized trial (n = 24) to assess the suitability of it as a training tool for implant orientation. Both groups had standard lectures before the experiment started. During the experiment, both groups had four training sessions, between which there was a 5- to 9-day interval. In each session, the study group used the HoloLens AR headset, while the control group had an expert surgeon who guided the training. The participants in the study group had a significantly lower error of target orientation (1° ± 1°) than those in the control group (6° ± 4°) as they confirmed the final orientation when the headset light turned from red to green ($p < 0.001$). The result showed significant improvements in both groups when compared the final assessment score with the pretest score correspondingly. There was no significant difference in accuracy between the two groups in the assessment ($p = 0.281$) and concealed the pelvic tilt test ($p = 0.301$). 11 of 12 participants stated that they would use the AR platform as a training tool for developing visuospatial skills and 10 of 12 for procedure-specific rehearsals. Most participants (11 of 12) stated that a combination of an expert trainer for learning and AR for unsupervised training would be

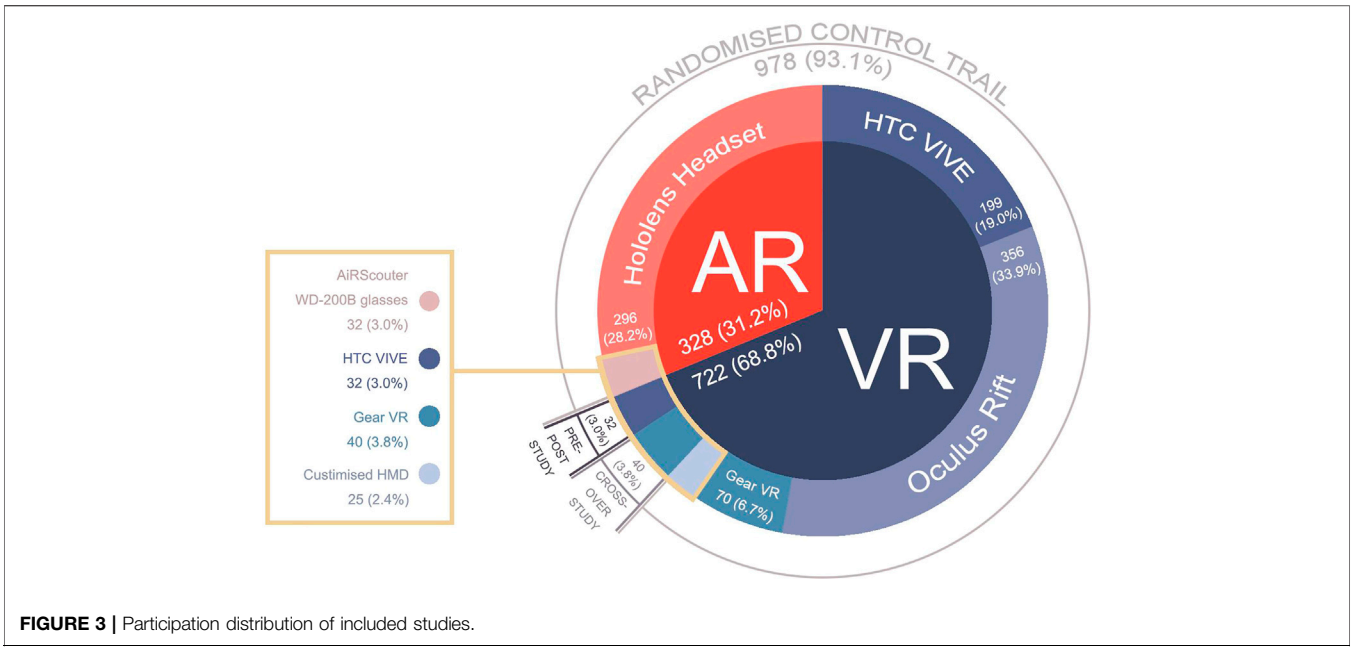

FIGURE 3 | Participation distribution of included studies.

TABLE 2 | Compared with the traditional method.

Category	Intervention	Authors	*n*	Results
Theoretical assessment	VR	Pulijiala et al. (2018)	91	Study group $p = 0.024$ Control group $p = 0.025$ Between groups $p > 0.05$
		Cai et al. (2019)	50	Between groups (study group performed better, $p = 0.023$
	AR	Jiang et al. (2019)	52	Between groups $p > 0.05$
		Wang H. et al. (2019)	120	Between groups (study group performed better, $p < 0.05$)
Surgical operation assessment				Between groups (study group performed better, $p < 0.05$)
Error rate				Between groups (study group performed better, $p < 0.05$)

preferred. This demonstrates an interesting result where AR learning as a self-evaluation tool could prove useful in the future.

Rojas-Muñoz et al. (2019) investigated the benefits of an AR HMD telementoring system when compared with a conventional telestrator in surgical guidance by conducting a comparative study ($n = 20$). The study group used the HoloLens AR HMD to receive the instructions during the operation, while the control group needed to watch on a nearby screen. The result showed significant differences between the study group and the control group in placement errors (Task 1: $p < 0.001$; Task 2: $p = 0.01$) and focus shifts (Task 1: $p < 0.001$; Task 2: $p = 0.0039$). In general, the study group used more time to complete each task. It was reported in this study that the use of the HMD caused discomfort, which can be very common with older three degrees of freedom (3DoF) HMDs and is still an issue with the current six degrees of freedom (6DoF) AR/VR HMDs but appears to be steadily improving with every generation.

2.2.4.3 Supportive Usage

Five studies used VR/AR HMD as additional training to support education (**Table 4**). Zackoff et al. (2020) conducted a randomized controlled prospective study ($n = 168$) to determine whether exposure to immersive VR-simulated pediatric respiratory distress environment improves students' emergency recognition. All participants received the standard curriculum with a subsequent high-fidelity mannequin simulation, while the study group underwent an additional VR curriculum using the Oculus Rift VR headset. The result showed a significant difference for consideration/interpretation of mental status ($p < 0.01$). The study group performed significantly better in the assignment of assessing appropriate respiratory status ($p < 0.01$) and recognizing a need for escalation of care. Meng et al. (2018) conducted a similar experiment ($n = 70$). Two senior lecturers taught both control and study groups by traditional teaching methods (CT image, slides, and daily operation observation). The study group used the Gear VR headset (only 3DoF compared to most VR HMD with 6DoF) to watch real operation 360° videos after the course. The postintervention test showed a significant difference between groups in the score ($p < 0.05$). Stepan et al. (2017) conducted a randomized controlled study ($n = 66$) using the Oculus Rift VR system as an additional training method to evaluate the effectiveness, satisfaction, and

TABLE 3 | Compared with the nontraditional method.

Category	Intervention	Authors	*n*	Results
Theoretical assessment	VR	Coulter et al. (2007)	25	Study group $p < 0.001$ Control group $p = 0.024$ Between groups $p > 0.05$
		Wang et al. (2019)	125	Study group $p < 0.05$ Control group $p < 0.05$ Between groups $p = 0.136$
Surgical operation Assessment				Between groups (study group performed better, $p = 0.027$)
		Logishetty et al. (2019a)	24	Between groups (study group performed better, $p < 0.001$)
	AR	Logishett et al. (2019b)	24	Between groups (study group performed better, $p = 0.301$)
Error rate	VR	Logishetty et al. (2019a)	24	Study group mean error 4° Control group mean error 16°
	AR	Logishett et al. (2019b)	24	Study group 1° ± 1° Control group 6° ± 4° Between groups $p < 0.001$
		Rojas-Muñoz et al. (2019)	20	Placement errors between groups Task 1: study group performed better, $p < 0.001$) Task 2: study group performed better, $p = 0.01$)

TABLE 4 | Supportive usage.

Category	Intervention	Authors	*n*	Results
Theoretical assessment	VR	Zackoff et al. (2020)	168	Between groups (study group performed better, $p < 0.05$)
		Meng et al. (2018)	70	Between groups (study group performed better, $p < 0.05$)
		Stepan et al. (2017)	66	Between groups (study group performed better in post-test, $p = 0.87$) Between groups (study group performed better in retention test, $p = 0.47$)
Surgical operation Assessment	AR	Alismail et al. (2019)	32	Between groups (study group performed better, $p < 0.001$)
		Chen et al. (2019)	80	Study group vs. control group = 93.3 vs. 42.5% Between groups $p < 0.05$ (first-pass rate)

motivation in teaching medical students neuroanatomy. Both groups used the same teaching materials, while the study group had access to a VR headset which allowed them to view virtual brain anatomy generated from CT and MRI data. Different from the two studies mentioned above, there was no significant difference in preintervention ($p = 0.86$), postintervention ($p = 0.87$), or retention test ($p = 0.47$) between the two groups. However, the experimental group performed significantly better in the instructional materials motivational survey with greater attention, relevance, confidence, and satisfaction ($p < 0.01$).

Alismail et al. (2019) presented a study ($n = 32$) to assess the effectiveness of using the AR headset as an assistance tool to perform intubation simulation procedure. All participants watched a video and then started the intubation procedure; in the meantime, those in the study group used the AiRScouter AR headset, from which they could see the slides as a guideline additionally. The result showed a significant difference in ventilation time (seconds) between the study and control groups (280 vs. 205; $p = 0.005$). Moreover, the study group had a higher percentage adherence to following the checklist ($p < 0.001$). Chen et al. (2019) conducted a randomized experiment ($n = 80$) to evaluate the mixed reality application in lateral ventricle puncture training. The study group and the control group were taught traditionally for one month, while the study group used HoloLens AR headset to train the puncture in a simulated environment. As a result, the study group had a significantly higher first-pass rate than the control group (93.3 vs. 42.5%, $p < 0.05$). In the meantime, the study group participants had significantly better 3D reconstruction and more confidence ($p < 0.05$).

2.2.4.4 Cognition and Emotion

Frederiksen et al. (2020) conducted an RCT ($n = 31$) to explore the cognitive load and performance changes after enhancing the immersion of laparoscopic surgery simulation training by using an HMD. The 360° videos were clipped into different stressor levels (calm, light, and severe). The study group and the control group used a conventional VR laparoscopic surgery simulator, while the study group used the Oculus Rift VR headset playing 360° videos of a real operating room in the meantime. The cognitive load was significantly different ($p < 0.001$) between the study group (15.2% in light stressor and 43.1% in severe stressor) and the control group (23.0%). The study group reported a significantly worse performance on most simulator metrics (time, blood loss, damage, and hand movement). The authors stated, "However, immersive VR offers some potential advantages over conventional VR such as more real-life conditions but we only recommend introducing immersive VR in surgical skills training after initial training in conventional VR."

2.2.4.5 Non-RCT

Harrington et al. (2018) designed a randomized crossover study (n = 40) to evaluate the efficiency of immersive 360° video in surgical education when compared with traditional 2D video. The participants were divided into two different groups randomly. One group attended the 360° video experiment using the Samsung Gear VR headset first and then attended the 2D video experiment, while the other group experimented with the same content in the opposite order. The result revealed a significantly higher engagement level ($p < 0.0001$) and a higher level of focusing ($p < 0.0001$) with the 360° immersive video. There was no significant difference in information retention between the two groups ($p = 0.143$).

Logishetty et al. (2020) designed a competency-based simulation curriculum study (n = 32) using a VR HMD to evaluate the skills measurement and visuospatial transfer performance. All the residents attended five consecutive VR training and assessment sessions. The outcome of each assessment was compared with four expert hip surgeons' performance in both a physical world assessment and a VR one-off assessment. The result showed that the residents progressively developed surgical skills in VR by practicing repeatedly, and it allowed them to match expert VR levels on nine out of the 10 metrics included in the study. In the preparation phase, the number of errors in instrument selection and usage errors ($p < 0.001$), number of prompts required ($p < 0.001$), and procedural time ($p < 0.001$) were reduced significantly. The performance of the residents in the VR assessments was significantly improved as the inclination error ($p < 0.005$) and anteversion error ($p < 0.001$) were reduced. In the physical world–simulated assessment, the errors in femoral osteotomy height ($p = 0.044$), in femoral osteotomy angle ($p = 0.002$), in acetabular cup inclination ($p < 0.001$), and in acetabular cup anteversion ($p < 0.001$) were significantly reduced, which indicated that the visuospatial skills were transferred from VR to the physical world successfully.

2.2.4.6 Secondary and Additional

The original proposed additional outcomes in the protocol (Section 1.1) are as follows: "side effects of applying HMD into medical education, such as headache, motion sickness, and claustrophobia," and "learning motivation improvement by HMD VR or AR." The measures of effect are questionnaires or interview subjective experience. Next, the additional outcomes found in the systematic review will be outlined in three aspects: motion sickness, limitations, and motivations.

The motion sickness symptoms can occur after the user uses the VR or AR HMDs, especially when the virtual space movement does not match the user's movement in reality or their mind, which can be highly dependent on the content (Saredakis et al., 2020). This is heightened if the experiences are on a device only capable of 3DoF (e.g., roll, pitch, and yaw) and not 6DoF (e.g., X, Y, Z, roll, pitch, and yaw). Other factors include frame lag or screen tearing caused by low device capability or bad software optimization, which may enhance such symptoms. Several studies included in this systematic review reported that some participants in the VR intervention group had motion sickness after the experiment (Meng et al., 2018; Cai et al., 2019; Wang H. et al., 2019). Furthermore, the limited field of view (FOV) and the imagery of the HoloLens AR headset may produce head discomfort and ocular fatigue (Rojas-Muñoz et al., 2019). However, in Frederiksen et al. (2020) 's experiment, there were no motion sickness cases reported. The possible reason might be "minimal head movements compared to immersive VR video games where motion sickness has been an issue."

Regarding the limitations of VR/AR HMDs summarized from the included articles, their price is generally too high to deploy in a class-scale teaching environment (Stepan et al., 2017; Pulijala et al., 2018; Wang H. et al., 2019). However, as the technology develops, the price of these devices are reducing (Logishetty et al., 2019a) and are cheaper than an orthopedic simulator, open surgical platforms, or synthetic hip models (Meng et al., 2018; Logishetty et al., 2019b, Logishetty et al., 2020). The above conclusions indicated the VR AR technologies are expensive to be applied in some cases; nevertheless, they have the potential to be a cost-effective teaching method compared with other simulators and be an alternative teaching method in the future. The other limitations reported are the lack of model details and haptic feedback (Cai et al., 2019; Logishetty et al., 2019a; Wang H. et al., 2019), the extra workload needed for the user to get familiar with the devices (Stepan et al., 2017; Jiang et al., 2019), and bad user experience caused by limited FOV or the weight of the devices (Jiang et al., 2019; Rojas-Muñoz et al., 2019; Wang H. et al., 2019). Last but not least, Wang P. et al. (2019) mentioned that as one HMD can only support one user, it is time-consuming to conduct an experiment or teaching mission and has potential health problems with devices sharing, which needs extra attention under the current COVID pandemic situation.

As for the motivation and confidence aspect, the included studies found that the usage of VR/AR HMDs could improve participants' learning motivations and self-confidence by the immersive environment and interactive teaching process. The more satisfied students are, the more engaging students are in the teaching process. Compared with the traditional teaching method, the study group participants showed significantly higher satisfaction and motivation to the teaching method than those in the control group (Cai et al., 2019; Jiang et al., 2019; Wang H. et al., 2019). The same effect also showed when the VR/AR techniques were compared with other simulators (Wang H. et al., 2019) or used as an additional teaching tool (Stepan et al., 2017). However, in Meng et al. (2018) 's experiment, there was no significant difference in the mean satisfaction score. The confidence level significantly increased in both groups in several studies, but the participants of the study group showed significantly higher self-confidence scores (Pulijala et al., 2018; Chen et al., 2019).

2.3 Risk of Bias Within Studies

To reduce the bias of language, this systematic review included English studies and Chinese studies. Among 15 RCTs, nine were in English and six were Chinese. All RCT studies' risk of bias was assessed by using the RoB 2.0 tool (**Figure 4**). Four English

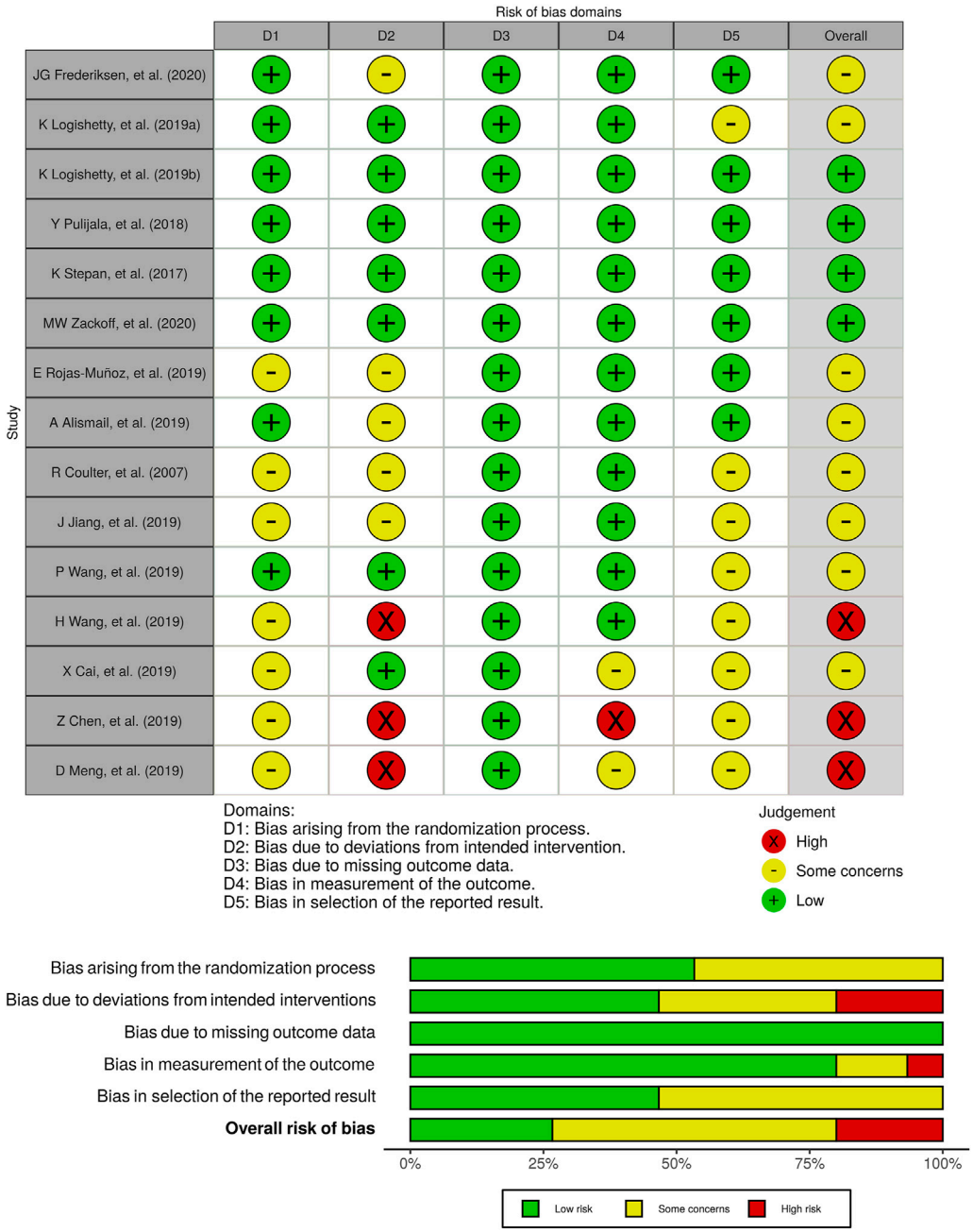

FIGURE 4 | Risk of bias analysis.

articles had a low risk of bias, while the other five English articles had some bias concerns. The primary concern was bias due to deviations from the intended intervention. For instance, they did not mention whether the participants or the data accessors were blinded to the random assignment. Some studies did not clarify whether the allocation sequence was random, and some did not mention a prespecified analysis plan for analyzing the result.

Meanwhile, none of the Chinese articles had a low risk of bias. Three studies had a high risk of bias due to deviations from intended intervention as there was no information about clarifying the assignment process or analysis after the assignment. All Chinese studies lacked a prespecified analysis plan or an experiment protocol, and the majority of them did not specify the detail of random sequences. Three studies in Chinese had a bias in the measurement of the reported result. Overall, four English studies had a low risk of bias, three studies in Chinese had a high risk of bias, and the rest studies had some concerns.

3 DISCUSSION

The review will first summarize the evidence for both advantages and disadvantages of VR/AR HMD application in medical training. These are the first two questions of this study while the third question will be answered in the proof subsection, demonstrating that these approaches do indeed increase the efficiency of teaching in medical education. Finally, the use of AR/VR as a support tool will be addressed in the final subsection. After summarizing all the evidence, the limitations of this review will be discussed.

3.1 Summary of Evidence

This systematic review focused on clinical educational studies related to VR/AR HMD application in medicine. It revealed that compared with traditional teaching media and other additional teaching methods, the application of VR and AR HMDs improved students' learning curve and motivation. The participant who used virtual HMDs had a better performance in the theory test and the operation examination. Furthermore, the HMDs also provided immersion for the simulated learning environment to increase students' cognition load, maintaining students' performance in the real-life study case.

3.1.1 When Compared with the Standard Teaching Method or Other Types of Simulators, What Are the Comparative Effectiveness of HMD VR or AR Usage in Medical Education?

The standard teaching method refers to the case that lecturers give out the course by using paper-based teaching materials, slides, and videos, while other types of simulators, in this review case, could be 3D print solid or silicon models, PC/phone educational software, and conventional simulator without HMD such as LapSim®[4]. This systematic review found three aspects of comparative effectiveness of VR and AR HMD application: motivation, learning efficiency, and space efficiency.

Firstly, this review has found that the immersive environment provided by the HMDs increases student's learning motivation and course satisfaction. The results of studies show that students who use HMD intervention during the study process are more satisfied and motivated (Stepan et al., 2017; Pulijala et al., 2018; Cai et al., 2019; Jiang et al., 2019; Wang H. et al., 2019). Furthermore, the simulation offers the residents a chance to experience the test or operational environment before entering an actual one. It increases residents' self-confidence in the knowledge they gained (Pulijala et al., 2018; Chen et al., 2019). With the mental status enhanced, the knowledge can be transferred more effectively (Zackoff et al., 2020). Secondly, the HMDs provide a stereoscopic view, which would potentially benefit the curriculum that needs students to reconstruct spatial information. According to the result of this review, residents who used VR or AR HMDs performed better in 3D reconstruction (Meng et al., 2018; Cai et al., 2019; Jiang et al., 2019) and had a better understanding of the new information (Harrington et al., 2018). Based on those benefits, the system can generate detailed operation replay or a high-quality 3D virtual model to support the learning process. The student will have unlimited chances to learn and practice without considering any waste of cadaver resources, which maximizes learning opportunities while cutting down the cost at the same time (Meng et al., 2018; Cai et al., 2019; Chen et al., 2019; Logishetty et al., 2019a). The interactive learning mode and hands-on learning experience can benefit student's learning curve (Meng et al., 2018; Cai et al., 2019; Jiang et al., 2019), because of which, the student can conduct unsupervised self-driven learning (Logishetty et al., 2019a; Logishetty et al., 2019b) with live feedback (Alismail et al., 2019; Logishetty et al., 2019b). Finally, the usage of VR or AR HMDs as teaching supportive material is space-efficient compared with actual 3D models and simulators and causes fewer collisions during the practice when compared with other media (Rojas-Muñoz et al., 2019).

3.1.2 What Are the Disadvantages of Using HMD VR or AR? Which One Has a Lower Side Effect?

The results of included studies (Section 2.2.4.6) give answers to the second question proposed in the protocol. This review discovered two disadvantageous aspects of HMDs usage in medical education: motion sickness and other limitations. Motion sickness symptoms cases were reported in several studies, while there was no specific figure to reflect the scale (Meng et al., 2018; Cai et al., 2019; Wang H. et al., 2019). The VR HMD motion sickness can be eased by minimizing head movements (Frederiksen et al., 2020). According to the result, AR HMD has a lower side effect as only one AR study reported head discomfort and ocular fatigue (Rojas-Muñoz et al., 2019).

Except for the potential motion sickness issue, state-of-the-art VR and AR HMDs have some other limitations. As commented in several included articles, the cost of VR HMDs is too high to apply in a class-scale teaching scenario (Stepan et al., 2017; Pulijala et al., 2018; Wang H. et al., 2019); however, the price of VR HMD is reducing when the technique is developing (Logishetty et al., 2019a), and the price of AR HMD is lower than an orthopedic simulator, open surgical platforms, or synthetic hip models (Meng et al., 2018; Logishetty et al., 2019b; Logishetty et al., 2020). The majority of studies that mentioned price limitations are those using VR HMDs intervention; however, this review cannot conclude that AR HMDs are easier to deploy as HoloLens AR HMD is not a commercial product and its price is much higher than a commercial VR HMD. One of the other limitations reported is the lack of model details and tactile feedback in the VR environment (Cai et al., 2019; Logishetty et al., 2019a; Wang H. et al., 2019). The AR devices may potentially have similar limitations due to their lower capacity in graphics computation. However, those limitations are not reported in the included articles. The reason might be the different functionalities between VR and AR applications. AR is generally used as a reference tool that provides extra information to the real object or person, while VR is more isolated so that the virtual environment detail affects the learning process directly. Moreover, the HMD design itself can lead to a bad user experience caused by limited

[4]https://surgicalscience.com/systems/lapsim/

FOV and the extra weight on a user's head (Jiang et al., 2019; Rojas-Muñoz et al., 2019; Wang H. et al., 2019); this issue as mentioned before is becoming less of an issue due to the rapid improvements in HMD design.

3.1.3 Is There a Definitive Advantage of HMD VR and AR When Used for Increasing the Efficiency of Teaching in Medical Education?

The third question proposed in the protocol is addressed in the outcome section (Section 2.2.4). Some of the VR or AR HMDs intervention groups performed significantly better than the control groups in the theoretical posttest (Meng et al., 2018; Cai et al., 2019; Wang H. et al., 2019), while some studies did not find a significant difference between the two groups in the theory test, but both groups had significant improvements and the study group performed better (Stepan et al., 2017; Pulijala et al., 2018; Jiang et al., 2019).

Regarding the actual or simulated surgical exam, the included articles' study groups had significantly higher scores (Alismail et al., 2019; Chen et al., 2019; Logishetty et al., 2019a; Wang H. et al., 2019; Wang P. et al., 2019; Zackoff et al., 2020) and a lower error rate than the control groups (Logishetty et al., 2019a; Logishetty et al., 2019b; Rojas-Muñoz et al., 2019; Wang H. et al., 2019). Even when the control group was guided by a surgical expert individually, the improvement of the study group using the AR HMD self-study system was still comparable (Logishetty et al., 2019b), which indicates the potential of using AR HMD in an alternative supportive teaching role. Besides the improvements to learning outcomes, the VR HMD intervention could aid in the development of more real-life skills as they have been shown to increase cognitive load due to the stress of experiencing a more realistic environment than other teaching methods. One example of this effect is in the study by Frederiksen et al., (2020), where the VR study group had significantly worse performance on most simulator metrics (time, blood loss, damage, and hand movement) due to the extra cognitive load when compared to the control. This was due to the real-life operational 360° video the participants were immersed in. This indicates that the usage of VR HMD could help guarantee the skill transfer from the simulators to a real-life case, but basic skills should still be taught more abstractly. This abstraction could still be taught in VR, and it is the power of this medium that allows changes to fidelity at will. Finally, by repeatedly practicing with the VR HMD operation simulator, the novice surgeons could gradually build up their skills until they performed as same as an expert level within the same VR assessment; nevertheless, the knowledge gain could also be transferred to the physical world–simulated assessment (Logishetty et al., 2020).

3.1.4 Do HMD VR and AR Have the Potential to be Support Tools for Medical Education?

The above evidence can also be used to answer the last question proposed as although current stage VR and AR HMDs have some limitations such as motion sickness and can still be relatively costly if an entire class needs access to multiple HMD's, they still have great potential to be supportive medical education tools (Stepan et al., 2017; Harrington et al., 2018; Logishetty et al., 2019a; Frederiksen et al., 2020).

With ongoing hardware development, the motion sickness issue should be eased and even completely avoided by making the headset lighter and smaller and increasing the rendering capacity. Looking ahead in terms of accessibility, the high-performance hardware's price is reducing and is getting similar to a high-end smartphone.

Several researchers in the included studies within this review pointed out that the VR and AR applications would never replace the traditional teaching method but could provide supportive teaching materials (Logishetty et al., 2019b; Wang H. et al., 2019). As the skills and knowledge gained in the virtual world can be successfully transferred to the physical world (Logishetty et al., 2020; Zackoff et al., 2020), the current medical and veterinary anatomy education challenges, such as the lack of anatomy cadaver resources, could be eased with the introduction by merging VR and AR technique into the teaching curriculum.

3.2 Limitations

The main limitations of this systematic review are the following three points:

- Language bias. The search strategy includes English and Chinese articles to reduce language bias. However, to avoid language bias, more languages need to be added to the search strategy.
- Risk of bias for the RCTs. According to the risk of bias analysis chart (**Figure 4**), over half of the included article has some bias concerns. Due to the publication format difference, most Chinese articles cannot meet the requirements of the RoB framework (Sterne et al., 2019), which makes three included Chinese articles high risk of bias.
- Abstracts covered. This systematic review did not include the articles or studies that only provided abstract because it is hard to judge whether the study meets all the inclusive criteria. However, this fact became one of the limitations in this review as it did not cover all the articles, including gray publications and clinical trial protocols.

3.3 Conclusion

VR and AR HMD's applications in medical training are moving slowly into the mainstream as with their reduced cost and increased availability, researchers have taken notice in their search to improve education efficiency. Compared with traditional teaching methods and other non-HMD VR simulators, VR and AR HMDs stimulate students' learning motivation, increase their satisfaction, and improve students' learning outcomes. The immersive VR-simulated environment prepares students' better mentally before dealing with emotionally challenging real-life medical situations, which can help guarantee the skill transferred from virtuality to reality. Motion sickness and some hardware limitations are reported in this review, but with every passing year, innovations in this field mean these limitations are either being reduced or becoming not existent.

The future study directions can be divided into two aspects: HMDs as tools to support students' theoretical knowledge gain in the curriculum and be simulators to training students' surgical skills. The current studies concentrate on developing theoretical knowledge. However, in future, these studies need to be expanded, and more extensive study groups are needed to evaluate the training efficiency to integrate the interventions into the traditional teaching process. In terms of skill training, the future VR/AR HMD intervention in medical education will be more commonly combined with actual surgical equipment to bridge the gap between simulation and reality. Thus, future studies can target the actual skill and knowledge transfer rate from virtuality to reality with larger intervention groups. As the included articles all focused on some particular scenario, more wide-ranging and longitudinal studies are needed to validate this type of intervention.

Due to the pandemic, remote learning, which already was on the rise before the crisis, has accelerated. It is not just in education as countries such as Ireland have passed laws to give the legal right to request home working. Working from home now has become part of society's fabric, in conjunction with the move to requiring continuous professional development for most professions. Research into alternatives to the traditional physical labs could be essential, not just for medical education but for all of education. VR and AR intervention can potentially be a supportive tool for lecturers' teaching, students' self-learning, and professional practitioners' self-evaluation.

Few studies evaluate remote learning using VR or AR interventions, so this is still an ongoing research area. The future experiment direction in this area should concentrate on how online remote teaching could increase the teaching efficiency in medical and veterinary education. The rise of the use of VR/AR within academia, even allowing remote conferences (MacIntyre 2020) to be held in VR, has helped demonstrate its future. Remote learning will still flourish after the pandemic is over, as this natural experiment has demonstrated that these approaches can be successful. With the increasing adoption of VR/AR within remote learning, these successes can be built upon. This trend complements the fact that VR/AR HMD's are also becoming more inexpensive, thus allowing for increasing equity and access to education across the world with these new technologies if the lessons from many of the experiments outlined in this review are heeded.

At this current stage, VR and AR intervention cannot replace actual cadaver learning material due to their lack of fidelity and lack of tactical feedback will affect students' cognition when faced with actual surgical cases. However, along with ongoing HMD development, the interventions will be more accessible and easier to blend into medical education in the future. Furthermore, the high-fidelity model and haptic innovations will blur the edge between virtuality and reality; but crucially, more experiments are needed to gauge educational efficiency gain and evaluate and verify whether the VR and AR simulators can be a possible replacement to cadavers, avoiding existing ethical problems and resource limitations. Medical education, in particular, has always suffered the problem of having more qualified applicants than places across the world due to resource limitations. Removing these resource limitations could significantly impact equity and access to medical education in the future.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

Conception and design, XX and AC; data collection, XX; article screening, XX, EM, and AC; analysis and interpretation, XX; writing the article, XX and AC; critical revision of the article, XX, EM, and AC; final approval of the article, EM and AC; overall responsibility: XX, EM, and AC.

FUNDING

Author XX has been supported by the China Scholarship Council (201908300021).

REFERENCES

Alaker, M., Wynn, G. R., and Arulampalam, T. (2016). Virtual Reality Training in Laparoscopic Surgery: A Systematic Review & Meta-Analysis. *Int. J. Surg.* 29, 85–94. doi:10.1016/j.ijsu.2016.03.034

Alismail, A., Thomas, J., Daher, N. S., Cohen, A., Almutairi, W., Terry, M. H., et al. (2019). Augmented Reality Glasses Improve Adherence to Evidence-Based Intubation Practice. *Amep* Vol. 10, 279–286. doi:10.2147/amep.s201640

Cai, X., Qin, J., Huang, S., and Yi, X. (2019). Application of Virtual Reality Technology in Clinical Teaching of Spinal Surgery. *China Contin. Med. Educ.* 11, 18. doi:10.3969/j.issn.1674-9308.2019.23.008

Chen, Z., Liu, Y., He, B., Huang, S., Hong, W., Liao, Z., et al. (2019). Application of Ventricle Puncture Training System Based on Mixed Reality in Medical Education and Training. *J. TRAUMA EMERG.* 7, 5. doi:10.16746/j.cnki.11-9332/r.2019.01.002

Coulter, R., Saland, L., Caudell, T., Goldsmith, T. E., and Alverson, D. (2007). The Effect of Degree of Immersion upon Learning Performance in Virtual Reality Simulations for Medical Education. *InMedicine Meets Virtual Reality* 15, 155.

Frederiksen, J. G., Sørensen, S. M. D., Konge, L., Svendsen, M. B. S., Nobel-Jørgensen, M., Bjerrum, F., et al. (2020). Cognitive Load and Performance in Immersive Virtual Reality versus Conventional Virtual Reality Simulation Training of Laparoscopic Surgery: a Randomized Trial. *Surg. Endosc.* 34, 1244–1252. doi:10.1007/s00464-019-06887-8

Harrington, C. M., Kavanagh, D. O., Wright Ballester, G., Wright Ballester, A., Dicker, P., Traynor, O., et al. (2018). 360° Operative Videos: A Randomised Cross-Over Study Evaluating Attentiveness and Information Retention. *J. Surg. Educ.* 75, 993–1000. doi:10.1016/j.jsurg.2017.10.010

Jensen, L., and Konradsen, F. (2018). A Review of the Use of Virtual Reality Head-Mounted Displays in Education and Training. *Educ. Inf. Technol.* 23, 1515–1529. doi:10.1007/s10639-017-9676-0

Jiang, J., Sun, J., Luo, X., Sun, X., Wang, Y., Xu, X., et al. (2019). Effect Analysis of Applying Mixed Reality Technology in Spinal Surgery Teaching. *China Med. Educ. Technol.* 34, 230. doi:10.13566/j.cnki.cmet.cn61-1317/g4.202002028

Larsen, C. R., Oestergaard, J., Ottesen, B. S., and Soerensen, J. L. (2012). The Efficacy of Virtual Reality Simulation Training in Laparoscopy: a Systematic

Review of Randomized Trials. *Acta obstetricia gynecologica Scand.* 91, 1015–1028. doi:10.1111/j.1600-0412.2012.01482.x

Logishetty, K., Gofton, W. T., Rudran, B., Beaulé, P. E., and Cobb, J. P. (2020). Fully Immersive Virtual Reality for Total Hip Arthroplasty. *JBJS* 102, e27. doi:10.2106/jbjs.19.00629

Logishetty, K., Rudran, B., and Cobb, J. P. (2019a). Virtual Reality Training Improves Trainee Performance in Total Hip Arthroplasty: a Randomized Controlled Trial. *Bone Jt. J.* 101-B, 1585–1592. doi:10.1302/0301-620x.101b12.bjj-2019-0643.r1

Logishetty, K., Western, L., Morgan, R., Iranpour, F., Cobb, J. P., and Auvinet, E. (2019b). Can an Augmented Reality Headset Improve Accuracy of Acetabular Cup Orientation in Simulated Tha? a Randomized Trial. *Clin. Orthop. Relat. Res.* 477, 1190–1199. doi:10.1097/corr.0000000000000542

MacIntyre, B. (2020). Remote conference participation in social virtual worlds.

McGuinness, L. A., and Higgins, J. P. T. (2021). Risk-of-bias VISualization (Robvis): An R Package and Shiny Web App for Visualizing Risk-of-bias Assessments. *Res. Syn Meth* 12, 55–61. doi:10.1002/jrsm.1411

Meng, D., Zhao, L., Ouyang, Y., Lin, H., and Huihao, C. (2018). Application of Visual Reality Technology in the Clinical Teaching of Orthopedics Interns. *Med. Inf.* 31, 17. doi:10.3969/j.issn.1006-1959.2018.22.006

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., and Group, P. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: the Prisma Statement. *Plos Med.* 6, e1000097. doi:10.1371/journal.pmed.1000097

Pringle, A., Campbell, A. G., Hutka, S., and Keane, M. T. (2018). "Using an Industry-Ready AR HMD on a Real Maintenance Task: AR Benefits Performance on Certain Task Steps More Than Others," in 2018 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), (Munich, Germany, 16–20. *October 2018* (IEEE). doi:10.1109/ISMAR-Adjunct.2018.00075

Pulijala, Y., Ma, M., Pears, M., Peebles, D., and Ayoub, A. (2018). Effectiveness of Immersive Virtual Reality in Surgical Training-A Randomized Control Trial. *J. Oral Maxillofac. Surg.* 76, 1065–1072. doi:10.1016/j.joms.2017.10.002

Rojas-Muñoz, E., Cabrera, M. E., Andersen, D., Popescu, V., Marley, S., Mullis, B., et al. (2019). Surgical Telementoring without Encumbrance. *Ann. Surg.* 270, 384–389. doi:10.1097/sla.0000000000002764

Saredakis, D., Szpak, A., Birckhead, B., Keage, H. A. D., Rizzo, A., and Loetscher, T. (2020). Factors Associated with Virtual Reality Sickness in Head-Mounted Displays: a Systematic Review and Meta-Analysis. *Front. Hum. Neurosci.* 14, 96. doi:10.3389/fnhum.2020.00096

Stepan, K., Zeiger, J., Hanchuk, S., Del Signore, A., Shrivastava, R., Govindaraj, S., et al. (2017). Immersive Virtual Reality as a Teaching Tool for Neuroanatomy. *Int. Forum Allergy Rhinol.* 7, 1006–1013. Wiley Online Library. doi:10.1002/alr.21986

Sterne, J. A. C., Savović, J., Page, M. J., Elbers, R. G., Blencowe, N. S., Boutron, I., et al. (2019). Rob 2: a Revised Tool for Assessing Risk of Bias in Randomised Trials. *bmj* 366, l4898. doi:10.1136/bmj.l4898

Wang, H.Ou, Y., Hu, P., Deng, Y., Zou, L., Deng, Q., et al. (2019). Application Effect of Mixed Reality in the Teaching of Hepatobiliary Surgery. *Chin. J. Med. Edu Res.* 18, 1230. doi:10.3760/cma.j.issn.2095-1485.2019.12.011

Wang, P.Yan, W., Wang, F., Wang, L., Cai, H., and Chen, W. (2019). The Application of Immersive Virtual Reality Technology in Experimental Teaching of Intravenous Injection. *Chin. Nurs. Manage.* 20, 176. doi:10.3969/j.issn.1672-1756.2020.02.006

Zackoff, M. W., Real, F. J., Sahay, R. D., Fei, L., Guiot, A., Lehmann, C., et al. (2020). Impact of an Immersive Virtual Reality Curriculum on Medical Students' Clinical Assessment of Infants with Respiratory Distress*. *Pediatr. Crit. Care Med.* 21, 477–485. doi:10.1097/pcc.0000000000002249

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Virtual Reality

ORIGINAL RESEARCH
published: 28 October 2021
doi: 10.3389/frvir.2021.692641

Application of Mixed Reality in Medical Training and Surgical Planning Focused on Minimally Invasive Surgery

Juan A. Sánchez-Margallo[1], Carlos Plaza de Miguel[2], Roberto A. Fernández Anzules[3] and Francisco M. Sánchez-Margallo[4]*

[1]Bioengineering and Health Technologies Unit, Jesús Usón Minimally Invasive Surgery Centre, Cáceres, Spain, [2]TREMIRS Project, Jesús Usón Minimally Invasive Surgery Centre, Cáceres, Spain, [3]Thoracic Surgery Unit, Cáceres University Hospital, Cáceres, Spain, [4]Scientific Direction, Jesús Usón Minimally Invasive Surgery Centre, Cáceres, Spain

OPEN ACCESS

Edited by:
Marientina Gotsis, University of Southern California, United States

Reviewed by:
Juan Manuel Jacinto-Villegas, National Council of Science and Technology (CONACYT), Mexico
Nancy Rodriguez, UMR5506 Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier (LIRMM), France

***Correspondence:**
Juan A. Sánchez-Margallo
jasanchez@ccmijesususon.com

Specialty section:
This article was submitted to Virtual Reality in Medicine, a section of the journal Frontiers in Virtual Reality

Received: *08 April 2021*
Accepted: *15 October 2021*
Published: *28 October 2021*

Citation:
Sánchez-Margallo JA, Plaza de Miguel C, Fernández Anzules RA and Sánchez-Margallo FM (2021) Application of Mixed Reality in Medical Training and Surgical Planning Focused on Minimally Invasive Surgery. Front. Virtual Real. 2:692641. doi: 10.3389/frvir.2021.692641

Introduction: Medical training is a long and demanding process, in which the first stages are usually based on two-dimensional, static, and unrealistic content. Conversely, advances in preoperative imaging have made it an essential part of any successful surgical procedure. However, access to this information often requires the support of an assistant and may compromise sterility in the surgical process. Herein, we present two solutions based on mixed reality that aim to improve both training and planning in minimally invasive surgery.

Materials and Methods: Applications were developed for the use of the Microsoft HoloLens device. The urology training application provided access to a variety of anatomical and surgical training contents. Expert urological surgeons completed a questionnaire to evaluate its use. The surgical planning solution was used during laparoscopic renal tumorectomy in an experimental model and video-assisted right upper lobectomy in an adult patient. Surgeons reported their experience using this preoperative planning tool for surgery.

Results: The solution developed for medical training was considered a useful tool for training in urological anatomy, facilitating the translation of this knowledge to clinical practice. Regarding the solution developed for surgical planning, it allowed surgeons to access the patient's clinical information in real-time, such as preoperative imaging studies, three-dimensional surgical planning models, or medical history, facilitating the surgical approach. The surgeon's view through the mixed reality device was shared with the rest of the surgical team.

Conclusions: The mixed reality-based solution for medical training facilitates the transfer of knowledge into clinical practice. The preoperative planning tool for surgery provides real-

Abbreviations: MIS, minimally invasive surgery; OR, operating room; VR, virtual reality; AR, augmented reality; MR, mixed realityl; 3D, three-dimensional; CT, computer tomography; MRI, magnetic resonance imaging; DICOM, digital imaging and communication on medicine.

time access to essential patient information without losing the sterility of the surgical field. However, further studies are needed to comprehensively validate its clinical application.

Keywords: mixed reality, medical training, surgical planning, minimally invasive surgery, laparoscopy

INTRODUCTION

Medical education is a long and demanding process that requires extensive theoretical knowledge, along with technical and non-technical skills. During the early stages of medical education, training methods are often based on static and non-realistic learning content. Currently, these methods are being replaced by new approaches based on the use of information and communication technologies (Langridge et al., 2018; Williams et al., 2020). Apprenticeship models in surgical training have rapidly evolved from traditional approaches based on an educational philosophy following the principle of "see one, do one, teach one" to more sophisticated surgical simulators aimed at increasing the number of simulations following the "see one, simulate many deliberately, do one" philosophy (Kerr and O'Leary, 1999; Scott et al., 2008), thus allowing a dramatic increase in the skills of medical professionals and the safety of patients (Viglialoro et al., 2021). There are some strategies for surgical training based on serious video games (Rosenberg et al., 2005; Goris et al., 2014), animal models (Daly et al., 2014; DeMasi et al., 2016), and cadavers (Jacobson et al., 2009; Zuckerman et al., 2009; Rocha e Silva et al., 2016). However, due to the economic and ethical issues involved in some of these solutions, surgical training has rapidly shifted toward the use of simulation-based systems (Forgione and Guraya, 2017).

Advances in preoperative imaging have allowed for its extensive application in surgical planning, which has thus become an essential part of any successful surgical procedure (Sánchez-Margallo et al., 2015). Specifically, when facing complex surgeries, surgical planning provides valuable information for predicting and reducing any potential risks during surgery, thereby improving its safety levels. However, preoperative imaging systems are often located outside the operating room (OR) and, thus, need to be accessed outside the surgical area, or their operation requires the help of an assistant. In addition, the devices available in the OR for surgical planning may entail the loss of sterility, mainly due to the manipulation of touch screens, keyboards, and other computer equipment. In this regard, new technologies such as virtual reality (VR), augmented reality (AR), and mixed reality (MR) have the potential to provide medical students with interactive and realistic training systems; furthermore, they can be valuable tools for surgeons to facilitate the planning of surgical interventions (Sadeghi et al., 2020).

Medical visualisations have already been widely exploited for supporting diagnosis in the form of X-rays, computed tomography (CT), and magnetic resonance imaging (MRI) scans (Smith et al., 2020). The use of three-dimensional (3D) representations of these data in immersive settings provides new ways to explore the data and to further enhance the tools available to medical professionals in several areas including medical training, surgical planning, and intraoperative guidance. This evolution is even more evident in the case of minimally invasive surgery (MIS), which often lacks adequate access to the patient's anatomy (Sánchez-Margallo et al., 2018a). In this context, technological advances have radically changed surgical training and planning (Lahanas et al., 2015; Jayender et al., 2018; Li et al., 2020; Sánchez-Margallo et al., 2021).

In surgical training, most simulation-based approaches have focused on traditional VR and AR technologies, which offer different degrees of immersive experience but are generally unable to interact with 3D information combined with the real-world environment. Recently, MR techniques have replaced these traditional technologies intending to combine the real working environment with virtual content so that users can interact with both simultaneously. MR surgical simulators and medical training applications are becoming an important part of the training process for physicians, as they allow for a training environment appropriate for recreating realistic and reproducible scenarios without putting the patient at risk (Sánchez-Margallo et al., 2018b; Sappenfield et al., 2018; Amparore et al., 2021).

The use of 3D models to estimate the size and shape before performing the surgical procedure has been effectively implemented for almost a decade (Hurson et al., 2007). The irruption of MR techniques can make an important difference in this field. This technique can generate personalised 3D models for each patient and visualise the internal anatomy in a fully immersive environment. This opens up new possibilities, such as preoperative simulations, to determine optimal procedures and to predict the final surgical outcomes. MR technology has already been successfully applied as a planning tool in different surgical scenarios, including urology (Li et al., 2020), thoracic surgery (Perkins et al., 2020), neurosurgery, colorectal, and bariatric surgery (Cartucho et al., 2020). These solutions allow for the inclusion of elaborate information such as holographic images or 3D objects that can be placed within the surgeon's field of view, thus avoiding the need to use alternative displays in the OR and facilitating a more precise alignment between virtual information and physical objects. This would reduce the need of awkward postures for the surgeon and provide new interactive experiences in surgical planning (Hu et al., 2019).

In the field of surgical assistance, the use of MR wearable devices such as the HoloLens (Microsoft; Redmond, Washington, United States), in combination with new emerging imaging technologies, can benefit the surgical process, especially in complex procedures. This technology facilitates the spatial localisation of anatomical structures and improves mental alignment, which simplifies preoperative planning (Lee et al., 2017). This technology has already been evaluated as an assistance tool during endoscopic procedures (Al Janabi et al., 2020), spine surgery (Liu et al., 2020), interventional radiology

procedures (Deib et al., 2018; Heinrich et al., 2019), and orthopaedic surgery (Gregory et al., 2018). In the latter, this technology was tested using the Holoportal MR application (TeraRecon; Durham, NC), as a proof of concept, in a real surgical environment during the implantation of a shoulder prosthesis (Gregory et al., 2018).

The main objective of this study was to describe and test a set of innovative MR-based solutions for the improvement of both training and planning in MIS. The proposed solutions will allow the use of new and more realistic scenarios for medical training, as well as access to different sources of preoperative patient information, to support the planning of surgical procedures. The software solution developed for medical training focused on urology. Surgical planning solutions have been tested in two different surgical scenarios, namely laparoscopic renal tumorectomy and video-assisted lobectomy.

MATERIALS AND METHODS

Two MR-based applications for MIS training and planning were developed and evaluated in this study. The information was displayed through interactive holograms controlled by hand gestures or voice commands. The view cursor (similar to mouse pointer) was controlled by the gaze of the user and the interactions were triggered by the gazed holograms followed by hand gestures or voice commands. The two main hand gestures integrated in both applications were "air-tap" (raise the index finger in front of the field of view and then tap by flexing the index finger down. Similar to mouse click) to interact with user interface buttons and "air-tap and hold gesture" (similar to holding down the mouse button while dragging it) to scale, position and rotate the 3D holograms. These gestures allow the surgeon to show or hide information, interact with the 3D models of the training application and navigate between the different axes of the preoperative studies of the surgical planning application, among other actions.

These applications allowed viewing the contents from the HoloLens glasses themselves (one user) or sharing their experiences with other devices via streaming (several users). They fostered communication between the surgical team and the transmission of knowledge to other people in real-time.

The first generation of HoloLens was used as the MR device. This wearable headset combines several types of sensors (an inertial measurement unit, four environmental understanding cameras, one depth camera, one high-definition video camera, four microphones, and an ambient light sensor) along with an Intel 32-bit architecture processor (Intel Corporation; Santa Clara, CA, United States) and a custom-built Microsoft holographic processing unit. The weight of the device is 579 g and the battery durability can reach 5.5 h. HoloLens creates visual information using the reflection of two high-definition 16:9 light engines onto each retina of the user (offering interpupillary automatic calibration), which does not interfere with the visual information of the surrounding environment.

Mixed Reality Framework

Unity (Unity Technologies, San Francisco, CA, United States) was selected as the development platform because it allows easy interaction with visual elements (both 2D or 3D) and integrates different plugins and libraries that greatly facilitate the implementation, allowing for the porting of applications to most extended platforms. Each application considered one or more scenes, which in turn were made up of objects structured in the form of a parent-child hierarchy. Each of the objects had a "*Transform*" component, a script that controls its position, rotation, and scale, fostering the possibility of adding many other different components even in the form of user-programmed scripts. Microsoft Visual Studio was adopted as the programming environment, with C# as the programming language, to deploy the applications on the device.

The applications were developed using the Mixed Reality Toolkit development kit for Unity, which has become the standard for developing any MR application with HoloLens devices. This has been used in many medical applications for rhinoplasty (Maasthi et al., 2020), arthroplasty training (Turini et al., 2018), and open abdominal surgery (Galati et al., 2020). Loading of preoperative imaging studies with Digital Imaging and Communication on Medicine (DICOM) format was performed using a customised version of the Fellow Oak DICOM toolkit (Al-Zu'bi et al., 2017).

The two MR-based applications were implemented using Microsoft Windows 10 operating system. The final products were two Universal Window Platform applications (Microsoft) implemented on Intel's x86 architecture (Intel Corporation).

Training in Urology

Medical training applications focused on the human pelvis. Specifically, a 3D anatomical model was developed based on the CT study of a real patient. The model included interactive information on the different anatomical systems of the pelvic cavity (vascular, nervous, muscular, bone, digestive, urinary, and reproductive systems), as shown in **Figures 1A,B**.

The developed MR solution allowed the visualisation of the 3D models, manipulating them spatially (scale, rotation, and translation), and being able to activate or deactivate them independently (**Figure 1A**). In addition, the user received real-time information about the anatomical element being pointed at or visualised, highlighting the anatomical structure and displaying its name and other relevant information next to it. Voice commands were available to show the different anatomical systems by saying "show/hide" followed by "muscular/bone/vascular/nervous/renal/reproductive" and ended by "system".

In addition to the interactive visualisation of the 3D anatomical model, the application allowed the holographic visualisation of videos about related surgical techniques (laparoscopic nephrectomy, prostatectomy, etc.), preoperative imaging studies with and without pathologies, and medical illustrations.

This new medical training solution was tested by expert urologists who attended a training activity at the Jesús Usón Minimally Invasive Surgery Centre (Cáceres, Spain). At the beginning of the session, participants received a brief explanation of the gestural and voice interaction methods, including different voice commands for

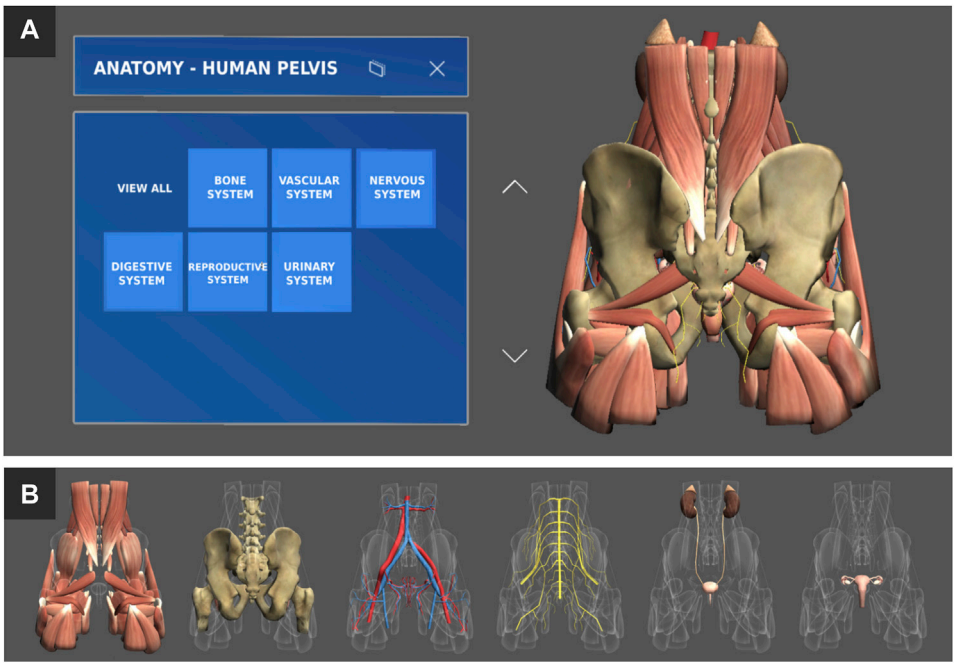

FIGURE 1 | (A) General interface of the training application with a 3D interactive anatomical model of the human pelvis. **(B)** Detail of the different systems available in the human pelvis model (from left to right): muscular, skeletal, vascular, nervous, renal, and reproductive systems.

TABLE 1 | Set of subjective parameters regarding the surgeon's experience with the urology training application developed for HoloLens.

Item	Category	Description
1	Ergonomics	Comfort to wear the mixed reality glasses
2	Intuitiveness	The gesture control method is easy to use
3	Intuitiveness	The voice control method is easy to use
4	Intuitiveness	The training application based on mixed reality is intuitive to use
5	Educational usefulness	The application facilitates interaction with the educational material compared to traditional methods
6	Educational usefulness	The application provides a useful tool for training in urological anatomy
7	Presentation of educational information	The way to visualise the 3D holographic models is clear
8	Presentation of educational information	The way in which the holographic 3D models are presented is useful
9	Presentation of educational information	The way in which the holographic 3D models are structured is orderly
10	Further applications	Additional utility of this technology for surgical assistance

the MR device. Next, participants were invited to interact with the functionalities of the interactive 3D anatomical model of the human pelvis and all its associated systems, reference videos of related surgical procedures, and medical illustrations. They used both gestural interactions and voice commands. To evaluate the user experience with the application and the use of the MR glasses, they completed a personalised questionnaire at the end of the session (**Table 1**). Each item is rated on a 5-point Likert scale. In addition, they were provided with space to indicate any additional comments.

Surgical Planning in Minimally Invasive Surgery

The general functionalities of this application included visualisation and interaction with preoperative imaging studies of the patient (CT or MRI studies), as shown in **Figure 2A**, and interaction with 3D anatomical models of the patient, generated from the preoperative studies (**Figure 2B**). In addition, it allowed for the visualisation of medical illustrations regarding the anatomical structures to be addressed during surgery (**Figure 2C**) and videos/tutorials regarding similar MIS procedures (**Figure 2E**), as well as providing access to the patient's medical history *in situ* (**Figure 2D**). All content was displayed in the form of holograms that the medical professional could move at will and position at the most appropriate location in the surgical work environment. Additionally, voice commands were available to show or hide the different tools by saying "show/hide" followed by "preoperative study/three-dimensional model/clinical history/medical illustration/surgical video".

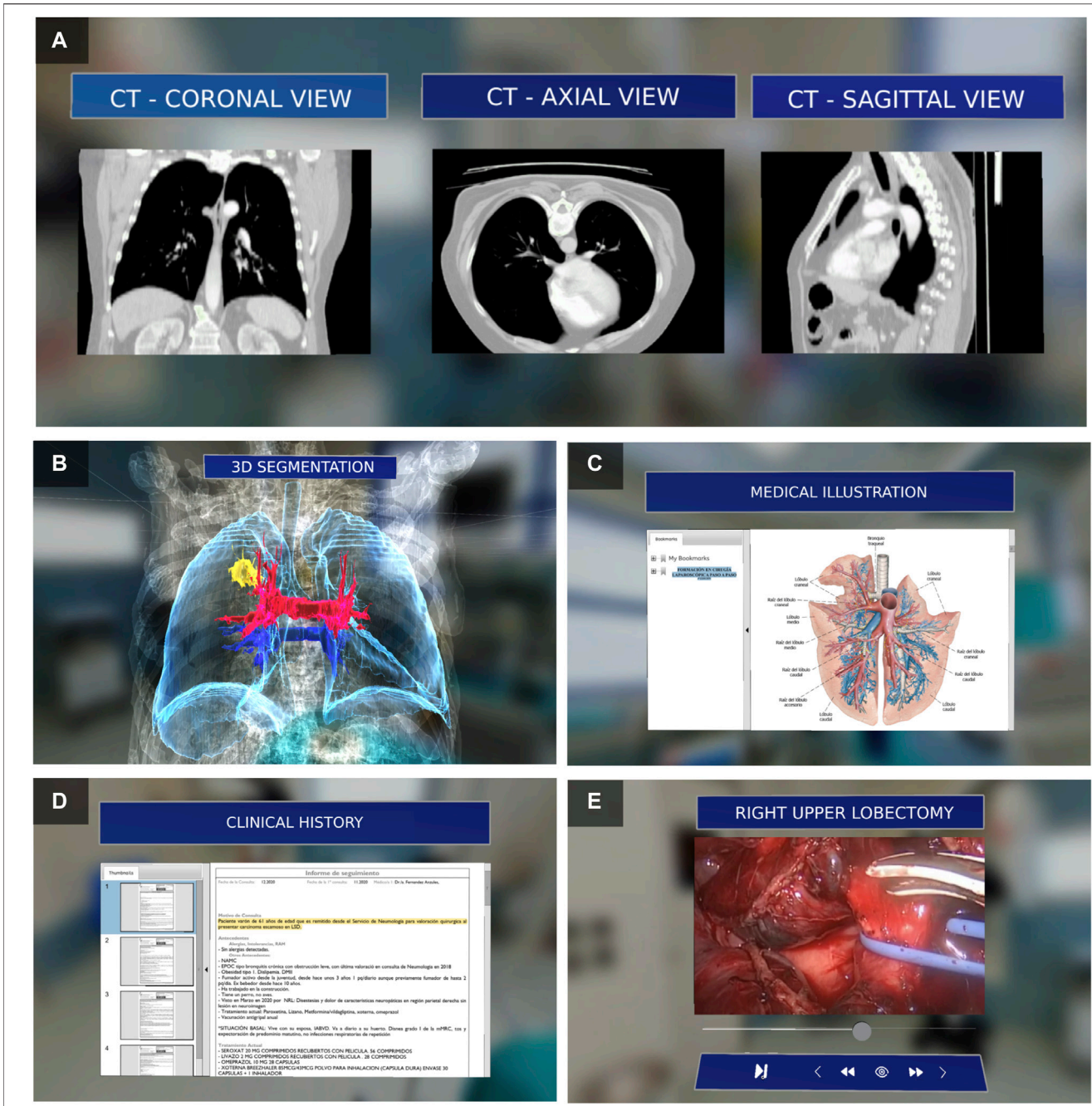

FIGURE 2 | Assistance content for surgical planning: **(A)** views of the preoperative imaging study, **(B)** 3D anatomical model of the patient, **(C)** medical illustration of the anatomy to be addressed during surgery, **(D)** medical history of the patient, and (C) reference video of the surgical procedure.

As a first step, the application loaded preoperative imaging studies (following the DICOM standard) from a local or remote file location. Each image view (axial, coronal, and sagittal) was displayed on an individual panel (**Figure 2A**), which allowed the user to navigate (forward or backward) within the set of available slices. These preoperative imaging studies have also been used to create 3D anatomical models. For this purpose, 3D Slicer (www.slicer.org), an open-source software package for medical image analysis, was used. The user can adjust the position, rotation, and scale of the 3D model to facilitate its visualisation. Interactions with the holograms were possible using gestures or voice control.

This application was tested during laparoscopic renal tumorectomy in a porcine model. This study was conducted in the experimental operating rooms of the Jesús Usón Minimally Invasive Surgery Centre in Cáceres (Spain) and was approved by the local animal welfare and ethics committee. Prior to surgery, an artificial renal tumour model was developed using a mixture of alginate and saline. Subsequently, a CT scan of the animal was

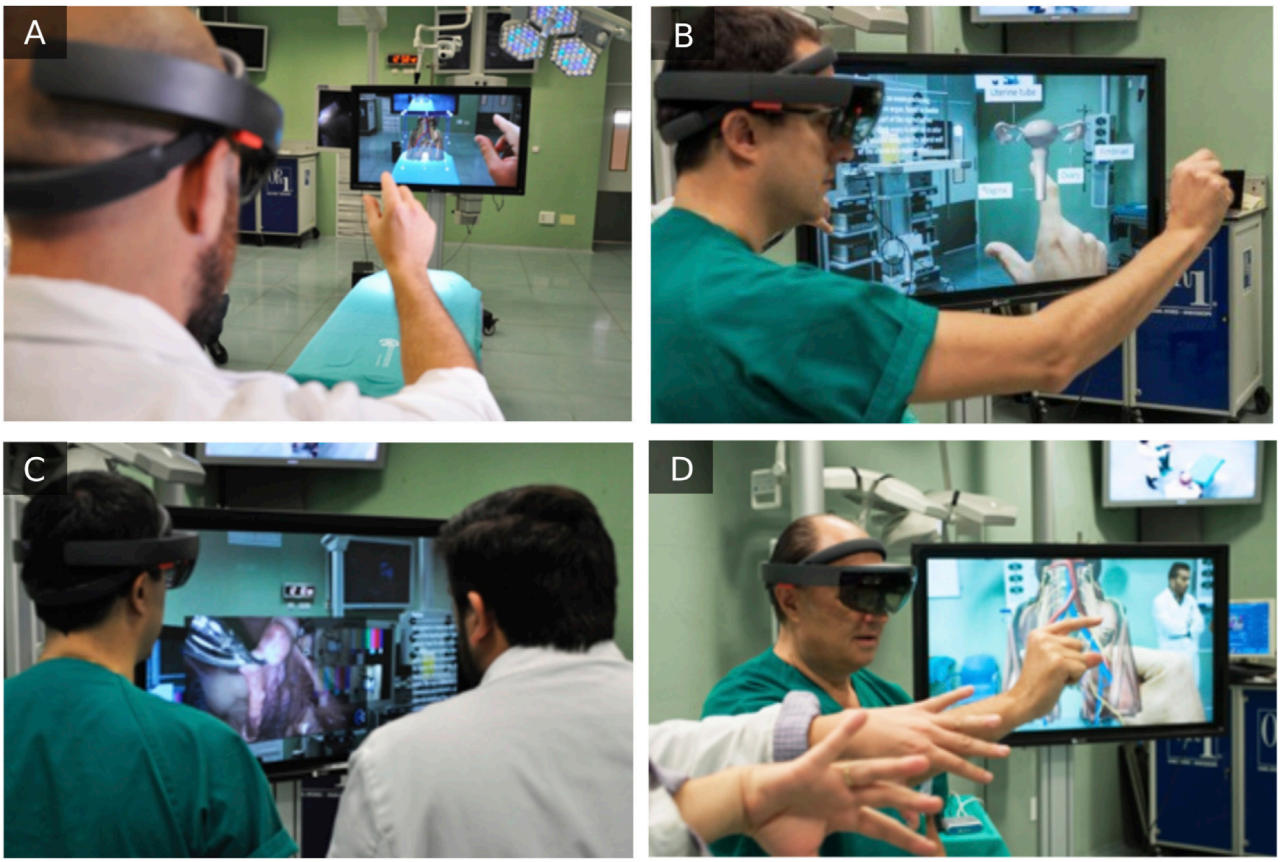

FIGURE 3 | Interaction with 3D anatomical models: the human pelvis **(A,D)** and uterus **(B)**. Visualisation of reference videos of related surgical procedures **(C)**.

obtained, and a 3D anatomical model from the preoperative study was created.

Finally, the MR application was used as a tool to assist in surgical planning during video-assisted right upper lobectomy, including systematic lymphadenectomy for squamous cell carcinoma in the right upper lobe. This procedure was performed at Cáceres University Hospital (Spain).

RESULTS

Training in Urology

A group of six surgeons, experienced in urology (>100 laparoscopic procedures performed), evaluated this application. All of them interacted with the different functionalities of the application: a 3D anatomical model of the human pelvis and its different anatomical systems (**Figures 3A,B,D**), reference videos of related surgical procedures, and medical illustrations (**Figure 3C**).

The surgeons found the MR solution to be a very useful tool for learning and studying the human pelvic anatomy and its application in urological surgeries, both individually and in groups (through broadcasting on external screens). They stated that the application facilitates the transfer of theoretical knowledge to actual practice and that this technology can potentially be useful for surgical planning and assistance during MIS.

Intuitively, the interactivity with preoperative imaging studies and the clarity and organisation of the 3D anatomical models were the most highly rated aspects by the surgeons (**Figure 4**). In contrast, the comfort of wearing the glasses obtained the lowest score.

Surgical Planning in Minimally Invasive Surgery

Two experienced laparoscopic surgeons (>100 laparoscopic procedures performed) tested the MR surgical planning application during laparoscopic renal tumorectomy in a porcine model. They were able to interact in the experimental OR with different views of the preoperative study (CT scan) and thus identify the lesion to be addressed during surgery (**Figure 5A**). In addition, interaction with the 3D anatomical model of the animal made it easier for the surgeons to plan the different phases of the surgical procedure, mainly in aspects related to the localisation of the renal artery and the planning of the tumour resection area. As support material for the surgical planning, the surgeons also had access to reference videos of the

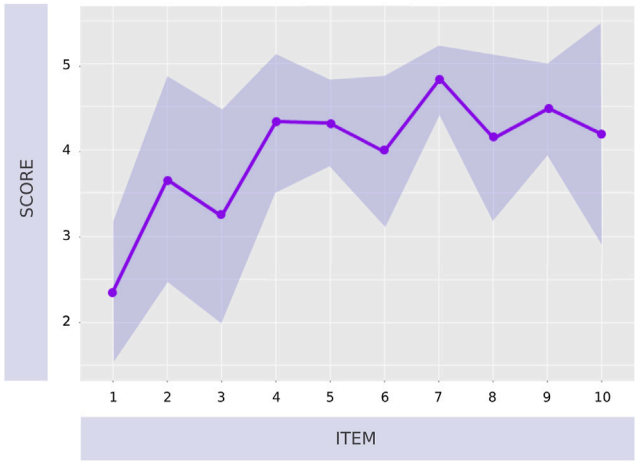

FIGURE 4 | Results of the subjective questionnaire regarding the usefulness and functionalities of the application and ergonomics of the MR device. Results are presented as mean values and standard deviations (shaded area around the main plot).

procedure to be performed and medical illustrations of the porcine anatomy (along with the different steps to be carried out during the renal tumorectomy).

Regarding the use of MR application for surgical planning in video-assisted right upper lobectomy, no complications were observed during surgery. Prior to surgery, the system allowed the surgeon to access the patient's medical history *in situ* and in real-time and to review the patients preoperative study (CT scan). The system also allowed the surgeon to readily visualise and manipulate a 3D model of the lung, with its respective vascular and bronchial elements, as well as the tumour to be addressed (**Figures 5B,C**).

The surgeon's view through the device was shared with the rest of the surgical team (**Figure 5D**). The surgeon placed the holographic models (with surgical planning information) behind the field of view of the operating table for possible consultation during the surgical procedure. As in the previous case, the surgeon reported some ergonomic aspects to be improved with regard to the MR device, such as the weight (579 g) and heat generated in the front side during its use.

DISCUSSION

In this study, we presented two applications based on mixed MR, oriented to surgical training and planning in MIS. The information was displayed using interactive holograms that were controlled by the user through hand gestures or voice commands. The contents could be viewed from the glasses themselves (one user) and could be shared with others via streaming, encouraging the exchange of information. Both applications allowed the user to choose the content to be displayed so that, once developed, they could be fed with specific training content or surgical planning content specific to each type of surgery. Some of the features shown in this study were foreseen in previous studies as promising applications of MR for surgical assistance (Gregory et al., 2018).

An important feature to consider in MR devices is that they overlap digital content with the real world. As a result, it is highly important to optimise the application so that the frame rate per second is as stable and as high as possible. Specifically, it is advisable to have a frame rate above 30 frames per second to avoid discomfort, nausea, or dizziness (Louis et al., 2019). These issues are not as crucial as in VR devices, but it is recommended to maintain these precautions for an optimal user experience when using MR applications.

During the development of the various modules that integrate the presented MR solutions, some aspects must be highlighted for further applications. For the visualisation module of the three views (axial, coronal, and sagittal) of the preoperative study, the content to be displayed did not present a high computational cost for the MR device. Although it internally processed the volumetric point cloud of the DICOM file, it simply rendered three planes, which did not increase the frame rate.

Regarding the module for generating the 3D model from the preoperative imaging study, it allowed the scaling, rotation, and positioning of the model to the user's preference for a better perception of the patient's anatomy. The generation of this content involved a certain computational cost depending on the model; therefore, caution should be exercised when segmenting the anatomical areas of interest, as well as in the subsequent reduction of polygons of the resulting mesh. In the cases described in this study, it was not necessary to reduce the mesh of the 3D model obtained. However, it is suggested to use standard materials provided by Microsoft's Mixed Reality ToolKit framework.

For the visualisation of reference surgical videos, it would be possible to temporarily label them and separate their content into chapters, thus allowing easier access to the different steps involved in the surgery. The maximum resolution for viewing videos on the device was 1,280 × 720 pixels. Therefore, although the application allowed videos to be played at higher resolutions, it was recommended to insert videos in this resolution to avoid overloading the performance and to maintain the desired frame rate (30 FPS).

Regarding the findings obtained from the experience of users with the developed MR application for training in urology, there were three surgeons who experienced a steep learning curve regarding the interaction with the MR device. This can be seen in items 2 and 3 of the subjective evaluation questionnaires (**Figure 4**). This type of technology introduces new concepts and methods of interaction for users that require some familiarization time (Hurson et al., 2007; Maasthi et al., 2020). The second generation of the HoloLens glasses (HoloLens v2) could help solving this issue due to its advanced features to enhance user interaction. Another aspect of interaction that is challenging for users is the use of voice commands (**Figure 4**, item 3). Since the commands have been implemented in English language, in order to facilitate the universality of the applications, it could lead to some complications for non-native English users (Hurson

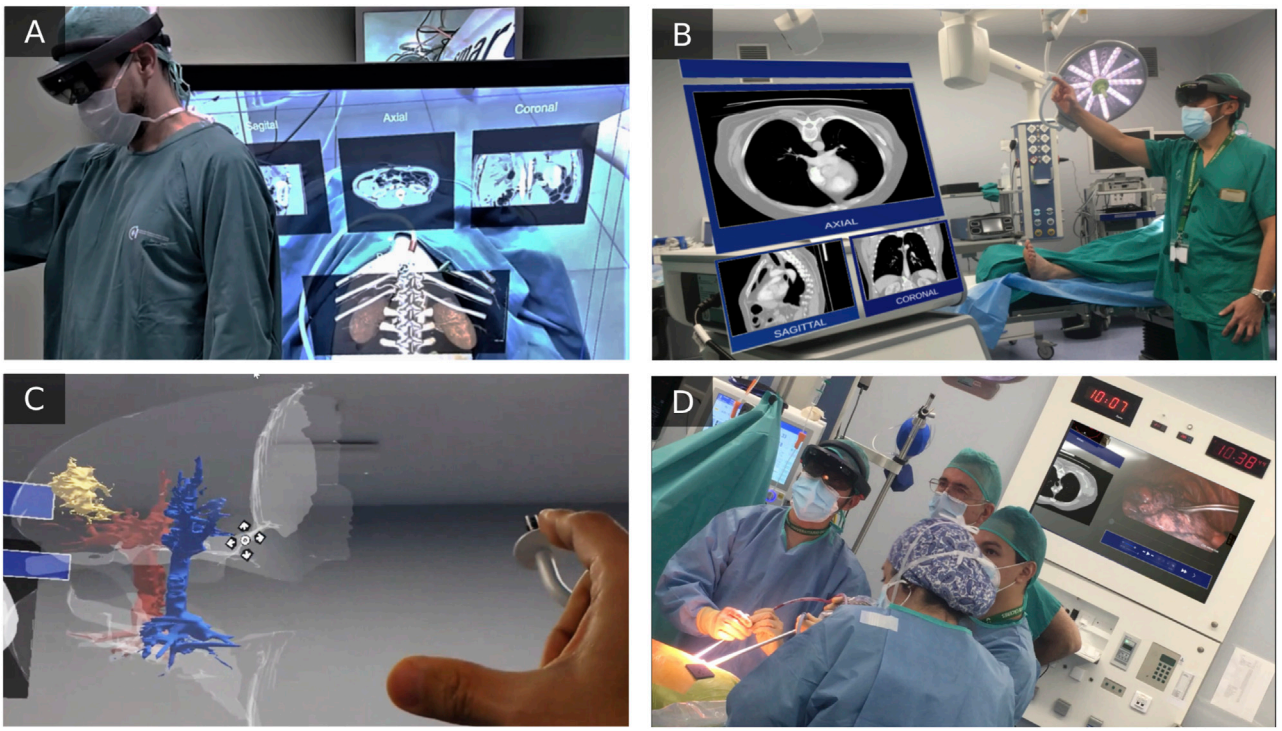

FIGURE 5 | (A) Interaction with preoperative study views and their corresponding 3D model for planning the renal tumorectomy in the porcine model. **(B)** Axial, coronal, and sagittal views of the preoperative study for planning the video-assisted lobectomy. **(C)** Interaction with the patient's anatomical 3D model with information on the vascular structures and the tumor to be addressed during the surgical procedure. The surgeon placed the model according to the patient's position **(D)** An additional screen shared the surgical planning assistance contents with the rest of the surgical team, such as the CT study and a reference video of the surgical procedure to be performed.

et al., 2007). As for the ergonomic aspects of the device, both the users of the training application (**Figure 4**, item 1) and those of the surgical planning application considered that this is a feature that needs to be improved (Turini et al., 2018). The mixed reality device is still relatively uncomfortable to wear, especially when used for a prolonged period of time, mainly due to its weight and the heat it may cause on the user's forehead. As future work, most surgeons proposed the extension of training models include additional anatomical structures such as the prostate. This would improve anatomical training and preparation for surgeries, such as laparoscopic prostatectomy. Some users experienced a steeper learning curve concerning the interaction with the MR device. The second generation of HoloLens glasses (HoloLens v2) could help solve this issue because of its advanced features to enhance user intuitiveness.

Few applications have been found for MIS training using MR technology. A study by Amparore et al. compared 3D virtual reconstruction with 3D printing of organs, such as the kidney and prostate, to determine which method was more suitable for visualisation and localisation of tumour lesions (Amparore et al., 2021). They concluded that MR is the preferred choice for surgical training and planning, with HoloLens MR glasses being considered the most adequate technology for surgical planning.

The MR solutions for surgical planning presented in this study were tested during two different MIS procedures, in which surgeons provided feedback on their experiences. This will allow us to make necessary improvements to enhance interaction and user experience in future applications. No complications were reported in either surgery group. In both cases, the MR solutions allowed navigation over the CT studies, as well as the visualisation of real 3D models of the patient's anatomy. In the porcine model, the renal anatomy was shown together with the artificial tumour to be excised. In the adult patient, the lung anatomy was shown in combination with the vascular system, bronchi, and the tumour to be treated. The surgeon also had access to reference surgical videos, as well as different documents with the patient's clinical history and reference anatomical illustrations. The surgeon's vision, together with the information in the form of holograms, was shared on the screens of the OR via streaming. It should be noted that this video streaming suffered a slight time delay of approximately one second during the entire retransmission using the Microsoft Windows Device Portal software.

The streaming option of our tool allowed all personnel inside and outside the OR to directly see what the surgeon saw. This feature was also reported by Gregory et al. during surgery for the implantation of a shoulder prosthesis (Gregory et al., 2018). This tool can also be used in videoconferences during live surgeries as a

method of immersion in the surgery. Another feature to point out is the possibility of recording the surgery from the surgeon's perspective. In the event of a complication during surgery, this would allow the surgeon to access the recording and review it in more detail.

As it has also been indicated for the training application, surgeons have reported some ergonomic aspects of MR devices that should be improved. Although it did not cause significant discomfort, they stated that the device (HoloLens v1) has a weight that can be uncomfortable if worn for several hours. The new model of this device (HoloLens v2) already has a lighter design. Additionally, surgeons indicated that the device generated heat in the forehead area. This can be solved by increasing the separation.

The most complete MR-based surgical planning solution found in scientific publications offers information in the form of interactive holograms of both a 3D model and images of the different MRI/CT views and even a component to display intraoperative information (e.g., intraoperative ultrasound) (Cartucho et al., 2020). However, this application was not used during any actual surgery as a surgical planning tool, but only a pilot study with a phantom was used to collect data through a survey. Other MR solutions (with less functionality) were used retrospectively as surgical planning tools in patients undergoing thoracic surgery (Perkins et al., 2020). This application allowed the visualisation of only one of the three views of the preoperative imaging study, as well as the manipulation of 3D models obtained from it. They used a simulation of lung motion by animating the 3D model, which facilitated the estimation of the tumour location. However, it does not allow displaying multiple views of preoperative imaging studies or other additional information to support surgical planning, such as the patient's medical history, medical illustrations, or videos of similar surgical interventions.

The largest study on the use of MR applications in assistance during laparoscopic surgeries has been in a comparative study of 50 laparoscopic nephrectomies with MR assistance versus 50 similar surgeries without it (Li et al., 2020). The results concluded that MR technology can improve the success rate in laparoscopic surgeries, as well as offer added value in clinical applications such as planning, navigation, consultation, teaching, and patient communication.

Other solutions made use of MR as a substitute for conventional screens in the OR, capturing the endoscopic video directly on the device in the form of a hologram, thus allowing the surgeon to act in a more comfortable position during surgery (Deib et al., 2018; Al Janabi et al., 2020). Some MR clinical applications seek to spatially reference 3D holograms on anatomical elements, thus being able to overlap virtual information with reality (Heinrich et al., 2019; Liu et al., 2020).

To the best of our knowledge, the present MR-based surgical planning solution is the first to be applied during video-assisted lobectomy. It is important to note the novelty of the inclusion of hologram visualisation of the volumetric point cloud of the 3D surgical planning model. The application was iteratively refined after its evaluation in an experimental model used by different surgeons, optimising the interaction and usability.

This study has some limitations to be taken into account for further research, such as the few cases in which the developed solutions have been applied, as well as the limited number of surgeons who have been able to test them. As reflected in the results, the learning curve of this technology is an aspect to be considered, as these MR devices are not common in the day-to-day work of surgeons. Although the method of interaction is optimal and allows the surgeon to maintain sterile conditions (since there is no real contact with the elements), the lack of tangible hardware devices to interact with, such as a computer mouse, joystick, or tablet, requires a more pronounced adaptation process.

Several future studies are required to improve the proposed solutions. One of our main objectives is to optimise the visualisation performance of volumetric point clouds in 3D models. To achieve this, different possible solutions will be analysed to improve the visualisation of the DICOM files in real-time and the performance of the MR device itself. In contrast, we propose the development of a customised method for retransmission of the surgeon's view together with holograms via streaming to overcome the latencies presented by the current method. This solution could be the first step toward using MR glasses as a monitor for the laparoscopic camera with a real-time video feed, improving ergonomics for the surgeon during surgery. Additionally, the solutions presented will be adapted for use with the HoloLens v2, so that its eye-tracking system can be used for interaction with holographic models. This allows the user to provide direct feedback about the element he/she is looking at. Similarly, these data can be analysed for the generation of heat maps with the areas most consulted by medical professionals compared to those consulted by medical students and residents. Once we have a final version of the applications for training and surgical planning in MIS, incorporating all the improvements and feedback obtained in this study, several specific aspects related to user experience could be validated. The mental and physical workload of users with regard to the use of these applications could be determined using a NASA-TLX (Task Load Index) questionnaire (Turini et al., 2018). Similarly, the ultimate system usability or user's interest/enjoyment could be analyzed by means of the System Usability Scale (SUS) (Gregory et al., 2018) or the Intrinsic Motivation Inventory (IMI) scale (Galati et al., 2020), respectively.

CONCLUSION

The MR-based solution for surgical training presented in this study is a useful tool for urological anatomy training, facilitating the transfer of this knowledge to actual clinical practice. The solution developed for assistance during surgical planning provides real-time access to essential patient information, such as preoperative imaging studies, the 3D surgical planning model, or the clinical history, without losing the sterility of the surgical act. This tool has been successfully tested during laparoscopic tumorectomy in an experimental model and video-assisted lobectomy. The surgeon's view can be shared for communication and learning purposes, as well as for a later

review of possible surgical complications. However, further studies are needed to validate its clinical application comprehensively.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding authors.

ETHICS STATEMENT

The animal study was reviewed and approved by the competent local Animal Welfare and Ethics Committee.

AUTHOR CONTRIBUTIONS

JS, CP, RF, and FS: Conceptualization, Methodology. CPM: Software. JS, CP, and RF: Investigation. JS, CP: Writing—Original Draft. RF, FS: Writing—Review ; Editing. FS: Supervision, Funding acquisition. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

This work has been partially funded by the Spanish Ministry of Science and Innovation, the European Regional Development Fund (FEDER) "A way to make Europe" and the Junta de Extremadura (Spain) (TA18023, GR18199, CPI-2019-33-1-TRE-14).

REFERENCES

Al Janabi, H. F., Aydin, A., Palaneer, S., Macchione, N., Al-Jabir, A., Khan, M. S., et al. (2020). Effectiveness of the HoloLens Mixed-Reality Headset in Minimally Invasive Surgery: a Simulation-Based Feasibility Study. *Surg. Endosc.* 34 (3), 1143–1149. doi:10.1007/s00464-019-06862-3

Al-Zu'bi, S., Al-Ayyoub, M., Jararweh, Y., and Shehab, M. A. (2017). Enhanced 3D Segmentation Techniques for Reconstructed 3D Medical Volumes: Robust and Accurate Intelligent System. *Proced. Comp. Sci.* 113, 531–538. doi:10.1016/j.procs.2017.08.318

Amparore, D., Pecoraro, A., Checcucci, E., De Cillis, S., Piramide, F., Volpi, G., et al. (2021). 3D Imaging Technologies in Minimally-Invasive Kidney and Prostate Cancer Surgery: Which Is the Urologists' Perception. *Minerva Urol. Nephrol.* 26 [Online ahead of print]. doi:10.23736/S2724-6051.21.04131-X

Cartucho, J., Shapira, D., Ashrafian, H., and Giannarou, S. (2020). Multimodal Mixed Reality Visualisation for Intraoperative Surgical Guidance. *Int. J. CARS* 15, 819–826. doi:10.1007/s11548-020-02165-4

Daly, S. C., Wilson, N. A., Rinewalt, D. E., Bines, S. D., Luu, M. B., and Myers, J. A. (2014). A Subjective Assessment of Medical Student Perceptions on Animal Models in Medical Education. *J. Surg. Educ.* 71 (1), 61–64. doi:10.1016/j.jsurg.2013.06.017

Deib, G., Johnson, A., Unberath, M., Yu, K., Andress, S., Qian, L., et al. (2018). Image Guided Percutaneous Spine Procedures Using an Optical See-Through Head Mounted Display: Proof of Concept and Rationale. *J. Neurointervent Surg.* 10, 1187–1191. doi:10.1136/neurintsurg-2017-013649

DeMasi, S. C., Katsuta, E., and Takabe, K. (2016). Live Animals for Preclinical Medical Student Surgical Training. *Edorium J. Surg.* 3 (2), 24–31. doi:10.5348/S05-2016-16-OA-6

Forgione, A., and Guraya, S. Y. (2017). The Cutting-Edge Training Modalities and Educational Platforms for Accredited Surgical Training: a Systematic Review. *J. Res. Med. Sci.* 22, 51. doi:10.4103/jrms.JRMS_809_16

Galati, R., Simone, M., Barile, G., De Luca, R., Cartanese, C., and Grassi, G. (2020). Experimental Setup Employed in the Operating Room Based on Virtual and Mixed Reality: Analysis of Pros and Cons in Open Abdomen Surgery. *J. Healthc. Eng.* 2020, 1–11. doi:10.1155/2020/8851964

Goris, J., Jalink, M. B., and Ten Cate Hoedemaker, H. O. (2014). Training Basic Laparoscopic Skills Using a Custom-Made Video Game. *Perspect. Med. Educ.* 3 (4), 314–318. doi:10.1007/s40037-013-0106-8

Gregory, T. M., Gregory, J., Sledge, J., Allard, R., and Mir, O. (2018). Surgery Guided by Mixed Reality: Presentation of a Proof of Concept. *Acta Orthopaedica* 89 (5), 480–483. doi:10.1080/17453674.2018.1506974

Heinrich, F., Schwenderling, L., Becker, M., Skalej, M., and Hansen, C. (2019). HoloInjection: Augmented Reality Support for CT-guided Spinal Needle Injections. *Healthc. Tech. Lett.* 6 (6), 165–171. doi:10.1049/htl.2019.0062

Hu, H., Shao, Z., Ye, L., and Jin, H. (2019). Application of Mixed Reality Technology in Surgery. *Int. J. Clin. Exp. Med.* 12 (4), 3107–3133.

Hurson, C., Tansey, A., O'Donnchadha, B., Nicholson, P., Rice, J., and McElwain, J. (2007). Rapid Prototyping in the Assessment, Classification and Preoperative Planning of Acetabular Fractures. *Injury* 38, 1158–1162. doi:10.1016/j.injury.2007.05.020

Jacobson, S., Epstein, S. K., Albright, S., Ochieng, J., Griffiths, J., Coppersmith, V., et al. (2009). Creation of Virtual Patients from CT Images of Cadavers to Enhance Integration of Clinical and Basic Science Student Learning in Anatomy. *Med. Teach.* 31 (8), 749–751. doi:10.1080/01421590903124757

Jayender, J., Xavier, B., King, F., Hosny, A., Black, D., Pieper, S., et al. (2018). A Novel Mixed Reality Navigation System for Laparoscopy Surgery. *Med. Image Comput. Comput. Assist. Interv.* 11703, 72–80. doi:10.1007/978-3-030-00937-3_9

Kerr, B., and O'Leary, J. P. (1999). The Training of the Surgeon: Dr. Halsted's Greatest Legacy. *Am. Surg.* 65 (11), 1101–1102.

Lahanas, V., Loukas, C., Smailis, N., and Georgiou, E. (2015). A Novel Augmented Reality Simulator for Skills Assessment in Minimal Invasive Surgery. *Surg. Endosc.* 29 (8), 2224–2234. doi:10.1007/s00464-014-3930-y

Langridge, B., Momin, S., Coumbe, B., Woin, E., Griffin, M., and Butler, P. (2018). Systematic Review of the Use of 3-Dimensional Printing in Surgical Teaching and Assessment. *J. Surg. Educ.* 75 (1), 209–221. doi:10.1016/j.jsurg.2017.06.033

Lee, S. C., Fuerst, B., Tateno, K., Johnson, A., Fotouhi, J., Osgood, G., et al. (2017). Multi-modal Imaging, Model-based Tracking, and Mixed Reality Visualisation for Orthopaedic Surgery. *Healthc. Technol. Lett.* 4, 168–173. doi:10.1049/htl.2017.0066

Li, G., Dong, J., Wang, J., Cao, D., Zhang, X., Cao, Z., et al. (2020). The Clinical Application Value of Mixed-reality-assisted Surgical Navigation for Laparoscopic Nephrectomy. *Cancer Med.* 9 (15), 5480–5489. doi:10.1002/cam4.3189

Liu, H., Wu, J., Tang, Y., Li, H., Wang, W., Li, C., et al. (2020). Percutaneous Placement of Lumbar Pedicle Screws via Intraoperative CT Image-Based Augmented Reality-Guided Technology. *J. Neurosurg. Spinespine* 32 (4), 1–6. doi:10.3171/2019.10.SPINE19969

Louis, T., Troccaz, J., Rochet-Capellan, A., and Bérard, F. (2019). "Is it Real? Measuring the Effect of Resolution, Latency, Frame Rate and Jitter on the Presence of Virtual Entities," in Proceedings of the 2019 ACM International Conference on Interactive Surfaces and Spaces (ISS '19), Soul, South Korea, November 2019 (New York, USA: Association for Computing Machinery), 5–16. doi:10.1145/3343055.3359710

Maasthi, M. J., Gururaj, H., Janhavi, V., Harshitha, K., and Swathi, B. (2020). "An Interactive Approach Deployed for Rhinoplasty Using Mixed Reality," in 2020 International Conference on COMmunication Systems & NETworkS (COMSNETS), Bengaluru, India, 7-11 Jan. 2020 (IEEE), 680–682. doi:10.1109/comsnets48256.2020.9027491

Perkins, S. L., Krajancich, B., Yang, C.-F. J., Hargreaves, B. A., Daniel, B. L., and Berry, M. F. (2020). A Patient-specific Mixed-Reality Visualization Tool for

Thoracic Surgical Planning. *Ann. Thorac. Surg.* 110 (1), 290–295. doi:10.1016/j.athoracsur.2020.01.060

Rocha e Silva, R., Lourenção, A., Jr, Goncharov, M., and Jatene, F. B. (2016). Low Cost Simulator for Heart Surgery Training. *Braz. J. Cardiovasc. Surg.* 31 (6), 449–453. doi:10.5935/1678-9741.20160089

Rosenberg, B. H., Landsittel, D., and Averch, T. D. (2005). Can Video Games Be Used to Predict or Improve Laparoscopic Skills. *J. Endourology* 19 (3), 372–376. doi:10.1089/end.2005.19.372

Sadeghi, A. H., Bakhuis, W., Van Schaagen, F., Oei, F. B. S., Bekkers, J. A., Maat, A. P. W. M., et al. (2020). Immersive 3D Virtual Reality Imaging in Planning Minimally Invasive and Complex Adult Cardiac Surgery. *Eur. Hear. J. - Digit Heal* 1 (1), 62–70. doi:10.1093/ehjdh/ztaa011

Sánchez-Margallo, F. M., and Sánchez-Margallo, J. A. (2015). "Computer-Assisted Minimally Invasive Surgery: Image-Guided Interventions and Robotic Surgery," in *Computer-Assisted Surgery*. Editor X. Chen (New York, NY: Nova Science Publishers, Inc.), 43–94.

Sánchez-Margallo, F. M., Sánchez-Margallo, J. A., Cristo, A., Rodríguez, A., and Suárez, M. (2018). Application of Mixed Reality Technology for Surgical Training in Urology. *Surg. Endosc.* 32, S655. doi:10.1007/s00464-019-06728-8

Sánchez-Margallo, F. M., Sánchez-Margallo, J. A., Suárez, M., Cristo, A., Rodríguez, A., and Moyano-Cuevas, J. L. (2018). Tecnologías de control gestual y realidad aumentada para la asistencia en cirugía de mínima invasión. *Cirugía Española* 96, 1. (Espec Congr).

Sánchez-Margallo, F. M., Durán Rey, D., Serrano Pascual, Á., Mayol Martínez, J. A., and Sánchez-Margallo, J. A. (2021). Comparative Study of the Influence of Three-Dimensional versus Two-Dimensional Urological Laparoscopy on Surgeons' Surgical Performance and Ergonomics: A Systematic Review and Meta-Analysis. *J. Endourology* 35 (2), 123–137. doi:10.1089/end.2020.0284

Sappenfield, J. W., Smith, W. B., Cooper, L. A., Lizdas, D., Gonsalves, D. B., Gravenstein, N., et al. (2018). Visualization Improves Supraclavicular Access to the Subclavian Vein in a Mixed Reality Simulator. *Anesth. Analgesia* 127, 83–89. doi:10.1213/ane.0000000000002572

Scott, D. J., Cendan, J. C., Pugh, C. M., Minter, R. M., Dunnington, G. L., and Kozar, R. A. (2008). The Changing Face of Surgical Education: Simulation as the New Paradigm. *J. Surg. Res.* 147 (2), 189–193. doi:10.1016/j.jss.2008.02.014

Smith, R. T., Clarke, T. J., Mayer, W., Cunningham, A., Matthews, B., and Zucco, J. E. (2020). Mixed Reality Interaction and Presentation Techniques for Medical Visualisations. *Adv. Exp. Med. Biol.* 1260, 123–139. doi:10.1007/978-3-030-47483-6_7

Turini, G., Condino, S., Parchi, P., Viglialoro, R., Piolanti, N., Gesi, M., Ferrari, M., and Ferrari, V. (2018). "A Microsoft HoloLens Mixed Reality Surgical Simulator for Patient-specific Hip Arthroplasty Training," in International Conference on Augmented Reality, Virtual Reality and Computer Graphics, Otranto, Italy, 14 July 2018 (New York City, USA: Springer, Cham), 201–210. AVR.

Viglialoro, R. M., Condino, S., Turini, G., Carbone, M., Ferrari, V., and Gesi, M. (2021). Augmented Reality, Mixed Reality, and Hybrid Approach in Healthcare Simulation: A Systematic Review. *Appl. Sci.* 11 (5), 2338. doi:10.3390/app11052338

Williams, M. A., McVeigh, J., Handa, A. I., and Lee, R. (2020). Augmented Reality in Surgical Training: a Systematic Review. *Postgrad. Med. J.* 96 (1139), 537–542. doi:10.1136/postgradmedj-2020-137600

Zuckerman, J. D., Wise, S. K., Rogers, G. A., Senior, B. A., Schlosser, R. J., and DelGaudio, J. M. (2009). The Utility of Cadaver Dissection in Endoscopic Sinus Surgery Training Courses. *Am. J. Rhinology allergy* 23 (2), 218–224. doi:10.2500/ajra.2009.23.3297

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Virtual Reality

MINI REVIEW
published: 08 November 2021
doi: 10.3389/frvir.2021.595771

Enhancing Upper Limb Rehabilitation of Stroke Patients With Virtual Reality: A Mini Review

Julie Bui[1,2,3,4]*, *Jacques Luauté*[1,2,3,4] *and Alessandro Farnè*[1,2,3,5]

[1]Integrative Multisensory Perception Action and Cognition Team - ImpAct, Lyon Neuroscience Research Center, INSERM U1028, CNRS U5292, Lyon, France, [2]University UCBL Lyon 1, University of Lyon, Lyon, France, [3]Hospices Civils de Lyon, Neuro-immersion, Mouvement et Handicap, Lyon, France, [4]Trajectoires, Lyon Neuroscience Research Center, INSERM U1028, CNRS U5292, Lyon, France, [5]Center for Mind/Brain Sciences (CIMeC), University of Trento, Trento, Italy

Upper limb motor impairment following stroke is a common condition that impacts significantly the independence and quality of life of stroke survivors. In recent years, scholars have massively turned to virtual reality (VR) to develop more effective rehabilitation approaches. VR systems are promising tools that can help patients engage in intensive, repetitive and task-oriented practice using new technologies to promote neuroplasticity and recovery. Multiple studies have found significant improvements in upper limb function for patients using VR in therapy, but the heterogeneity of methods and tools employed make the assessment of VR efficacy difficult. Here we aimed to assess the potential of VR as a therapy tool for upper limb motor impairment and to provide initial assessment of what is the added value of using VR to both patients and clinicians. Our mini-review focuses the work published since the Cochrane review (2017) and suggests that VR may be particularly effective when used in combination to conventional rehabilitation approaches. We also highlight key features integrated in VR systems that appear to influence rehabilitation and can help maximizing therapy outcomes, if exploited properly. We conclude that although promising results have already been gathered, more focused research is needed to determine the optimal conditions to implement VR in clinical settings in order to enhance therapy and to better define and leverage the true potential of VR. The rapid pace of technological development and increasing research interest toward VR-based therapy will help providing extensive knowledge and lead to rapid advancements in the near future.

Keywords: stroke, upper limb, rehabilitation, virtual reality therapy, naturalistic

OPEN ACCESS

Edited by:
Marientina Gotsis, University of Southern California, United States

Reviewed by:
Belén Rubio Ballester, Institute for Bioengineering of Catalonia (IBEC), Spain
Fazel Naghdy, University of Wollongong, Australia

***Correspondence:**
Julie Bui
buijulie15@gmail.com

Specialty section:
This article was submitted to Virtual Reality in Medicine, a section of the journal Frontiers in Virtual Reality

Received: *17 August 2020*
Accepted: *21 October 2021*
Published: *08 November 2021*

Citation:
Bui J, Luauté J and Farnè A (2021) Enhancing Upper Limb Rehabilitation of Stroke Patients With Virtual Reality: A Mini Review. Front. Virtual Real. 2:595771. doi: 10.3389/frvir.2021.595771

INTRODUCTION

Stroke is the second leading cause of death and the third most common cause of disability worldwide (Feigin et al., 2017). The interruption of blood supply to the brain occurring during stroke can cause several physical and cognitive impairments that may highly affect patients' participation in activities of daily living (ADL) and their quality of life. In particular, hemiplegia represents the most prevalent impairment for stroke patients, resulting in impaired arm and hand movements, with deficits in motor control and grip strength. Upper limb motor abilities often remain affected after a stroke and become a chronic condition. In stroke patients with complete initial hemiplegia, longitudinal observational studies showed a very low recovery rate for the upper-limb, and the absence of functional recovery when the impairment remains complete after a delay of 3 weeks (Wade and |

Hewer, 1987). Recent modeling showed that the probability to recover upper-limb motricity remains extremely low after 12 weeks post-onset (van der Vliet et al., 2020). Upper limb function plays a major role when performing ADL as many activities require the coordinated use of both hands (Ekstrand et al., 2016), and is strongly associated with the quality of life of stroke survivors (Nichols-Larsen et al., 2005). Thus, rehabilitation of upper limbs represents a major need and challenge in stroke management and motor rehabilitation is recommended to be initiated early in order to enhance the recovery process (Duncan et al., 2005).

Given the urgent need for effective approaches, innovative tools are currently being investigated as new treatment methods. A number of new technologies have emerged in the recent years and are becoming more accessible to rehabilitation clinics. In particular, virtual reality (VR) is being regarded as a promising treatment tool, and presents characteristics that may be beneficial for therapists' intervention and for the functional recovery of stroke patients (Bohil et al., 2011; Massetti et al., 2018). VR can be defined as "the use of interactive simulations created with computer hardware and software to present users with opportunities to engage in environments that appear and feel similar to real world objects and events" (Weiss et al., 2004, Introduction section, para. 2). Users can interact with a virtual environment using controllers, joysticks or a computer mouse to manipulate virtual objects. They can also be represented by an avatar within the virtual environment, whose movements will match those of the users by means of motion capture technology (Bohil et al., 2011). More particularly, VR systems may help stroke survivors engage in a virtual environment with sensory stimulations in multiple forms such as visual, auditory or haptic that can simulate real-life situations and help the practice of goal-oriented tasks in environments similar to the real world (Klinger, 2008).

For these reasons, the exploration of VR usages for clinical applications is increasing rapidly, with an ever-growing number of publications in the past few years (Garrett et al., 2018). In stroke rehabilitation research, the use of VR is often compared with conventional therapy (CT) delivered by physical therapists and occupational therapists. The updated Cochrane review by Laver et al. concluded in 2017 that VR-based therapy was not more beneficial than CT for improving upper limb function. Specifically, they report that VR "may be beneficial in improving upper limb function and activities of daily living function when used as an adjunct to usual care (to increase overall therapy time)." (Laver et al., 2017, p. 2). Nowadays, VR is indeed used in clinical settings for rehabilitation purposes alongside CT and associated technologies have become more and more accessible and widespread.

These factors lead to frequent updates in rehabilitation research regarding the use of VR, its efficacy in motor recovery, and how it may be implemented in clinical settings. The aims of this mini-review are thus to 1) assess the current results regarding efficacy of VR therapy in upper limb rehabilitation following stroke and 2) start identifying potential characteristics of VR-based therapy that can be beneficial for upper limb rehabilitation for both clinicians and patients. As VR-based therapy is being extensively investigated, we aim to specifically provide a brief update on the growing state of research for upper limb rehabilitation, in order to inform on the recent developments on VR in rehabilitation but also to provide insights on how this field may progress in the future.

Current Evidence Regarding the Efficacy of Virtual Reality-Based Therapy

Traditional methods for the rehabilitation of the upper limb in clinical centers are usually provided by physical and occupational therapists, including ADL training. Recent studies have reached the same conclusions as the Cochrane systematic review (Laver et al., 2017). Investigating a VR system specifically designed for upper limb rehabilitation and VR as a stand-alone therapy, Schuster-Amft et al. (2018) found that chronic stroke patients in both the experimental group and the control group improved their hand dexterity, arm function and independence in ADL after a 4-week treatment, with no between-group differences after the same amount of therapy. In line with the results of Laver et al.s' review, Hung et al. (2019) observed that a VR-based training combined with CT also did not lead to different results when compared with CT only, for the same amount of therapy and with similar training contents. Brunner et al. (2017) compared improvements in upper limb motor function after additional VR training with additional conventional rehabilitation, both provided as an adjunct to standard therapy, but did not observe significant differences between the two modalities, although they both led to significant improvement of all outcomes for subacute stroke patients, further suggesting that VR training is simply as effective as CT in upper limb rehabilitation.

However, several authors recently found conflicting results. Significantly greater improvements in upper limb motor recovery and gross manual dexterity were observed in several studies in either subacute or chronic stroke patients who benefited from VR training in addition to conventional treatments, as compared to patients who only had CT (Aşkın et al., 2018; Ikbali Afsar et al., 2018; Lee et al., 2018; Rogers et al., 2019). These improvements could at least in part result from the increased therapy time, as observed in Aşkın et al.'s study where patients in the experimental group had one more hour of therapy every day than the control group, but these studies may also suggest that VR-based therapy is an effective tool, especially when combined with CT. Importantly, Wang et al. (2017) and Kiper et al. (2018) compared stroke patients undergoing VR training along with CT with patients who had CT only, for the same amount of therapy in both groups. They observed a significantly greater improvement on motor function in the experimental group where VR training was added to canonical therapy. The recent work of Ain et al. (2021) also indicate improvements on upper extremity function in favor of the experimental group who underwent Xbox Kinect-based training and CT for the same duration.

The conflicting results observed in these recent studies could result from differences in their experimental protocols, which differ in number and frequency of the training sessions. We note

that in studies where no difference was observed between VR and CT groups, patients received two to five rehabilitation sessions a week. In studies where a greater improvement was observed for VR groups, all patients received a more intensive therapy with sessions on 5 days a week. The various training frequencies thus resulted in different training time. For example, patients in Schuster-Amft et al.'s study (2018) received a total of 12 h of VR-based training during 4 weeks, with no between-groups differences observed, while patients in Wang et al.'s study (2017) received a total of 45 h of therapy during 4 weeks, with greater improvements for the VR group. These results may indicate a dose-effect relationship in VR therapy that needs further investigation to determine more precisely how many hours of VR per week are needed to make VR-based therapy effective in upper limb rehabilitation and how the dose impacts the outcomes.

Besides differing in intensity, different VR systems were also used in the studies, possibly concurring to explain part of the reported outcome differences. In addition to hand movement tracking, some of the systems presented distinct features, such as enhanced feedback and enriched virtual environment (Kiper et al., 2018; Rogers et al., 2019), the use of a sensorized real object (Kiper et al., 2018), hand-held objects (Rogers et al., 2019), the use of a controller (Lee et al., 2018), or an avatar hand of the patient's movements appearing on the screen (Wang et al., 2017). Thus, groups of patients interacted differently with the virtual environments during their training depending on the system they used. Along with the development of VR technologies and features, we deem important that future studies will investigate if and what specific features of VR, such as augmented feedback, or the use of physical objects that patients grasp, may favour rehabilitation outcomes.

We additionally note that the studies included patients in the subacute phase, or in the chronic phase of stroke. Considering the delay since stroke, no stringent difference was noted between studies showing an additional beneficial effect of VR as compared to CT vs studies showing no additional effect of VR compared to CT. Interestingly four out of seven studies showing an additional effect of VR concerned patients in the sub-acute phase (Ikbali Afsar et al., 2018; Lee et al., 2018; Rogers et al., 2019; Wang et al., 2017), two studies concerned chronic patients (delay since stroke above 6 months) (Ain et al., 2021; Aşkın et al., 2018), and one concerned both subacute and chronic patients (Kiper et al., 2018). These results suggest that the likelihood to observe an additional gain provided by VR is increased at the sub-acute phase but a further improvement induced by VR therapy may also occur at the chronic phase.

Despite conflicting observations, these recent results contribute increasing evidence that VR therapy is not to be overlooked in upper limb rehabilitation as it may be concretely beneficial to patients' recovery. The more consolidate findings so far suggest that VR could enhance CT and increase the rehabilitation potential. Rather than relying on one method, multiplying therapeutic approaches to include VR therapy in existing rehabilitation programs appears to be an effective way to further advance stroke rehabilitation outcome.

Effects of neuroplasticity as a direct result of VR therapy is also being investigated, but evidence is still modest (Laver et al., 2017). In their study on the combined use of VR and CT, compared with CT alone, Wang et al. (2017) evaluated the neural reorganization in sub-acute stroke patients with fMRI before and after training with a Leap-Motion based VR system. Patients were asked to perform movements where they had to use the thumb of their impaired hand to touch their opposite palm. They observed a shift in the sensorimotor cortex activation from ipsilateral to contralateral regions and an increased activation in the contralateral cortex in both the experimental and control group. Yet, this change was significantly greater in the experimental VR group. In addition, the experimental VR group also displayed larger improvement in the experimental group using the Wolf motor function test (WMFT), used to assess patients' upper limb motor function. These findings suggest that repeated exercises with the affected limb and task-oriented practice in a virtual environment can facilitate neural reorganization to a larger extent compared to CT alone, promoting motor recovery of the affected upper limbs. Future neuroimaging studies will hopefully help better characterizing VR training dependent effects and thus guiding the development of VR as a therapy tool.

Benefits of Integrating Virtual Reality as a Therapy Tool for Therapists

VR systems developed in the recent years display features that therapists can exploit for their expert intervention. The large number of studies conducted help provide more insights on which among those characteristics may come into play in VR-based rehabilitation, and how they may influence rehabilitation outcomes and/or the therapeutic protocols that can be conducted.

It has been widely documented that VR systems offer the ability to provide an intensive training with a high number of movement repetitions per session (Perez-Marcos et al., 2017). It is suggested repetition of movement and duration of training are factors that may optimize motor rehabilitation outcome and ability to perform ADL, although dose-response effects and difficulty level of each task should be assessed to ensure an optimal therapy dosing (Baniña et al., 2020; Dromerick et al., 2009; Kleim & Jones, 2008). VR systems are believed to help increasing the rehabilitation dosage and to provide significant amounts of therapy to patients thus enabling simulated practice of functional tasks (Laver et al., 2017). Perez-Marcos et al. (2017) and Baniña et al. (2020) reported that training with a VR-based motor rehabilitation system was indeed feasible and could provide high rehabilitation doses, with a high number of repetitions per session and active training time for more efficient training sessions. In Perez-Marcos et al.'s study (2017), various shoulder, arm and wrist exercises were proposed and integrated into functional tasks, like grasping or pointing at virtual objects, and led to significant improvements in upper limb function of chronic stroke patients.

It is also suggested that VR systems can help increase the dosage of therapy without needing to increase staffing levels (Laver et al., 2017). VR systems can be equipped with a tracking functionality,

allowing therapists to monitor their patients' progression without the need for physical supervision at all time. Using the VR system and following an exercise program predefined by their therapist, patients can then participate in a more intensive and frequent training without increasing staffing and achieve positive results in their upper limb recovery (Norouzi-Gheidari et al., 2019). If successfully implemented, VR could then become a cost-effective rehabilitation tool.

In particular, these apparent benefits of VR technologies open new perspectives and opportunities for tele-rehabilitation, an emerging solution which allows patients to have access to a home-based therapy following discharge from the stroke and rehabilitation units and to extend patients' therapy duration, with remote monitoring from therapists (Allegue et al., 2020; Laver et al., 2020). Self-administered treatment at home, through technology-based training and conventional exercises, has been previously found to be accepted by chronic stroke patients (Nijenhuis et al., 2017). A few VR systems have been specifically designed for home-based use like the Neurofenix platform (Kilbride et al., 2018), aiming to encourage stroke patients to exercise independently at home, in their environment, and with minimal therapist supervision. Feasibility studies reported that patients, trained at their own home during 4 weeks, have gained significant improvements in bilateral upper limb function, grasp strength and motor control (Burdea et al., 2019; Thielbar et al., 2020). Findings from a recent study also suggest that VR-based training taking place at home can induce cortical reorganization and is associated with upper limb functional gains (Ballester et al., 2017). All these recent developments add to suggest VR is a technology of interest to spread the development of tele-rehabilitation for patients suffering from upper limb impairment and these positive results strongly encourage to conduct further studies on the use of VR at home, to determine the effectiveness of the intervention but also help guide therapists on how to effectively conduct their intervention remotely using these technologies. We argue that facilitating access to therapy in a remote location and an increased treatment period may turn out to be major arguments in favor of the use of VR in rehabilitation.

Potential Ingredients That Render Virtual Reality-Based Therapy Effective

The above reviewed recent studies revealed several factors inherent to VR therapy that may highly enhance neurorehabilitation and participate in the significant improvement observed so far in upper limb function. We advance that it would be particularly interesting to accurately identify what those factors are, to help optimizing VR systems developed in the future for rehabilitation purposes.

It has been highlighted that the distinction between specialized or non-specialized VR systems might be an important factor in regards to efficacy (Aminov et al., 2018). Specialized systems are VR systems that were specifically developed for upper limb rehabilitation. Examples of specialized systems include SaeboVR, MindMotion Pro or Bi-Manu Trainer. Non-specialized systems refer to off-the-shelf systems and commercial gaming systems, such as the Nintendo Wii or Microsoft Xbox 360 consoles, often designed originally for recreational purposes. Thus far, both types of VR systems have been exploited in different studies investigating the efficacy of VR in upper limb rehabilitation (Subramanian et al., 2020). Some studies have also adapted commercial gaming systems and specifically added games that were designed for rehabilitation of stroke patients (Aşkın et al., 2018). In this respect, it has recently been suggested that the type of systems used may greatly influence the results of motor recovery. In their meta-analysis, Maier et al. (2019) concluded that therapy with specialized VR systems leads to a higher beneficial impact on recovery, body function and on activity than CT, whereas non-specialized systems do not render the same outcome. Tailor-made systems designed to be used by patients with upper limb impairments appear to be a more viable tool to deliver effective motor rehabilitation, compared to off-the-shelf systems that were designed for healthy users.

The literature also suggests that VR systems, in particular specialized ones, can integrate multiple principles of neurorehabilitation in the therapeutic protocols and help manipulating practice conditions, in order to optimize motor learning and neuroplasticity processes. More specifically, task-specific practice, increase of difficulty level, variety of tasks with different goals, avatar representation or promoted use of the affected limb are key principles that can be particularly exploited for VR therapy (Maier et al., 2019). They can also contribute to the development of novel techniques for upper limb rehabilitation like the Reinforcement-Induced Movement Therapy that includes a VR-based training and aims to promote the use of the paretic limb for motor recovery (Ballester et al., 2016).

One major feature of VR systems is that they can typically deliver explicit and implicit feedback during therapeutic training, to a larger extent than in CT (Maier et al., 2019). Feedbacks can be delivered in different forms and provide information to patients on their movements, their performance and their results in real time, while they interact with the virtual environment during entrainment. Examples of multisensory feedbacks include: an on-screen avatar representing the patient's arms and hands, display of scores and records attained, or acoustic signals to provide information on the correct execution of a movement (Kiper et al., 2018; Rogers et al., 2019). As an example, a virtual environment with reinforced and frequent feedback was reported to have an added therapeutic effect as compared to CT, with a better motor recovery outcome in stroke patients (Kiper et al., 2018). Recent VR systems developed for rehabilitation purposes, such as the Elements system, were designed to specifically provide augmented feedback. Rogers et al. (2019) observed that patients receiving therapy with the Elements system experienced greater improvements in upper limb function than controls. Providing more feedbacks in order for patients to have more knowledge on their results and their performance during a single session, in real time, may help promoting motor learning in upper limb rehabilitation. As feedback can be provided simultaneously when using VR, it may also induce a more active participation from patients, associated with an increased motivation to succeed in the activities (Kiper et al., 2018; Rogers et al., 2019).

Increased participants' motivation is a recurring observation in studies. VR-based therapy appears to be more appealing to stroke patients. Several studies have included safety and

technology acceptance evaluations in their protocols and found positive assessments in regards to acceptance and motivation, with patients reporting augmented motivation and willingness to pursue VR training at hospital, or at home (Burdea et al., 2019; Perez-Marcos et al., 2017; Warland et al., 2019). The qualitative substudy conducted by Pallesen et al. (2018) highlighted multiple factors that influenced patients' motivation in Brunner et al. (2017) clinical trial: the playful nature of the activities, the ability to progress in the games depending on their abilities, as well as the reward and feedback systems integrated in the VR solution. These factors may contribute to patients' motivation as they make therapy sessions more challenging and the perception of their improvements is facilitated throughout the treatment duration.

Patients' satisfaction is also often reported as very high after VR sessions (Demers et al., 2019; Lee et al., 2020). More precisely, the variety of activities in virtual environments, performing exercises in the form of games and the possibility of training in an enriched environment make VR-based therapy enjoyable to patients and possibly more engaging than CT (Wang et al., 2017; Hung et al., 2019; Rogers et al., 2019). Importantly, motivation and engagement in rehabilitation are related to compliance and adherence to therapy (Perez-Marcos et al., 2017). As a result, higher levels of motivation induced by VR-based therapy are likely to positively influence rehabilitation outcomes and lead to significant improvements in upper limb function. Also, therapists can establish a rehabilitation program that matches their patients' needs and preferences using VR systems settings (Kim et al., 2018; Hung et al., 2019). We conclude that since VR-based therapy has been established to be motivating to patients and associated with high adherence to therapy, it constitutes a viable tool to strongly encourage patients to exercise independently and frequently in the hospital and upon discharge, at home.

Virtual Reality: A Patient-Centered Tool

VR systems are now often equipped with motion capture technologies such as the Leap Motion hand tracking device (Wang et al., 2017) or the Microsoft Kinect (Aşkın et al., 2018), which can track patients' movements and be used to gather data in regards to performance, kinematics and help provide an analysis of movement quality (Perez-Marcos et al., 2017). While this allows for clinicians to be able to track their patient's performance throughout entrainment, which is particularly advantageous if we consider a home-based rehabilitation, motion tracking solution also offers the opportunity to further individualize the clinician's intervention. Data can be used by therapists for a better assessment of each patient's abilities, to track their progression and more importantly, to adapt their intervention at every step of the rehabilitation process to ensure it matches the patient's needs and their goals. Adjusting the difficulty level of exercises is possible within multiple VR systems, making it possible to offer an intervention that is tailored to each patient's abilities, preferences and motor function level when training with VR (Hung et al., 2019; Kim et al., 2018). Using an artificial intelligence (AI) module, the novel BrightBrainer VR system used in Burdea et al.'s study (2019) changed game difficulties based on a patient's prior performance during tele-rehabilitation. Depending on the system used, therapists can specifically choose the focus of the exercises. Manipulation, hand grasping, whole-arm movement, pronation-supination or bimanual coordination are among the movements that can be selected by therapists to tailor each patient's exercise program (Brunner et al., 2017; Kiper et al., 2018; Schuster-Amft et al., 2018). VR systems designed for rehabilitation can also integrate modules that automatically adjust the difficulty of a task according to a patient's performance, as in the Rehabilitation Gaming System (RGS) (Cameirão et al., 2011). By capturing specific features of a user's upper limb, the system can adapt a task's parameters to an individual's abilities, allowing for further individualization of the therapy (Cameirão et al., 2010).

When focusing more particularly on how rehabilitation is conducted, VR therapy could also prove to be a safe tool for patients. The practice of ADL is possible in virtual environments, with a wide range of ADL proposed such as grocery shopping or crossing a street (Adams et al., 2018). Thus, VR systems allow therapists to propose tasks that would possibly be unsafe if performed in the real world (Laver et al., 2017). For example, practicing a cooking activity in a virtual environment would remove the risk of burns.

Ecological Validity of Virtual Reality in Therapy

In VR-based therapy, patients interact within a virtual environment that can simulate daily life situations and reproduce, to different levels of realism, the real world. As a result, VR-based training may offer an almost naturalistic, ecologically-valid environment. Owing to these features, VR systems may facilitate accessibility to the practice of ADLs during patients' stay at the hospital, as it is not always feasible within a hospital facility. The effectiveness of ADL-focused therapy is already established in upper limb rehabilitation but VR may add advantages to ADL-focused interventions (Legg et al., 2007). Specific VR systems enable task-oriented practice in virtual worlds in order to reacquire functional skills through different activities such as cooking, gardening or grocery shopping in a virtual world (Adams et al., 2018; Aşkın et al., 2018). As an example, practicing virtual ADL with the SaeboVR system designed for upper limb rehabilitation was associated with significant improvements in motor function measures of chronic stroke patients in Adams et al.'s study (2018). In addition, the practice of ADL that are particularly meaningful and relatable to the patient can contribute to an increased adherence to the treatment and increased motivation to pursue rehabilitation (Adams et al., 2018).

However, even if the practice of ADL in a virtual environment is now feasible, with significant improvements observed, it has yet to be determined if gains of VR training do translate to improved performance of real-life activities in the long term. Evidence of the transfer of VR training effects to ADL for patients who suffered a stroke is still limited (Aminov et al., 2018). Long-term follow-up studies are necessary to assess more carefully the effect of VR-based therapy on independence in ADL following discharge. Effects of VR-based tele-rehabilitation on ADL also remain to be evaluated.

Limits of Using Virtual Reality in Rehabilitation

There are some limitations that have been noted concerning VR. Using VR as a rehabilitation tool may be accompanied by some relatively minor adverse effects that may stem from the equipment used and prolonged exposure to a screen while doing different exercises and movements. A few cases of motion sickness, headaches or soreness have been reported by patients in studies (Hung et al., 2019; Perez-Marcos et al., 2017). However, these are rare, and most patients who participated were not subject to any major adverse event over the course of their treatment (Aşkın et al., 2018; Norouzi-Gheidari et al., 2019; Wang et al., 2017), even when using a head-mounted display for a fully immersive VR experience during multiple sessions (Lee et al., 2020).

VR-based therapy is also very dependent on the proper functioning of the equipment. Frequent device malfunctions such as screen freezing, inaccuracy of movement tracking or communication problems can occur in the middle of an activity and be associated with frustration or decrease in motivation, which may reduce the benefits of the treatment (Burdea et al., 2019; Pallesen et al., 2018).

Suggestions Regarding the Use of Virtual Reality in Clinical Settings

Following this review, we suggest some recommendations can be made regarding the use of VR for the rehabilitation of the upper limb for stroke patients. While VR appears to be a suitable tool for rehabilitation, using VR as an adjunct, combined with conventional occupational and physical therapy, may be more beneficial for the recovery of upper limb function rather than relying on VR alone (Kiper et al., 2018; Wang et al., 2017). In addition, therapists should exploit specific VR systems features such as augmented feedback, gamified and motivating activities, movement tracking, practice of virtual ADLs and the possibility of training in an environment similar to the real world as they may help enhancing functional outcomes in upper limb rehabilitation and optimize their intervention (Adams et al., 2018; Maier et al., 2019; Pallesen et al., 2018; Rogers et al., 2019). Also, current evidence comfort us in suggesting that specialized VR systems, specifically designed for upper limb rehabilitation, are to be preferred (Aminov et al., 2018; Maier et al., 2019). Specialized VR systems are indeed more effective and offer flexible patient-based tailoring to therapists. But specialized VR equipment is not yet widely available in clinical settings and its expensive cost may, in some circumstances, constitute a barrier to the development of its use in rehabilitation clinics.

Perspectives on the Future of Virtual Reality in Clinical Settings and Research

One major development regarding the use of VR in the recent years is the immersive feature of some VR systems. Immersion refers to the sensorimotor coupling between the user and the virtual environment provided by the system, and determines the potential of a VR system to effectively isolate a user from the real world (Mestre, 2015). Fully immersive systems place users in an environment that integrates 3D images and objects, where they have no access to the real world and are only exposed to sensory feedbacks coming from the system itself. In contrast, non-immersive VR systems generally display a virtual environment on a screen that users interact with using devices such as keyboards, controllers or joysticks, letting users experience both the real and virtual world at the same time (Huang et al., 2019; Kilbride et al., 2018). Non-immersive systems are more common in rehabilitation settings and have been predominant in VR studies until recently. However, with head-mounted display technologies becoming more and more popular, immersive VR can now become more widespread. It is suggested the level of immersion of a given VR system might play a role in motor recovery, although it is still unclear how exactly (Adams et al., 2018). Fully immersive VR therapy may enhance the feeling of immersion, enabling an even more engaging experience and facilitating patients' performance when executing movements with their impaired upper limb, especially as they can provide more realistic virtual environments. More specifically, immersive properties of virtual environments are associated with the notion of presence, that refers to the feeling of being inside the virtual world. The feeling of presence enables participants to behave in the virtual environment as if it was the real world (Mestre, 2015). Only few studies have investigated fully immersive VR with the use of head-mounted displays such as the HTC Vive, whose spatiotemporal resolutions complies with this sort of behavioral applications (Verdelet et al., 2019). They have reported significant improvements in upper limb function and performance in ADL after multiple therapy sessions (Ögün et al., 2019; Lee et al., 2020; Mekbib et al., 2020). More studies are nevertheless needed to assess the effectiveness of immersive VR-based therapy for the recovery of upper limb motor function, to determine if fully-immersive systems are more effective than non-immersive ones at short- and long-term. Future studies on immersive VR systems may provide better insights into how the level of immersion influences neuroplasticity and cortical reorganization in stroke patients, what mechanisms are at work, and how to better integrate VR in upper limb rehabilitation for stroke patients (Ahmed et al., 2020). As this technology is becoming more widespread, it is likely that immersive VR systems will take on an important part in future rehabilitation.

When considering research on VR as a rehabilitation tool more globally, additional clinical studies are needed with larger samples of patients in order to gather stronger evidence of VR efficacy. It is also necessary to further investigate effects of VR-based therapy in the longer term. Results of follow-up studies will give a better understanding on the retention of the motor learning acquired during treatment with VR.

For VR to become a viable therapy tool, it is also important that research focuses on identifying what the "ingredients" for effective VR are, as well as the conditions whereby VR can be best used, to maximize its potential. Studies investigating effectiveness of VR have applied different experimental designs in terms of frequency (ranging from two to five sessions a week), duration of training sessions (30–60 min) and length of treatment (from 4 to

12 weeks). Since it has been suggested that a higher dose of training volume is preferable, with more than 15 h of total intervention time (Laver et al., 2017), future studies are needed to determine if the dose of VR-therapy does have a significant effect on motor rehabilitation outcomes and if so, which dosage has to be applied when implementing VR in therapy. Future studies will also help specify the effects of timing of VR interventions on functional outcomes and thus, may help determine the optimal timing during which VR interventions can lead to significant improvements in stroke rehabilitation (Merians et al., 2020).

There are also several open questions concerning the patients who can use and benefit from VR therapy, regarding factors such as the severity of the motor impairment or the lesion topography. Kiper et al. (2018) observed that a VR intervention was effective after both hemorrhagic and ischemic stroke, suggesting that stroke etiology does not influence therapy outcomes differently. But further studies are still necessary to determine the population that can best benefit from VR therapy. In addition, active rehabilitation is recommended early, in the subacute phase of stroke recovery, in order to maximize motor recovery gains. In VR research, few studies have been conducted with patients in subacute stage and chronic stage although improvements have been observed in both populations (Aminov et al., 2018), hence it is still necessary to identify the time window for applying VR therapy.

Not last, stroke patients can suffer from cognitive impairments on top of their motor deficits. Patients with severe cognitive impairment were often excluded from previous studies (Aşkın et al., 2018; Brunner et al., 2017; Kiper et al., 2018; Norouzi-Gheidari et al., 2019; Perez-Marcos et al., 2017; Schuster-Amft et al., 2018). However, there are now VR systems intended for rehabilitation of both cognitive and motor functions for stroke patients (Rogers et al., 2019), which broaden the target population and illustrate further the potential of VR for the treatment of major stroke sequelae.

CONCLUSION

Severity of upper limb impairment following stroke is a predictor of poor functional hand ability (Wade et al., 1983; Lai et al., 2002) and a predictor of poor quality of life (Nichols-Larsen et al., 2005). Effective rehabilitation approaches are needed to enhance motor and functional recovery. Since VR has emerged as a suitable rehabilitation tool, VR interventions have shown to offer patients with intensive, repetitive and task-specific entrainment tools in naturalistic virtual environments. Recent evidence show that VR-based therapy combined with CT produce significant improvements in upper limb motor function in stroke patients. Beyond evidence of efficacy, VR systems appear to offer highly engaging and motivating activities to patients, in virtual environments that may be similar to the real world. They also present peculiar features such as movement tracking and the integration of key principles of neurorehabilitation including reinforced feedback. These elements may be advantageous to patients and clinicians, in order to enhance rehabilitation treatments but also to improve therapists' intervention and optimize single patient's tailored care, in the hospital and at a patient's home. Further studies are needed to maximize the potential offered by VR and to ensure it is used effectively as a therapy tool.

AUTHOR CONTRIBUTIONS

JB and AF contributed to conception and organization of the research work. JB collected the literature materials and wrote the first draft of the manuscript. AF and JL critically revised the manuscript and added sections of the manuscript. All authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

JB, JL and AF were supported by a grant of the IHU CeSaMe ANR-10-IBHU-0003 and it was performed within the framework of the LABEX CORTEX (ANR-11-LABX-0042) of Université de Lyon.

ACKNOWLEDGMENTS

We thank JL Borach, S Alouche, S Terrones for administrative support.

REFERENCES

Adams, R. J., Lichter, M. D., Ellington, A., White, M., Armstead, K., Patrie, J. T., et al. (2018). Virtual Activities of Daily Living for Recovery of Upper Extremity Motor Function. *IEEE Trans. Neural Syst. Rehabil. Eng.* 26 (1), 252–260. doi:10.1109/TNSRE.2017.2771272

Ahmed, N., Mauad, V. A. Q., Gomez-Rojas, O., Sushea, A., Castro-Tejada, G., Michel, J., et al. (2020). The Impact of Rehabilitation-Oriented Virtual Reality Device in Patients with Ischemic Stroke in the Early Subacute Recovery Phase: Study Protocol for a Phase III, Single-Blinded, Randomized, Controlled Clinical Trial. *J. Cent. Nerv Syst. Dis.* 12, 117957351989947. doi:10.1177/1179573519899471

Ain, Q. U., Khan, S., Ilyas, S., Yaseen, A., Tariq, I., Liu, T., et al. (2021). Additional Effects of Xbox Kinect Training on Upper Limb Function in Chronic Stroke Patients: A Randomized Control Trial. *Healthcare* 9 (3), 242. doi:10.3390/healthcare9030242

Allegue, D. R., Kairy, D., Higgins, J., Archambault, P., Michaud, F., Miller, W., et al. (2020). Optimization of Upper Extremity Rehabilitation by Combining Telerehabilitation with an Exergame in People with Chronic Stroke: Protocol for a Mixed Methods Study. *JMIR Res. Protoc.* 9 (5), e14629. doi:10.2196/14629

Aminov, A., Rogers, J. M., Middleton, S., Caeyenberghs, K., and Wilson, P. H. (2018). What Do Randomized Controlled Trials Say about Virtual Rehabilitation in Stroke? A Systematic Literature Review and Meta-Analysis of Upper-Limb and Cognitive Outcomes. *J. Neuroengineering Rehabil.* 15 (1), 29. doi:10.1186/s12984-018-0370-2

Aşkın, A., Atar, E., Koçyiğit, H., and Tosun, A. (2018). Effects of Kinect-Based Virtual Reality Game Training on Upper Extremity Motor Recovery in Chronic

Stroke. *Somatosensory Mot. Res.* 35 (1), 25–32. doi:10.1080/08990220.2018.1444599

Ballester, B. R., Maier, M., San Segundo Mozo, R. M., Castañeda, V., Duff, A., and M. J. Verschure, P. F. (2016). Counteracting Learned Non-use in Chronic Stroke Patients with Reinforcement-Induced Movement Therapy. *J. Neuroengineering Rehabil.* 13 (1), 74. doi:10.1186/s12984-016-0178-x

Ballester, B. R., Nirme, J., Camacho, I., Duarte, E., Rodríguez, S., Cuxart, A., et al. (2017). Domiciliary VR-Based Therapy for Functional Recovery and Cortical Reorganization: Randomized Controlled Trial in Participants at the Chronic Stage post Stroke. *JMIR serious games* 5 (3), e15. doi:10.2196/games.6773

Baniña, M. C., Molad, R., Solomon, J. M., Berman, S., Soroker, N., Frenkel-Toledo, S., et al. (2020). Exercise Intensity of the Upper Limb Can Be Enhanced Using a Virtual Rehabilitation System. *Disabil. Rehabil. Assistive Tech.*, 1–7. doi:10.1080/17483107.2020.1765421

Bohil, C. J., Alicea, B., and Biocca, F. A. (2011). Virtual Reality in Neuroscience Research and Therapy. *Nat. Rev. Neurosci.* 12 (12), 752–762. doi:10.1038/nrn3122

Brunner, I., Skouen, J. S., Hofstad, H., Aßmus, J., Becker, F., Sanders, A.-M., et al. (2017). Virtual Reality Training for Upper Extremity in Subacute Stroke (VIRTUES). *Neurology* 89 (24), 2413–2421. doi:10.1212/WNL.0000000000004744

Burdea, G. C., Grampurohit, N., Kim, N., Polistico, K., Kadaru, A., Pollack, S., et al. (2019). Feasibility of Integrative Games and Novel Therapeutic Game Controller for Telerehabilitation of Individuals Chronic post-stroke Living in the Community. *Top. Stroke Rehabil.* 27, 321–336. doi:10.1080/10749357.2019.1701178

Cameirão, M. S., Badia, S. B. i., Oller, E. D., and Verschure, P. F. (2010). Neurorehabilitation Using the Virtual Reality Based Rehabilitation Gaming System: Methodology, Design, Psychometrics, Usability and Validation. *J. Neuroengineering Rehabil.* 7 (1), 48. doi:10.1186/1743-0003-7-48

da Silva Cameirão, M., Bermúdez i Badia, S., Duarte, E., and Verschure, P. F. M. J. (2011). Virtual Reality Based Rehabilitation Speeds up Functional Recovery of the Upper Extremities after Stroke: a Randomized Controlled Pilot Study in the Acute Phase of Stroke Using the Rehabilitation Gaming System. *Restorative Neurol. Neurosci.* 29 (5), 287–298. doi:10.3233/RNN-2011-0599

Demers, M., Chan Chun Kong, D., and Levin, M. F. (2019). Feasibility of Incorporating Functionally Relevant Virtual Rehabilitation in Sub-acute Stroke Care: Perception of Patients and Clinicians. *Disabil. Rehabil. Assistive Tech.* 14 (4), 361–367. doi:10.1080/17483107.2018.1449019

Dromerick, A. W., Lang, C. E., Birkenmeier, R. L., Wagner, J. M., Miller, J. P., Videen, T. O., et al. (2009). Very Early Constraint-Induced Movement during Stroke Rehabilitation (VECTORS): a Single-center RCT. *Neurology* 73 (3), 195–201. doi:10.1212/WNL.0b013e3181ab2b27

Duncan, P. W., Zorowitz, R., Bates, B., Choi, J. Y., Glasberg, J. J., Graham, G. D., et al. (2005). Management of Adult Stroke Rehabilitation Care: a Clinical Practice Guideline. *Stroke* 36 (9), e100–43. doi:10.1161/01.STR.0000180861.54180.FF

Ekstrand, E., Rylander, L., Lexell, J., and Brogårdh, C. (2016). Perceived Ability to Perform Daily Hand Activities after Stroke and Associated Factors: a Cross-Sectional Study. *BMC Neurol.* 16 (1), 208. doi:10.1186/s12883-016-0733-x

Feigin, V. L., Norrving, B., and Mensah, G. A. (2017). Global Burden of Stroke. *Circ. Res.* 120 (3), 439–448. doi:10.1161/CIRCRESAHA.116.308413

Garrett, B., Taverner, T., Gromala, D., Tao, G., Cordingley, E., and Sun, C. (2018). Virtual Reality Clinical Research: Promises and Challenges. *JMIR Serious Games* 6 (4), e10839. doi:10.2196/10839

Huang, Q., Wu, W., Chen, X., Wu, B., Wu, L., Huang, X., et al. (2019). Evaluating the Effect and Mechanism of Upper Limb Motor Function Recovery Induced by Immersive Virtual-Reality-Based Rehabilitation for Subacute Stroke Subjects: Study Protocol for a Randomized Controlled Trial. *Trials* 20 (1), 104. doi:10.1186/s13063-019-3177-y

Hung, J.-W., Chou, C.-X., Chang, Y.-J., Wu, C.-Y., Chang, K.-C., Wu, W.-C., et al. (2019). Comparison of Kinect2Scratch Game-Based Training and Therapist-Based Training for the Improvement of Upper Extremity Functions of Patients with Chronic Stroke: A Randomized Controlled Single-Blinded Trial. *Eur. J. Phys. Rehabil. Med.* 55, 542–550. doi:10.23736/S1973-9087.19.05598-9

Ikbali Afsar, S., Mirzayev, I., Umit Yemisci, O., and Cosar Saracgil, S. N. (2018). Virtual Reality in Upper Extremity Rehabilitation of Stroke Patients: A Randomized Controlled Trial. *J. Stroke Cerebrovasc. Dis.* 27 (12), 3473–3478. doi:10.1016/j.jstrokecerebrovasdis.2018.08.007

Kilbride, C., Scott, D. J. M., Butcher, T., Norris, M., Ryan, J. M., Anokye, N., et al. (2018). Rehabilitation via HOMe Based Gaming Exercise for the Upper-Limb post Stroke (RHOMBUS): Protocol of an Intervention Feasibility Trial. *Bmj Open* 8 (11), e026620. doi:10.1136/bmjopen-2018-026620

Kim, W.-S., Cho, S., Park, S. H., Lee, J.-Y., Kwon, S., and Paik, N.-J. (2018). A Low Cost Kinect-Based Virtual Rehabilitation System for Inpatient Rehabilitation of the Upper Limb in Patients with Subacute Stroke. *Medicine* 97 (25), e11173. doi:10.1097/MD.0000000000011173

Kiper, P., Szczudlik, A., Agostini, M., Opara, J., Nowobilski, R., Ventura, L., et al. (2018). Virtual Reality for Upper Limb Rehabilitation in Subacute and Chronic Stroke: A Randomized Controlled Trial. *Arch. Phys. Med. Rehabil.* 99 (5), 834–842. doi:10.1016/j.apmr.2018.01.023

Kleim, J. A., and Jones, T. A. (2008). Principles of Experience-dependent Neural Plasticity: Implications for Rehabilitation after Brain Damage. *J. Speech Lang. Hear. Res.* 51 (1), S225–S239. doi:10.1044/1092-4388(2008/018)

Klinger, É. (2008). *Apports de la réalité virtuelle à la prise en charge du handicap*. Saint-Denis: Le traitement du signal et ses applications.

Lai, S.-M., Studenski, S., Duncan, P. W., and Perera, S. (2002). Persisting Consequences of Stroke Measured by the Stroke Impact Scale. *Stroke* 33 (7), 1840–1844. doi:10.1161/01.STR.0000019289.15440.F2

Laver, K. E., Adey-Wakeling, Z., Crotty, M., Lannin, N. A., George, S., and Sherrington, C. (2020). Telerehabilitation Services for Stroke. *Cochrane Database Syst. Rev.* 11, CD008349. doi:10.1002/14651858.CD010255.pub3

Laver, K. E., Lange, B., George, S., Deutsch, J. E., Saposnik, G., and Crotty, M. (2017). Virtual Reality for Stroke Rehabilitation. *Cochrane Database Syst. Rev.* 2018. doi:10.1002/14651858.CD008349.pub4

Lee, M. M., Lee, K. J., and Song, C. H. (2018). Game-Based Virtual Reality Canoe Paddling Training to Improve Postural Balance and Upper Extremity Function: A Preliminary Randomized Controlled Study of 30 Patients with Subacute Stroke. *Med. Sci. Monit.* 24, 2590–2598. doi:10.12659/MSM.906451

Lee, S. H., Jung, H. Y., Yun, S. J., Oh, B. M., and Seo, H. G. (2020). Upper Extremity Rehabilitation Using Fully Immersive Virtual Reality Games with a Head Mount Display: A Feasibility Study. *PM&R* 12 (3), 257–262. doi:10.1002/pmrj.12206

Legg, L., Drummond, A., Leonardi-Bee, J., Gladman, J. R. F., Corr, S., Donkervoort, M., et al. (2007). Occupational Therapy for Patients with Problems in Personal Activities of Daily Living after Stroke: Systematic Review of Randomised Trials. *BMJ* 335 (7626), 922. doi:10.1136/bmj.39343.466863.55

Maier, M., Rubio Ballester, B., Duff, A., Duarte Oller, E., and Verschure, P. F. M. J. (2019). Effect of Specific over Nonspecific VR-Based Rehabilitation on Poststroke Motor Recovery: A Systematic Meta-Analysis. *Neurorehabil. Neural Repair* 33, 112–129. doi:10.1177/1545968318820169

Massetti, T., da Silva, T. D., Crocetta, T. B., Guarnieri, R., de Freitas, B. L., Bianchi Lopes, P., et al. (2018). The Clinical Utility of Virtual Reality in Neurorehabilitation: A Systematic Review. *J. Cent. Nerv Syst. Dis.* 10, 117957351881354. doi:10.1177/1179573518813541

Mekbib, D. B., Zhao, Z., Wang, J., Xu, B., Zhang, L., Cheng, R., et al. (2020). Proactive Motor Functional Recovery Following Immersive Virtual Reality-Based Limb Mirroring Therapy in Patients with Subacute Stroke. *Neurotherapeutics* 17, 1919–1930. doi:10.1007/s13311-020-00882-x

Merians, A. S., Fluet, G. G., Qiu, Q., Yarossi, M., Patel, J., Mont, A. J., et al. (2020). Hand Focused Upper Extremity Rehabilitation in the Subacute Phase Post-stroke Using Interactive Virtual Environments. *Front. Neurol.* 11, 1449. doi:10.3389/fneur.2020.573642

Mestre, D. R. (2015). "On the Usefulness of the Concept of Presence in Virtual Reality Applications," in Proceedings Volume 9392, The Engineering Reality of Virtual Reality 2015, San Francisco, California, United States, 17 March 2015 (Bellingham: International Society for Optics and Photonics), 93920J. doi:10.1117/12.2075798

Nichols-Larsen, D. S., Clark, P. C., Zeringue, A., Greenspan, A., and Blanton, S. (2005). Factors Influencing Stroke Survivors' Quality of Life during Subacute Recovery. *Stroke* 36 (7), 1480–1484. doi:10.1161/01.STR.0000170706.13595.4f

Nijenhuis, S. M., Prange-Lasonder, G. B., Stienen, A. H., Rietman, J. S., and Buurke, J. H. (2017). Effects of Training with a Passive Hand Orthosis and Games at home in Chronic Stroke: a Pilot Randomised Controlled Trial. *Clin. Rehabil.* 31 (2), 207–216. doi:10.1177/0269215516629722

Norouzi-Gheidari, N., Hernandez, A., Archambault, P. S., Higgins, J., Poissant, L., and Kairy, D. (2019). Feasibility, Safety and Efficacy of a Virtual Reality Exergame System to Supplement Upper Extremity Rehabilitation Post-Stroke: A Pilot Randomized Clinical Trial and Proof of Principle. *Ijerph* 17 (1), 113. doi:10.3390/ijerph17010113

Ögün, M. N., Kurul, R., Yaşar, M. F., Turkoglu, S. A., Avci, Ş., and Yildiz, N. (2019). Effect of Leap Motion-Based 3D Immersive Virtual Reality Usage on Upper Extremity Function in Ischemic Stroke Patients. *Arq. Neuro-psiquiatr.* 77 (10), 681–688. doi:10.1590/0004-282X20190129

Pallesen, H., Andersen, M. B., Hansen, G. M., Lundquist, C. B., and Brunner, I. (2018). Patients' and Health Professionals' Experiences of Using Virtual Reality Technology for Upper Limb Training after Stroke: A Qualitative Substudy. *Rehabil. Res. Pract.* 2018, 1–11. doi:10.1155/2018/4318678

Perez-Marcos, D., Chevalley, O., Schmidlin, T., Garipelli, G., Serino, A., Vuadens, P., et al. (2017). Increasing Upper Limb Training Intensity in Chronic Stroke Using Embodied Virtual Reality: a Pilot Study. *J. Neuroengineering Rehabil.* 14 (1), 119. doi:10.1186/s12984-017-0328-9

Rogers, J. M., Duckworth, J., Middleton, S., Steenbergen, B., and Wilson, P. H. (2019). Elements Virtual Rehabilitation Improves Motor, Cognitive, and Functional Outcomes in Adult Stroke: Evidence from a Randomized Controlled Pilot Study. *J. Neuroengineering Rehabil.* 16 (1), 56. doi:10.1186/s12984-019-0531-y

Schuster-Amft, C., Eng, K., Suica, Z., Thaler, I., Signer, S., Lehmann, I., et al. (2018). Effect of a Four-Week Virtual Reality-Based Training versus Conventional Therapy on Upper Limb Motor Function after Stroke: A Multicenter Parallel Group Randomized Trial. *PLOS ONE* 13 (10), e0204455. doi:10.1371/journal.pone.0204455

Subramanian, S. K., Cross, M. K., and Hirschhauser, C. S. (2020). Virtual Reality Interventions to Enhance Upper Limb Motor Improvement after a Stroke: Commonly Used Types of Platform and Outcomes. *Disabil. Rehabil. Assistive Tech.*, 1–9. doi:10.1080/17483107.2020.1765422

Thielbar, K. O., Triandafilou, K. M., Barry, A. J., Yuan, N., Nishimoto, A., Johnson, J., et al. (2020). Home-based Upper Extremity Stroke Therapy Using a Multiuser Virtual Reality Environment: A Randomized Trial. *Arch. Phys. Med. Rehabil.* 101 (2), 196–203. doi:10.1016/j.apmr.2019.10.182

Verdelet, G., Salemme, R., Desoche, C., Volland, F., Farnè, A., Coudert, A., Hermann, R., Truy, E., Gaveau, V., and Pavani, F. (2019). "Assessing Spatial and Temporal Reliability of the Vive System as a Tool for Naturalistic Behavioural Research," in 2019 International Conference on 3D Immersion (IC3D), Brussels, Belgium, 11-11 Dec. 2019 (IEEE). doi:10.1109/IC3D48390.2019.8975994

van der Vliet, R., Selles, R. W., Andrinopoulou, E. R., Nijland, R., Ribbers, G. M., Frens, M. A., et al. (2020). Predicting Upper Limb Motor Impairment Recovery after Stroke: a Mixture Model. *Ann. Neurol.* 87 (3), 383–393. doi:10.1002/ana.25679

Wade, D. T., and Hewer, R. L. (1987). Functional Abilities after Stroke: Measurement, Natural History and Prognosis. *J. Neurol. Neurosurg. Psychiatry* 50 (2), 177–182. doi:10.1136/jnnp.50.2.177

Wade, D. T., Langton-Hewer, R., Wood, V. A., Skilbeck, C. E., and Ismail, H. M. (1983). The Hemiplegic Arm after Stroke: Measurement and Recovery. *J. Neurol. Neurosurg. Psychiatry* 46 (6), 521–524. doi:10.1136/jnnp.46.6.521

Wang, Z.-R., Wang, P., Xing, L., Mei, L.-P., Zhao, J., and Zhang, T. (2017). Leap Motion-Based Virtual Reality Training for Improving Motor Functional Recovery of Upper Limbs and Neural Reorganization in Subacute Stroke Patients. *Neural Regen. Res.* 12 (11), 1823–1831. doi:10.4103/1673-5374.219043

Warland, A., Paraskevopoulos, I., Tsekleves, E., Ryan, J., Nowicky, A., Griscti, J., et al. (2019). The Feasibility, Acceptability and Preliminary Efficacy of a Low-Cost, Virtual-Reality Based, Upper-Limb Stroke Rehabilitation Device: a Mixed Methods Study. *Disabil. Rehabil.* 41 (18), 2119–2134. doi:10.1080/09638288.2018.1459881

Weiss, P. L., Rand, D., Katz, N., and Kizony, R. (2004). Video Capture Virtual Reality as a Flexible and Effective Rehabilitation Tool. *J. Neuroengineering Rehabil.* 1 (1), 12. doi:10.1186/1743-0003-1-12

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Big Data

SYSTEMATIC REVIEW
published: 14 July 2022
doi: 10.3389/fdata.2022.850383

Ethics of AI in Radiology: A Review of Ethical and Societal Implications

Melanie Goisauf[*†] *and Mónica Cano Abadía*[*†]

ELSI Services and Research, BBMRI-ERIC, Graz, Austria

Artificial intelligence (AI) is being applied in medicine to improve healthcare and advance health equity. The application of AI-based technologies in radiology is expected to improve diagnostic performance by increasing accuracy and simplifying personalized decision-making. While this technology has the potential to improve health services, many ethical and societal implications need to be carefully considered to avoid harmful consequences for individuals and groups, especially for the most vulnerable populations. Therefore, several questions are raised, including (1) what types of ethical issues are raised by the use of AI in medicine and biomedical research, and (2) how are these issues being tackled in radiology, especially in the case of breast cancer? To answer these questions, a systematic review of the academic literature was conducted. Searches were performed in five electronic databases to identify peer-reviewed articles published since 2017 on the topic of the ethics of AI in radiology. The review results show that the discourse has mainly addressed expectations and challenges associated with medical AI, and in particular bias and black box issues, and that various guiding principles have been suggested to ensure ethical AI. We found that several ethical and societal implications of AI use remain underexplored, and more attention needs to be paid to addressing potential discriminatory effects and injustices. We conclude with a critical reflection on these issues and the identified gaps in the discourse from a philosophical and STS perspective, underlining the need to integrate a social science perspective in AI developments in radiology in the future.

Keywords: artificial intelligence, ethics, radiology, explainability, trustworthiness, bias

OPEN ACCESS

Edited by:
Martin Schlather,
University of Mannheim, Germany

Reviewed by:
Pekka Ruusuvuori,
University of Turku, Finland
Rabia Saleem,
University of Derby, United Kingdom

***Correspondence:**
Melanie Goisauf
melanie.goisauf@bbmri-eric.eu
Mónica Cano Abadía
monica.cano.abadia@bbmri-eric.eu

†These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Big Data

Received: *07 January 2022*
Accepted: *13 June 2022*
Published: *14 July 2022*

Citation:
Goisauf M and Cano Abadía M (2022) Ethics of AI in Radiology: A Review of Ethical and Societal Implications. Front. Big Data 5:850383. doi: 10.3389/fdata.2022.850383

INTRODUCTION

Artificial Intelligence (AI) is seen as a promising innovation in the medical field. The term AI encompasses the ability of a machine to imitate intelligent human behavior (Tang et al., 2018). Machine learning (ML) is a subfield of AI which is widely applied to medical imaging (Pesapane et al., 2018a) and includes deep learning (DL), which produces data with multiple levels of abstraction (LeCun et al., 2015). These technologies have been developed to help improve predictive analytics and diagnostic performance, and specifically to improve their accuracy and ability to support personalized decision-making, as researchers have demonstrated that they can "outperform humans" when conducting medical image analysis (McKinney et al., 2020). Many researchers have also expressed the hope that they can help improve the provision of healthcare, and especially by enabling more rapid diagnosis, in coping with the workload resulting from an increase in screening (Mudgal and Das, 2020, 6), and in advancing health equity. Overall, AI systems are expected to have a significant impact in radiology.

"Artificial Intelligence (AI) is the talk of the town" (Ferretti et al., 2018, 320). Developments in AI are progressing rapidly in the medical field. This is reflected, for instance, in the enormous increase in publications on the development of AI systems in radiology, i.e., from about 100–150 per year in 2007–2008 to 700–800 per year in 2016–2017 (Pesapane et al., 2018a). However, the progress in development of the discourse on these technologies has not corresponded with the progress in the implementation of these technologies in healthcare. In other words, "The state of AI hype has far exceeded the state of AI science, especially when it pertains to validation and readiness for implementation in patient care" (Topol, 2019, 51). For example, few radiological AI systems have been implemented in the NHS, but several are awaiting approval (Mudgal and Das, 2020).

In the evolving field of Ethics of AI, investigations are carried out on the far-reaching consequences of AI in several areas of society. AI is steadily gaining importance in the medical field; as a result, researchers and practitioners are carefully considering the ethical and societal implications of AI use in order to avoid harmful consequences for individuals and groups, and especially those for the most vulnerable populations. Without a doubt, AI will have a profound impact in the field of radiology: It will affect end users and will introduce far-reaching challenges into clinical practice. While the introduction of AI is changing the role of the "radiologists-in-the-loop," patients and other societal groups are being confronted with complex questions concerning the scope of informed consent, biases that may result in inequality, and risks associated with data privacy and protection, as well as open questions regarding responsibility and liability. These questions are accompanied by concerns that AI systems could perpetuate or even amplify ethical and societal injustices. Based on key ethical values such as respect, autonomy, beneficence, and justice (Beauchamp and Childress, 2001), several guiding principles and recommendations have been formulated to tackle these issues (Currie et al., 2020; Ryan and Stahl, 2020)—Such principles and recommendations have also been communicated on EU level (High-Level Expert Group on Artificial Intelligence, 2019), and initiatives such as FUTURE-AI (Lekadir et al., 2021) have been started, which have been developed to ensure that advances in AI systems and advances in AI ethics do not contradict one another.

In this paper, we contribute to the discourse on ethics of AI in radiology by reviewing the state-of-the-art literature and discussing the findings from a philosophical and social science perspective. We consider the comment made by (Mittelstadt and Floridi, 2016, 468), namely, that "reviewing literature is a first step to conduct ethical foresight, in the sense that it allows one to distinguish between issues and implications that are currently under consideration, and those that are not yet acknowledged or require further attention." In our review, we highlight underexplored ethical and societal aspects and point out the necessary future research directions in the field. Our analysis was guided by two key research questions: (1) What types of ethical issues are raised by the use of AI in medicine and biomedical research, and (2) how are these issues being tackled in radiology, especially in the case of breast cancer? In the next section of this article, we describe the methods used, then present the outcomes of the review. We conclude the article with a critical discussion of the findings, highlight the identified gaps and indicate future directions.

METHODS

Search and Eligibility Criteria

We performed a comprehensive review of ethical and societal issues that have already been identified and discussed, as well as how these issues have been addressed in the context of AI. To do so, we carried out a systematic review of state-of-the art academic literature between July and December 2021. Five search engines were used (Google Scholar, Microsoft Academic, PubMed, Scopus, and Web of Science) to identify relevant articles on these issues. Twelve search strings were created that included terms relevant to the research questions (e.g., "AI," "ethics," "radiology," "imaging," "oncology," "cancer," "predictive medicine," "trustworthy," "explainable," "black box," and "breast cancer"). Hence, the selection of the search terms aimed at including both key aspects in the general discussion of AI in medicine and biomedical research, as well as specific approaches for radiology and oncology. In addition, breast cancer was included in the search as a specific case to analyze in-depth the societal implications associated with social categories such as gender, race, and socioeconomic background. The different levels were expected to allow us to better situate the topic in the broader Ethics of AI discourse. "IT" was defined as an exclusion criterion to refine the search and limit it to the ethical and societal aspects related to AI. All search strings were applied to the five search engines using the "Publish or Perish" app, introducing some minor differences in punctuation to adapt to the internal logic of different search engines. The search was limited to articles written in English language and to papers published after 2017. Outputs consisted mainly of peer-reviewed journal articles, but also included literature in the form of commentaries, reports, and book chapters. These sources were not excluded from the sample, as they are also seen as contributing to the discourse on the ethics of AI use in radiology.

Data Analysis

All identified records were imported into Microsoft Excel spreadsheets for further analysis. The subsequent screening procedure was conducted in two major steps. First, we scanned paper titles and abstracts to identify papers that included discussions on ethical and societal aspects of AI. Duplicates and papers that did not match the inclusion criteria were removed from the sample, as well as articles that were identified by the search engines because they contained an ethics statement. Second, the full texts of the resulting sample items ($n = 56$) were analyzed using thematic analysis (Terry et al., 2017).

Guided by our research questions, we coded each article in the final sample to develop overarching themes or patterns. Semantic codes were generated, on the one hand, to deductively assign terms also used in the search strings to the material (i.e., terms that are commonly used in the AI ethics discourse, such as "explainability," "trustworthy," and "black box") and, on the other hand, inductively developed from the data. Reviewers coded independently on paper and by using the Atlas.ti software

package. In the next analytical phase, the codes and coded text segments were collated to identify themes across the sample. Each theme and the corresponding text segments were analyzed to determine their specific content and depth, but also scrutinized to identify conceptual gaps.

RESULTS

The results of the systematic literature review are included in an adapted PRISMA flow diagram (**Figure 1**) (Page et al., 2021).

After screening the records identified by the search engines to determine their eligibility for inclusion, the full texts of a final sample of 56 papers were reviewed. We observed that sources that placed an explicit focus on the ethical issues of AI and breast cancer were rare in the sample (two); some papers on breast cancer identified in the search were excluded, as these placed a primary focus on technical issues, were determined to be irrelevant for the aim of this study or were identified due to the presence of an ethics statement in the article. In terms of content, we concluded that no article in the sample placed an explicit focus on the ethics of AI use in radiology and breast cancer.

The review results show that the application of medical and radiological AI systems is widely discussed in the scientific discourse. As mentioned in the introduction, these technologies are accompanied by hypes and hopes regarding their potential to improve predictive analytics, diagnostic performance, and eventually patient outcomes, as well as challenges that arise due to the (potential) real-world application. During the analysis, certain topics were identified as especially important, which are mainly organized around approaches and principles. Guided by our research questions (i.e., what types of ethical issues are raised by medical AI and how these are tackled in radiology and the case of breast cancer in particular), we analyzed the key themes regarding their claims about ethical and societal implications.

In the next sections, we organize the key themes identified in the literature review as follows: First, we map the expectations regarding the application of AI systems in the medical field, as these are important indicators of their imagined innovative potential and how the discourse is framed, and then enumerate

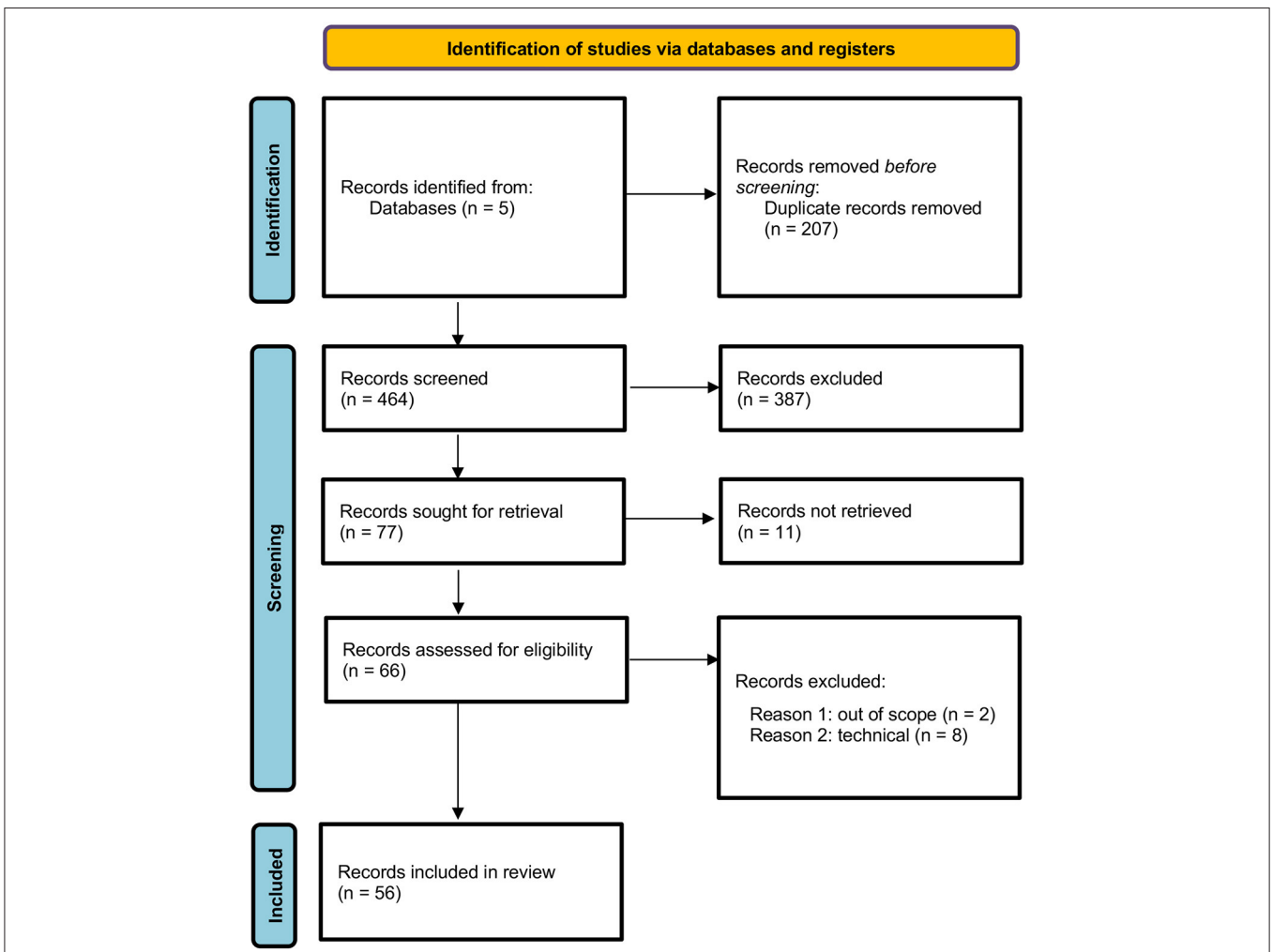

FIGURE 1 | PRISMA flow diagram showing the results of the systematic literature review.

the key challenges. Second, we describe the ethical principles addressed in the literature, such as explainability, interpretability, trust/trustworthiness, responsibility and accountability, justice, and fairness. These sections are followed by a critical discussion of the ethical and societal implications of the results.

Expectations

The analysis of the literature shows that the application of AI systems in healthcare has been welcomed, according to the expectations associated with this application. Thereby, we identified three main areas in which changes are expected, and especially areas in which improvements are expected: better analytical performance, benefits for patients, clinicians and society, and a change in the professional role of radiologists.

The expectation that AI will significantly improve diagnostics and patient care is a key assumption that is expressed throughout the sample. Choy et al. (2018) describe current AI applications to help with case triage, maximize image quality, detect and interpret findings automatically, perform automated processes related to treatment (e.g., in radiotherapy) and points out that these applications can support the personalization of treatment *via* predictive analytics by making scheduling easier. Hosny et al. (2018) identify three radiological tasks in which AI can play a significant role: in the detection of abnormalities, their characterization, and in monitoring changes. Other authors note that further applications of AI are expected to increase analytical power (i.e., to perform analyses more rapidly than humans and to minimize human error) and to identify as-yet unknown relationships (Pesapane et al., 2018a; Brady and Neri, 2020). Eventually, such applications can be used to detect diseases earlier and to provide proper treatment with fewer unnecessary procedures, better cost efficiency, and lower inter- and intra-reader variability (Mazurowski, 2020).

These advances in medical diagnostics are also expected to benefit the end users. For instance, Kelly et al. (2019) identify a quadruple aim for the application of AI systems in healthcare: improving experience of care, improving the health of populations, reducing per capita costs of healthcare, and improving the work life of healthcare providers. Ryan and Stahl (2020) highlight the ethical principle of beneficence and emphasize the supposition that AI should benefit societies and support the social as well as the common good.

The expectations regarding radiologists are ambiguous, with some scholars highlighting the fact that AI will outperform clinicians and be able to diagnose more rapidly and accurately. Bjerring and Busch (2021, 350) state "We can at least with some warrant adopt the assumption that AI systems will eventually outperform human practitioners in terms of speed, accuracy, and reliability when it comes to predicting and diagnosing central disease types such as cancer, cardiovascular diseases, and diseases in the nervous system." On the other hand, some scholars consider that AI will not be able to perform all tasks that health practitioners currently do without any human intervention. Naqa et al. (2020) propose what they consider to be a realistic vision that keeps "humans-in-the-loop." According to this perspective, AI systems will serve as physicians' partners, enabling them to deliver improved healthcare by combining AI/ML software with the best human clinician knowledge. This partnership would allow the delivery of healthcare that outperforms what either can deliver alone, thus improving both credibility and performance.

Challenges

Despite the hype surrounding AI implementation in healthcare and radiology, numerous authors highlight the challenges this can raise. One key challenge concerns the data that are used to train AI, such as the lack of labeled (i.e., annotated) data. Massive amounts of data are needed to train algorithms (Matsuzaki, 2018) and training images must be annotated manually. This challenge is accompanied by a secondary challenge: the fact that the amount of radiological imaging data continues to grow at a disproportionate rate as compared to the number of available, trained readers (Hosny et al., 2018). While some authors propose using a model that has been developed to keep "humans-in-the loop," others consider that the availability of human validation will limit the promises of AI. Tizhoosh and Pantanowitz (2018) comment that "The pathologist is the ultimate evaluation if AI solutions are deployed into clinical workflow. Thus, full automation is neither possible, it seems, nor wise as the Turing test postulates."

The affordability of required computational expenses poses another challenge (Tizhoosh and Pantanowitz, 2018). Geis et al. (2019: 330) point out that AI could increase imbalances in the distribution of resources, creating a gap between institutions that have more and less "'radiology decision-making' capabilities." Small or resource-poor institutions may find it difficult to allocate the necessary resources to manage complex AI systems, especially those that are proprietary. These authors (Geis et al., 2019, 332) emphasize that "Almost certainly some radiology AI will be proprietary, developed by large academic or private health care entities, insurance companies, or large companies with data science expertise but little historical radiology domain knowledge. This may exacerbate disparities in access to radiology AI."

The scarcity of resources is also closely connected to or could result in some form of bias, and in particular automation bias, which is the "tendency for humans to favor machine-generated decision, ignoring contrary data or conflicting human decisions" (Neri et al., 2020, 519). Geis et al. (2019, 332) argue that automation bias can lead to errors of omission, i.e., humans might fail to notice or might disregard the failure of AI tools. This could clash with the need identified in the literature to take a "human-in-the-loop" approach, as "risks may be magnified in resource-poor populations because there is no local radiologist to veto the results." (Geis et al., 2019, 332).

Some doubts have been voiced in the literature regarding the possibility of implementing AI into daily clinical practice, as real-world deployments are still rare, and only a few algorithms have been clinically tested or implemented (Kelly et al., 2019; Mudgal and Das, 2020). In this regard, it is questioned if this implementation is a realistic goal and that it is not clear how to effectively integrate AI systems with human decision-makers (Tizhoosh and Pantanowitz, 2018). Other authors (Gaube et al., 2021) noted that, in the few cases where systems have been

implemented, no proof of improved clinical outcomes has been provided, while Kelly et al. (2019) mention different challenges associated with the use of such systems, including logistical difficulties, quality control, human barriers, and algorithmic interpretability claims.

The challenges outlined so far highlight specific institutional and resource-related issues that may influence the further development of radiological AI; however, these issues also determine how ethical and social implications are reflected upon and manifested in addressing these challenges. This becomes tangible when examining the two major recurring themes identified in the reviewed sample: black box and bias.

Black Box

Many ML algorithms, and especially DL algorithms, are often referred to as operating in a "black box." This black box is defined in the literature as "an apparatus whose inner-workings remain opaque to the outside observer" (Quinn et al., 2021, 2), as "oracular inference engines that render verdicts without any accompanying justification" (Watson et al., 2019, 2), or as "systems [that] are often unable to provide an audit trail for how a conclusion or recommendation is reached because of its convolutional nature" (Smith and Bean, 2019, 25). While some authors only indicate that black boxes generate challenges without going into further detail (Choy et al., 2018; Tizhoosh and Pantanowitz, 2018; Naqa et al., 2020; Kim et al., 2021), others have taken a more specific approach to address the consequences of applying black boxes in medicine. For instance, Bjerring and Busch (2021) apply Price's (2018) concept of black-box medicine: a subtype of AI-informed medicine where opaque or transparent AI systems play an essential role in decision-making. These definitions imply that opacity, intelligible justifications, and recommendations are key issues that need to be discussed when considering ethical requirements and the practitioner-patient relationship.

Ferretti et al. (2018) frame the problem of black boxes in medicine by applying the concept of opacity, which can be differentiated into three types: (1) lack of disclosure, (2) epistemic opacity, and (3) explanatory opacity. The (1) lack of disclosure is defined as a lack of transparency regarding the use of data. The patients' privacy and awareness of the use of their data, their consent (Mudgal and Das, 2020), and ownership of the data (Krupinski, 2020) appear as associated concerns. Larson et al. (2020) also address this issue, providing examples of partnerships between hospitals and data science companies that raised concerns about whether these companies are profiting from the use of patient data, often without their consent. Mudgal and Das (2020) also warn against the risks of defining the value of data on the basis of its face value. To mitigate this risk, "radiology's goal should be to derive as much value as possible from the ethical use of AI, yet resist the lure of extra monetary gain from unethical uses of radiology data and AI" (Geis et al., 2019, 330). To ensure the ethical use of data and to address a lack of disclosure, "patients should know who has access to their data and whether (and to what degree) their data has been de-identified. From an ethical perspective, a patient should be aware of the potential for their data to be used for financial benefit to others and whether potential changes in legislation increase data vulnerability in the future, especially if there is any risk that the data could be used in a way that is harmful to the patient" (Currie et al., 2020, 749). In this sense, regulations for safety, privacy protection, and ethical use of sensitive information are needed (Pesapane et al., 2018b). (2) Epistemic opacity is the lack of understanding of how the AI system works and, for Ferretti et al. (2018) it is caused by procedural darkness (the rules that the AI system is following are not available) or procedural ignorance (the rules are available, but it is impossible to understand them). (3) Explanatory opacity, on the other hand, is the lack of a clinical explanation: A system might find patterns that do not have a clinical explanation with the current medical knowledge.

Deep learning conflicts with ethical requirements: The lack of understanding and transparency regarding how an AI system reaches a decision presents a major ethical concern. However, a more explainable system diminishes the power of DL (Brady and Neri, 2020; Currie et al., 2020; Quinn et al., 2021). This conflict is related to the question of whether "high stakes" institutions, such as healthcare, should use black-box AI (Brady and Neri, 2020; Bjerring and Busch, 2021; Quinn et al., 2021). Bjerring and Busch (2021) note that AI introduces some obvious differences, but also point out that black boxes do not present a fundamentally new epistemic challenge, as opaque decision-making is already common in non-AI-based medicine. By keeping the "practitioner-in-the-loop," however, at least some knowledge available to support informed decision-making. In the case of black-box medicine, "there exists no expert who can provide practitioners with useful causal or mechanistic explanations of the systems' internal decision procedures" (Bjerring and Busch, 2021, 17). Furthermore, some of the consequences of black-box medicine are epistemic in nature: Black-box medicine may lead to a loss of knowledge, and specifically to a loss of medical understanding and explanation and, thus, medical advances.

These challenges are associated with considerations about the impact of black-box AI on validity and the potential harm it presents patients. An opaque system makes it difficult to keep humans in the loop and enable them to detect errors and to identify biases. Such a system can have negative effects on underrepresented or marginalized groups and can also fail in clinical settings (Quinn et al., 2021). In addition, it can pose certain risks for radiologists, who are expected to validate something that they cannot understand (Neri et al., 2020), open them to adversarial attacks (Tizhoosh and Pantanowitz, 2018; Geis et al., 2019), or intensify the clash between black-box medicine and the duty of care, presuming that the radiologists have the ability to understand the technology, its benefits, and potential risks (Geis et al., 2019; Currie et al., 2020). The latter is also associated with depriving the patients of the ability to make decisions based on sufficient information and justifications, which contradicts the ethical requirement for the patients to exercise autonomy by giving their informed consent (Quinn et al., 2021). This type of medicine cannot be described as "patient-centered medicine" (Bjerring and Busch, 2021), and may have negative effects on the relationship of trust that is established between the patient and clinician.

Bias

The lack of transparency inherent in black-box AI tools is also a problem associated with bias. This lack is difficult to detect, measure, or correct unless the person using the tool has transparent access to the reasoning of the algorithm or the epistemic tools to understand this reasoning (Quinn et al., 2021). Radiology AI may also be biased by clinically confounding attributes such as comorbidities and by technical factors such as data set shifts and covariate shifts due to subtle differences in the raw and post-processed data that arise from the use of different scanning techniques (Geis et al., 2019), AI systems used in healthcare might have both a racial and a gender bias (Rasheed et al., 2021), but the reviewed literature discusses mainly racial bias. Many algorithms in medicine have been shown to encode, reinforce, and even exacerbate inequalities within the healthcare system (Owens and Walker, 2020) and can worsen the outcomes for vulnerable patients (Quinn et al., 2021). Such biases are introduced due to the data used to train an algorithm and the labels given to these data, which may be laden with human values, preferences, and beliefs (Geis et al., 2019). The generated outputs will thus eventually reflect social and political structures, including injustices and inequalities. Consequently, AI systems cannot provide entirely unbiased or objective outcomes based on incomplete or unrepresentative data; instead, they mirror the implicit human biases in decision-making (Balthazar et al., 2018; Pesapane et al., 2018b; Ware, 2018; Abràmoff et al., 2020). This has effects that extend beyond training, an aspect underlined by Quinn et al. (2021, 4), who point out that "most training data are imperfect because learning is done with the data one has, not the sufficiently representative, rich, and accurately labeled data one wants. [. . .] even a theoretically fair model can be biased in practice due to how it interacts with the larger healthcare system." According to Abràmoff et al. (2020) this "algorithmic unfairness" stems from model bias, model variance, or outcome noise. Model bias arises when models are selected to best represent the majority but not the unrepresented groups; model variance is caused by insufficient data from minorities, while outcome noise is caused by interference between unobserved variables and the model predictions. The latter can be avoided by broadening the scope of data to include underrepresented groups and minimize the possibility of unobserved variables interfering.

Common sources of bias that potentially promote or harm group level subsets are based on gender, sexual orientation, ethnic, social, environmental, or economic factors, as well as on unequal access to healthcare facilities and geographical bias. Referring to existing research, Owens and Walker (2020) and Quinn et al. (2021) point out racial bias that stems from the seemingly effective proxies for health needs (such as health costs) in algorithms that do not use race as a predictor for the models. Health costs are not a race-neutral proxies for health needs; this implies a need for a concerted and deep understanding of the social mechanisms of structural discrimination. Furthermore, biases in AI tools have a strong tendency to affect groups more strongly that are already suffering from discrimination based on these factors. Furthermore, AI biases have a strong tendency to affect groups more that are already suffering from discrimination based on these factors: "Blind spots in ML can reflect the worst societal biases, with a risk of unintended or unknown accuracies in minority subgroups, and there is fear over the potential for amplifying biases present in the historical data." Kelly et al. (2019, 4) note that "Blind spots in ML can reflect the worst societal biases, with a risk of unintended or unknown accuracies in minority subgroups, and there is fear over the potential for amplifying biases present in the historical data." The authors clearly illustrate this by providing the example of underperformance regarding the classification of images of benign and malignant moles on dark-skinned patients, because the algorithms are trained with data from predominantly fair-skinned patients. AI systems are often developed by companies in western countries and tested on Caucasian data, generating imbalances of representation in the datasets. "When the algorithm is trained on data that inherit biases or do not include under-represented population characteristics, existing disparities can be reinforced" (Akinci D'Antonoli, 2020, 504).

"Fairness and equality are not AI concepts" (Geis et al., 2019, 331). This statement indicates that AI tools cannot correct this type of bias on their own, but researchers developing such tools and companies providing such tools can. One solution described in the literature is to ensure diversity when collecting data and to address bias in the design, validation, and deployment of AI systems. Algorithms should be designed with the global community in mind, and clinical validation should be performed using a representative population of the intended deployment population. Careful performance analyses should be performed on the basis of population subgroups, including age, ethnicity, sex, sociodemographic stratum, and location. Understanding the impact of a new algorithm is particularly important; this means that, if the disease spectrum detected using the AI system differs from that identified using current clinical practice, then the benefits and harms of detecting this different disease spectrum must be evaluated (Kelly et al., 2019, 4–5). Owens and Walker (2020) emphasize the fact that making analyses "race neutral" is not enough and advocate taking a proactive, explicitly anti-racist approach; they even suggest that failing to recognize and anticipate structural bias in datasets or the social implications of AI systems should be considered as scientific misconduct. They urge readers to introduce a culture shift that would contribute to alleviating inequities stemming from unreflective algorithmic design. For this purpose, education on racial justice is needed at all levels, as researchers and providers often do not have the expertise to identify or address structural factors. Balthazar et al. (2018) suggest that active engagement with small population data sets is needed to consider social determinants of health and to promote access to data from underprivileged populations. Learning to identify these biases can promote "algorithmic fairness," and ML approaches might be used to correct them (Abràmoff et al., 2020). Geis et al. (2019, 331) propose certain questions that can be asked to identify bias to advance toward algorithmic fairness: How and by whom are labels generated? What kinds of bias may exist in the datasets? What are the

possible risks that might arise from those biases? What steps have we taken to mitigate these risks?

Guiding Principles

Approaches that can be taken to meet the expectations described and to tackle the challenges are often formulated as principles in the literature. This reflects an understanding of "bioethics as a scholarly discipline and its methodological approaches, with focus on the so-called "principlism" and the widely known four principles, namely beneficence, non-maleficence, autonomy, and justice" (Rasheed et al., 2021, 15). The proliferation of guidelines and recommendations makes it difficult for developers and users of AI systems to decide which ethical issues to address (Ryan and Stahl, 2020). These principles are often developed to provide guidance for many different stakeholder groups and lack specificity, presenting concepts that are often too abstract and broad and are difficult to adopt to address practical issues (Ryan and Stahl, 2020).

Explainability and Interpretability

To manage the risks inherent in the use of medical black boxes and the resulting bias, the requirement is often posed that the way an AI system arrives at its decision must be transparent and sufficiently understandable for the "human-in-the-loop" to improve patient safety and to gain the patient's trust. For that reason, "explainability" has become a key principle in the area of AI ethics, and especially in the context of healthcare.

The discourse has developed such that explainability and interpretability have become two closely associated concepts, and these concepts are often used synonymously by different authors of the reviewed literature. However, these concepts express two different directions of thought: Interpretability refers to how well one can understand how an AI system works, while explainability refers to how well one can explain what happens in AI decision-making in understandable terms (Brady and Neri, 2020; Rasheed et al., 2021). The conceptual constellation revealed in the review of the literature overlaps, often without clarity, with the concepts of interpretability, explainability, intelligibility, understandability, transparency, trustworthiness, agency, accountability, reliability, explicability, communication, and disclosure. And some authors define one term by using another. For example, explainability is defined as "AI's capacity for transparency and interpretability" and "designing explainability into AI tools is essential if they are to be trusted and if their users are to be able to exercise agency when making decisions, whether they be professional or lay users. In other words, AI must be accountable to users for the ways in which they behave" (Procter et al., 2020, 2). In other papers, explainability is associated with transparency, as in the comment "if an algorithm fails or contributes to an adverse clinical event, one needs to be able to understand why it produced the result that it did and how it reached a decision. For a model to be transparent, it should be both visible and comprehensible to outside viewers. How transparent a model should be is debatable" (Geis et al., 2019, 331). And transparency is then related, in turn, to accountability, as illustrated by Akinci D'Antonoli's comment (2020, 509) that "Transparency and accountability principles can be brought under the explicability principle. Artificial Intelligence systems should be auditable, comprehensible and intelligible by "natural" intelligence at every level of expertise, and the intention of developers and implementers or AI systems should be explicitly shared."

Overall, transparency is one of the most widely discussed principles in the AI ethics debate and is becoming one of the defining characteristics. Nevertheless, some scholars still question how much transparency AI systems should have without leaving them open to malicious attacks or intellectual property breaches (Ryan and Stahl, 2020) or enabling their misuse for harmful purposes outside the clinical context (Watson et al., 2019). Brady and Neri (2020) point out that the more explainable an AI model is, the less it can utilize the power of DL. Thus, some authors consider that transparency and explainability should be placed in a human context, as humans are often also unable to fully explain their decisions and the outcomes of their reasoning. Watson et al. (2019, 3) specifically mention that "clinicians are not always able to perfectly account for their own inferences, which may be based more on experience and intuition than explicit medical criteria." Even without the intervention of AI, complex diagnoses can be difficult to explain to other professionals or to patients. Even without the intervention of AI, complex diagnosis can be difficult to explain to other professionals or to patients. If this perspective is taken, the expectation for AI should be that "AI can explain itself at least as well as human explain their own actions and reasonings, systems would demonstrate transparency and honesty" (Ware, 2018, 21).

The issue of interpretability and explainability has interesting ramifications with reference to contestability, which is understood as the capacity of individuals (patients or medical staff) to contest and counter medical decisions (Sand et al., 2021). In line with this, the European General Data Protection Regulation (GDPR) has emphasized the patient's right to receive an explanation as a top priority in ML research. The right to an explanation encompasses the right to receive an explanation about the outputs of the algorithm, especially when decisions need to be made that significantly affect an individual. Ferretti et al. (2018, 321) explain that "the idea of a right to explanation stems from the value of transparency in data processing and it is intended to counterbalance the opacity of automated systems." Individuals have a right to protect themselves against discrimination; to do so, they have a right to know how decisions that affect them are made. In the case of AI applications in healthcare, individuals should have a right to contest (suspected) bias in the diagnostic process or the treatment selection process.

Trust and Trustworthiness

"Trust is such a fundamental principle for interpersonal interactions and is a foundational precept for society to function" (Ryan and Stahl, 2020, 74) and, thus, it is a key requirement for the ethical use of AI. As such, it has been chosen as one of the guiding principles by the High-Level Expert Group of the European Commission (2019) and identified as the defining paradigm for their ethics guidelines.

The review enabled us to find some consensus in the literature that black boxes and the lack of interpretability and explainability

can lead to a lack of trust (worthiness) in and acceptance of AI systems by clinicians and patients (Ware, 2018; Quinn et al., 2021). This aspect requires special consideration, as AI involves an element of uncertainty and risk for the vulnerable patient. Therefore, explainability is key to encouraging trust in an AI system, i.e., because people trust what they can understand (Larasati and DeLiddo, 2020). Similarly, (Spiegelhalter, 2020, 8) connects trust with explainability when proposing a series of questions about trustworthiness that include "Could I explain how it works (in general) to anyone who is interested? Could I explain to an individual how it reached its conclusion in their particular case?" Transparency becomes a fundamental factor: AI systems should be transparent enough that those using them can have access to the processes that govern them and be able to explain them. This requires access to accessible, intelligible, and usable information that can be effectively evaluated. In turn, a lack of explainability, lack of transparency, and lack of human understanding of how AI systems work will inevitably result in clinicians failing to trust decisions made by AI, as well as failing to trust the reliability and accuracy of such systems (Larasati and DeLiddo, 2020; Bjerring and Busch, 2021).

Given the fact that trust is repeatedly emphasized in the literature as a key ethical principle and mentioned as a prerequisite for the successful implementation of AI systems in medical practice, it is surprising that the authors of the reviewed papers preserve a relative silence regarding the need for an in-depth analytical approach with trust as a concept, although they echo the value of such trust. However, Quinn et al. (2021, 3) note that "the medical profession is built on various forms of trust"—and these forms of trust, its conditions, and social and institutional contexts would require a deeper analysis.

Responsibility and Accountability

AI's lack of transparency also has an impact on matters of responsibility and accountability. Ryan and Stahl (2020, 74) specifically point out that "End users should be able to justly trust AI organizations to fulfill their promises and to ensure that their systems function as intended [...]. Building trust should be encouraged by ensuring accountability, transparency and safety of AI." In that sense, "criminal liability, the tort of negligence, and breach of warranty must be discussed before utilizing AI in medicine" (Matsuzaki, 2018, 268). Neri et al. (2020) pose the question of who is responsible for benefits and harms resulting from the use of AI in radiology, and, like Akinci D'Antonoli (2020), claim that radiologists remain responsible for the diagnosis when using AI, even if they might be validating something unknown that is based on black boxes and possible automation bias. Therefore, radiologists should be taught how to use AI tools appropriately and familiarized with the guiding principles for increasing trust in AI. (Geis et al., 2019, 333) underlined this point effectively by stating that "Radiologists will remain ultimately responsible for patient care and will need to acquire new skills to do their best for patients in the new AI ecosystem."

Sand et al. (2021) argue that the kind of accountability and responsibility that is being pursued in medical AI is connected to liability and blame. As an alternative, they propose a "forward-looking responsibility," which "can be understood as a safeguard to decrease the risk of harm in cases of cognitive misalignment between the physicians and the AI system—when an AI output cannot be confirmed (verified or falsified)" (Sand et al., 2021, 3). Accordingly, the authors list the following responsibilities of clinicians: the duty to report uncertainty (sensitivity/specificity rates) to the patients; to understand and critically assess whether AI outputs are reasonable given a certain diagnostic procedure; to know and understand the input data and its quality; to have an awareness of their own experience and decline in skills; to have an awareness and understanding of the specificity of the task; and to assess, monitor, and report the output development over time.

One of the challenges of AI application in healthcare is the role of private companies who own the AI systems. Ryan and Stahl (2020, 71) mention the risk that companies try to "obfuscate blame and responsibility." This lack of transparency regarding who is truly responsible and accountable further complicates issues of liability and undermines the ability of clinicians to act with integrity. Mudgal and Das (2020, 7) note that this lack of transparency and the subsequent problems that arise could be solved by maintaining a "human-in-the-loop" perspective, keeping the liability and responsibility within the field of responsibility of the radiologist and their employer.

Justice and Fairness

Justice is one of the four principles of bioethics: autonomy, beneficence, non-maleficence, and justice (Beauchamp and Childress, 2001). Some of the reviewed sources refer to some extent to which these four principles apply to AI (Akinci D'Antonoli, 2020; Currie et al., 2020; Rasheed et al., 2021). Justice is also one of the three principles proposed in the Belmont Report (United States National Commission for the Protection of Human Subjects of Biomedical Behavioral Research, 1978), one of the most widely recognized standards for biomedical ethics. In this report, justice refers to the idea that the benefits and costs of research and medical care should be distributed fairly (Larson et al., 2020).

Along with trust, transparency, accountability, and other principles, "diversity, non-discrimination and fairness" are principles that were proposed by the High-Level Experts Group on Artificial Intelligence of the European Commission in 2018. As Neri et al. (2020, 519) state, "the group recommended that the development, deployment and use of AI systems should adhere to the ethical principles of respect for human autonomy, prevention of harm, fairness/equity and explicability." The principle of justice often appears to be associated with beneficence and non-maleficence, as the unfair distribution of resources leads to discrimination and can cause harm. (Geis et al., 2019, 330) pointed out that it is necessary to "inspire radiology AI's builders and users to enhance radiology's intelligence in humane ways to promote just and beneficial outcomes while avoiding harm to those who expect the radiology community to do right by them." The association between injustice, discrimination, and unfair decisions made by AI systems has been also linked to bias in the reviewed literature, as "discrimination and unfair outcomes stemming from algorithms has become a hot topic within the

media and academic circles" (Ryan and Stahl, 2020, 67). Biased AI systems lead to unfair, discriminatory behavior or mistaken decisions (Morley et al., 2020) and to the aforementioned "algorithmic unfairness" (Abràmoff et al., 2020).

Integrating AI systems in medicine incurs the risk of replicating discriminations that already exist in society; therefore, "the development of AI should promote justice while eliminating unfair discriminations, ensuring shareable benefits, and preventing the infliction of new harm that can arise from implicit bias" (Akinci D'Antonoli, 2020, 508–509). AI tools can decide in favor of one group of patients due to implicit biases rather than prioritizing a real emergency in radiology, underlining the necessity for everybody involved in the process to adhere to ethical guidelines that promote justice.

DISCUSSION

This literature review was carried out to identify ethical issues discussed in the recent academic literature associated with the use of AI in healthcare and to determine how these are being tackled in view of biomedical research, and especially in radiology and oncology imaging. This review enabled us to identify key themes which place a focus on expectations about medical AI, challenges posed by the use of this technology, and approaches that can be taken to ensure ethical AI use. Most of these themes are formulated by the authors as principles. In this section of this article, we critically discuss our findings from an ethical and social science perspective.

Several expectations are expressed in the literature regarding the potential for medical AI use to improve diagnostic performance and patient outcomes, but the socio-technological conditions under which these expectations can be met, and, at the same time, challenges can be managed are not clearly defined. We previously quoted that "the state of AI hype has far exceeded the state of AI science, especially when it pertains to validation and readiness for implementation in patient care" (Topol, 2019, 51). This statement illustrates an important gap: The contexts in which medical AI tools are being implemented have not been thoroughly explored. Considering the results of our review, this holds particularly true regarding the close connection between AI algorithms and societal structures. Although some scholars have discussed the fact that AI use "can increase systemic risks of harm, raise the possibility of errors with high consequences, and amplify complex ethical and societal issues" (Geis et al., 2019, 330), few studies have clearly defined exactly how AI tools interact with pre-existing systemic harm, how they can contribute to this harm, or how complex ethical and societal issues might be amplified through the use of such tools. In the reviewed literature, we identified a need for profound, specific, and interdisciplinary conversations about how firmly AI is embedded in systemic structures and power relations that intersect with identity traits (e.g., gender, race, class, ability, education) and about the implications of private ownership and the role of corporations, profit-making, and geopolitical structures.

Bias

In that sense, we have observed that bias has not been framed in the context of power relations and societal conditions, nor has it been referenced to the existing body of research on, e.g., how gender and race shapes and affects biomedicine and healthcare practice (Roberts, 2008; Schiebinger and Schraudner, 2011; Oertelt-Prigione, 2012; Kaufman, 2013) or how gender and racial bias in algorithms could have a negative impact in certain areas of society (e.g., O'Neil, 2016; Noble, 2018). Bias has been shown to affect every stage of data processing (i.e., in generating, collecting, and labeling data that are used to train AI tools) and to affect the variables and rules used by the algorithms. Hence, AI tools can be taught to discriminate, reproduce social stereotypes, and underperform in minority groups, an especially risky proposition in the context of healthcare (Char et al., 2018; Wiens et al., 2019).

In the analyzed sample, little attention was given to sex and gender bias in AI systems used in healthcare. Nonetheless, research has already been done to analyze in detail how sex and gender bias is generated, how it affects patients and society, and how its effects can be mitigated. Using sex- and gender-imbalanced datasets to train deep-learning-based systems may affect the performance of pathology classification with minority groups (Larrazabal et al., 2020). Other authors also show that these social categories could influence the diagnosis although there is no direct link to the disease, and that potentially missed detection of breast cancer at mammography screening was greater among socioeconomically disadvantaged groups (Rauscher et al., 2013). Unfortunately, most of the currently used biomedical AI technologies do not account for bias detection, and most algorithms ignore the sex and gender dimensions and how these contribute to health and disease. In addition, few studies have been performed on intersex, transgender, or non-binary individuals due to narrow and binary background assumptions regarding sex and gender (Cirillo et al., 2020). Ignoring how certain identity traits affect the application of AI systems in healthcare can lead to the production of skewed datasets and harm certain minority people and groups. Applying feminist standpoint theory (Haraway, 1988; Hekman, 1997), some authors argue that all knowledge is socially situated and that the perspectives of oppressed groups are systematically excluded from general knowledge and practices that ignore the specific identity traits of certain individuals. Based on this argument, knowledge must be presented in a way that enables people to be aware of intersecting power relations that influence its production. The results of our literature review indicate that, rather than ignoring sex, gender, or race dimensions, close attention must be paid to these dimensions in datasets (Zou and Schiebinger, 2018; Larrazabal et al., 2020), even to the extent of introducing an amount of desirable bias to counteract the effects of undesirable biases that result in unintended or unnecessary discrimination (Cirillo et al., 2020; Pot et al., 2021).

Diversity in the datasets becomes an increasingly important point that is being addressed by researchers to counteract bias that can be potentially harmful (Leavy, 2018). Nonetheless, ensuring diversity in and of itself is not enough (Li et al.,

2022); more research is needed to understand how discrimination intersects with socioeconomic factors to keep bias from being introduced into healthcare algorithms through structural inequalities in society (Quinn et al., 2021). Anticipating structural bias in datasets and understanding the social implications of using AI systems before their implementation is considered best practice; some authors in the sample even propose that failing to do so should be qualified as scientific misconduct (Owens and Walker, 2020). This will require reflecting on how social categories are constructed in big data-driven research and on how the underlying social classification and categorization systems are incorporated into and reproduced in the knowledge produced from analyzing the existing datasets (Goisauf et al., 2020).

Lack of Analytic Accuracy

We observed that explainability and interpretability were often used interchangeably with other terms such as understandability and even transparency in our sample, as clear definitions of and analytic distinction between the terms are lacking. The lack of analytical precision that can be observed in the ethics of AI literature often leads to a lack of specificity and vague assumptions that do not enable scholars to reach the core of certain issues that are associated with epistemic justice (Fricker, 2007). The GDPR, for instance, states that subjects have a right to understand their lived experiences, especially experiences of injustice. Although research addresses the problem of how this right to an explanation is outlined in the legislation (Edwards and Veale, 2017), we argue that the lack of knowledge about why and how certain decisions that impact (negatively) our lives are made constitutes a specific wrongful act, i.e., epistemic injustice (Fricker, 2007). This injustice results in someone being wronged specifically in their capacity as a possessor of knowledge; they are wronged, therefore, in a capacity essential to human value. The opacity of AI and the implications of the use of AI tools makes it difficult for patients to exercise their autonomy. This inability is consequently also reflected in their practical limitation to give their informed consent and affects their capacity to contest decisions. To address epistemic injustices, knowledge must be made available to people affected by the decisions made by AI technology.

In our sample literature, the possibility of making information available and understandable is often treated as a technical feature of AI. It may then seem as though these issues are technical problems that can be solved by applying technical solution that deal with black boxes. Again, we have observed a need to take a social sciences perspective and to achieve a broader understanding of how our epistemic capabilities are also intertwined with power relations. In "AI ethics, technical artifacts are primarily seen as isolated entities that can be optimized by experts so as to find technical solutions for technical problems. What is often lacking is a consideration of the wider contexts and the comprehensive relationship networks in which technical systems are embedded" (Hagendorff, 2020, 103). It will be necessary to carefully consider the structures that surround the production and distribution of knowledge by performing further analyses of the ethics of AI in healthcare.

Trust

Trust was often mentioned as an important factor in the reviewed literature, and trustworthiness has become a key principle regarding ethical AI. As we have shown, a clear definition and deeper understanding of the complexities of trust in AI are lacking. In the reviewed literature, for example, we found that trustworthiness is conflated with acceptance (Gaube et al., 2021) or explainability (Larasati and DeLiddo, 2020). Some authors have mentioned that "a possible imbalance in the data should be considered when developing the model to ensure the trustworthiness of the model" (Alabi et al., 2020, 7). However, for a model to be considered worthy of trust, more than simple technical solutions that even out technical "imbalances" in the training phase are needed, and especially when a risk of gender or racial bias exists. This is a more complex issue that will need to be addressed. Also, while it is important to encourage trust in technology, trust is built on the foundation of social relations. Healthcare practitioner-patient relationships are based on trust and empathy (Morley et al., 2020), and decision-making in the medical context, and especially in connection to technology, is often based on "gut feelings" (Goisauf and Durnová, 2018).

Previous research has shown that trust cannot be understood as unidirectional. Instead, trust needs to be understood as a complex, situated, context-dependent, and relational concept that involves several trustor/trustee relationships, such as trust in persons (e.g., scientists who trust each other, patients who trust scientists and clinicians), technology, and institutions (Wyatt et al., 2013; Bijker et al., 2016). Trust involves "the willingness to accept vulnerability based on positive expectations about another's intentions or behaviors [...] Trust makes decision making more efficient by simplifying the acquisition and interpretation of information. Trust also guides action by suggesting behaviors and routines that are most viable and beneficial under the assumption that the trusted counterpart will not exploit one's vulnerability" (McEvily et al., 2003, 92–93). In building trust, embodied experience matters, and this experience occurs as an emotional reaction, e.g., in the form of the aforementioned "gut feelings" (Goisauf and Durnová, 2018). Trust or more precisely trusting relationships are fragile and require continuous work, which means that they need to be actively established and sustained. This includes trustworthiness (i.e., the idea that a person or object is worthy of being trusted), which is a key requisite for the sustainability of a trusting relationship (McEvily et al., 2003). To ensure trustworthiness, researchers must understand how trusting relationships are constituted *via* the social process, how trust in technologies is established and sustained, and under what conditions AI can be deemed trustworthy.

This discussion places an emphasis on trusting relationships between a practitioner and patient regarding medical AI use, the expectations and brings the needs of these actors into focus. Unfortunately, this is rarely the case in the reviewed literature, as relatively little attention is paid to the patients' and radiologists' perspectives, with only a few exceptions (e.g., Balthazar et al., 2018). However, (Ferretti et al., 2018, 331) stated that "more research is needed to understand patients'

and physicians' attitudes toward opacity in AI systems." Patients clearly want to be informed about how their health data are used (also a requirement of the GDPR), and the engagement of members of the public, patients, practitioners, and those developing the technology will be crucial to build trust and ensure both public and professional support.

CONCLUSIONS AND FUTURE DIRECTIONS

Performing this literature review, we have looked back at how current discourses revolve around the ethical and societal issues related to AI use in radiology. We have identified imaginaries of science and technology as aspects that are neutral, universal, and detached from societal structures, imaginaries that have already been described in the philosophy of science and STS fields (Haraway, 1988; Longino, 1990; Fox Keller, 1996).

We have observed that the current literature discourse does not delve into the broader origins and implications of bias, especially when bias is treated only as a technical problem with a technical solution. We believe that integrating a social science perspective into the analysis of ethical and societal issues associated with AI use in radiology is crucial to understanding the scope of these issues. To thoroughly address the topic of ethical AI use in radiology, a perspective must be taken to analyze how science is situated in a certain socioeconomic context and to understand the application of AI systems in medicine as a situated practice. Understanding the socioeconomic context is a fundamental step that will enable scholars to gain this perspective. In the future, inter- and trans-disciplinary research should be carried out to help situate knowledge production and its ethical and societal implications. In this sense, it will be necessary to shift from DL about to a deep understanding of the societal implications, and in particular to an understanding of the interactions of social values and categories with scientific knowledge production, of the relations between knowledge and societal trust that affects how science functions in society, and especially of how new technologies are perceived and accepted in society.

This review and the ensuing discussion also enabled us to identify a lack of precision regarding the use of terms for principles that have been proposed to apply AI technology more ethically in the future. Terms such as trustworthiness, transparency, or trust are extensively used in the literature, often without clearly defining specifically how they are meant or used. Researchers working in the field of ethics of AI in medicine will need to strive for accuracy and precision by providing clear definitions for these concepts in this specific context and by situating them within a broader context. In order to do this, interdisciplinary research with social scientists but also with clinicians in order to incorporate clinical concepts (Lekadir et al., 2021, 31) will be crucial.

More interdisciplinary and concrete research will deepen our understanding of biases in radiology. Adopting an intersectional perspective that takes into consideration how different traits of our identity intersect will be crucial, especially in the case of breast cancer. As previous research has shown, other factors that intersect with gender contribute to the formation of bias, such as ethnicity, skin color, socioeconomics, geography or breast density (Lekadir et al., 2021). In this regard, the issue of gender bias in female-only datasets requires a more detailed analysis. Considering breast cancer in connection to gender can lead to the abridged conclusion that gender bias could not have a significant impact. However, this reflects a one-dimensional understanding of gender as a social category, since gender is never isolated, but occurs at the intersection with other categories. Therefore, women cannot be assumed to be a homogeneous group, but are differentiated along other categories such as age, race, and socioeconomic background, which, as has been shown, could have an influence on breast cancer diagnosis.

In conclusion, the value of AI for radiology would increase by integrating a more precise and interdisciplinary consideration of the societal context in which AI is being developed to generate more just outcomes and allow all members of society equal access to the benefits of these promising applications.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary materials, further inquiries can be directed to the corresponding author/s.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work and approved it for publication.

FUNDING

This study was funded by EuCanImage, a project that has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 952103.

REFERENCES

Abràmoff, M. D., Tobey, D., and Char, D. S. (2020). Lessons learned about autonomous AI: finding a safe, efficacious, and ethical path through the development process. *Am. J. Ophthalmol.* 214, 134–142. doi: 10.1016/j.ajo.2020.02.022

Akinci D'Antonoli, T. (2020). Ethical considerations for artificial intelligence: an overview of the current radiology landscape. *Diagn. Interv. Imaging* 26, 504–511. doi: 10.5152/dir.2020.19279

Alabi, R. O., Vartiainen, T., and Elmusrati, M. (2020). "Machine learning for prognosis of oral cancer: what are the ethical challenges?" in *Proceedings of the Conference on Technology Ethics 2020 – Tethics 2020: CEUR Workshop Proceedings)*.

Balthazar, P., Harri, P., Prater, A., and Safdar, N. M. (2018). Protecting your patients' interests in the era of big data, artificial intelligence, and predictive analytics. *J. Am. Coll. Radiol.* 15, 580–586. doi: 10.1016/j.jacr.2017.11.035

Beauchamp, T. L., and Childress, J. F. (2001). *Principles of Biomedical Ethics.* Oxford, New York, NY: Oxford University Press.

Bijker, E. M., Sauerwein, R. W., and Bijker, W. E. (2016). Controlled human malaria infection trials: how tandems of trust and control construct scientific knowledge. *Soc. Stud. Sci.* 46, 56–86. doi: 10.1177/0306312715619784

Bjerring, J. C., and Busch, J. (2021). Artificial intelligence and patient-centered decision-making. *Philos. Technol.* 34, 349–371. doi: 10.1007/s13347-019-00391-6

Brady, A. P., and Neri, E. (2020). Artificial intelligence in radiology—ethical considerations. *Diagnostics* 10, 231. doi: 10.3390/diagnostics10040231

Char, D. S., Shah, N. H., and Magnus, D. (2018). Implementing machine learning in health care – addressing ethical challenges. *N. Engl. J. Med.* 378, 981–983. doi: 10.1056/NEJMp1714229

Choy, G., Khalilzadeh, O., Michalski, M., Do, S., Samir, A. E., Pianykh, O. S., et al. (2018). Current applications and future impact of machine learning in radiology. *Radiology* 288, 318–328. doi: 10.1148/radiol.2018171820

Cirillo, D., Catuara-Solarz, S., Morey, C., Guney, E., Subirats, L., Mellino, S., et al. (2020). Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *NPJ Digit. Med.* 3, 81. doi: 10.1038/s41746-020-0288-5

Currie, G., Hawk, K. E., and Rohren, E. M. (2020). Ethical principles for the application of artificial intelligence (AI) in nuclear medicine. *Eur. J. Nucl. Med. Mol. Imaging* 47, 748–752. doi: 10.1007/s00259-020-04678-1

Edwards, L., and Veale, M. (2017). Slave to the algorithm: why a right to an explanation is probably not the remedy you are looking for. *Duke Law Technol. Rev.* 16, 18. doi: 10.31228/osf.io/97upg

Ferretti, A., Schneider, M., and Blasimme, A. (2018). Machine learning in medicine: opening the new data protection black box. *Eur. Data Prot. Law Rev.* 4, 320. doi: 10.21552/edpl/2018/3/10

Fox Keller, E. (1996). *Reflections on Gender and Science: Tenth Anniversary Paperback Edition*. New Haven, CT: Yale University Press.

Fricker, M. (2007). *Epistemic Injustice: Power and the Ethics of Knowing.* Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780198237907.001.0001

Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S. J., Lermer, E., et al. (2021). Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ Digit. Med.* 4, 31. doi: 10.1038/s41746-021-00385-9

Geis, J. R., Brady, A. P., Wu, C. C., Spencer, J., Ranschaert, E., Jaremko, J. L., et al. (2019). Ethics of artificial intelligence in radiology: summary of the Joint European and North American Multisociety Statement. *Can. Assoc. Radiol. J.* 70, 329–334. doi: 10.1016/j.carj.2019.08.010

Goisauf, M., Akyüz, K., and Martin, G. M. (2020). Moving back to the future of big data-driven research: reflecting on the social in genomics. *Humanit. Soc. Sci. Commun.* 7, 55. doi: 10.1057/s41599-020-00544-5

Goisauf, M., and Durnová, A. (2018). From engaging publics to engaging knowledges: enacting "Appropriateness" in the Austrian Biobank Infrastructure. *Public Underst. Sci.* 28, 275–289. doi: 10.1177/0963662518806451

Hagendorff, T. (2020). The ethics of AI ethics: an evaluation of guidelines. *Minds Mach.* 30, 99–120. doi: 10.1007/s11023-020-09517-8

Haraway, D. (1988). Situated knowledges: the science question in feminism and the privilege of partial perspective. *Fem. Stud.* 14, 575–599. doi: 10.2307/3178066

Hekman, S. (1997). Truth and method: feminist standpoint theory revisited. *Signs: J. Women Cult. Soc.* 22, 341–365. doi: 10.1086/495159

High-Level Expert Group on Artificial Intelligence (2019). *Ethics Guidelines for Trustworthy AI*. Brussels: European Commission.

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., and Aerts, H. J. W. L. (2018). Artificial intelligence in radiology. *Nat. Rev. Cancer* 18, 500–510. doi: 10.1038/s41568-018-0016-5

Kaufman, J. S. (2013). "Chapter 4. "Ethical Dilemmas in statistical practice: the problem of race in biomedicine," in *Mapping "Race": Critical Approaches to Health Disparities Research*, eds. E. G. Laura, and L. Nancy (Ithaca: Rutgers University Press), 53–66. doi: 10.36019/9780813561387-007

Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., and King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* 17, 195. doi: 10.1186/s12916-019-1426-2

Kim, J., Kim, H. J., Kim, C., and Kim, W. H. (2021). Artificial intelligence in breast ultrasonography. *Ultrasonography* 40, 183–190. doi: 10.14366/usg.20117

Krupinski, E. A. (2020). An ethics framework for clinical imaging data sharing and the greater good. *Radiology* 295, 683–684. doi: 10.1148/radiol.2020200416

Larasati, R., and DeLiddo, A. (2020). "Building a trustworthy explainable AI in healthcare," in *Human Computer Interaction and Emerging Technologies: Adjunct Proceedings from the INTERACT 2019 Workshops*, eds. F. Loizides, M. Winckler, U. Chatterjee, J. Abdelnour-Nocera, and A. Parmaxi (Cardiff: Cardiff University Press), 209–214. doi: 10.18573/book3.ab

Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H., and Ferrante, E. (2020). Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc. Nat. Acad. Sci. U. S. A.* 117, 12592–12594. doi: 10.1073/pnas.1919012117

Larson, D. B., Magnus, D. C., Lungren, M. P., Shah, N. H., and Langlotz, C. P. (2020). Ethics of using and sharing clinical imaging data for artificial intelligence: a proposed framework. *Radiology* 295, 675–682. doi: 10.1148/radiol.2020192536

Leavy, S. (2018). "Gender bias in artificial intelligence: the need for diversity and gender theory in machine learning," in: *Proceedings of the 1st International Workshop on Gender Equality in Software Engineering* (Gothenburg: Association for Computing Machinery). doi: 10.1145/3195570.3195580

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Lekadir, K., Osuala, R., Gallin, C., Lazrak, N., Kushibar, K., Tsakou, G., et al. (2021). FUTURE-AI: guiding principles and consensus recommendations for trustworthy artificial intelligence in medical imaging. *arXiv preprint arXiv:*2109.09658.

Li, J., Bzdok, D., Chen, J., Tam, A., Ooi, L. Q. R., Holmes, A. J., et al. (2022). Cross-ethnicity/race generalization failure of behavioral prediction from resting-state functional connectivity. *Sci. Adv.* 8, eabj1812. doi: 10.1126/sciadv.abj1812

Longino, H. E. (1990). *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton: Princeton University Press. doi: 10.1515/9780691209753

Matsuzaki, T. (2018). Ethical issues of artificial intelligence in medicine. *Calif. West. Law Rev.* 55, 255–273.

Mazurowski, M. A. (2020). Artificial Intelligence in radiology: some ethical considerations for radiologists and algorithm developers. *Acad. Radiol.* 27, 127–129. doi: 10.1016/j.acra.2019.04.024

McEvily, B., Perrone, V., and Zaheer, A. (2003). Trust as an organizing principle. *Organ. Sci.* 14, 91–103. doi: 10.1287/orsc.14.1.91.12814

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., et al. (2020). International evaluation of an AI system for breast cancer screening. *Nature* 577, 89–94. doi: 10.1038/s41586-019-1799-6

Mittelstadt, B. D., and Floridi, L. (2016). The ethics of big data: current and foreseeable issues in biomedical contexts. *Sci. Eng. Ethics* 22, 303–341. doi: 10.1007/s11948-015-9652-2

Morley, J., Machado, C. C. V., Burr, C., Cowls, J., Joshi, I., Taddeo, M., et al. (2020). The ethics of AI in health care: a mapping review. *Soc. Sci. Med.* 260, 113172. doi: 10.1016/j.socscimed.2020.113172

Mudgal, K. S., and Das, N. (2020). The ethical adoption of artificial intelligence in radiology. *BJR Open* 2, 20190020. doi: 10.1259/bjro.20190020

Naqa, I. E., Haider, M. A., Giger, M. L., and Haken, R. K. T. (2020). Artificial intelligence: reshaping the practice of radiological sciences in the 21st century. *Br. J. Radiol.* 93, 20190855. doi: 10.1259/bjr.20190855

Neri, E., Coppola, F., Miele, V., Bibbolino, C., and Grassi, R. (2020). Artificial intelligence: who is responsible for the diagnosis? *Radiol. Med.* 125, 517–521. doi: 10.1007/s11547-020-01135-9

Noble, S. U. (2018). Algorithms of oppression. New York, NY: New York University Press. doi: 10.2307/j.ctt1pwt9w5

Oertelt-Prigione, S. (2012). The influence of sex and gender on the immune response. *Autoimmun. Rev.* 11, A479–A485. doi: 10.1016/j.autrev.2011.11.022

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York, NY: Crown.

Owens, K., and Walker, A. (2020). Those designing healthcare algorithms must become actively anti-racist. *Nat. Med.* 26, 1327–1328. doi: 10.1038/s41591-020-1020-3

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372, n71. doi: 10.1136/bmj.n71

Pesapane, F., Codari, M., and Sardanelli, F. (2018a). Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine. *Eur. Radiol. Exp.* 2, 35. doi: 10.1186/s41747-018-0061-6

Pesapane, F., Volonté, C., Codari, M., and Sardanelli, F. (2018b). Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States. *Insights Imaging* 9, 745–753. doi: 10.1007/s13244-018-0645-y

Pot, M., Kieusseyan, N., and Prainsack, B. (2021). Not all biases are bad: equitable and inequitable biases in machine learning and radiology. *Insights Imaging* 12, 13. doi: 10.1186/s13244-020-00955-7

Price, W. N. (2018). "Medical malpractice and black-box medicine," in *Big Data, Health Law, and Bioethics*, eds. I. Cohen, H. Lynch, E. Vayena, and U. Gasser (Cambridge: Cambridge University Press), 295–306. doi: 10.1017/9781108147972.027

Procter, R., Rouncefield, M., and Tolmie, P. (2020). Accounts, accountability and agency for safe and ethical AI. *arXiv preprint arXiv:*2010.01316.

Quinn, T. P., Jacobs, S., Senadeera, M., Le, V., and Coghlan, S. (2021). The three ghosts of medical AI: can the black-box present deliver? *Artif. Intell. Med.* 102158. doi: 10.1016/j.artmed.2021.102158

Rasheed, K., Qayyum, A., Ghaly, M., Al-Fuqaha, A., Razi, A., and Qadir, J. (2021). Explainable, trustworthy, and ethical machine learning for healthcare: a survey. *Preprint*. 1–26. doi: 10.36227/techrxiv.14376179

Rauscher, G. H., Khan, J. A., Berbaum, M. L., and Conant, E. F. (2013). Potentially missed detection with screening mammography: does the quality of radiologist's interpretation vary by patient socioeconomic advantage/disadvantage? *Ann. Epidemiol.* 23, 210–214. doi: 10.1016/j.annepidem.2013.01.006

Roberts, D. E. (2008). Is race-based medicine good for us?: African American approaches to race, biomedicine, and equality. *J. Law Med. Ethics* 36, 537–545. doi: 10.1111/j.1748-720X.2008.302.x

Ryan, M., and Stahl, B. C. (2020). Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. *J. Inf. Commun. Ethics Soc.* 19, 61–86. doi: 10.1108/JICES-12-2019-0138

Sand, M., Durán, J. M., and Jongsma, K. R. (2021). Responsibility beyond design: physicians' requirements for ethical medical AI. *Bioethics*. 36, 1–8. doi: 10.1111/bioe.12887

Schiebinger, L., and Schraudner, M. (2011). Interdisciplinary approaches to achieving gendered innovations in science, medicine, and engineering 1. *Interdiscip. Sci. Rev.* 36, 154–167. doi: 10.1179/030801811X13013181961518

Smith, M. J., and Bean, S. (2019). AI and ethics in medical radiation sciences. *J. Med. Imaging Radiat. Sci.* 50, S24–S26. doi: 10.1016/j.jmir.2019.08.005

Spiegelhalter, D. (2020). Should we trust algorithms? *Harvard Data Sci. Rev.* 2, 1–12. doi: 10.1162/99608f92.cb91a35a

Tang, A., Tam, R., Cadrin-Chênevert, A., Guest, W., Chong, J., Barfett, J., et al. (2018). Canadian Association of Radiologists White Paper on artificial intelligence in radiology. *Can. Assoc. Radiol. J.* 69, 120–135. doi: 10.1016/j.carj.2018.02.002

Terry, G., Hayfield, N., Clarke, V., and Braun, V. (2017). "Thematic analysis," in *The SAGE Handbook of Qualitative Research in Psychology*, eds C. Willig, and W. Stainton Rogers (London: Sage), 17–37. doi: 10.4135/9781526405555.n2

Tizhoosh, H. R., and Pantanowitz, L. (2018). Artificial intelligence and digital pathology: challenges and opportunities. *J. Pathol. Inform.* 9, 1–6. doi: 10.4103/jpi.jpi_53_18

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* 25, 44–56. doi: 10.1038/s41591-018-0300-7

United States National Commission for the Protection of Human Subjects of Biomedical Behavioral Research (1978). *The Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research*. New York, NY: The Commission.

Ware, A. B. (2018). Algorithms and automation: fostering trustworthiness in artificial intelligence. *Honors Theses Capstones* 416, 1–37.

Watson, D. S., Krutzinna, J., Bruce, I. N., Griffiths, C. E., McInnes, I. B., Barnes, M. R., et al. (2019). Clinical applications of machine learning algorithms: beyond the black box. *BMJ* 364, l886. doi: 10.1136/bmj.l886

Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., et al. (2019). Do no harm: a roadmap for responsible machine learning for health care. *Nat. Med.* 25, 1337–1340. doi: 10.1038/s41591-019-0548-6

Wyatt, S., Harris, A., Adams, S., and Kelly, S. E. (2013). Illness online: self-reported data and questions of trust in medical and social research. *Theory Cult. Soc.* 30, 131–150. doi: 10.1177/0263276413485900

Zou, J., and Schiebinger, L. (2018). AI can be sexist and racist—it's time to make it fair. *Nature* 559, 324–326. doi: 10.1038/d41586-018-05707-8

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers | Frontiers in Virtual Reality

TYPE Hypothesis and Theory
PUBLISHED 04 August 2022
DOI 10.3389/frvir.2022.894162

OPEN ACCESS

EDITED BY
Marientina Gotsis,
University of Southern California,
United States

REVIEWED BY
Andrea Chirico,
Sapienza University of Rome, Italy
Ali Fardinpour,
Wise Realities Institute for Healthcare Emerging Technologies Research, Australia

*CORRESPONDENCE
Hélène Buche,
buchehelene@gmail.com
Aude Michel,
aude.michel@univ-montp3.fr

SPECIALTY SECTION
This article was submitted to Virtual Reality in Medicine, a section of the journal Frontiers in Virtual Reality

RECEIVED 14 April 2022
ACCEPTED 06 July 2022
PUBLISHED 04 August 2022

CITATION
Buche H, Michel A and Blanc N (2022), Use of virtual reality in oncology: From the state of the art to an integrative model.
Front. Virtual Real. 3:894162.
doi: 10.3389/frvir.2022.894162

Use of virtual reality in oncology: From the state of the art to an integrative model

Hélène Buche[1]*, Aude Michel[1,2]* and Nathalie Blanc[1]

[1]Univ Paul Valéry Montpellier 3, Epsylon Ea 4556, Montpellier, France, [2]Montpellier Institut Du Sein, Clinique Clémentville, Montpellier, France

Over the past 20 years, virtual reality (VR) has been the subject of growing interest in oncology. More and more researchers are studying the effects of virtual environments to contribute to current thinking on technologies likely to support patients undergoing oncological treatment. Recent research highlights how VR can divert attention while reducing anxiety in stressful healthcare situations through its multisensory and participative nature. VR appears to be a promising tool capable of reducing cancer-related anxiety symptoms, improving treatment adherence, and increasing satisfaction with oncology care. While the literature reports these positive effects in the therapeutic management of cancer, few studies have focused on theoretical models capable of explaining the psychological benefits of virtual immersion. This literature review provides a theoretical framework combining results from all relevant empirical work in oncology. The review can help researchers identify the optimal conditions for using VR in oncology and bridge the gap between divergent devices, modalities, and practices (e.g., headmounted displays, environments, interactivity, immersion time).

KEYWORDS

cancer, anxiety, pain, immersion, presence, interaction, equipment

1 Introduction

For the past 30 years, the number of new cancer cases has been steadily increasing. The National *Cancer* Institute (Institut National du Cancer, 2019) reported 328,000 diagnoses in metropolitan France in 2018 compared to 320,000 in 2005. The most common cancers in men were prostate cancer (48,427 new cases in 2013), followed by lung (32,500 cases) and colorectal (24,000 cases). In women, breast cancer was the most frequent (59,000 cases), followed by colon-rectal cancer (21,000 cases) and lung cancer (17,000 cases) (Defossez et al., 2019). Many stress factors have been identified at different times in cancer management, including diagnosis, treatment, and long-term management of the disease (Chirico et al., 2015). Among patients treated for cancer, 55% met clinical criteria for an anxiety disorder (O'Connor et al., 2010), with an increase to 77% in patients who received chemotherapy (Nikbakhsh et al., 2014). In addition, the prevalence of cancer-related pain was 39.3% in patients who received curative treatment, increasing to 55% in patients undergoing cancer treatment and reaching 71% in advanced or metastatic

cancer (Van den Beuken-van Everdingen et al., 2016; Alawneh et al., 2017). Many stress agents and physical symptoms can cause increased emotional distress (Arrieta et al., 2013).

In this context, virtual reality (VR) is the object of interest and curiosity in cancerology. More and more researchers are studying the effects of VR to improve the conditions of oncological treatments (Pittara et al., 2020). Most studies have highlighted the benefits of VR, which, thanks to its distraction power, can divert attention while reducing the anxiety and pain of patients facing particularly distressing care situations (Chirico et al., 2016; Ahmad et al., 2020). Although the literature focuses on the positive effects of this tool in the context of cancer treatment, few studies have focused on the theoretical models of cognitive science that explain and try to understand the benefits of VR. Rather than viewing it as a technical medium, in-depth research based on an appropriate theoretical framework is needed to explore the complexity of virtual environments (de Loor and Tisseau, 2011). Only these foundations can give scientific legitimacy to this technological revolution (de Loor and Tisseau, 2011) and provide us with elements of knowledge on the mechanisms that promote patients' emotional wellbeing. Let us note that beyond understanding the mechanisms, these foundations could be used as support to design specialized interfaces adapted to different clinical situations.

VR became more accessible for consumer use after 2016 (Tsaï, 2016). It is "the application that allows the user to navigate and interact in real-time with a three-dimensional environment generated by a computer" (Pratt et al., 1995). This artificial environment is usually made possible using a computer screen that responds to the individual's head movements by providing synthetic sensory stimuli such as images of real or imaginary landscapes, spatialized sounds, and sometimes tactile or olfactory feedback (Chirico et al., 2016; Chirico et al., 2019). VR equipment also includes devices that allow action in the virtual world, such as a mouse, keyboard, or more sophisticated game controllers (Pittara et al., 2020; Indovina et al., 2018). In other words, different systems offer users different sensations and levels of involvement.

The development of high-performance virtual reality devices accelerates innovation focused on health to facilitate the realization of cancer care by offering a quality immersive device allowing patients to escape from their distress and painful medical situations (Pittara et al., 2020; Ahmad et al., 2020). Immersion in a virtual environment is considered both as a distractor (reducing anxiety and pain) (Chirico et al., 2019; Bani Mohammad and Ahmad, 2018; Ahmadpour et al., 2020) and as a tool for emotional regulation (reducing negative emotions, inducing positive emotions), allowing improvement in care tolerance (Pizzoli et al., 2019).

The benefits of VR were first observed in oncology during chemotherapy sessions. The results were encouraging (Oyama et al., 1999; Schneider and Workman, 1999), promoting a decrease in anxiety, an improvement in mood as well as an underestimation of care time (Schneider and Workman, 1999; Schneider et al., 2003), (Schneider et al., 2004; Schneider and Hood, 2007). Today, the distractive power of VR is of interest in a range of oncology situations ranging from palliative care (Niki et al., 2019; Johnson et al., 2020) to the support of hospitalized patients during various medical procedures (Pittara et al., 2020; Ahmad et al., 2020; Zeng et al., 2019).

Although several studies have emphasized the effectiveness of VR distraction in oncology, the virtual reality devices used are wildly divergent in terms of content, intervention strategies, and technological qualities. It is thus necessary to go beyond the wonder and attractiveness that VR arouses to resist this technological hype toward rethinking and resituating its use within our knowledge of the human. This literature review aims to take stock of the benefits of using VR as a distraction tool for anxiety and pain management in oncology. To this end, the results known to date are listed, and their analysis is considered according to the methodology used. This literature review aims to bring out the points of consensus and the methodological divergences in the research while emphasizing that few interventional studies are theoretically anchored. Based on this review of the available literature, recommendations will be made to enable the research community to move towards common methodological choices and thus improve clinical practice. Another aim of this literature review is to leverage the theoretical foundations identified toward a theoretical model that will allow us to think about the contributions of VR in oncology, especially the cognitive and emotional processes involved.

2 Method

2.1 Data source and search method

Based on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) method, we proceeded stepwise using six computerized databases: Google Scholar, PubMed, PsychInfo, Academic Search Premier, Ebsco, and Sciencedirect to search for relevant studies. We limited the search to 10 years (2011–2021). In each database, we used the same search terms: virtual reality and cancer, virtual reality and oncology, virtual reality and anxiety, virtual reality and cancer and anxiety, virtual reality and pain, virtual reality and cancer and pain. We also manually searched bibliographic references of included studies and previously published systematic reviews.

2.2 Study selection

Our inclusion criteria incorporated studies explicitly examining the effectiveness of VR as a distraction tool in oncology. In this sense, we excluded all studies that were

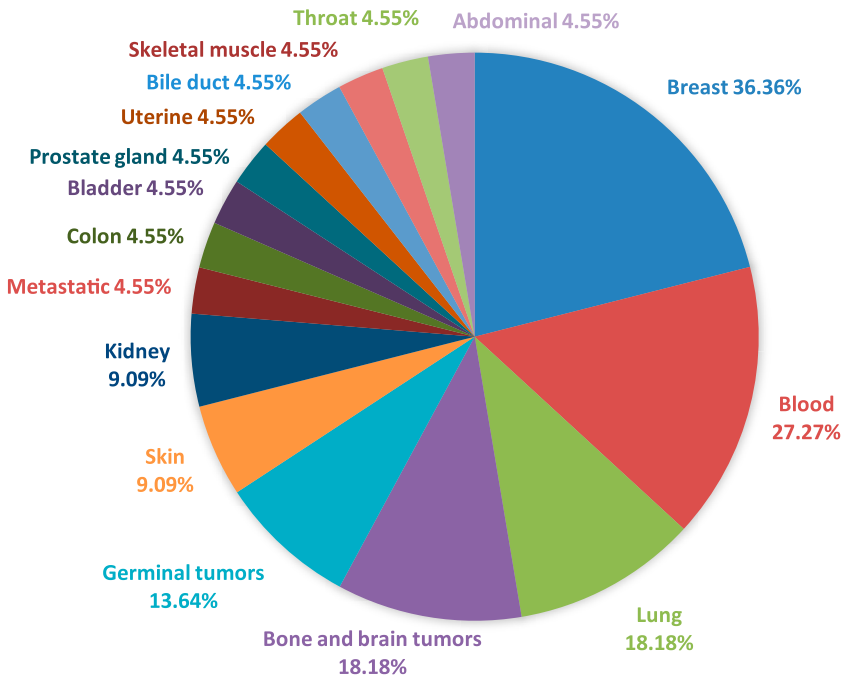

FIGURE 1
Type of cancer in which VR has been proposed.

unrelated to cancer and all research conducted with cancer populations whose purpose was not associated with distraction to improve emotional state and decrease pain.

2.3 Data collection

To collect the data, we extracted all relevant information from the selected articles into an Excel file: characteristics of the study population sample, type of cancer, psychological variables, VR equipment, environments, immersive tasks, methodology, objectives of the studies, medical context, stated theoretical frameworks and main results, as well as current limitations of VR and its future direction.

2.4 Data analysis

The selected articles were subjected to a literature review to exploit and classify the results according to recurrent characteristics that allowed the different studies to be compared. The selected characteristics included VR equipment, immersive modalities, environments, effectiveness of VR in oncology, theoretical basis for the benefits of VR, limitations, and future direction of VR distraction to decrease pain intensity and anxiety in clinical situations.

3 Results

3.1 State of the art presentation

3.1.1 Characteristics of the studies

3.1.1.1 Population

Nearly three-quarters of the selected studies evaluating the intervention of VR during the management of cancer patients (1,153 participants aged 6–85 years) were conducted with adults (72, 73%, 16/22 studies) (Chirico et al., 2019; Bani Mohammad and Ahmad, 2018; Pizzoli et al., 2019; Niki et al., 2019; Johnson et al., 2020; Schneider et al., 2011; Espinoza et al., 2012; Baños et al., 2013; Li et al., 2016; Glennon et al., 2018; Gupta and Hande, 2019; Higgins et al., 2019; Garrett et al., 2020; Gerçeker et al., 2020; Scates et al., 2020; Buche et al., 2021). The remaining studies were conducted in pediatric oncology (27, 27%, 6/22 studies) (Li et al., 2011; Atzori et al., 2018; Birnie et al., 2018; Semerci et al., 2020; Sharifpour et al., 2020; Tennant et al., 2020).

3.1.1.2 Type of cancer in which virtual reality has been proposed

The qualitative analysis of these studies reveals a clear diversity in the medical context for evaluating the effects of VR according to the type of cancer. As shown in Figure 1, more than a third of the studies were performed during the treatment of breast cancer (36.36%) (Chirico et al., 2019; Bani Mohammad

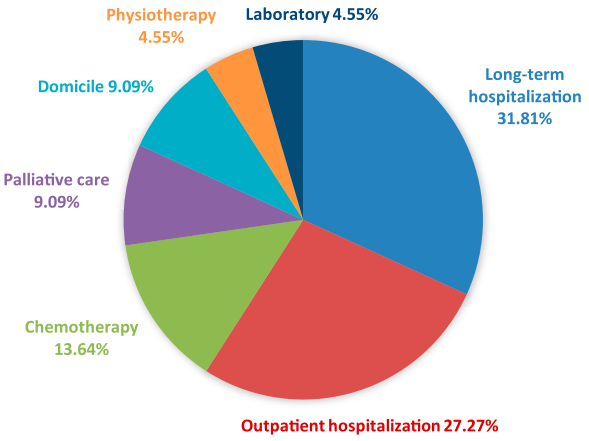

FIGURE 2
Context of VR intervention in oncology.

and Ahmad, 2018; Pizzoli et al., 2019; Schneider et al., 2011; Espinoza et al., 2012; Gupta and Hande, 2019; Garrett et al., 2020; Buche et al., 2021) and almost a third during the management of blood cancer (e.g., leukemia) and/or lymphatic system (e.g., lymphoma) (27.27%) (Glennon et al., 2018; Garrett et al., 2020; Li et al., 2011; Atzori et al., 2018; Birnie et al., 2018; Tennant et al., 2020). A few studies have examined the effects of VR during treatment of lung cancer (Niki et al., 2019; Schneider et al., 2011; Espinoza et al., 2012; Garrett et al., 2020), bone cancer and brain tumors (Li et al., 2011; Birnie et al., 2018; Sharifpour et al., 2020; Tennant et al., 2020) (18.18% each). Few studies included patients with germ cell tumors (13.64%) (Li et al., 2011; Sharifpour et al., 2020; Tennant et al., 2020), skin cancer (9.09%) (Higgins et al., 2019; Tennant et al., 2020), or kidney cancer (9.09%) (Niki et al., 2019; Garrett et al., 2020), while some types of cancer were invoked only once in VR applicability (4.55% each) (see Figure 1: Type of cancer in which VR has been proposed) (Niki et al., 2019; Schneider et al., 2011; Espinoza et al., 2012; Baños et al., 2013; Garrett et al., 2020; Sharifpour et al., 2020).

3.1.2 Context of virtual reality intervention in oncology

In addition to the types of cancer, studies have evaluated the benefits of VR according to the context of VR use (see Figure 2: Contexts of Use). In the context of long-term hospitalization, VR is used as a distraction tool to promote emotional and physical well-being (31.81%, 7/22 studies) (Bani Mohammad and Ahmad, 2018; Espinoza et al., 2012; Baños et al., 2013; Gupta and Hande, 2019; Higgins et al., 2019; Li et al., 2011; Tennant et al., 2020). In the context of day hospitalization, it is proposed in particular when patients have to undergo a painful medical procedure (i.e., catheter port placement, venipuncture, IV station, bone marrow aspiration and biopsy) to reduce acute pain (27.27%, 6/22 studies) (Glennon et al., 2018; Gerçeker et al., 2020; Scates et al., 2020; Atzori et al., 2018; Birnie et al., 2018; Semerci et al., 2020). Its application in oncology is no longer limited to chemotherapy sessions (13.64%, 3/22 studies) (Chirico et al., 2019; Schneider et al., 2011; Sharifpour et al., 2020). Distraction under VR is now used in palliative care (9.09%, 2/22 studies) to relieve symptoms in terminally ill patients (Niki et al., 2019; Johnson et al., 2020) and at home (9.09%, 2/22 studies) to manage patients' chronic pain (Garrett et al., 2020), alleviate symptoms of psychological distress and promote patient empowerment (Li et al., 2016). In physiotherapy, this distraction strategy has recently been proposed during post-mastectomy scar massage sessions by comparing participative and contemplative distraction (4.55%, 1/22 studies) (Buche et al., 2021). Finally, only one study went outside the medical context to test the first virtual laboratory experiment measuring the effects of VR associated with two different relaxation techniques (i.e., breath control vs. Body Scanning Procedure) on breast cancer patients (4.55%, 1/22 studies) (Pizzoli et al., 2019).

3.1.3 Benefits of virtual reality in oncology

Distraction is a non-pharmacological technique increasingly used by healthcare professionals to alleviate anxiety and pain related to medical procedures (Bani Mohammad and Ahmad, 2018; Gold et al., 2007). The underlying mechanism of the power of distraction relies on the limited cognitive resources of an individual's attention (Arane et al., 2017). An engaging and attractive distractor diverts the patients' attention and hinders their ability to process external negative stimuli, decreasing anxiety, and pain (Gold et al., 2007; Kleiber and McCarthy, 2006). Two forms of distraction can be distinguished: a passive form (e.g., watching television, listening to music) and an active form (e.g., electronic games) (Arane et al., 2017; Koller and Goldman, 2012). Thus, using a distractor is a cognitive strategy that can passively redirect the patients' attention or actively involve them in a task (Gold et al., 2007; Kleiber and Harper, 1999). VR is a powerful distractor as it can offer several degrees of involvement by immersing the patient in a contemplative or participative environment that mobilizes several senses (Chirico et al., 2019; Ahmadpour et al., 2020; Buche et al., 2021). The multimodal aspect of VR induces a subjective feeling of being present in the environment (Chirico et al., 2019).

On the one hand, the effectiveness of VR lies in the intensity of this multisensory immersion called the sense of presence (Tennant et al., 2020), that is, the subjective experience of being in another place than the one where the individual is physically located (Witmer and Singer, 1998). On the other hand, its effectiveness depends on the patients' sensory, cognitive, and emotional involvement as well as the level of acceptability of this tool (Garrett et al., 2020). The degree of engagement and interactivity are closely related to the sense of presence and increased attention to distraction, leading to an increase in the positive effects of VR (Birnie et al., 2018).

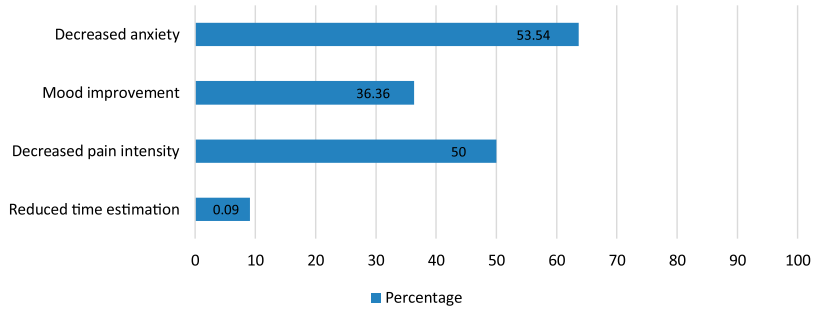

FIGURE 3
Percentage of studies evaluating the effects of VR in oncology.

3.1.3.1 Anxiety/stress

The benefits of VR have been shown to affect anxiety in cancer patients (Chirico et al., 2019; Bani Mohammad and Ahmad, 2018; Pizzoli et al., 2019; Niki et al., 2019; Espinoza et al., 2012; Baños et al., 2013; Gupta and Hande, 2019; Higgins et al., 2019; Garrett et al., 2020; Gerçeker et al., 2020; Scates et al., 2020; Li et al., 2011; Tennant et al., 2020; Buche et al., 2021). Two-thirds of the selected studies focused on anxiety relief (14/22 studies, see Figure 3: Percentage of studies evaluating the effects of VR in oncology) (Chirico et al., 2019; Bani Mohammad and Ahmad, 2018; Niki et al., 2019; Johnson et al., 2020; Schneider et al., 2011; Li et al., 2016; Glennon et al., 2018; Gupta and Hande, 2019; Higgins et al., 2019; Gerçeker et al., 2020; Scates et al., 2020; Li et al., 2011; Tennant et al., 2020; Buche et al., 2021). In most cases, the application of VR as a distraction tool promotes a significant decrease in anxiety during chemotherapy sessions (Chirico et al., 2019), during hospitalization (Bani Mohammad and Ahmad, 2018; Niki et al., 2019; Gupta and Hande, 2019; Higgins et al., 2019; Tennant et al., 2020), during painful procedures (Gerçeker et al., 2020; Scates et al., 2020) and physiotherapy rehabilitation (Buche et al., 2021). Participative VR seems to be a more effective distractive strategy than music for improving emotional wellbeing (Chirico et al., 2019). Distraction is defined by Lazarus and Folkman's (1984) stress and coping model (Lazarus et al., 1984) as a coping strategy, namely the set of cognitive and behavioral efforts intended to control, reduce, or tolerate an aversive situation (Chirico et al., 2019). Distraction under VR regulates patients' emotional responses related to distressing medical procedures through selective attention that focuses attention on pleasant stimuli in the virtual environment. Thus, using participative VR is an active "vigilant" strategy, while listening to music is a distractive strategy that requires only passive attentional engagement on the part of patients.

Moreover, immersion in a natural environment significantly enhances the power of distraction by, among other things, leading to increased feelings of peace and relaxation in patients (Scates et al., 2020). Scates et al. (2020) support Kaplan and Kaplan's (1989) (Kaplan and Kaplan, 1989) attention restoration theory that natural environments can refocus attention but also Ulrich et al.'s (1991) psychophysiological stress recovery theory (Ulrich et al., 1991) where positive distractions involving natural elements (e.g., trees, flowers, streams, etc.) help individuals combat stress. Beyond natural content, (Niki et al., 2019) speculate that retrieval of episodic memories involving the medial temporal lobe may promote decreased anxiety and depression (Ramirez et al., 2015). Thus, they suggest that the hippocampal region is particularly involved in the biological mechanisms by which a VR simulating a pleasant place already visited by the individual in the real world would alleviate anxiety and depression.

3.1.3.2 Mood improvement

As for the studies focused on mood improvement (8/22 studies), they generally show that VR can promote the emotional wellbeing of patients (Pizzoli et al., 2019; Niki et al., 2019; Buche et al., 2021; Li et al., 2011) by increasing positive emotions such as joy or happiness and decreasing negative emotions such as fear (Gerçeker et al., 2020), sadness (Espinoza et al., 2012; Baños et al., 2013) and anger (Tennant et al., 2020). Baños et al. (2013) refer to the broaden-and-build theory proposed by Fredrickson et al. (2001), which is based on positive psychology. According to this theory, the promotion and experience of positive emotions expand individuals' momentary repertoires of thought-action. The ability to experience positive emotions can create and strengthen lasting personal resources that are useful for coping with difficult times during cancer management.

3.1.3.3 Perception of pain

Half of the studies presented in Table 1 focused on the reduction of pain intensity in oncology (11/22 studies) (Bani Mohammad and Ahmad, 2018; Niki et al., 2019; Johnson et al., 2020; Glennon et al., 2018; Garrett et al., 2020; Gerçeker et al., 2020; Atzori et al., 2018; Birnie et al., 2018; Semerci et al., 2020; Sharifpour et al., 2020; Tennant et al., 2020). The different

TABLE 1 Studies on the benefits of virtual reality.

Study	Objectives	Procedure	Theoretical framework	Results
Schneider et al. (2011)	To decrease anxiety and reduce perceived treatment time	Chemotherapy	The pacemaker– accumulator cognitive model of time perception Burle and Casini, (2001); Wittmann and Paulus, (2008); Droit-Volet and Gil, (2009)	Reduction of the perceived time during the intervention
Li et al. (2011)	To evaluate the benefits of therapeutic VR games to help children cope with hospital anxiety and depression	Hospitalization		Decrease in depression
Espinoza et al. (2012)	To induce positive emotions and improve emotional wellbeing	Hospitalization		Improvement of distress and happiness level; Increase of positive emotions (joy, relaxation); Decrease of negative emotions (sadness, anxiety)
Baños et al. (2013)	To induce positive emotions and improve the emotional wellbeing of patients with metastatic cancer	Hospitalization	Fredrickson's theory (2001) broaden-and-build theory; Fredrickson, (2001)	Increase in positive emotions (joy, relaxation); Decrease in negative emotions (sadness, anxiety)
Li et al. (2016)	To alleviate symptoms of psychological distress and promote patient autonomy through low-cost VR distraction	At home		Relaxing environment for most participants
Atzori et al. (2018)	To control pain in young patients during venipuncture with VR distraction	Painful procedure Venipuncture	The Eccleston and Crombez's (1999) Attention Pain Theory; Eccleston and Crombez, (1999)	Decrease in pain
Birnie et al. (2018)	To manage pain (pain management) in young patients using distraction in VR	Painful procedure: Implantable Venous Access (IVAD)		Fun and enjoyable pain management; Interactivity, engagement, and pleasure influence the sense of presence resulting in a decrease in the intensity of acute pain
Glennon et al. (2018)	To determine the effects of VR on pain and anxiety	Painful procedure: Bone marrow aspiration and biopsy		No significant effects on pain and anxiety
Bani Mohammad and Ahmad, (2018)	To decrease pain intensity and anxiety	Hospitalization		Improvement of morphine analgesia; Decreased anxiety
Chirico et al. (2019)	To relieve psychological distress through distraction and improve treatment tolerance	Chemotherapy	The Lazarus and Folkman's stress and coping model (1984) Lazarus et al. (1984)	Decreased anxiety after VR and music therapy; More effective than music therapy in decreasing anxiety (NS), depression and fatigue
Gupta and Hande, (2019)	To decrease hospital anxiety	Hospitalization after surgery (mastectomy)		Decreased anxiety and depression
Higgins et al. (2019)	To minimize feelings of anxiety or pain	Ambulatory surgery		Significant improvement in patient anxiety and satisfaction with VR, no decrease in pain intensity
Niki et al. (2019)	To improve the various symptoms of terminal cancer patients	Palliative		Decreased all cancer-related symptoms in both conditions, but NS for the "Places desired to visit but never visited" group
Pizzoli et al. (2019)	To promote emotional wellbeing through two relaxation exercises in VR	Laboratory		Soothing and pleasant state after each relaxation exercise under VR, but more relaxation after the body scan
Sharifpour et al. (2020)	To evaluate the effect of VR therapy on chemotherapy-related pain	Chemotherapy	The gate control theory of pain, Reduction of attentional bias related to pain; Melzack and Wall, (1996)	Improvement in pain intensity, anxiety, catastrophizing and self-efficacy; The positive effect of VR remained constant in the 1st and 2nd follow-up period
Garrett et al. (2020)	To manage chronic pain (chronic pain management) through daily VR therapy	At home		Immersive VR distraction facilitated a sense of presence, drawing attention away from pain; Improved sleep quality and emotional state
Gerçeker et al. (2020)	Distraction under VR: to decrease pain intensity, fear and anxiety related to Huber's needle	Painful procedure Port access		Decreased pain intensity, fear, and needle anxiety in pediatric hematology-oncology patients

(Continued on following page)

TABLE 1 (*Continued*) Studies on the benefits of virtual reality.

Study	Objectives	Procedure	Theoretical framework	Results
Johnson et al. (2020)	To examine the utility of VR for terminal cancer patients	Palliative		Pleasant, useful and globally well tolerated; Tendency to improve pain, fatigue, drowsiness, depression and anxiety (NS)
Scates et al. (2020)	To determine if distraction by immersion in a natural virtual environment can decrease pain intensity and anxiety	Painful procedure: port access, venipuncture, IV station	Kaplan and Kaplan's (1989) attention restoration theory; Kaplan and Kaplan, (1989), psychophysiological stress recovery theory Ulrich et al. (1991)	Increased relaxation and feelings of peace, considerable distraction, reduced frustration
Semerci et al. (2020)	To decrease pain intensity with VR distraction	Painful procedure: Port access		Decrease in pain intensity; Can be considered as a complementary intervention
Tennant et al. (2020)	To determine the effects of VR on psychophysiological symptoms by comparing them to the effects of the iPad	Hospitalization		Decrease in negative symptoms more important with VR; Positive mood regardless of content; Decrease in pain more important with natural content; Decrease in anger more important after high immersion
Buche et al. (2021)	To compare two immersive modalities (participatory vs. contemplative) to listening to music and the presence of a practitioner to improve emotional state after breast surgery	Physiotherapy		Increase in positive emotions (i.e., joy and happiness) and decrease in anxiety regardless of the proposed accompaniment; More intense spatial presence with participatory VR; Reduction in perceived time with VR

Note. NS, Non-Significant.

results show that immersion in an artificial world is associated with an analgesic effect (Bani Mohammad and Ahmad, 2018; Niki et al., 2019; Garrett et al., 2020; Gerçeker et al., 2020; Atzori et al., 2018; Birnie et al., 2018; Semerci et al., 2020; Sharifpour et al., 2020; Tennant et al., 2020). VR is a pleasant and effective distraction strategy used to reduce pain during medical procedures that can be painful for patients, such as venipuncture (Atzori et al., 2018) or veinous port access (Gerçeker et al., 2020; Birnie et al., 2018; Semerci et al., 2020). The immersive and participative experience can significantly reduce the acute pain associated with treatments (Bani Mohammad and Ahmad, 2018; Birnie et al., 2018; Sharifpour et al., 2020; Tennant et al., 2020) and reduce chronic pain (Niki et al., 2019; Garrett et al., 2020). According to Eccleston and Crombez's (1999) Attention Pain Theory (Eccleston and Crombez, 1999), the illusion of being in an artificial world and the patients' interaction with objects in the virtual environment may reduce the amount of attention available to deal with painful stimuli, thus decreasing the perception of conscious pain (Atzori et al., 2018). Within the theory of Melzack and Wall (1960) (Melzack and Wall, 1996) entitled "Gate Control Theory of Pain," the nervous system contains a neurological gateway controlled by the cortex that could either block the ascending and descending pain signals or allow their transmission to the brain to continue (Sharifpour et al., 2020). For example, attention and negative emotions such as fear and sadness can open this gateway, increasing pain perception. In contrast, distraction and positive emotions such as joy and calmness can close this gateway, decreasing pain perception. When the gateway is open, nociceptive messages are allowed to reach the brain; when it is closed, nociceptive messages are inhibited. Based on this model, distraction under VR can alleviate pain by decreasing negative emotions and favoring positive emotions, thus inducing a decrease in pain perception. In other words, virtual reality generates a slower reaction to pain reporting by acting on attention, emotion, and in a broader sense, cognition (Gold et al., 2007), (Arane et al., 2017).

3.1.3.4 Temporal perception

In the past 10 years, few studies have addressed the issue of time perception in oncology (Schneider et al., 2003; Schneider et al., 2004), (Schneider and Hood, 2007). One study (Schneider et al., 2011), based on the simulation-accumulation cognitive model (Burle and Casini, 2001; Wittmann and Paulus, 2008; Droit-Volet and Gil, 2009), explains the effects of distraction intervention on the perception of time. It seems that time spent under virtual immersion passes more quickly due to the decrease in heart rate and negative stimuli of the stressful context, thus diverting attention from processing temporal information.

3.1.4 Technological diversity

Although the literature has identified the advantages of distraction under VR in oncology (Michel et al., 2019a), the

variety of the tools and methods used should be highlighted to define the optimal conditions for using VR and propose interfaces adapted to support cancer patients.

3.1.4.1 Hardware used

The vast majority of studies examined in this literature review take advantage of fully immersive devices through an HMD headset (86.36%, 19/22 studies) (Chirico et al., 2019; Bani Mohammad and Ahmad, 2018; Pizzoli et al., 2019; Niki et al., 2019; Johnson et al., 2020; Schneider et al., 2011; Li et al., 2016; Glennon et al., 2018; Gupta and Hande, 2019; Higgins et al., 2019; Garrett et al., 2020; Gerçeker et al., 2020; Scates et al., 2020; Buche et al., 2021; Atzori et al., 2018; Birnie et al., 2018; Semerci et al., 2020; Sharifpour et al., 2020; Tennant et al., 2020), while a minority (9.09%, 2/22 studies) use a device that researchers describe as "non-immersive" virtual reality for clinical purposes in oncology via a 32-inch LCD television screen connected to a computer, keyboard, mouse, and headset (Espinoza et al., 2012; Baños et al., 2013). Overall, the immersive devices used are smartphone VR headsets with the distinction of being low-cost systems (68.42%, 13/19 studies) (Chirico et al., 2019; Bani Mohammad and Ahmad, 2018; Pizzoli et al., 2019; Johnson et al., 2020; Li et al., 2016; Gupta and Hande, 2019; Higgins et al., 2019; Gerçeker et al., 2020; Scates et al., 2020; Buche et al., 2021; Semerci et al., 2020; Sharifpour et al., 2020; Tennant et al., 2020). In some cases, smartphones VR headsets are accompanied by headphone (Bani Mohammad and Ahmad, 2018; Birnie et al., 2018), or earphones (Pizzoli et al., 2019; Atzori et al., 2018), and joysticks (hand controllers) (Li et al., 2016; Birnie et al., 2018). Few researchers opt for systems as high-tech as the HCT VIVE headset (Niki et al., 2019; Higgins et al., 2019; Garrett et al., 2020), ez Vision X4 (Gupta and Hande, 2019) or Oculus Go (Buche et al., 2021) (26.31%, 5/19 studies). One study exploited a particular VR system (PlayMotion) (4.55%, 1/22 studies) in a playroom of a pediatric oncology department. This system has the particularity of increasing the immersive space by transforming the room into a totally intuitive and participative virtual environment since it does not require a headset or a controller. The software responds to patients' actions by analyzing the shadows of moving limbs projected on the walls thanks to sensors.

3.1.4.2 Immersive environments

Regarding the content of virtual environments, a consensus emerges around natural relaxing environments (90.91% or 20/22 studies) (Chirico et al., 2019; Bani Mohammad and Ahmad, 2018; Pizzoli et al., 2019; Niki et al., 2019; Johnson et al., 2020; Schneider et al., 2011; Espinoza et al., 2012; Baños et al., 2013; Li et al., 2016; Glennon et al., 2018; Higgins et al., 2019; Garrett et al., 2020; Gerçeker et al., 2020; Scates et al., 2020; Atzori et al., 2018; Birnie et al., 2018; Semerci et al., 2020; Sharifpour et al., 2020; Tennant et al., 2020; Buche et al., 2021) rather than urban ones (13.64%, or 3/22 studies) (Espinoza et al., 2012; Baños et al., 2013), (Li et al., 2011). Thanks to the extent of research, we now have a range of natural environments that correspond to the demand of patients (Michel et al., 2019b). On the one hand, the environments are built with synthetic images such as sea worlds (Schneider et al., 2011; Glennon et al., 2018; Higgins et al., 2019; Gerçeker et al., 2020; Buche et al., 2021; Birnie et al., 2018; Sharifpour et al., 2020), forests (Pizzoli et al., 2019; Espinoza et al., 2012; Baños et al., 2013; Garrett et al., 2020; Gerçeker et al., 2020; Buche et al., 2021) paradise islands (Chirico et al., 2019; Bani Mohammad and Ahmad, 2018; Sharifpour et al., 2020; Buche et al., 2021) and mountains (Chirico et al., 2019) with waterfalls (Pizzoli et al., 2019), and on the other hand, the environments are created with images captured in 360° of real world destinations (Niki et al., 2019; Johnson et al., 2020; Gerçeker et al., 2020; Tennant et al., 2020). For some of them, this natural component is complemented by playful content (50%, 11/22 studies) which includes, for example, roller coaster simulations (Johnson et al., 2020; Gerçeker et al., 2020; Semerci et al., 2020) or space travel (Johnson et al., 2020; Garrett et al., 2020). Some studies include educational (Bani Mohammad and Ahmad, 2018; Gerçeker et al., 2020), enigmatic (Schneider et al., 2011; Garrett et al., 2020), creative (Higgins et al., 2019; Li et al., 2011), cultural (Schneider et al., 2011; Tennant et al., 2020), musical (Garrett et al., 2020), or sports games (Li et al., 2011) environments. Some studies are not standardized and vary accordingly to content by integrating mixed environments (i.e., playful and relaxing) (18.18%, 4/22 studies) (Johnson et al., 2020; Higgins et al., 2019; Garrett et al., 2020; Gerçeker et al., 2020) with still images while others involve videos (Johnson et al., 2020).

3.1.4.3 Interactivity

The diversity of the devices also concerns the levels of sensorimotor interactivity. Contemplative VR inviting patients to observe the virtual environment (45.45%, 10/22 studies) (Pizzoli et al., 2019; Niki et al., 2019; Espinoza et al., 2012; Baños et al., 2013; Glennon et al., 2018; Garrett et al., 2020; Buche et al., 2021; Semerci et al., 2020; Sharifpour et al., 2020; Tennant et al., 2020), is opposed to participative VR, called participative VR, which offers patients the possibility to act as an actor in the virtual world (27.27%, 6/22 studies) (Chirico et al., 2019; Li et al., 2016; Buche et al., 2021; Li et al., 2011; Atzori et al., 2018; Birnie et al., 2018). Almost a third of the studies do not control for this participative variable that involves patients to different degrees in immersive experiences (27.27%, 6/22 studies) (Bani Mohammad and Ahmad, 2018; Johnson et al., 2020; Schneider et al., 2011; Higgins et al., 2019; Garrett et al., 2020; Gerçeker et al., 2020) or do not report on the sensorimotor interaction between patients and the virtual device (4.55%, 1/22 studies) (Gupta and Hande, 2019). Contemplative immersions consist of passive observation of virtual environments (Bani Mohammad and Ahmad, 2018; Glennon et al., 2018) with sometimes the possibility of navigating (Espinoza et al., 2012; Baños et al., 2013; Buche et al., 2021;

Tennant et al., 2020) or performing meditation (Johnson et al., 2020; Garrett et al., 2020) and relaxation exercises such as the control of breathing frequencies (Espinoza et al., 2012; Baños et al., 2013) or the focusing of attention on physical sensations to improve emotional wellbeing (Pizzoli et al., 2019). Participatory immersions offer multiple possibilities of actions such as participative explorations by body limb movements (Chirico et al., 2019; Bani Mohammad and Ahmad, 2018; Gerçeker et al., 2020; Buche et al., 2021; Li et al., 2011) or educational ones by information retrieval (Bani Mohammad and Ahmad, 2018; Li et al., 2011). Explorations require solving mysteries by strategically choosing different options to advance in the scenario (Schneider et al., 2011; Garrett et al., 2020). Others consist in modifying objects in the environment (Li et al., 2016; Buche et al., 2021) or in painting one's environment in three dimensions (Higgins et al., 2019; Li et al., 2011). Finally, target games allow the patient to aim at characters or objects present in the environment by pointing with the use of game controllers (Johnson et al., 2020; Birnie et al., 2018) or a computer mouse and keyboard (Atzori et al., 2018).

3.1.4.4 Audio and sound

Apart from the visual contents and their participative potentialities, there is a form of consensus on the need to solicit the auditory sensory modality (Bani Mohammad and Ahmad, 2018; Pizzoli et al., 2019; Johnson et al., 2020; Espinoza et al., 2012; Baños et al., 2013; Li et al., 2016; Glennon et al., 2018; Garrett et al., 2020; Gerçeker et al., 2020; Scates et al., 2020; Buche et al., 2021; Semerci et al., 2020; Sharifpour et al., 2020). This auditory component is thought to favor the immersive experience that increases the intensity of the sense of presence in the virtual world. However, we notice a certain heterogeneity regarding the aural characteristics of the proposed devices. Some immersions are enhanced by a background sound related to the virtual environment (e.g., nature sounds, sound feedback, educational narration) (Bani Mohammad and Ahmad, 2018; Gerçeker et al., 2020; Scates et al., 2020; Buche et al., 2021; Sharifpour et al., 2020), whereas others are accompanied by soothing musical stimuli (Li et al., 2016; Glennon et al., 2018; Garrett et al., 2020; Buche et al., 2021; Semerci et al., 2020) associated with guided relaxation (Johnson et al., 2020; Espinoza et al., 2012; Baños et al., 2013) with the help of a qualified yoga and mindfulness instructor (Pizzoli et al., 2019).

3.1.5 Methodological diversity

3.1.5.1 Experimental design

Beyond the technological diversity, there are differences in the scientific methodologies used. These differences can be observed in terms of the comparison of experimental methods. Almost half of the studies do not compare distraction under VR to a control group or to another form of distraction (40.91%, 9/22 studies) (Pizzoli et al., 2019; Niki et al., 2019; Johnson et al., 2020; Espinoza et al., 2012; Baños et al., 2013; Li et al., 2016; Gupta and Hande, 2019; Higgins et al., 2019; Birnie et al., 2018). As for the control groups, they consist of apprehending the medical act without distraction (50%) (Chirico et al., 2019; Bani Mohammad and Ahmad, 2018; Niki et al., 2019; Schneider et al., 2011; Glennon et al., 2018; Gerçeker et al., 2020; Scates et al., 2020; Buche et al., 2021; Atzori et al., 2018; Semerci et al., 2020; Sharifpour et al., 2020). Thus, the difference between the groups could be due to using a distractive device rather than the specific use of VR. Only 22.73% of the research (5/22 studies) compared the virtual device to another distractive mode, either by presenting the same content through another medium (i.e., computer, television, or tablet: 13.64%) (Glennon et al., 2018; Garrett et al., 2020; Tennant et al., 2020) or by comparing VR to music (9.09%) (Chirico et al., 2019; Buche et al., 2021). According to the reported results, VR was more conducive to reducing negative symptoms with a greater decrease in anger levels after more intense immersion (Tennant et al., 2020). A gender effect was found with a higher increase in positive mood with VR than with iPad in young females (Tennant et al., 2020). Therefore, VR may be a more powerful form of distraction than tablet games by facilitating a sense of presence in a new environment diverting attention from pain (Garrett et al., 2020). Similarly, VR has been shown to be more effective than music therapy in relieving depression and fatigue (Chirico et al., 2019). VR was also more effective than listening to classical music in reducing estimated care time regardless of whether the immersion was participative or passive (Buche et al., 2021).

3.1.5.2 Familiarization

Only six out of twenty-two studies implemented a familiarization phase before starting the real immersive experience (27.27%), (Chirico et al., 2019; Johnson et al., 2020; Gupta and Hande, 2019; Gerçeker et al., 2020; Atzori et al., 2018; Scates et al., 2020). The studies that implemented this familiarization phase in their research protocol showed significant results in reducing anger, pain, and anxiety (83.33%, 5/6 studies). This step might be necessary to decrease the surprise effect and the naive attractiveness of the patients to obtain a more accurate measure of their emotional states associated with the virtual immersion (Buche et al., 2021). These familiarization phases nevertheless present methodological differences. The most frequent method consists of the experimenter accompanying the patients to guide them during their first manipulations (50%, 3/6 studies), (Chirico et al., 2019; Johnson et al., 2020; Atzori et al., 2018). In comparison, others consist in viewing a handholding video during which the patient can practice (16.66%, 1/6 studies) (Tennant et al., 2020) or start the immersion a few minutes before the medical procedure (16.66%, 1/6 studies) (Gerçeker et al., 2020). In daily VR exposures, this familiarization phase can result in a short immersion of 10 min on the first day of experimentation with a progressive increase in immersion time going up to 30 min per day (16.66%, 1/6 studies) (Gupta and Hande, 2019).

TABLE 2 Nature of measure and instruments in studies evaluating the effects of virtual reality in oncology.

Study	Nature of measure	Instruments
Schneider et al. (2011)	Anxiety Fatigue Temporality	State Anxiety Inventory (STAI) Piper Fatigue Scale (PFS) Oral questions
Li et al. (2011)	Anxiety Mood state	Chinese Version of the State Anxiety Scale for Children (CSAS-C) Center for Epidemiologic Studies Depression Scale for Children (CES-DC)
Espinoza et al. (2012)	Anxiety - Depression Mood state Pain - Fatigue	Hospital Anxiety and Depression Scale (HADS) Fordyce Questionnaire, Visual Analogical Scales (VAS) Mood VAS Physical Discomfort
Baños et al. (2013)	Mood state Pain - Fatigue Cyber Sickness Virtual experience	VAS Mood VAS Physical Discomfort Open-ended questions about side effects VAS Satisfaction, Open ended questions on the level of engagement, the difficulties encountered, the immersive experience
Li et al. (2016)	Anxiety Cyber Sickness Virtual experience	Semi-structured interview Motion Sickness Susceptibility Questionnaire (MSSQ Short Version) Semi-structured interview for the VR interface
Atzori et al. (2018)	Pain Cyber Sickness Virtual experience	VAS Pain VAS Nausea VAS quality and pleasure of the VR experience
Birnie et al. (2018)	Anxiety Pain Cyber Sickness Virtual experience	Numerical Rating Scale (NPS) anxiety Numerical Pain Scale (NRS) NRS Nausea Semi-structured interview on the immersive experience, acceptability, feelings
Glennon et al. (2018)	Anxiety Pain Physiology	Likert-type scale: anxiety NPS Blood pressure, pulse rate, respiration, temperature, oxygen saturation percentage in oxygen
Bani Mohammad and Ahmad, (2018)	Anxiety Pain Cognitive function	STAI VAS Pain Mini-Mental State Examination (MMSE)
Chirico et al. (2019)	Anxiety Mood state Cyber Sickness	STAI Short Version of Profile of Mood States (SV-POM) Mood stateCyber Sickness Questionnaire (VRSQ)
Gupta and Hande, (2019)	Anxiety - Depression	HADS
Higgins et al. (2019)	Anxiety Pain Virtual experience	Beck Anxiety Inventory (BAI) 10-point scale 10-point scale
Niki et al. (2019)	Palliative symptoms Cyber Sickness Virtual experience	Edmonton Symptom Assessment System (ESAS) Japanese version NRS in 11 points: Dizziness and headaches NRS in 11 points: Pleasure of the experience
Pizzoli et al. (2019)	Mood state Sense of presence	Self-Assessment Manikin (SAM), VAS relaxation VAS sense of presence
Sharifpour et al. (2020)	Pain	Pain Anxiety Symptoms Scale (PASS), Pain Catastrophizing Scale (PCS), Pain Self-Efficacy Questionnaire (PSEQ), McGill Pain Questionnaire (MPQ)
Garrett et al. (2020)	Chronic pain Virtual experience	Focus group and semi-structured interview Focus group and semi-structured interview: effectiveness of VR, mode of action, usability, technical aspects
Gerçeker et al. (2020)	Anxiety Pain Fear	The Children's Anxiety Meter-State (CAM-S) Wong-Baker Faces (WBS) Pain Rating Scale The Child Fear Scale (CFS)
Johnson et al. (2020)	Palliative symptoms	Revised Edmonton Symptom Assessment Scale (ESAS-r)
Scates et al. (2020)	Anxiety Pain Virtual experience	Likert-type scale Likert-type scalle Open ended questions about the feeling and the immersive experience
Semerci et al. (2020)	Pain	WBS Pain Rating Scale

(Continued on following page)

TABLE 2 (*Continued*) Nature of measure and instruments in studies evaluating the effects of virtual reality in oncology.

Study	Nature of measure	Instruments
Tennant et al. (2020)	Anxiety	VAS, Child-report Spence Children's Anxiety Scale (SCAS) short form
	Mood state	VAS
	Pain	VAS
	Sense of presence	Child-report Adapted version of the Total Immersion subscale of the Augmented Reality Immersion (ARI) questionnaire
	Cyber sickness	Child Simulation Sickness Questionnaire (CSSQ)
	Physiology	Puls
	Quality of life	Parent-proxy report Pediatric Quality of Life Inventory™ *Cancer* Module (PedsQL)
Buche et al. (2021)	Anxiety	STAI
	Mood state	SAM
	Temporality	VAS
	Sense of presence	Independent Television Commision – Sens of Presence Inventory (ITC-SOPI)
	Cyber sickness	Questionnaire on Cyber sickness (CQ)
	Virtual experience	Multiple choice questions

3.1.5.3 Duration of immersion

The duration varies mainly according to the duration of the medical act. For short painful procedures such as catheter insertion or venipuncture, immersion varies from 3 to 18 min (Glennon et al., 2018; Buche et al., 2021; Scates et al., 2020; Atzori et al., 2018; Birnie et al., 2018; Semerci et al., 2020). When the context allows for a longer immersion, as is the case in chemotherapy, during long-term hospitalization or on return home, VR is proposed between 10 and 63 min (Schneider et al., 2011; Higgins et al., 2019; Tennant et al., 2020; Buche et al., 2021), although in 60% of cases (i.e., 9/15 studies), the immersion time mainly applied by the experimenters corresponds to 30 min (Niki et al., 2019; Johnson et al., 2020; Espinoza et al., 2012; Baños et al., 2013; Li et al., 2016; Gupta and Hande, 2019; Garrett et al., 2020; Li et al., 2011; Sharifpour et al., 2020). Most virtual immersions lasting 30 min reported positive effects (88.88%, 8/9 studies) (Niki et al., 2019; Espinoza et al., 2012; Baños et al., 2013; Li et al., 2016; Gupta and Hande, 2019; Garrett et al., 2020; Li et al., 2011; Sharifpour et al., 2020). According to the diversity of the medical act in which VR is proposed, there is no strong consensus on the most favorable duration of immersion.

In terms of measurement tools (See Table 2: Nature of measures and instruments in studies evaluating the effects of VR in oncology), most studies collected quantitative data (21/22, 95.45%). Only one study used a qualitative inductive approach using the interpretive description method to explore participants' experiences (Garrett et al., 2020). Seven studies collected qualitative data (31.82%) (Baños et al., 2013; Li et al., 2016; Higgins et al., 2019; Scates et al., 2020; Buche et al., 2021; Birnie et al., 2018; Sharifpour et al., 2020). Only two studies (9.09%) collected physiological data such as blood pressure, pulse rate, respiration, temperature, and percent oxygen saturation using an oximeter (Glennon et al., 2018; Tennant et al., 2020).

Regarding measures reflecting emotional state, anxiety was mainly measured using the State Anxiety Inventory (STAI) (Chirico et al., 2019; Bani Mohammad and Ahmad, 2018; Schneider et al., 2011; Buche et al., 2021) and depression using the Hospital Anxiety Depression Scales (HADS). Mood states were most often assessed using the Visual Analogical Scales (VAS) (Pizzoli et al., 2019; Espinoza et al., 2012; Baños et al., 2013; Tennant et al., 2020) and the Self-Assessment Manikin (SAM) (Pizzoli et al., 2019; Buche et al., 2021). Concerning pain, most researchers have opted for scales (see Table 2: Nature and measurement tools in studies evaluating the effects of VR in oncology) (Bani Mohammad and Ahmad, 2018; Espinoza et al., 2012; Baños et al., 2013; Glennon et al., 2018; Higgins et al., 2019; Scates et al., 2020; Atzori et al., 2018; Birnie et al., 2018; Semerci et al., 2020; Tennant et al., 2020), while others have used specific questionnaires to measure several components of pain such as pain anxiety, catastrophizing, self-efficacy and intensity (Sharifpour et al., 2020). In addition, the Edmonton Symptom Assessment System (ESAS) questionnaire has been used to assess the various symptoms of palliative cancer (Niki et al., 2019; Johnson et al., 2020). The question of temporality was asked orally (Schneider et al., 2011) or by using a VAS from 0 to 40 min with a 5-min interval (Buche et al., 2021).

Semi-structured interviews (Li et al., 2016; Garrett et al., 2020; Birnie et al., 2018) accompanied by various scales (Higgins et al., 2019) (Baños et al., 2013) and supplemented by open-ended (Baños et al., 2013; Scates et al., 2020) or multiple-choice questions (Buche et al., 2021) were conducted to examine the virtual experience with patients. Discomfort that could be caused by the virtual device was monitored through different questionnaires (Chirico et al., 2019; Li et al., 2016; Tennant et al., 2020; Buche et al., 2021), scales (Niki et al., 2019; Birnie et al., 2018; Atzori et al., 2018) and open-ended questions (Baños et al., 2013). Only three studies (13.64%) assessed the subjective feeling of presence in the virtual world using questionnaires (Tennant et al., 2020; Buche et al., 2021) or a VAS (Pizzoli et al., 2019).

In addition, uncommon measures in VR in oncology were collected: one study assessed cognitive function to screen for cognitive impairment in hospitalized adults and determine patients' ability to manipulate the virtual device (Bani

Mohammad and Ahmad, 2018). Another assessed quality of life in young patients (Tennant et al., 2020).

3.2 Research recommendations

Based on the twenty-two studies selected, this third part aims to optimize the methodological choices made in the studies by encouraging the use of practices that are comparable from one study to another for a more rigorous comparison of the reported effects. From a strictly methodological point of view, it seems promising to continue the reflection already initiated at several levels: the degree of interactivity of the devices to be proposed to the patients; the contents to be preferred; the duration of the distractive session; the context of use.

Given the literature, it seems that having access to dynamic feedback from our actions in the virtual environment is a primary criterion for giving patients the feeling of being immersed inside this environment (Chirico et al., 2019; Buche et al., 2021). Participatory immersion can provide better experiential quality than contemplative immersion by actively engaging patients (Garrett et al., 2020). Future studies should evaluate the links between immersive quality and distraction power benefits under VR to leverage this finding. It is worth noting that the auditory component contributes to the immersion of patients in the virtual world (Michel et al., 2019b) as this is notably the case of natural environments enhanced with background sound, relaxing music, or guided relaxation. These results are more convincing when the technology allows a qualitative VR experience. Devices with high technological quality promote the feeling of presence (Cummings and Bailenson, 2016) and the quality of the distraction. In summary, the better the technical quality, the more intense the transport into the virtual environment.

If there is a consensus on the need to present patients with natural and high-definition sound content, developing new and constantly renewed content is essential to overcome the phenomenon of habituation. A regularly updated system could preserve the awe of this innovative device and continue to captivate patients even after repeated immersions. The exploitation of future software should further engage the patient in the immersive task mobilizing his cognitive resources at different levels ranging from distraction to concentration or skill reinforcement (Ahmadpour et al., 2020).

Although VR is a promising technology, there are still some limitations to applying this distractive tool in oncology. To date, it is difficult to recommend an immersive duration most conducive to patients' emotional comfort. It would be interesting to evaluate the differential effects of time immersed in the virtual environment (Tennant et al., 2020). Immersion time seems to be determined by the nature and duration of medical procedures and not by the relaxation/distraction needs of the patient. Thus, devices that adapt the duration of immersion to individual patient needs and preferences would be a considerable asset to enhance the benefits of distraction.

The context of VR use essentially conditions the duration of immersion. However, specific methodological recommendations can be retained. Given the observed results, an extended hospitalization allows a progressive increase in immersion time, allowing the patients to become a little more familiar with the virtual device each day (Gupta and Hande, 2019). To ensure the benefits of VR, it would be preferable that the virtual experience not exceed 30 min per day during a long-term hospitalization (Niki et al., 2019; Johnson et al., 2020; Espinoza et al., 2012; Baños et al., 2013; Gupta and Hande, 2019; Li et al., 2011). During an outpatient hospitalization involving short, painful procedures such as port access or venipuncture, it would seem appropriate that the immersion starts 2–5 min before the medical act (familiarization phase) and continues until the end of the procedure (experimental phase) (Buche et al., 2021; Atzori et al., 2018). In chemotherapy, following Chirico et al. (2019), a familiarization phase of 5–10 min could be introduced to optimize the effects of the virtual experience. As for the duration of the immersive experience during the administration of chemotherapy, there is currently no consensus in the literature to propose a recommendation (Chirico et al., 2019; Schneider et al., 2011; Sharifpour et al., 2020).

When examining the benefits using VR in oncology, it is regrettable to note the absence of a control condition in nearly half of the studies (Pizzoli et al., 2019; Niki et al., 2019; Johnson et al., 2020; Espinoza et al., 2012; Baños et al., 2013; Li et al., 2016; Gupta and Hande, 2019; Higgins et al., 2019; Birnie et al., 2018). In the future, researchers should design randomized controlled studies that compare medical care using VR for distraction with the same care without using distraction (i.e., a control condition) as well as this same care using other distractive strategies (i.e., different conditions) to reveal in a more rigorous comparison setting the true effectiveness of virtual immersion in oncology.

In addition, some measurement tools are not systematically used. Assessing the risk of side effects from virtual devices is helpful to ensure that VR is well tolerated by patients. This also allows us to distinguish between the physical discomfort of treatment and those that VR may cause. Future studies evaluating the effects of VR through physiological variables such as heart rate, oxygenation rate, or skin conductance could refine the assessment of patients' emotional states (Chirico et al., 2019). The measurement of the subjective sense of presence in the virtual environment should be systematized in oncology, knowing that this feeling is closely linked to the sensation of escape (Tennant et al., 2020; Witmer and Singer, 1998).

Furthermore, the effectiveness of VR depends on personal acceptance (Garrett et al., 2020) and patient interest in the device (Lessiter et al., 2001). It may be that the positive results reported in the literature are partly a result of the acceptance rate at recruitment and the predisposition of patients to the virtual experience. Patients who prefer to maintain control and observe the routine of care may be more likely to decline the experience, while patients who are more open to the device may already be in a favorable emotional state to use VR. Future research should not neglect to assess patient

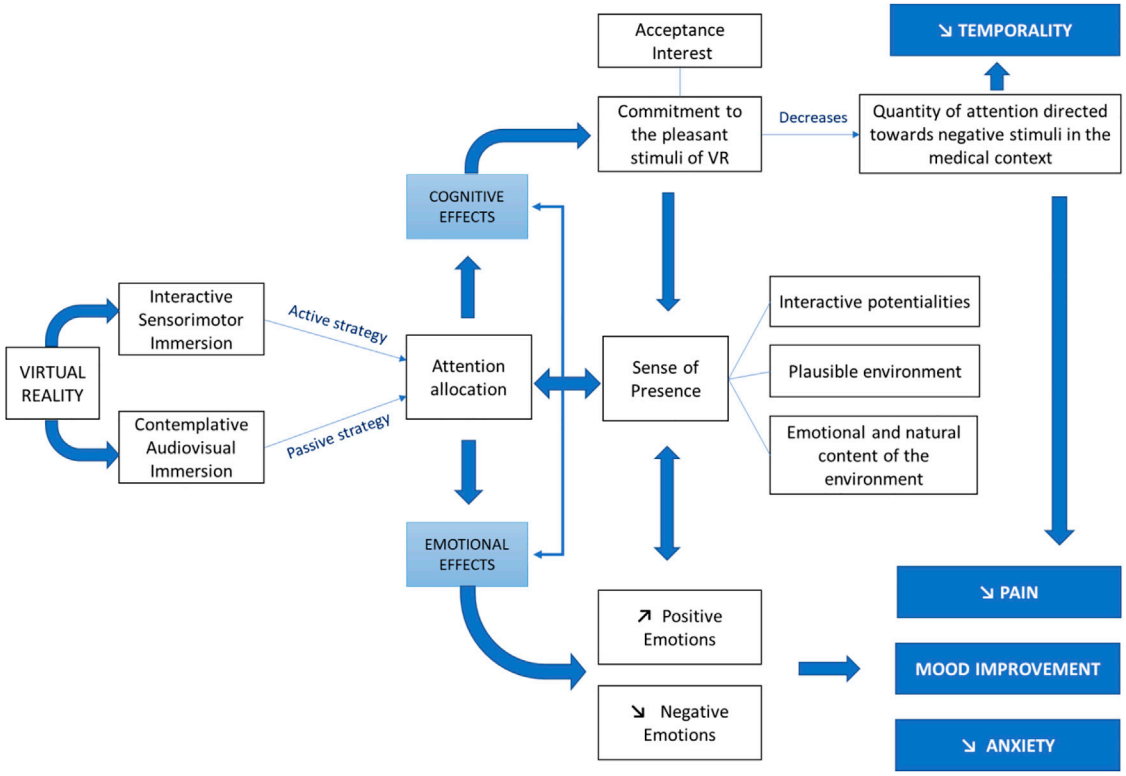

FIGURE 4
Model of the mechanisms involved in VR distractive experience and underlying its benefits.

enjoyment and motivation to engage in the immersive experience to consider their level of involvement in the immersion. Like (Bani Mohammad and Ahmad, 2018), future studies would benefit from considering the patient's ability to process sensorimotor information from VR through the measurement of cognitive impairments to ensure an optimal relaxing experience.

Finally, this device is considered an effective distractive strategy when it fulfills certain conditions according to the medical context, but even more so according to the patient's needs at the time. As we have seen in the study by Buche et al. (2021), VR can be used not only to distract from anxiety-provoking or painful experiences during treatment but also to compensate for the monotonous nature of the treatments. It can also be used when the practitioners are unavailable, for instance, if they have other things to manage than the patient relationship.

3.3 Theoretical model

The richness of the available scientific literature and the exploitation of the state of the art allow us to think of an integrative theoretical model that considers the effects of VR on both the cognitive and emotional levels. Articulating the cognitive and emotional sides will enable us to envisage a valid and robust schematic representation aligned with the benefits reported in oncology and the theories mentioned (see Table 1: Studies of the Benefits of VR). Based on this careful exploitation of the current state of knowledge and the methodological and theoretical choices made by the community, we propose an explanatory model of the effects of exposure to VR (see Figure 4: Model of the mechanisms involved in VR distractive experience and underlying its benefits) to contribute to the understanding of the processes leading to the emergence of the positive effects of virtual immersion with cancer patients during medical interventions. This model is based on an ideal situation where the use of VR as a distractive tool has been preceded by a familiarization phase (i.e., when the handling of the device is no longer likely to hinder the relaxing experience).

The theoretical basis of our model is mainly based on the allocation of attentional resources related to the limited cognitive capacities of human beings (Arane et al., 2017). As stated above, we consider that VR can offer several levels of immersion involving different senses simultaneously (Chirico et al., 2019). The immersive technologies employed can mobilize active or passive cognitive strategies that aim to reduce attention to the physical environment. The first effect of multimodal immersion is to spontaneously draw attention to pleasurable VR stimuli by passively or actively engaging the patient in the virtual experience.

Engagement or involvement is a state of strong concentration in which the patient no longer directs their conscious attention towards external negative stimuli and forgets the medical context in which they are situated. This results in a feeling of presence, that is, the impression that the patient is escaping into a world other than their real-world (Witmer and Singer, 1998). According to the presence model (Lessiter et al., 2001), the immersive task depends mainly on the individual's interest in the experience. However, these authors underline that the device's immersive qualities and participative potentialities are likely to awaken or hinder the interest in VR.

It should be noted that the immersive qualities determine the credibility of the experience by recreating the perceptive attributes (e.g., tracking level, stereoscopy, and field of view) that a person can find in physical reality (Cummings and Bailenson, 2016). Thus, the level of engagement, interactivity, and plausible environment influence the prevalence of presence which focuses attention on immersion. Since fantasy environments can also benefit patients (Pourmand et al., 2018), most oncology studies used believable natural environments. In addition to inducing a sense of presence, the cognitive resources mobilized modify the perception of temporality by giving the impression that time is passing more rapidly within the virtual environment. Furthermore, the attentional engagement in the immersive task affects the cognitive evaluation of pain by reducing the amount of attention available to process the painful information, thus attenuating the pain felt (Eccleston and Crombez, 1999; Atzori et al., 2018).

Moreover, the cognitive effects maintain a virtuous circle with the emotional effects generated by this distractive strategy. VR is a medium capable of increasing positive emotions and decreasing negative emotions thanks to immersion in a natural environment (Scates et al., 2020), which carries positive emotions (Baños et al., 2013). Riva et al. (2007) have demonstrated the bidirectional relationship between emotions and presence: A relaxing environment generates a higher sense of presence than a neutral environment, and once the sense of presence is established, positive emotions are felt more intensely (Bouvier, 2009). This emotional induction not only decreases anxiety and improves mood by inducing joy and calmness but also influences pain perception. Attention focused on positive emotions inhibits the nociceptive message conveyed by the nervous system, which leads to a decrease in the intensity of the pain felt (Sharifpour et al., 2020; Melzack and Wall, 1996).

4 Discussion

Based on the accumulated results, which primarily convey a positive image of VR, there is no doubt today that the use of this technology is of major interest. However, the beneficial effects regularly reported must be understood in terms of the characteristics of the technology used and according to the particularities of the patients and their immersion preferences.

The objective of this article is twofold, given the converging and diverging points highlighted in this literature review. The first is identifying avenues for harmonizing the procedures and tools used in future research. This analysis of the current state of practice in measuring the effects of VR in oncology synthesizes the data accumulated over the past decade on the distractive power of VR in oncology. Based on this analysis, the scientific community has the means to move towards a more substantial consensus to encourage more rigorous reflection by clarifying methodological regularities. Secondly, this article invites the scientific community to consider more systematically the need for a theoretical foundation that contributes to consolidating the understanding of the processes at work in the results reported in the scientific literature used in this article. While some authors have attempted to explain the psychological phenomena that underlie the benefits of VR, few of them have articulated their approach to theoretical models of reference. The theoretical model proposed in this article considers the available knowledge and provides a promising framework for future studies that aim to deepen the cognitive and emotional processes at stake during the use of VR. Our framework describes the broader impact of VR benefits concerning cognitive and emotional regulation. The medical context (cancer) from which our theoretical model has emerged could be applied broadly where pain and anxiety reduction are critical (e.g., child dental care (Du et al., 2022), wound care and rehabilitation after burns (Czech et al., 20221028), skin prick testing (Stassart and Giebels, 2022). Also, other sectors beyond healthcare can substantially contribute to testing the validity of our theoretical framework. Indeed, there is no doubt that the potentialities offered by our framework would benefit from being considered outside the medical context to ensure the robustness and generalizability of its articulation between emotion and cognition.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding authors.

Author contributions

All authors listed have made a substantial direct and intellectual contribution to the workand approved it for publication.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Ahmad, M., Bani Mohammad, E., and Anshasi, H. A. (2020). Virtual reality technology for pain and anxiety management among patients with cancer: A systematic review. *Pain Manag. Nurs.* doi:10.1016/j.pmn.04.002

Ahmadpour, N., Keep, M., Janssen, A., Rouf, A. S., and Marthick, M. (2020). Design strategies for virtual reality interventions for managing pain and anxiety in children and adolescents: Scoping review. *JMIR Serious Games* 8 (RM), e14565. doi:10.2196/14565

Alawneh, A., Anshasi, H., Khirfan, G., Yaseen, H., and Quran, A. (2017). Symptom prevalence of patients with cancer in a tertiary cancer center in Jordan. *Gulf J. Oncol.* 1, 37–43.

Arane, K., Behboudi, A., and Goldman, Ran. D. (2017). Virtual reality for pain and anxiety management in children. *Can. Fam. Physician* 63, 932–934.

Arrieta, O., Angulo, L., Núnez-Valencia, C., Dorantes-Gallareta, Y., Macedo, E., Martínez-Lopez, D., et al. (2013). As-sociation of depression and anxiety on quality of life, treatment adherence, and prognosis in patients with advanced non-small cell lung cancer. *Ann. Surg. Oncol.* 20, 1941–1948. doi:10.1245/s10434-012-2793-5

Atzori, B., Hoffman, H. G., Vagnoli, L., Patterson, D. R., Alhalabi, W., Messeri, A., et al. (2018). Virtual reality analgesia during venipuncture in pediatric patients with onco-hematological diseases. *Front. Psychol.* 9, 2508. doi:10.3389/fpsyg.2018.02508

Bani Mohammad, E., and Ahmad, M. (2018). Virtual reality as a distraction technique for pain and anxiety among patients with breast cancer: A randomized control trial. *Palliat. Support. Care* 17, 29–34. doi:10.1017/s1478951518000639

Baños, R. M., Espinoza, M., García-Palacios, A., Cervera, J. M., Esquerdo, G., Barrajón, E., et al. (2013). A positive psychological intervention using virtual reality for patients with advanced cancer in a hospital setting: A pilot study to assess feasibility. *Support. Care Cancer* 21, 263–270. doi:10.1007/s00520-012-1520-x

Birnie, K. A., Kulandaivelu, Y., Jibb, L., Hroch, P., Positano, K., Robertson, S., et al. (2018). Usability testing of an interactive virtual reality distraction intervention to reduce procedural pain in children and adolescents with cancer. *J. Pediatr. Oncol. Nurs.* 35, 406–416. doi:10.1177/1043454218782138

Bouvier, P. (2009). *La présence en réalité virtuelle, une approche centrée utilisateur. Thèse de 699 doctorat en informatique*. Paris: Université de Paris-Est.

Buche, H., Michel, A., Piccoli, C., and Blanc, N. (2021). Contemplating or acting? Which immersive modes should Be favored in virtual reality during physiotherapy for breast cancer rehabilitation. *Front. Psychol.* 12, 631186. doi:10.3389/fpsyg.2021.631186

Burle, B., and Casini, L. (2001). Dissociation between activation and attention effects in time estimation: Implications for internal clock models. *J. Exp. Psychol. Hum. Percept. Perform.* 27, 195–205. doi:10.1037/0096-1523.27.1.195

Chirico, A., Lucidi, F., De Laurentiis, M., Milanese, C., Napoli, A., Giordano, A., et al. (2016). Virtual reality in health system: Beyond entertainment. A mini-review on the efficacy of VR during cancer treatment. *J. Cell. Physiol.* 231, 275–287. doi:10.1002/jcp.25117

Chirico, A., Lucidi, F., Mallia, L., D'Aiuto, M., and Merluzzi, T. V. (2015). Indicators of distress in newly diagnosed breast cancer patients. *PeerJ* 3, 1107. doi:10.7717/peerj.1107

Chirico, A., Maiorano, P., Indovina, P., Milanese, C., Giordano, G. G., Alivernini, F., et al. (2019). Virtual reality and music therapy as distraction interventions to alleviate anxiety and improve mood states in breast cancer patients during chemotherapy. *J. Cell. Physiol.* 235, 5353–5362. doi:10.1002/jcp.29422

Cummings, J. J., and Bailenson, J. N. (2016). How immersive is enough? A meta-analysis of the effect of immersive technology on user presence. *Media Psychol.* 19 (2), 272–309. doi:10.1080/15213269.2015.1015740

Czech, O., Wrzeciono, A., Baťalík, B., Szczepańska-Gieracha, S. G., Malicka, I., Rutkowski, S., et al. (202210283). Virtual reality intervention as a support method during wound care and rehabilitation after burns: A systematic review and meta-analysis. *Complementary Ther. Med.* 68. doi:10.1016/j.ctim.2022.102837

de Loor, P., and Tisseau, J. (2011). Réalité Virtuelle et énaction. *J. de l'Association Française de Réalité Virtuelle* 3. ffhal-00603993f.

Defossez, G., Le Guyader-Peyrou, S., Uhry, Z., Grosclaude, P., Colonna, M., Dantony, E., et al. (2019). *Estimations nationales de l'incidence et de la mortalité par cancer en France métropolitaine 35 entre 1990 et 2018. Volume 1 – tumeurs solides. Saint-Maurice (Fra)*. Saint-Maurice: Santé publique France, 372.

Droit-Volet, S., and Gil, S. (2009). The time–emotion paradox. *Phil. Trans. R. Soc. B* 364, 1943–1953. doi:10.1098/rstb.2009.0013

Du, Q., Ma, X., Wang, S., Zhou, S., Luo, C., Tian, K., et al. (2022). A digital intervention using virtual reality helmets to reduce dental anxiety of children under local anesthesia and primary teeth extraction: A randomized clinical trial. *Brain Behav.* 12, e2600. doi:10.1002/brb3.2600

Eccleston, C., and Crombez, G. (1999). Pain demands attention: A cognitive– affective model of the interruptive function of pain. *Psychol. Bull.* 125, 356–366. doi:10.1037/0033-2909.125.3.356

Espinoza, M., Baños, R. M., Garcia-Palacios, A., Cervera, J. M., Esquerdo, G., Barrajon, E., et al. (2012). Promotion of emotional wellbeing in oncology inpatients using VR. *Stud. Health Technol. Inf.* 181, 53–57.

Fredrickson, B. (2001). The role of positive emotions in positive psychology. The broaden-and-build theory of positive emotions. *Am. Psychol.* 56, 218–226. doi:10.1037/0003-066x.56.3.218

Garrett, B. M., Tao, G., Taverner, T., Cordingley, E., and Sun, C. (2020). Patients perceptions of virtual reality therapy in the management of chronic cancer pain. *Heliyon* 6, e03916. doi:10.1016/j.heliyon.2020.e03916

Gerçeker, G. Ö., Bektaş, M., Aydınok, Y., Ören, H., Ellidokuz, H., Olgun, N., et al. (2020). The effect of virtual reality on pain, fear, and anxiety during access of a port with huber needle in pediatric hematology-oncology patients: Randomized controlled trial. *Eur. J. Oncol. Nurs.* 50, 101886. doi:10.1016/j.ejon.2020.101886

Glennon, C., McElroy, S., Connelly, L., MischeLawson, L., Bretches, A., Gard, A., et al. (2018). Use of virtual reality to distract from pain and anxiety. *Oncol. Nurs. Forum* 45, 545–552. doi:10.1188/18.onf.545-552

Gold, J. I., Belmont, K. A., and Thomas, D. A. (2007). The neurobiology of virtual reality pain attenuation. *CyberPsychol. Behav.* 10, 536–544. doi:10.1089/cpb.2007.9993

Gupta, N., and Hande, D. (2019). Is virtual reality program is effective in reducing anxiety in post mastectomy among breast cancer patient? *Int. J. Multidiscip. Res. Dev.* 6, 32–35.

Higgins, S., Feinstein, S., Hawkins, M., Cockburn, M., and Wysong, A. (2019). Virtual reality to improve the experience of the mohs patient-A prospective interventional study. *Dermatol. Surg.* doi:10.1097/DSS.0000000000000001854

Indovina, P., Barone, D., Gallo, L., Chirico, A., De Pietro, G., Giordano, A., et al. (2018). Virtual reality as a distraction intervention to relieve pain and distress during medical procedures. *Clin. J. Pain* 34, 858–877. doi:10.1097/ajp.0000000000000599

Institut National du Cancer (2019). *Les cancers en France : l'essentiel des faits et des chiffres*. https://www.oncorif.fr/wpcontent/uploads/2019/02/Cancers_en_FranceEssentiel_Faits_et_chiffres-2018.pdf.

Johnson, T., Bauler, L., Vos, D., Hifko, A., Garg, P., Ahmed, M., et al. (2020). Virtual reality use for symptom management in palliative care: A pilot study to assess user perceptions. *J. Palliat. Med.* 23, 1233–1238. doi:10.1089/jpm.2019.0411

Kaplan, R., and Kaplan, S. (1989). *The experience of nature: A psychological perspective*. Cambridge University Press.

Kleiber, C., and Harper, D. C. (1999). Effects of distraction on children's pain and distress during medical procedures: A metaanalysis. *Nurs. Res.* 48, 44–49. doi:10.1097/00006199-199901000-00007

Kleiber, C., and McCarthy, A. M. (2006). Evaluating instruments for a study on children's responses to a painful procedure when parents are distraction coaches. *J. Pediatr. Nurs.* 21, 99–107. doi:10.1016/j.pedn.2005.06.008

Koller, D., and Goldman, R. D. (2012). Distraction techniques for children undergoing procedures: A critical review of pediatric research. *J. Pediatr. Nurs.* 27, 652–681. doi:10.1016/j.pedn.2011.08.001

Lazarus, R. S., and Folkman, S. (1984). "Stress, appraisal, and coping," in *Behaviour research and therapy*. Editor S. P. Company. doi:10.1016/0005.7967(85)90087.7

Lessiter, J., Freeman, J., Keogh, E., and Davidoff, J. (2001). A cross-media presence questionnaire: The ITC-Sense of Presence Inventory. *Presence. (Camb).* 10, 282–297. doi:10.1162/105474601300343612

Li, W. H., Chung, J. O., and Ho, E. K. (2011). The effectiveness of therapeutic play, using virtual reality computer games, in promoting the psychological wellbeing of children hospitalised with cancer. *J. Clin. Nurs.* 20, 2135–2143. doi:10.1111/j.1365-2702.2011.03733.x

Li, Z., Han, X. G., Sheng, J., and Ma, S. J. (2016). Virtual reality for improving balance in patients after stroke: A systematic review and metaanalysis. *Clin. Rehabil.* 30, 432–440. doi:10.1177/0269215515593611

Melzack, R., and Wall, P. D. (1996). Pain mechanisms: A new theory. *Pain Forum* 5, 3–11. doi:10.1016/s1082-3174(96)80062-6

Michel, A., Brigaud, E., Cousson-Gélie, F., Vidal, J., and Blanc, N. (2019). La réalité virtuelle chez les femmes âgées suivies pour un cancer du sein : Intérêts et attentes. *Geriatr. Psychol. Neuropsychiatr. Vieil.* 17, 415–422. doi:10.1684/pnv.2019.0832

Michel, A., Vidal, J., Brigaud, E., Sokratous, K., and Blanc, N. (2019). Dessine-moi une réalité plus belle : La réalité virtuelle vue par les patientes atteintes d'un cancer du sein. *Psycho-Oncol.* 13, 69–78. doi:10.3166/pson-2019-0087

Nikbakhsh, N., Moudi, S., Abbasian, S., and Khafri, S. (2014). Prevalence of depression and anxiety among cancer patients. *Casp. J. Intern. Med.* 5, 167–170. doi:10.1016/j.jpainsymman.2015.12.340

Niki, K., Okamoto, Y., Maeda, I., Mori, I., Ishii, R., Matsuda, Y., et al. (2019). A novel palliative care approach using virtual reality for improving various symptoms of terminal cancer patients: A preliminary prospective, multicenter study. *J. Palliat. Med.* 22, 702–707. doi:10.1089/jpm.2018.0527

O'Connor, M., White, K., Kristjanson, L., Cousins, K., and Wilkes, L. (2010). The prevalence of anxiety and depression in palliative care patients with cancer in Western Australia and New South Wales. *Med. J. Aust.* 193, S44–S47. doi:10.5694/j.1326-5377.2010.tb03927.x

Oyama, H., Ohsuga, M., Tatsuno, Y., and Katsumata, N. (1999). Evaluation of the psycho-oncological effectiveness of the bedside wellness system. *CyberPsychol. Behav.* 2, 81–84. doi:10.1089/cpb.1999.2.81

Pittara, M., Matsangidou, M., Stylianides, K., Petkov, N., and Pattichis, C. S. (2020). Virtual reality for pain management in cancer : A comprehensive review. *IEEE Access* 8, 225475–225489. doi:10.1109/access.2020.3044233

Pizzoli, S. F. M., Mazzocco, K., Triberti, S., Monzani, D., Alcañiz Raya, M. L., Pravettoni, G., et al. (2019). User-centered virtual reality for promoting relaxation: An innovative approach. *Front. Psychol.* 10, 479. doi:10.3389/fpsyg.2019.00479

Pourmand, A., Davis, S., Marchak, A., Whiteside, T., and Sikka, N. (2018). Virtual reality as a clinical tool for pain management. *Curr. Pain Headache Rep.* 22 (8), 53. doi:10.1007/s11916-018-0708-2

Pratt, D. R., Zyda, M., and Kelleher, K. (1995). Virtual reality: In the mind of the beholder. *IEEE Comput.* 28, 17–19.

Ramirez, S., Liu, X., MacDonald, C. J., Moffa, A., Zhou, J., Redondo, R. L., et al. (2015). Activating positive memory engrams suppresses depression-like behaviour. *Nature* 522, 335–339. doi:10.1038/nature14514

Riva, G., Mantovani, F., Capideville, C. S., Preziosa, A., Morganti, F., Villani, D., et al. (2007). Affective interactions using virtual reality: The link between presence and emotionsectiveinteractions using virtual reality : The link between presence and emotions. *CyberPsychol. Behav.* 10, 45–56. doi:10.1089/cpb.2006.9993

Scates, D., Dickinson, J. I., Sullivan, K., Cline, H., and Balaraman, R. (2020). Using nature-inspired virtual reality as a distraction to reduce stress and pain among cancer patients. *Environ. Behav.* 58, 895–918. doi:10.1177/0013916520916259

Schneider, S. M., Ellis, M., Coombs, W. T., Shonkwiler, E. L., and Folsom, L. C. (2003). Virtual reality intervention for older women with breast cancer. *Cyberpsychol. Behav.* 6, 301–307. doi:10.1089/109493103322011605

Schneider, S. M., and Hood, L. E. (2007). Virtual reality: A distraction intervention for chemotherapy. *Oncol. Nurs. Forum* 34 (1), 39–46. doi:10.1188/07.ONF.39-46

Schneider, S. M., Kisby, C. K., and Flint, E. P, (2011). Effect of virtual reality on time perception in patients receiving chemotherapy. *Support. Care Cancer* 19, 555–564. doi:10.1007/s00520-010-0852-7

Schneider, S. M., Prince-Paul, M., JoAllen, M., Silverman, P., and Talaba, D. (2004). Virtual reality as a distraction intervention for women receiving chemotherapy. *Oncol. Nurs. Forum* 31, 81–88. doi:10.1188/04.onf.81-88

Schneider, S. M., and Workman, M. L. (1999). Effects of virtual reality on symptom distress in children receiving chemotherapy. *CyberPsychology Behav.* 2, 125–134. doi:10.1089/cpb.1999.2.125

Semerci, R., Akgün Kostak, M., Eren, T., and Avci, G. (2020). Effects of virtual reality on pain during venous port access in pediatric oncology patients: A randomized controlled study. *J. Pediatr. Oncol. Nurs.* 38, 142–151. doi:10.1177/1043454220975702

Sharifpour, S., Manshaee, G., and Sajjadian, I. (2020). Effects of virtual reality therapy on perceived pain intensity, anxiety, catastrophising and self-efficacy among adolescents with cancer. *Couns. Psychother. Res.* 21, 218–226. doi:10.1002/capr.12311

Stassart, C., and Giebels, K. (2022). Effectiveness of virtual reality for pediatric pain and anxiety management during skin prick testing. *Open J. Med. Psychol.* 11 (03), 89–102. doi:10.4236/ojmp.2022.113007

Tennant, M., Youssef, G. J., McGillivray, J. A., Clark, T.-J., McMillan, L., McCarthy, M. C., et al. (2020). Exploring the use of immersive virtual reality to enhance psychological wellbeing in pediatric oncology: A pilot randomized controlled trial. *Eur. J. Oncol. Nurs.* 48, 101804. doi:10.1016/j.ejon.2020.101804

Tsaï, F. (2016). La réalité virtuelle, un outil pour renouer avec la sensorialité. *Hermes (Wiesb).* 74, 188. doi:10.3917/herm.074.0188

Ulrich, R. S., Simons, R. F., Losito, B. D., Fiorito, E., Miles, M. A., Zelson, M., et al. (1991). Stress recovery during exposure to natural and urban environments. *J. Environ. Psychol.* 11, 201–230. doi:10.1016/s0272-4944(05)80184-7

Van den Beuken-van Everdingen, M. H. J., Hochstenbach, L. M. J., Joosten, E. A. J., Tjan-Heijnen, V. C. G., and Janssen, D. J. A. (2016). *J. Pain Symptom Manage.* 51 (6), 1070–1090.e9. e9. doi:10.1016/j.jpainsymman.2015.12.340

Witmer, B. G., and Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence. (Camb).* 7, 225–240. doi:10.1162/105474698565686

Wittmann, M., and Paulus, M. P. (2008). Decision making, impulsivity and time perception. *Trends Cogn. Sci.* 12, 7–12. doi:10.1016/j.tics.2007.10.004

Zeng, Y., Zhang, J., Cheng, A., Cheng, H., and Wefel, J. (2019). Meta-analysis of the efficacy of virtual reality–based interventions in cancer-related symptom management. *Integr. Cancer Ther.* 18, 153473541987110. doi:10.1177/1534735419871108

frontiers | Frontiers in Virtual Reality

TYPE Review
PUBLISHED 09 February 2023
DOI 10.3389/frvir.2023.1065863

OPEN ACCESS

EDITED BY
Jean Botev,
University of Luxembourg, Luxembourg

REVIEWED BY
Błażej Cieślik,
Jan Długosz University, Poland
Orly Lahav,
Tel Aviv University, Israel
Andrea Stevenson Won,
Cornell University, United States

*CORRESPONDENCE
Despina Michael-Grigoriou,
despina.grigoriou@cut.ac.cy

This article was submitted to Virtual Reality in Medicine, a section of the journal Frontiers in Virtual Reality

RECEIVED 10 October 2022
ACCEPTED 27 January 2023
PUBLISHED 09 February 2023

CITATION
Hadjipanayi C, Banakou D and Michael-Grigoriou D (2023), Art as therapy in virtual reality: A scoping review. *Front. Virtual Real.* 4:1065863. doi: 10.3389/frvir.2023.1065863

Art as therapy in virtual reality: A scoping review

Christos Hadjipanayi [1], Domna Banakou [1,2] and Despina Michael-Grigoriou [1]*

[1]GET Lab, Department of Multimedia and Graphic Arts, Cyprus University of Technology, Limassol, Cyprus, [2]Arts and Humanities Division, New York University Abu Dhabi, Abu Dhabi, United Arab Emirates

This scoping review focuses on therapeutic interventions, which involve the creation of artworks in virtual reality. The purpose of this research is to survey possible directions that traditional practices of art therapy and therapeutic artmaking could take in the age of new media, with emphasis on fully immersive virtual reality. After the collection of papers from online databases, data from the included papers were extracted and analyzed using thematic analysis. The results reveal that virtual reality introduces novel opportunities for artistic expression, self-improvement, and motivation for psychotherapy and neurorehabilitation. Evidence that artmaking in virtual reality could be highly beneficial in therapeutic settings can be found in many aspects of virtual reality, such as its virtuality, ludicity, telepresence capacity, controlled environments, utility of user data, and popularity with digital natives. However, deficiencies in digital literacy, technical limitations of the current virtual reality devices, the lack of tactility in virtual environments, difficulties in the maintenance of the technology, interdisciplinary concerns, as well as aspects of inclusivity should be taken into consideration by therapy practitioners, researchers, and software developers alike. Finally, the reported results reveal implications for future practice.

KEYWORDS

virtual reality, art, psychotherapy, therapy, rehabilitation, wellbeing, new media

1 Introduction

Given the rapid technological advancements, the steady decrease in prices of technological apparatus, and the continuous permeation of information technology in various disciplines, Adams et al. (2008) predict that, during the current millennium, digital art will be transcending its aesthetic role by adapting in multiple applications as a transdisciplinary medium. The present review touches upon the ubiquity of digital art and focuses on the fields of mental and physical healthcare. In this regard, of special interest is Virtual Reality (VR), a multipurpose communication medium (Biocca and Levy, 2013) that intersects with the domains of both artistic expression (Kim and Lee, 2021) and wellbeing (Alqahtani et al., 2017; Montana et al., 2020). During the introductory part of this review, key concepts, which later contribute to a holistic understanding of the topic, are introduced. The present review approaches the integration of VR in practices relevant to art and therapy by presenting the state of the art, gaps in current knowledge, and potential future directions.

A VR system is characterized by its level of immersion, also referred to as sensory immersion, that has been described in the literature as the degree to which natural sensorimotor contingencies can be supported and engaged by the virtual simulation (Kim and Biocca, 2018; Slater, 2018; Berkman and Akan, 2019). Immersion depends on the characteristics of the apparatus and optimizing immersion levels is supported by various VR studies, which conclude that high immersion is an antecedent of the sensation of "being

present" in a Virtual Environment (VE) (Slater et al., 2009; Slater and Sanchez-Vives, 2016).Eliciting *presence* is a crucial element of a VR experience as it indicates how natural sensorimotor contingencies ostensibly are and therefore how likely it is for the user to act in the VR environment as if in real life (Slater and Sanchez-Vives, 2016), so that the experience becomes "organic, user-driven, and different for everyone" (Bailenson, 2018, p. 223) An example relevant to drawing/writing would be the visual stimuli of a pen matching the motor action of grasping it in the perceivably proper way (Farmer et al., 2018) The naturalness in which the user's body is perceived to be moving in order to form the action of grasping that pen, also contributes to the sensorimotor contingencies. Here, when the user embodies a virtual whole body or a body part, which they experience from a first-person perspective and onto which movements are mapped in real-time and in synchrony with the user's real movements, this gives rise to the illusion of body ownership (Maselli and Slater, 2013; Christofi et al., 2020). In VR storytelling, presence and embodiment together have been previously described as "narrative storyliving" (Vallance and Towndrow, 2022).

Interestingly, the virtual bodies (or avatars) that act as surrogate bodies during VR embodiment are found to have capabilities which go beyond their technical capabilities. In cases when immersive VR applications allow for bodily customization, a psychosocial layer is added to VR embodiment, which may enhance the sense of body ownership by enabling role-play and the free expression of various behaviors and attitudes (Bertrand et al., 2021). This influence on behaviors and attitudes which stems from the dispositional characteristics of the embodied avatar, is known as the "*Proteus* Effect" (Yee and Bailenson, 2007) and it has been shown to even lead to higher cognitive changes, such as, for example, embodying a child body causing adult VR users to overestimate the size of virtual objects (Banakou et al., 2013), embodiment in a different race body leading to changes in implicit racial attitudes (Maister et al., 2015; Banakou et al., 2020), or embodying a stereotypically empathic woman instigating empathy (Hadjipanayi and Michael-Grigoriou, 2022). Embodying avatars in fully immersive VR can also lead to the acquirement of soft-skills, where for instance, embodying the avatar of Sigmund Freud was found to help VR users offer more sound counselling advice compared to embodying the avatar of a therapy client (Osimo et al., 2015; Slater et al., 2019).

Contrary to fully immersive VR, non-immersive VR systems only offer a window on the virtual world, without this essentially being a disadvantage (Alqahtani et al., 2017). A Window on World (WoW) type of VR is commonly projected through regular monitor screens. Non-immersive VR systems are less expensive and easier to use than immersive VR systems (Bamodu and Ye, 2013). Desktop video games that include procedurally generated environments are classic examples of non-immersive VR (Alqahtani et al., 2017). Semi-immersive VR systems are hybrid systems that aim to maintain the simplicity and low cost of non-immersive VR systems while emulating the advantages in sensorimotor contingencies that are successfully achieved by fully immersive VR systems (Bamodu and Ye, 2013). This type of system occupies a portion of the physical environment and virtually transforms it to serve a specific purpose. For example, the therapeutic system for motor and cognitive rehabilitation that is introduced by Maggio et al. (2022) transforms an empty room into a virtual playground that can extend to the floor and the walls of the room to match the preferred design of specific rehabilitation exercises.

VR is a transdisciplinary medium that intertwines with subject areas relevant to both artistic creation and healthcare. However, artistic creation in VR is more commonly associated with aspects of creativity, entertainment, and culture rather than wellbeing (Rubio-Tamayo et al., 2017; Pissini, 2020). This scoping review relies on gathering literature pieces for which all the relevant subject areas of "VR," "art," and "therapy" intersect. This is particularly challenging due to the broad terminology surrounding "art", which is colloquially defined as any form of self-expression. For this reason, it is imperative that art-related aspects are further contextualized, while also keeping in mind that in this scoping review art is viewed through the lens of therapy and healing. in other words, the population included in this review is "VR users," the concept is "the creation of visual artworks in immersive VR," and the context is "therapy." Notably, this review focuses on the visual aspect of artmaking, as it is favored by VR technology, and to-date there is a gap in the VR literature regarding non-visual therapeutic artmaking projects. In order to address the multi-dimensional role of art as a therapeutic practice, it is essential to first address the differences between 1) expressive arts therapy and art therapy and 2) therapeutic artmaking and art therapy.

Four distinct categories (disciplines) of expressive arts therapy are widely recognized among practitioners and are known as 1) art therapy, 2) music therapy, 3) dramatherapy, and 4) dance movement therapy (Song et al., 2019). Disciplines of expressive arts therapy can sometimes be used adjunct to other affective, cognitive, or psychomotor approaches, forming tailored therapeutic interventions that meet the diverse needs of each cohort group (Malchiodi, 2020). As an example, Mishina et al. (2017) introduced a set of playground activities into an expressive arts therapy intervention with troubled adolescents whose emotional states improved after the intervention. These playground activities included rhythmic movements and image creation, among other activities, that formed a multimodal expressive arts therapy approach. In the cognitive domain, art therapy elements are commonly combined with cognitive-behavioral therapy (CBT) to provide trauma-based treatment. In many cases, drawing or symbolically reenacting a traumatic memory under the guidance of a therapist helped clients come to terms with themselves regarding their feelings of anger, helplessness, and self-blame (Pifalo, 2007; Sarid and Huss, 2010). Regarding the psychomotor domain, an improvement of psychomotor development in children with speech pathologies was observed after introducing a finger puppet theater approach (including the creation of puppets) in their correctional pedagogy training (Arkhipova and Lazutkina, 2022). It is apparent that psychomotor therapy is well complemented by expressive arts therapy, as they share the element of active participation into activities that promote kinesthetic abilities, cognitive processes, and personal development (Haeyen et al., 2021b; Arkhipova and Lazutkina, 2022). In order to limit the scope of this review on the healing qualities of creating a visual artwork, this review focuses solely on art therapy. Art therapy includes practices such as drawing, coloring, painting, collaging, sculpting, and allows the use of any media and materials that can be utilized to create visual artworks of symbolic value (Moon, 2011).

There are various definitions of art therapy as different schools of thought in psychology attribute different definitions and sometimes different psychotherapeutic goals to it. Three of the most notable approaches to art therapy are: humanistic, psychodynamic, and cognitive-behavioral. Humanistic art therapy is founded on the active participation of the client and the facilitation of the therapist

in the exploration of the client's artwork and its underlying narratives (Farokhi, 2011). The psychodynamic approach in art therapy focuses on the unconscious of the human mind and borrows concepts from analytical and archetypal psychology which is associated with symbolic images (Malchiodi, 2011). Cognitive-behavioral art therapy focuses on the attitude change of the client through visually externalizing problematic situations and identifying coping strategies (Rosal, 2018). Overall, art therapy, as an expressive arts therapy discipline, most commonly refers to a triangular psychotherapeutic relationship between therapist, client, and artwork. Schaverien (2000) proposes that what is referred to as "art therapy" *per se* in the field of psychology, encompasses two other distinct forms of therapy, namely "art psychotherapy" and "analytical art psychotherapy." The differences between these forms lie in the dynamic within the triangular psychotherapeutic relationship. Schaverien (2000) distinguishes "art therapy" as a process, in which the relationship between client and artwork is the main focus, whereas in "art psychotherapy" it is the relationship between client and therapist that is most emphasized, and in "analytical art psychotherapy" all three components of the relationship constellate equally.

Therapeutic artmaking, which can also be found in literature as "Art as Therapy," contrary to art therapy, is a low-intensity intervention for which the involvement of a therapist is considered a non-prerequisite. The inherently beneficial properties of artmaking and the inclination of patients to turn artmaking into a coping mechanism against illness have been evident for centuries but scientific research on the healing aspects of art is fairly recent (Farokhi, 2011). Immersive engagement with artmaking stimulates the senses, directs the artist's mind to the present time, and employs multiple cognitive processes, such as problem-solving, differentiation, and decision-making (Rosal, 2018). Also, contrary to art therapy, therapeutic artmaking can be characterized as a recreational experimentation with art materials, in which the overarching goal for the client is the creation of visually appealing artworks (Angheluta and Lee, 2011; Worden, 2020).

Some argue that therapeutic artmaking is a form of art therapy in which the psychoanalytic value of creating art is attenuated in favor of focusing on the inherent healing qualities of creating art (Czamanski-Cohen, et al., 2014). Others believe that art therapy and therapeutic artmaking should be considered as two completely different practices because of ethical considerations, as art therapy in contrast to therapeutic artmaking, requires the guidance and expertise of healthcare professionals (Angheluta and Lee, 2011). Despite the differences between the two therapeutic approaches, therapeutic artmaking and art therapy share a similar therapeutic intent. As Worden (2020) elucidates, self-expression in art therapy opts for making an individual able to work through past traumas among other psychological issues, whereas therapeutic artmaking opts for evoking a feeling of catharsis, encouraging socialization, honing technical skills that cater to visual self-expression, and increasing self-esteem. All the above are positive outcomes that affect wellbeing. Admittedly, therapeutic artmaking and artistic expression as a psychotherapy, rehabilitation, or counseling intervention, are used complementary to each other to various degrees (Farokhi, 2011; Malchiodi, 2020). Despite the main focus of the present scoping review being VR and therapeutic artmaking, the discipline of art therapy in this context cannot be omitted, because of the indicated interweaving between therapeutic artmaking and art therapy.

The topics of VR in therapeutic artmaking and art therapy remain vastly understudied, even though the groundwork that underscores potential uses of VR in different forms of psychotherapy has been laid during previous decades (Riva, 2005). For instance, virtual reality exposure therapy (VRET), which refers to the systematic habituation of patients to stimuli reminiscent of traumatic memories through VR, is the most studied form of psychological VR intervention and is widely endorsed as a valid alternative to traditional psychotherapy (Deng et al., 2019). The efficacy of VRET was evident since the infancy of VR (Hodges et al., 1995), despite the low quality of graphics or level of human-computer interaction. Nowadays VRET is deemed as an equally effective treatment to *in vivo* interventions for a variety of disorders, such as specific phobias and anxiety disorders (Carl et al., 2019; Mozgai et al., 2020) among others. With the exception of VRET, the number of studies addressing VR in psychotherapy is still limited (Frewen et al., 2020) and considering the continuous changes in the technological landscape, definitive conclusions about the efficacy of art therapy interventions in VR cannot yet be drawn.

Furthermore, multiple commercial applications for wellbeing can be found across platforms and devices, but their efficacy is under scrutiny. Wagener et al. (2021), after conducting a systematic application review, raised some critical points about the mismatch between the well-grounded theoretical background behind VR wellbeing applications and their subpar outcomes when it comes to practice. As they point out, most VR wellbeing applications unilaterally focus on specific wellbeing aspects and lack the flexibility of customization as well as opportunities for individual expression. Additionally, VR applications support users in identifying and reflecting upon their affective states, but they do so to a minimal degree. This reveals the need for further discourse on the potential and current limitations of practices about VR for wellbeing, which should derive insight from multiple disciplines for better understanding the intricacies of VR wellbeing. To this end, a scoping review is the most appropriate type of knowledge synthesis because it is most efficient in conveying the breadth of a variety of practices in a particular research area (Brien et al., 2010) and can help clarifykey interdisciplinary concepts and definitions, as well as identify types of evidence and knowledge gaps in the literature.

Presently, and to the best of our knowledge, no literature reviews specifically focusing on the subject of immersive VR and art as a therapeutic practice exist. Pissini (2020), who focuses on VR as a medium for artmaking, acknowledges the importance of studying VR practices in relation to the healthcare field, even though Pissini's indicated work is deliberately directed towards other interesting aspects of immersive VR artmaking, such as embodied creativity. One literature review most relevant to the present paper titled "Technology use in art therapy practice: 2004 and 2011 comparison" by Orr (2012), reviewed art therapy practice to every available technology at the time (between 2004 and 2011), and the author concluded that there was a gap between technological advancements and art therapy training, which further bred ethical limitations to the use of advanced technology in art therapy. Aspects of the ever-changing technology should be taken into account for the effective renewal of therapeutic practice, as old models of practice show signs of eventually becoming obsolete (Salles et al., 2020). Especially after the COVID-19 pandemic and the mobility restrictions imposed on therapists and clients alike, therapists were forced to reevaluate their methods and find ways to best utilize technology to mitigate the negative effects of the pandemic on the

normality of treatment procedures (Feniger-Schaal et al., 2022). Long-distance VR interventions are deemed as suitable alternatives to face-to-face interventions in the case of treatments that require more "acting" instead of "talking," such as psychomotor therapy, because of the experiential nature of using VR (Haeyen et al., 2021b). This scoping review revisits the technological gap indicated by Orr but purely focusing on VR, with the objective to assess the therapeutic utility of art-related practices in VR and provide guidelines for future research.

The question sought to be answered is twofold: 1) how has VR been integrated into practices relevant to art therapy and therapeutic artmaking? This review seeks to analyze how relevant studies define and juxtapose VR and artmaking in the context of therapy. Answering this first inquiry, while bearing in mind the overarching goals that each relevant study implies, can shed some light on the different forms in which advancing technology and artistic expression can manifest in a therapeutic setting; 2) how applicable are VR interventions for achieving therapeutic goals in relation to traditional art therapy and therapeutic artmaking practices? Through this inquiry, VR art-related interventions are being explored and contrasted to the well-established interventions that are traditionally used in art therapy and therapeutic artmaking. This inquiry is viewed from both the perspective of the client/user and the therapist/researcher to identify possible limitations and challenges.

2 Research methodology

2.1 Eligibility criteria

2.1.1 Inclusion criteria

The inclusion criteria were 1) academic manuscripts published between 2011 and the end of the data collection process, which ended in November 2022. VR and computer graphics have changed drastically during the last decade hence academic articles published before 2011 were omitted. The reason for this drastic change in the past decade is the sudden introduction of cost-effective VR head-mounted displays (Harley, 2020). 2) Academic manuscripts on the topic of VR, wellbeing, art therapy and therapeutic artmaking. Out of the four disciplines of expressive arts therapy mentioned above, the present scoping review focuses only on the discipline of art therapy, as defined in the introduction. Concepts that are universal to all expressive arts disciplines are also considered to be relevant. Under the scope of the present review, both art therapy and artmaking are accepted as relevant interventions. The relevance of the interventions is appraised based on the inclusion of artistic expression that results in the creation of visual artworks *via* visuomotor integration. 3) Peer-reviewed academic manuscripts (research articles, quantitative and qualitative studies, opinion pieces, and essays).

2.1.2 Exclusion criteria

The exclusion criteria were 1) Manuscripts which focus solely on VR therapy techniques. One of the prime examples of VR therapy that is distinctively different from VR art therapy is VRET. 2) VR-related manuscripts that focus on the sensory aspect of art when the virtual experience allows only a passive participation of virtual reality users (i.e., watching or observing a VE). Being exposed to a VR simulation that is purposefully designed for inducing wellbeing outcomes is a method that is typically employed for psychological healing through aestheticism (Moller et al., 2020) and spirituality (Pendse et al., 2016). Therefore, the combination of animated images, ambient sounds, and other stimuli, which is used to provide a therapeutic experience without involving psychomotor activity, is more related to VR therapy than the practices of art therapy or artmaking. 3) VR-related manuscripts which refer to VR as its non-immersive counterpart (WoW) only, or demonstrate digital applications with the premise of exocentric navigation only. Non-immersive VR interventions are excluded due to technical qualities that result in differences in the overall experience compared to fully immersive VR systems (Rubio-Tamayo et al., 2017; Slater, 2018). 4) VR-related manuscripts which include art therapy or therapeutic art-making practices but do not draw any associations between VR and art therapy or VR and therapeutic artmaking other than being parts of a holistic therapeutic approach.

2.2 Information sources

The searching process was carried out mainly through 4 databases, namely Scopus (138), Web Of Science (80 in the topic of "virtual reality and art therapy"), PubMed (67), ScienceDirect (263 in the subject areas of psychology, social sciences, computer science, neuroscience, and nursing and health professions). All 548 papers were sorted and then a second search from Google Scholar was carried out. From Google Scholar, 7 additional papers were found that were identified as potentially relevant. Within the first 6 search result pages of Google Scholar, only 4 potentially relevant papers were found, as most relevant articles had already been detected by previous search engines and had already been included for screening. Therefore, the total number of publications that potentially met the required criteria was finalized at 555. No additional articles were retrieved afterwards using incremental searching.

2.3 Search methods

The keywords used for the search engines to retrieve the 555 publications were "virtual reality art therapy." In the stage of the publication retrieval and screening process, the usage of the keyword "art" required additional caution. From the first few searching sessions, it was apparent that the keyword "art" was inviting a multitude of irrelevant publications in the search results. The main reason for this outcome was that "art" in scientific literature is an excessively used term, with its most common usage found in the phrase "state of the art." Therefore, the exclusion of the phrase "state of the art" from the search engines has significantly increased the quality and decreased the load of search results. For example, the search strategy for the Google Scholar search engine that was performed on 09 November 2022 included the search query: ((virtual reality AND art* AND therapy) -"state of the art"). Another issue caused by the keyword "art" is that the acronym "ART" is also found to be excessively used in areas of healthcare and science, technology, engineering, arts, mathematics education (STEAM). This issue has been resolved through the screening process, where the relevance of the retrieved publications has been appraised.

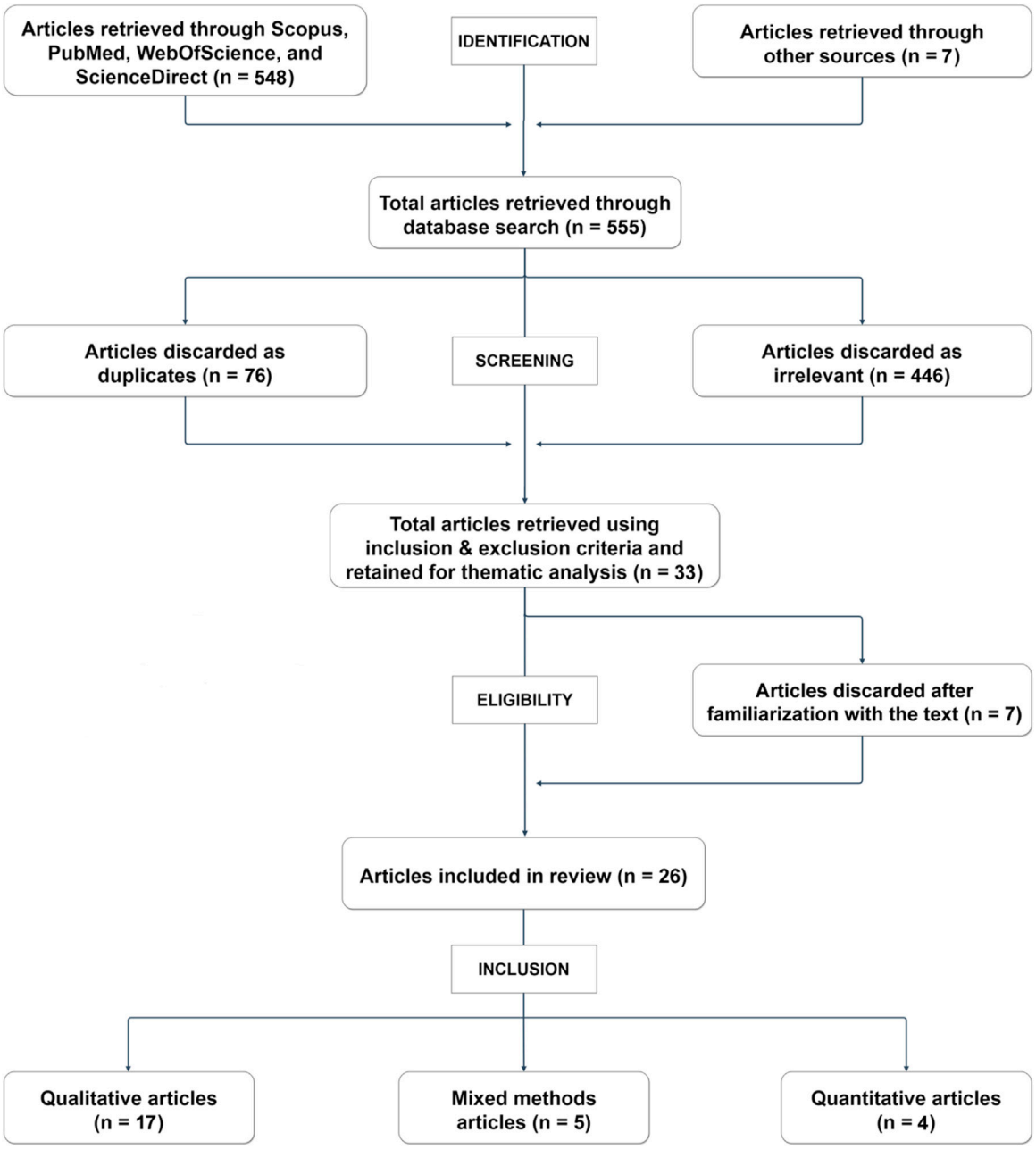

FIGURE 1
Flow diagram of the article retrieval process according to the PRISMA statement methodology.

2.4 Selection of sources of evidence

All 555 papers were screened by title and abstract. The screening and selection of the manuscripts was carried out on Rayyan (www.rayyan.ai), a web tool designed to help researchers carry out knowledge synthesis projects (Ouzzani et al., 2016). After all the sources were uploaded to the web tool, the 555 publications were manually categorized as duplicates (76), irrelevant (446), and relevant (33) according to the criteria and processes discussed. Some aspects of the inclusion and exclusion criteria were easier to pinpoint with precision once the authors were familiarized with the 33 selected papers. After a thorough examination of the theoretical background and if applicable the employed apparatus and methodology of the 33 papers, it became apparent that the relevance of 7 papers should be reconsidered. In 3 out of the 7 papers the human-computer interaction involved a 2D virtual interface instead of a VE, but their intervention was labelled as "virtual reality" regardless that VEs are integral components of immersive VR. In 2 out of 7 papers, the references to VR and the technology used during interventions were too vague and could be exclusively concerning a WoW approach or even be irrelevant to VR as defined in this review. In another paper, the study intervention involved solely mixed reality (MR) technology. The remaining 1 out of 7 papers describes a human-computer interaction model that

TABLE 1 Empirical studies included for analysis.

Author	N	Participants	Research Field	Research Design
Brahnam (2014)	N/A	Drama students and professional actors	Psychotherapy	Qualitative (Descriptive study)
Marks et al. (2017)	20	Clients from an art-based ethnographic framework	Psychotherapy	Qualitative (Pilot study)
Ying-Chun and Chwen-Liang (2018)	8	Art-therapists with clinical treatment experience	Psychotherapy	Qualitative (Pilot study)
Kaimal et al. (2020a)	20	Healthy, college-educated adults, aged 18-65	Psychotherapy	Qualitative (Pilot study)
Frewen et al. (2020)	36	University Students	Psychotherapy	Mixed Methods (Experimental Study)
Kaimal et al. (2020b)	24	Healthy adults aged 18-54	Psychotherapy	Mixed Methods (Pilot Study)
Baron et al. (2021)	16	Non-clinical volunteers	Rehabilitation	Mixed Methods (Experimental Study)
Richesin et al. (2021)	44	Undergraduate students, over 18 years old	Psychotherapy	Quantitative (Experimental Study)
Alex et al. (2021)	14	Stroke survivors aged 55–84	Rehabilitation	Mixed Methods (Field Study)
Haeyen et al. (2021a)	7	Art and psychomotor therapists aged 23–63	Psychotherapy	Mixed Methods (Action Research)
Iosa et al. (2021)	4	Stroke Survivors (time of acute event longer than 3 months ago)	Rehabilitation	Quantitative (Experimental Research)
Hacmun et al. (2021)	7	Expert female art therapists, aged 42-75	Psychotherapy	Qualitative (Action Research)
Shamri Zeevi (2021)	2	A boy, aged 16, and a girl, aged 13, who suffered from anxiety	Psychotherapy	Qualitative (Case Study)
Zhang et al. (2021)	10	Healthy subjects	Rehabilitation	Quantitative (Experimental Research)
Kaimal et al. (2022)	24	Healthy subjects aged 18–54	Psychotherapy	Quantitative (Experimental Research)

TABLE 2 Academic articles that synthesize empirical evidence and are included for analysis.

Author	Research field	Description
Carlton (2014)	Psychotherapy	Draws attention to the digital divide among art therapy practitioners and proposes ways to bridge the gap
Brahnam & Brooks (2014)	Psychotherapy	Proposes innovative practices in healthcare and encourages the investigation of the proposed ideas across disciplines
Hacmun et al. (2018)	Psychotherapy	Explores the potential clinical applications of VR, as a medium, under the framework of art therapy
Song et al. (2019)	Psychotherapy	Demonstrates the BioFlockVR bioresponsive system and its initial setup
Salles et al. (2020)	Psychotherapy	Addresses the issues that come with the digitization of art therapy and discusses possible solutions
Gatto et al. (2020)	Psychotherapy	Sets goals for developing an art therapy XR application for combating health hazards associated with social isolation
Jin et al. (2020)	Rehabilitation	Describes current tests used for the screening of various neurocognitive disorders and the possible applications of VR
Cheng et al. (2021)	Psychotherapy	Offers a perspective on the development of art therapy
Liu et al. (2021)	Psychotherapy, Communication, Computer Science	Summarizes the theme category and research hotspots as well as the application of art therapy aided health and wellbeing based on a bibliometric analysis
Li & Shen (2022)	Psychotherapy	Summarizes the clinical psychology research of expressive art therapy based on VR and reviews the current state of the art
Baldursson et al. (2022)	Psychotherapy	A brief demonstration of the Nebula VR application that seeks to expand the notion of somaesthetic appreciation through artistic creation

generates artworks solely based on the user's affective state, without any psychomotor activity being evidently required. With these 7 papers abiding to the exclusion criteria, the 26 remaining papers consisted of mixed methods (5), qualitative (17), and quantitative (4) research publications (Figure 1); 15 of the 26 included publications consist of empirical studies (Table 1). The rest of the papers consist of academic articles that have aims other than the production of novel empirical knowledge (Table 2).

TABLE 3 Summary of the advantages and disadvantages, challenges, and limitations in digitizing art therapy and therapeutic artmaking by using the VR medium.

Advantages of VR over traditional media					
Expanding the current limitations of artistic ability	Gaining novel insights of therapeutic value	Overcoming the restraints of physical space	Ludicity and Motivation	Digital natives in the therapeutic setting	Facilitating client data management
VE customization	Increased attention to the present time	Comfortability and freedom of physical transportation	Enhanced psychological support for completing treatments	Digital natives tend to be drawn to technology	Body movement tracking
Creation of "safe spaces"	Attitude change through presence and embodiment	Illusion of being transported to another world	Source of inspiration	More likely to foster trust towards therapist when digital natives are involved	High accuracy in physiological measurements
Manipulation of virtual physics	Perspective change and Replay Value	Space- efficient storage of artworks			
Disadvantages of VR over traditional media					
Tactility	Digital Literacy	Inclusivity	Interdisciplinary concerns	Technical Issues	Affordability and Maintenance
Lack of valid tactile input can lead to less emotional engagement	Lack of guidance regarding new media for therapists	Favors tech-nologically savvy teenagers and adults, who suffer from mild disorders	Cooperation between software developers and art therapists is essential	Less tools and materials for creating artworks	Relatively less affordable, renewable, and accessible to clients and therapists
	Lack of evaluation regarding VR approaches	Some experience cybersickness		VR artmaking sometimes feels alienating	
	Technophobia	Some experience escapist tendencies		Impeded physical movement and communication cues	
	Confidentiality concerns	Non-friendly VR applications for therapists			

2.5 Data charting process

After familiarization with the texts and the transcription process, the extracted transcriptions were recorded in a table format based on the content of the papers and more specifically, the advantages and disadvantages of VR over traditional media in the context of therapeutic artmaking and art therapy. Whenever applicable, the primary results, main conclusion, as well as methodological approaches, including apparatus, study purpose, data collection and analysis were also collected. The tables were tested by the reviewer team for refinements and to ensure that all relevant data were gathered. Therefore, data pertaining to the different approaches towards art and VR technology were also included. An inductive thematic analysis was used for the collected data to define the most common recurring themes. The data charting process was manually carried out while two reviewers were permitted to simultaneously edit the transcribed data on the shared tables. The emergent themes became apparent after one of the reviewers used color coding on the transcribed data that helped in discerning patterns and aggregating the data into meaningful categories.

2.6 Data items

The generated codes regarding the methodological approach of studies revealed that there are two main treatment categories, namely psychological and neurorehabilitation treatments. The generated codes regarding advantages and disadvantages of VR over traditional media in the context of therapeutic artmaking and art therapy revealed 6 main themes for each. The occurred themes regarding advantages were recognized as:

A.i. "Expanding the current limitations of artistic ability and expression in clinical settings through VR"

A.ii. "Gaining novel insights of therapeutic value through VR."

A.iii. "Overcoming the restraints of physical space"

A.iv. "Ludicity and motivation"

A.v. "Digital natives in the therapeutic setting"

A.vi. "Facilitating client data management"

The occurred themes regarding disadvantages were recognized as:

D.i. "Tactility",

D.ii. "Digital literacy"

D.iii. "Inclusivity"

D.iv. "Interdisciplinary concerns"

D.v. "Technical limitations"

D.vi. "Affordability and maintenance."

2.7 Synthesis of results

Thematic synthesis was used to formulate a descriptive analysis of the findings. In the results, reports on study logistics and the identified approaches regarding the use of art therapy and therapeutic artmaking in VR can be found. Also, advantages and disadvantages, limitations, and challenges of digitizing art therapy with the use of VR are presented in both narrative and table format (Table 3). This

scoping review was synthesized based on the PRISMA ScR checklist guidelines (Tricco et al., 2018).

3 Results

An overview of the 26 publications indicates that the concept of using VR as an artmaking tool in therapy context has been nascent in the past decade (2011–2022 included) and it has recently grown in trend, with 73.07% (n = 19) of the included papers being published between 2020 and 2022. A bibliometric analysis, which is conducted by Liu et al. (2021) and is spanning over a period of 75 years, confirms-through the co-occurrence of keywords used in the context of art therapy-that VR technology is becoming increasingly relevant with art therapy aided health and wellbeing research. Also, 57.69% (n = 15) of the included papers involve experimental research with human participants, with the remaining 42.30% (n = 11) consisting of opinion pieces, demos, and essays. All studies made use of a VR apparatus, most commonly mentioned being the Oculus, the HTC VIVE, and the Windows Mixed-Reality VR/MR headsets. Additionally, some of these studies included hardware complementary to VR, such as motion capture devices (MOCAP) for body tracking. Regarding the software, both custom and commercial VR drawing applications were used, with the most widely used one being the commercial application Google Tilt Brush. Participants fell under the category of either "patient" who suffers from psychological or physical conditions, "therapist," "university student," or "healthy subject," and the research objectives suggest high heterogeneity among studies.

3.1 Approaches to art therapy and therapeutic artmaking

This subsection constitutes general observations drawn from the reviewed papers, which do not necessarily conform to the acceptable definitions of "art therapy" or "therapeutic artmaking" as presented in the introductory section. The presentation of these observations provides a brief overview of the therapeutic approaches and the state of the art. In the reviewed papers, the terms of art therapy and therapeutic artmaking, sometimes clearly distinguished and other times used interchangeably, are used to describe a wide variety of treatments. In general, all the reviewed treatments operate on the basis that creative endeavors can generate emotions and incentives for both mental and physical wellbeing. Art, and more specifically art materials and media, are viewed as intermediaries between the realms of ideas and reality, which are experienced through the individual's senses. Specifically for VR, therapeutic activities such as painting, drawing, coloring, collaging, and sculpting, take a different form in the 3D environments, which could also vary (e.g., digital twins of an art therapy room or ostensibly infinite 3D canvases).

The authors identify two main categories, namely psychological and neurorehabilitation treatments, in which theoretical approaches to art diverge.

3.1.1 Psychological treatment approach

Most of the reviewed papers that are relevant to the field of psychology define their psychotherapeutic treatments in terms of art therapy, as a mental health profession and, more specifically, an expressive arts discipline. Art therapy is applied as a dynamic emotional therapy where art materials, the creative process, and the produced artwork serve as means of self-exploration and self-expression, in order to create personal change. In this context, the most widely mentioned theoretical influence is the psychodynamic perspective (Jungian psychology). This is often put in practice as depth psychology-based psychotherapy, in which the unconscious is brought to the surface thanks to the symbolic potential of artistic self-expression and, as a result, suppressed feelings are being revealed (Song et al., 2019). During this process, creating art and the psychotherapeutic relationship are elements of higher psychotherapeutic value than the final product of artistic creation. In analytical art psychotherapy, the patient's artwork is examined by the therapist to better understand the unconscious mind (Schaverien, 2000; Cheng et al., 2021).

Principles from other schools of thought are also adapted to the theoretical framework of the reviewed papers. Through the lens of CBT, art facilitates the communication of the individual's conceptual structures in a different way than verbal communication does, thus often providing alternative and illuminating perspectives to both the individual and the therapist. Within the triangular psychotherapeutic relationship, the artwork represents a subjective experience that is externalized by the client, in a way that the visualized mental relations of the client become more explorable and often reveal conflicting perspectives (Hacmun et al., 2021). The "open studio approach" to art therapy is an approach that many of the reviewed papers find befitting of VR art therapy because of the ludic nature of current VR art-related applications. The creation of the "safe space" is a prominent practice in art therapy. Drawing the "self," the "problem," and "coping mechanisms," such as a "sanctuary," is found to significantly alleviate psychological trauma (Frewen et al., 2020).

A strong implication that is pointed out is that artistic expression, even without being accompanied by verbal reflection or any psychotherapeutic intervention, could still be a source of psychological healing. Artmaking is viewed as an innate human characteristic and one of the most primitive forms of self-expression, which is continuously evolving as a result of technological advancement by encompassing new expressive capabilities (Hacmun et al., 2018). Art is deployed as a source of solidarity, inspiration, and a sensation of security during times of crisis (Gatto et al., 2020).

3.1.2 Neurorehabilitation treatment approach

Regarding rehabilitation treatments, it is revealed that VR drawing applications are often used by patients who suffer from post-stroke motor impairments or minor neurocognitive disorders. Under this framework, digital artmaking works as an effective treatment for rehabilitation because it values the enjoyability of the treatment, as it activates reward pathways of the brain, while supporting physical, cognitive, and emotional healing (Kaimal et al., 2020a). It also allows patients (e.g., stroke survivors) who encounter difficulties with speaking to express themselves non-verbally (Alex et al., 2021; Zhang et al., 2021). The most recent findings in the field of VR artmaking suggest that different approaches to artmaking activate different brain regions (Kaimal et al., 2022). Comparisons of prefrontal cortex activations between a visual tracing task of a drawing and creative self-expressive artmaking indicated significant differences. Distinctively, the implication is that creative self-expression, contrary to the tracing task, induces transient

hypofrontality, a state of the brain that is associated with relaxation and the inhibition of self-reflective processes. This suggests that different artmaking approaches could be used for achieving specific treatment goals.

From the perspective of neuroaesthetics, the field which engages with the perceptual, cognitive, and emotional aspects involved during an aesthetic experience, the element of the artwork is an apt addition in the neurorehabilitation practice. A reason for this is that creation, or even mere observation, of artworks and the practice of rehabilitation are both tightly associated with sensorimotor activity, which is found to be cognitively concomitant to the emotive expressions of painted figures (e.g., the figures in "The Creation of Adam" by Michelangelo) (Iosa et al., 2021). Artworks are found to neurobiologically induce motivation and affective arousal, which are fundamental aspects of neurorehabilitation, along with active participation and treatment intensity.

3.2 Advantages of VR over traditional media in therapeutic artmaking and art therapy

3.2.1 Expanding the current limitations of artistic ability and expression in clinical settings through VR

The authors of the reviewed papers appear to advocate for a potential revolutionization in the field of art therapy, because of the advent of VR. Especially in psychotherapy, VR's increasing repertoire of tools for creative self-expression enables clients to better convey their conceptual structures to therapists and researchers by transforming and customizing the virtual environments where the therapeutic process takes place. The ability to tailor virtual environments according to the client's psychological disorder and the psychotherapeutic approach of the therapist could more easily provide both clients and therapists or researchers a common ground of communication (Hacmun et al., 2018).

The most notable contributions of artistic expression in VR psychotherapy are found to be the evocation of familiarity and safety in clients, as well as enhanced self-reflection and meditation. A common practice that is observed is the creation of a "safe space," which is created by the client according to the client's personal preference to serve as an emotional refuge. Safe spaces, which are usually in the form of houses or caves, have been used in psychotherapy practice long before VR but the ability to step into your artworks, which is exclusive to VR technology, expands the frontiers of this practice (Frewen et al., 2020). VR safe spaces are most beneficial for people suffering with trauma and post-traumatic stress, as safe spaces allow them to gather and sort their thoughts and feelings out while being at one with themselves (Brahnam, 2014). However, the creative ways the clients can express themselves through VR could go far beyond the concept of safe spaces and drastically vary, as clients continue to experiment with VR as a creative outlet.

In VR neurorehabilitation practices that involve artistic creation, as derived from the reviewed papers, artistic technique with emphasis on precision of movement seems to have an equally prominent, if not more prominent role, to that of artistic expression. Rehabilitation practices through VR technology are applicable because VR technology has evolved to allow a sufficiently high precision of movement, compared to the physical world. VR has the capacity of allowing patients who suffer from impaired mobility to make bold and expansive brush strokes in the virtual world by making simple gestures in the physical world (Baron et al., 2021). This result can be achieved by accordingly adjusting the movement translation of the virtual avatar to the patient's range of motion. In similar ways, the laws of physics in VR environments can be "adjusted" to the needs and comfort of the patients. In this sense, the VR medium offers a high level of independence to the user (Alex et al., 2021).

3.2.2 Gaining novel insights of therapeutic value through VR

Throughout the reviewed papers, the concept of experiential discovery through VR is prevalent. From the client's perspective, VR is a medium that is assumed to be effective in inducing emotional responses and stimulating cognition. Exposure in immersive VEs is found to decrease distracting thoughts (mind wandering) and increase properties required for eliciting attention, awareness, and self-reflectiveness. Some of these properties are presence and embodiment, which are important factors for changing implicit attitudes (Hacmun et al., 2018; Gatto et al., 2020). From the therapist's perspective as well as that of the researcher, VR enables the exploration of the client's mind as an equivalent of exploring virtual worlds (Marks et al., 2017). During the analysis, two of the most notable VR qualities that were indicated as salient contributors in gaining novel insights of therapeutic value in VR are enhanced perspective change and replay value (or playback).

The affordance of perspective change encompasses the user's perspective but also the ability of virtual object manipulation. In this instance, "virtual objects" refers to the clients' artistic creations, which play the role of externalized concepts. VR offers the ability of viewing objects from any angle, including from within the artwork, and the ability of scaling the size of objects (Hacmun et al., 2018; Li and Shen, 2022). The viewing of such objects from different vantage points is a practice that is already employed in psychotherapy and VR can drastically enhance this practice. Through coming across these different perspectives, clients are given the opportunity to deconstruct and reconstruct their conceptual structures, such as the concept of the self (Hacmun et al., 2018).

The ability to replay a recorded psychotherapy session and attempt to reassess the client's behavior within the virtual environment allow the psychotherapist or researcher and the client to have a more accurate and clearer understanding of the client's therapy progress (Brahnam, 2014). In other words, the digital affordance of replaying each digital brushstroke during art therapy in VR can enhance reflection. Replay value is one of the many VR qualities that require ongoing study (Carlton, 2014).

3.2.3 Overcoming the restraints of physical space

The reviewed papers often focus on the use of VR headsets that can be used in a patient's daily life, even outside of the care services. The portability of VR allows the patients to engage in therapy in the comfort of their home, where no time or transportation restrictions apply (Baron et al., 2021). However, even when it is mandatory for patients to either visit a clinical facility or remain hospitalized, VEs of immersive VR have the ability to transfer patients outside of the sterile physical environment of a clinical setting and "place" them somewhere more idyllic, thus providing inspiration and positive emotions. Furthermore, VEs are designed for the exploration of imaginal worlds and their design is in accordance with the central tenets of the creative processes in art therapy (Gatto et al., 2020). This transportability through immersive VR allows the stimulation of

the proprioceptive and vestibular senses without the need of a sizeable physical space and thus distinguishes immersive VR from other digital media. Equally important is the fact that VR artworks can be efficiently stored and retrieved for further editing without occupying physical space, unlike materials and artworks produced by traditional art therapy practices (Baron et al., 2021).

3.2.4 Ludicity and motivation

The reviewed papers support the idea that ludic play and gamification models, which are compatible with VR technology, assist in establishing therapeutic interventions that drive the patient's engagement while maintaining autonomy. Physical therapy requires time commitment while the therapeutic process is often arduous for the patient. Even so, the inherent qualities of VR applications motivate patients to keep exercising and instill in them the willingness to gain mastery over the new media (Baron et al., 2021). As no hard-and-fast rules for artmaking exist, VR artmaking is often viewed as an activity for relaxation and recuperation with no substantial impact on the physical world and free from the fear of failure or committing mistakes (Li and Shen 2022). A study that focused on the physiological measures of VR users during artmaking in a 3D virtual space, found a reduction in anxiety and negative affect (Richesin et al., 2021). Also, the same study suggests that the aspect of having an end goal during a VR simulation, such as completing an artwork, plays an important role when aiming for specific wellbeing related outcomes. Another aspect worth mentioning is the element of inspiration. Guided imagery, as an art therapy practice, requires imagination, which patients sometimes lack. VR immersive environments may be able to provide the inspiration necessary for unimaginative patients to evoke concrete ideas and possibilities more easily while sparking the interest in further exploring these ideas (Kaimal et al., 2020a; Li and Shen, 2022).

3.2.5 Digital natives in the therapeutic setting

As technology has permeated every facet of current society, the group of digital natives is continuously increasing due to the succession of generations. Therapists and researchers from the reviewed papers suggest that most digital natives are accustomed to interacting with technology, including VR, and interaction with technology is intuitive and enjoyable to them. As digital natives grow up in a technologically abundant environment, their minds become wired towards best utilizing the technological resources at their disposal (Marks et al., 2017). This is the main reason digital natives are often alienated by traditional media (i.e., art materials), which they often find too messy or even obsolete. In a psychotherapeutic setting, digital natives tend to feel more comfortable expressing themselves through means other than a conversation or a paper-and-pencil drawing. VR interventions are found to be great alternative options of therapy in cases of clients rejecting traditional therapeutic methods. Therefore, VR assists in building rapport between clients, especially younger ones, and their therapists, by enriching the psychotherapeutic relationship (Shamri Zeevi, 2021; Li and Shen, 2022). The use of new media, such as VR, in art therapy paves the way for further cultural exploration of digital natives and their interaction with technology (Carlton, 2014).

3.2.6 Facilitating client data management

The suitability and assistance of VR technology in data collection and data representation is occasionally mentioned in the reviewed papers. VR is widely portrayed as a technology that allows easier tracking of body movements, which on one hand is necessary for creating art and at the same time constitutes a crucial element of art therapy (Ying-Chun and Chwen-Liang, 2018). Additionally, body movement is often viewed as a form of expression in and of itself. Especially for VR neurocognitive tests, the high ecological validity that is provided by VR, in comparison to their paper-and-pencil counterparts, makes the data collected through VR arguably more valid. This is because VR provides the possibility of safely reenacting activities of daily living in VR as if in real life (Jin et al., 2020). Also, digital technologies can reach a large audience of patients and gather patient data that could lead to more informed decisions by both healthcare professionals and researchers. Data from multiple VR trials can be easily gathered and compiled, leading to reassessment and optimization of VR tests (Jin et al., 2020; Salles et al., 2020).

3.3 Disadvantages of VR over traditional media in therapeutic artmaking and art therapy

3.3.1 Tactility

With the field of VR haptics still evolving, the reviewed papers point out that VR is insufficient in providing a similar level of sensory stimulation and tactility as traditional media in art therapy. VR technology replaces tangible art materials with virtual ones, which could be characterized as orderly and often unfamiliar, and this is especially true for clients who perceive traditional art therapy materials as more intuitive and easier to use (Alex et al., 2021). The joy of "holding my completed artwork in my hands," is a quality that seems to be exclusive to physical artworks (Hacmun et al., 2018). The deprivation of sensory cues through the absence of sufficient tactility stimulation in digital media often brought up in the reviewed papers, is one of the main factors that prevent some art therapists from considering the adoption of not only VR but also other digital media in their psychotherapeutic treatments (Haeyen et al., 2021a).

3.3.2 Digital literacy

Most art therapy practitioners lack training in VR technology and their resources for employing VR are sparse. In addition, art therapy practitioners tend to consider traditional media in art therapy as more therapeutic than new media, even though there is no clear evidence of this belief (Carlton, 2014). According to the reviewed papers that addressed these issues, the lack of systematic training in new media for art therapists discourages the use of VR technology in art therapy practice. Specifically, art therapists are hesitant about the use of VR technology, as they acknowledge the lack of technological expertise in the field and have no clear direction in how to incorporate digital materials in their psychotherapeutic treatments, which are often highly experiential and active by nature (Brahnam, 2014; Haeyen et al., 2021a). Consequently, the lack of technological expertise results in a lack of evaluation regarding the psychotherapeutic efficiency of the digital tools available and many practitioners arrive to the arbitrary conclusion that new media are inefficient compared to traditional media (Salles et al., 2020). Some of the authors use the term "technophobia" to describe this phenomenon of repulsion towards new media. Technophobia is observed in practitioners and clients alike, as many of them admittedly perceive digital media and digital art

as lesser than their traditional counterparts (Jin et al., 2020). An important factor that caters to technophobia in clients is the uncertainty regarding confidentiality and privacy of the client's data accumulated during VR sessions (Marks et al., 2017). All things considered, and especially the obscurity of new media in art therapy graduate programs, there is also the issue that art therapy practitioners who are unfamiliar with new media may never come across the possibility of adopting VR in their psychotherapeutic practice. However, even when practitioners overcome any possible bias and consider the possibility of adopting VR, they are often intimidated by the steep learning curve and other limitations.

3.3.3 Inclusivity

From the analysis of the reviewed papers, it can be concluded that art therapy in VR is less inclusive than traditional art therapy. First and foremost, there is a digital divide among art therapy clients and those who are technology-savvy are more likely to find benefit in VR interventions (Shamri Zeevi, 2021). Secondly, the use of VR by children who are below the age of twelve or thirteen is not recommended for safety reasons, according to policies of VR headset manufacturers (Ying-Chun and Chwen-Liang, 2018). This age restriction specifically applies to VR gaming, so children younger than twelve could potentially make a healthy use of VR headsets when supervised by adults. Even so, this age restriction implies that VR usage requires the user to have a level of cognitive development that is higher than the one required for the usage of traditional art therapy media. Also, VR interventions are unsuited for people with major neurocognitive disorders, acute motor and vestibular issues, and those who are prone to headache and nausea, as the phenomenon of cybersickness seems to be a glaring problem. In addition, VR interventions are unsuited for people who suffer from hallucinations and those who struggle to distinguish between reality and fantasy (Kaimal et al., 2020a; Jin et al., 2020). The restrictions mentioned so far do not apply to traditional art therapy and even when clients seem to qualify for the use of VR, the opposite could be proven during therapy. For example, the client could be prone to distraction by the VR intervention. In this case, the client could easily diverge from the course of therapeutic practice, especially if the art therapy facilitator is negligent or unfamiliar with new media (Carlton, 2014). It is difficult to predict a client's response to a VR intervention before the beginning of the intervention as VR qualities are experienced differently by each user. Some users do not experience the illusion of presence–being in a different place than the physical one when using VR—but others experience this illusion too intensely. Regarding the latter case, patients may use VR interventions for unhealthy escapism instead of coping with real life problems (Kaimal et al., 2020a). This adds an extra layer of complexity in deciding whether VR interventions are benefactory to all clients. Issues with inclusivity can be encountered from the side of the art therapy practitioner too. This view stems from the observation that most applications of VR in psychotherapy are used in the context of CBT, while other approaches, such as the humanistic approach and their practitioners are obscured. Similarly, VR could be considered as impractical for some groups of art therapy practitioners who endorse art therapy practices other than the ones offered by most of the available VR applications (Brahnam, 2014). It should be mentioned that practitioners often develop VR-applicable techniques based on various concepts and strategies (e.g., technical eclecticism, Ludic Engagement Designs for All) that could potentially cater to both the expertise of each practitioner and the unique needs of each client (Brahnam and Brooks, 2014; Frewen et al., 2020).

3.3.4 Interdisciplinary concerns

Despite the attempts to digitize art therapy and therapeutic artmaking, more research is needed to ensure the satisfaction of therapists and clients alike regarding the efficacy of digital interventions. This seems to be true for both psychological and rehabilitation treatments in VR, according to the reviewed papers. Physical rehabilitation in its traditional form has long been proven as an effective method of regaining functionality, whereas more studies are needed to determine the extent to which VR rehabilitation is efficient (Baron et al., 2021). Digital applications are notorious for widespread misinformation and, with applications in the healthcare industry being no exception, clinicians point out the significant risk of applications dictating therapy through digital means (Salles et al., 2020). Clinicians stipulate that digital applications relevant to therapy should be flexible enough to be customized for each client instead of being adapted to the developers' process of working. For example, art therapy is often misunderstood as the notion of simply making art for psychological healing or the notion that the completed artwork of the patient is a solid projection of the patient's psychopathology. When these misconceptions are transferred to the digital realm, there is the danger of excluding the framework that makes art therapy therapeutic, such as aspects of the triangular psychotherapeutic relationship and the subtle expressions during the construction of an artwork (Salles et al., 2020).

3.3.5 Technical limitations

Admittedly, the current state-of-the-art VR software for artmaking offers less sophisticated artistic capabilities for creative expression than traditional art therapy media. The range of tools and materials, which are available for drawing, painting, sculpting, and collage in VR, is comparatively more limited than the physical gamut of art tools and materials (Kaimal et al., 2020a).). Also, it can prove difficult for the client to convey some of the intentional or unintentional messages to the therapist or researcher due to the physical form of VR technology. By drawing examples from the reviewed papers, a common problem is that VR head-mounted displays hide the facial characteristics of the client and disallow eye-contact between the client and the physical environment, including the therapist (Shamri Zeevi, 2021). In the case of physical rehabilitation, the rigidity, agelessness, and multi-perspective angles of 3D digital strokes often alienate patients, whose location tends to remain constant, due to them being inert (Alex et al., 2021). A common issue for therapists and researchers is that they can only have a glimpse of the therapeutic process through a 2D projection of a computer screen, which makes monitoring the VR user (Shamri Zeevi, 2021). Overall, the reviewed papers suggest that it is easier to evaluate social cues and initiate social interactions during traditional methods rather than VR interventions.

3.3.6 Affordability and maintenance

According to Carlton (2014), there is a lack of affordability, renewability, and accessibility to new media compared to traditional media in art therapy. The issue of high cost in VR equipment and the development of specialized software applications, compared to traditional art therapy media and materials, is found in many of the reviewed papers. Specialized VR

systems are usually available to psychotherapists only if provided by healthcare organizations that have their own IT departments where limitations, such as cost and maintenance, are most viably mitigated (Brahnam, 2014).

4 Discussion

The findings indicate that the progress of VR technology is facilitating the use of VR in therapeutic practices relevant to the creation of artworks to a degree that allows a plethora of possibilities for innovation. In the field of psychotherapy, 3D digital brush strokes are employed to facilitate communication between client and therapist or researcher but also to act as building blocks for "safe spaces," work as colorful mood regulators for reducing anxiety, and unravel empirical insight for self-improvement. The 3D digital strokes of Tilt Brush were only an instance of how therapeutic artmaking and art therapy manifest in VR. New forms of artistic expression are beginning to emerge through VR, such as generating graphics using biomarkers or simple gestures (Song et al., 2019; Baldursson et al., 2022). Some of these art forms are made available thanks to the combination of VR with other technologies, such as the Brain-Computer Interfaces (BCI), where brain activity data can be easily collected and decoded to create control signals for virtual objects (Coogan and He, 2018). One of the least expected findings was the emergence of digital drawing techniques in the field of neurorehabilitation. Rehabilitation strategies that involved the creation of traditional artworks were proposed by Skinner and Nagel (1996), however the literature on the subject has been scarce for over 2 decades. Recent studies suggest that new media alleviate physical constraints from in-therapy motor-impaired patients to the point of allowing them to paint digital artworks and even "recreate" classical masterpieces in the form of a simulation (Iosa et al., 2021; Zhang et al., 2021). For example, researchers used VR technology to simulate the illusion of painting classical art masterpieces, designed for the neuroaesthetic stimulation of patients with an affected upper limb, and found a "Michelangelo effect" arising (Iosa et al., 2021). These imitations are neurologically comparable to performing the observed activity, akin to a virtual simulation, hence the term "embodied simulation" could be used (Buk, 2009; Finisguerra et al., 2021). It is argued that the capabilities of computer simulations in inducing neuroplasticity are best utilized by the technology of VR, which provides patients with interactive, stimulatory and ecologically valid VEs (Cheung et al., 2014). The sense of presence, which is induced within an immersive VR simulation, accounts for a bountiful allocation of cognitive resources that are relevant to motor control and is estimated to be one of the main factors that make VR technology especially suitable for rehabilitation (Slobounov et al., 2015).

VR art therapy and therapeutic artmaking seem to be promising future interventions for wellbeing. Artmaking in immersive VR was found to lessen the participants' insecurities about their skill in artmaking, allowing them to be more creative and focused on their therapeutic goals (Kaimal et al., 2020a). The malleability of VEs and their ability to adapt to the needs of clients as well as the illusion of presence are found to be some of the main contributors in facilitating healing. The symbolic, explorative, and controlled nature of VR art therapy allows personalized experiences that can be observed through different distances and points of view. These qualities allow individuals to create order from the fragmentary aspects of life and make sense of their emotions (Malchiodi, 2002). VR could induce motivation and inspiration in clients, especially in those who are receptive towards new media, such as digital natives, whose life typically operates in both the physical and the digital world. Some digital natives choose to allocate their energy resources more in the digital world than the physical one and this phenomenon seems relevant to sociobiological factors and data ubiquity. Taking into consideration the above findings and the conclusions of the systematic review by Wagener et al. (2021), it could be argued that the lack of applicability of holistic wellbeing approaches, customization, and self-expression in VR is prominent possibly because substantial steps to elevate the state of the art in the direction of art therapy and therapeutic artmaking have yet to be made.

The reviewed papers often indicate disadvantages of VR over traditional art media while implying that, nowadays, most of the highlighted issues are surmountable enough. Most of the explored issues, such as interdisciplinary concerns and the lack of digital literacy, are likely to move towards resolution the more they become adequately addressed. Other issues, such as the lack of cost-effective solutions for VR ownership and development, as well as tactility absence in VR during artmaking, can be more nuanced. The lack of tactility experienced through digital art therapy initially used to be one of the main sources of skepticism in art therapists regarding the adoption of digital art therapy practices. However, it has been observed that the lack of tactility could be an uncanny feeling for some but also a trivial matter for others, whose modalities combine inside the VR environment to create an illusion of tactility (Hacmun et al., 2021). Currently, electrostimulation-based techniques are employed to tackle both the issue of tactility absence and cybersickness in VR (Li et al., 2020; Vizcay et al., 2021).

All the reviewed papers are unanimous regarding the appropriateness of VR in art-based therapeutic practices, even though some researchers from the reviewed papers but also from the broader literature challenge the notion that fully immersive VR favors interventions for which the connection between client and therapist is deliberately distant (Gatto et al., 2020; Hacmun et al., 2021). Xiong et al. (2022) argue that Augmented Reality (AR) could be a more suitable technology for art-based rehabilitation interventions than fully immersive VR because of the increased possibilities for social interaction in AR, among other reasons. As Alex et al. (2021) argue, sociability in VR applications, even though it exists, needs to be a more prominent feature. The accessibility increase of psychotherapeutic practices to collaborative VR spaces through telepresence could mitigate some of the interdisciplinary and technical problems which VR psychotherapy sustains, such as the limited accommodation of the triangular psychotherapeutic relationship to the VE. Importantly, the role of the therapist or the researcher, who is cut off from observing the implicit actions of the client/user during an artistic fully immersive VR intervention, is bound to be degraded due to the observer's constrained ability to derive accurate conclusions. This implies that current VR technology favors the forms of art therapy in which the role of the therapist is peripheral to the psychotherapeutic relationship and this limitation is arguably detrimental to the overall usability of fully immersive VR as a therapeutic tool.

However, the authors by no means suggest that fully immersive VR is deleterious to the role of the therapist. Despite the drawbacks

that fully immersive VR has in store regarding the therapist's role, VR is also found to be especially useful for building trust between client and therapist, which is an important aspect of the psychotherapeutic relationship (Frewen et al., 2020; Shamri Zeevi, 2021). The term "collaborative VR" refers to social VR platforms that allow user-generated content and synchronous communication in 3D virtual spaces *via* telepresence (Saffo et al., 2021). Collaborative VR platforms where both client(s) and therapist(s) can simultaneously occupy the same virtual space through telepresence exist in a nascent stage (i. e., VRChat and the "Metaverse"), which are likely to become more of the norm in the years to come and prominent therapeutic spaces (Rzeszewski and Evans, 2020; Hacmun et al., 2021). These platforms can employ eye tracking and real-time facial expression mapping techniques for avatars, which could be a solution to the problem of communicating emotional cues through VR (Joachimczak et al., 2022). Nevertheless, a challenge that needs to be addressed regarding real-time facial expression mapping is that facial expressions of avatars need to mitigate their levels of perceived uncanniness and this challenge mostly concerns photorealistic avatars (Kumarapeli et al., 2022). As fully immersive VR is progressively becoming more geared towards commercial use, the development of collaborative VR is more likely to gain momentum and subsequently elevate the role of the therapist and the possibilities offered regarding art-based therapy treatments.

5 Future directions

Despite the topic of VR in therapy being relatively new and the challenges being many, there is promising evidence regarding the therapeutic use of art-based VR interventions. The variables that constitute an effective digital application for art therapy are already evident (Marks et al., 2017) but the transformation of theoretical knowledge into effective therapeutic practices, especially in the case of new media, needs further experimentation. As suggested by the reviewed papers, future research could pertain to the transfer and optimization of neurocognitive tests in VR with emphasis on drawing and visuospatial reasoning (Jin et al., 2020). VR also poses an opportunity for studying the impact of artmaking on the Autonomic Nervous System (ANS) from a theoretical standpoint that derives from pieces of research focusing on artmaking and anxiety disorders (Sandmire et al., 2012; Sandmire et al., 2016). It is already known that artmaking practices help in reducing stress and anxiety, but VR technology could elevate our understanding of artmaking even further, thanks to the facilitation of data management and the ecological validity offered by VR interventions (Richesin et al., 2021).

Even though many novel approaches to therapy have been described in this review, the full potential of immersive VR technology in therapeutic treatments seems to remain underutilized (Geraets et al., 2021) and art-based treatments are no exception. One of the most underutilized affordances of VR in art therapy and therapeutic artmaking is that of the embodied expression *via* virtual avatars. The concepts of embodied cognition, virtual avatars, and embodiment are commonly found in the literature of fully immersive VR (Kyrlitsias and Michael-Grigoriou, 2022), but the role of the avatars in the reviewed papers was given little to no attention. Reviewed studies that deployed fully immersive VR were found to be limited to the obligatory motion capture of the hands, provided by the controllers or other motion-capture techniques, which tend to make the VR user feel like a body-less ghost with visible hands. Arguably, the embodiment of virtual avatars, and its subsequent sensations of body ownership and body agency, should be considered as crucial elements of art-based therapeutic treatments in VR because of the indispensable role of kinesthetic and sensorimotor activity, as well as spatial awareness, during art therapy interventions. As Malchiodi (2020) notes, the most compelling reason for using any expressive arts therapy intervention is probably the sensory nature of the arts themselves, which cultivate cognitive and emotional awareness but also the awareness of somatic sensations that contributes to body-kinesthetic intelligence. Given that the facilitation of the communication between mind and body is one of the tenets of art therapy, when the VR user has no visual affirmation of having and controlling a body, the possibility that the lack of the expected visuo-proprioceptive stimuli downplays the efficacy of art therapy becomes prevalent. On a psychosocial level, embodied expression *via* avatars in VR could organically integrate into art-based therapeutic practices and enhance therapeutic experiences because of the possible influence of VR embodiment and presence in changing implicit attitudes. Adopting methods such as, the "*Proteus* Effect" could prove to be useful in the context of art therapy in numerous creative ways, for example using avatar-based emotional priming interventions (e.g., for attitude change) or aiming for avatar-assisted cultivation of psychomotor skills.

Continuing in the lines of embodied cognition, another area of interest for future research could be that of embodied simulation. The Michelangelo effect is a good example of the utilization of the mirror neuron system in motor-impaired individuals for neurorehabilitation. An inference from the study, in which the term "Michelangelo effect" hails (Iosa et al., 2021), could be that artmaking, even if subconsciously practiced *via* embodied simulation, activates visual-motor mirror neurons to the degree of facilitating neurorehabilitation. More research is needed to assess the level of significance of the association between VR artmaking (combined with shared body states of virtual humans) and neurorehabilitation.

Finally, psychotherapy interventions should go beyond "building safe spaces" when it comes to externalizations of mental representations in VR. Creative work in VR could often lead to creating a bridge between the physical world and emotions (Shamri Zeevi, 2021). Externalizations help in understanding where the person stands in relation to a problem, and what they might need in order to gain control over it, but it also provides understanding of the nature and the scale of the problem as it is already evident through enhanced perspective change and replay value (Marks et al., 2017). Intrusive mental images of distinct shapes and forms could also become available for constructive interaction through externalization. The case study of Walker et al. (2016) corroborates that bringing tormenting intrusive images to "the light of day" through artmaking allowed a sufferer to deconstruct and ultimately vanquish these reoccurring images. However, the efficacy of externalizations for the treatment of intrusive images on a large population with diverse experiences and levels of severity is still in question, since the exact underlying mechanisms of this treatment are unclear. VR technology, through its ecological validity, controllability of environments, and creative applications, provides an adequate opportunity for experimentation on externalizations of mental representations and intrusive mental imagery, while making it

possible to generalize results in a larger population. Rosal (2018) points out the necessity of clinical psychology research to focus on the study of intrusive mental images because they are related to many disorders for which our knowledge on their treatment is still insufficient.

6 Conclusion

The aim of this scoping review was to provide comprehensive information for therapy practitioners, application developers, and researchers, who could make use of the presented information to update current practices, and help elevate the state of the art in psychotherapy and rehabilitation. Knowledge regarding VR artmaking and art therapy in the area of wellbeing is already reported in various papers in a fragmented fashion and the scope of the present review was to congregate all the relevant information in a cohesive manner. The unique properties of VR and their significance to the area of art therapy and therapeutic artmaking were detailed and contrasted with traditional therapeutic practices. Further, this review provided the opportunity to focus on underexplored areas of VR practice in psychotherapy and rehabilitation, identify knowledge gaps in the literature and discuss potential future directions in the field of VR.

Author contributions

CH researched literature and conceived the review. CH wrote the first draft of the manuscript and substantially contributed to the creation of the final version. CH and DM-G were involved in data collection and CH was involved in data analysis and synthesis. DB substantially contributed to the revision of the manuscript and provided constructive feedback. DM-G substantially contributed to the critical evaluation of the manuscript and she also supervised the whole work from its conception to the final manuscript.

Funding

This work has been funded through the scholarship of academic excellence granted to CH for his doctoral studies and the research and other activities budget ED-DESPINA MICHAIL-300155-310200-3319 of the Cyprus University of Technology.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

R. Adams, S. Gibson, and S. M. Arisona (Editors) (2008). *Transdisciplinary digital art: Sound, vision and the new screen* (Springer Science & Business Media), 7.

Alex, M., Wünsche, B. C., and Lottridge, D. (2021). Virtual reality art-making for stroke rehabilitation: Field study and technology probe. *Int. J. Human-Computer Stud.* 145, 102481. doi:10.1016/j.ijhcs.2020.102481

Alqahtani, A. S., Daghestani, L. F., and Ibrahim, L. F. (2017). Environments and system types of virtual reality technology in stem: A survey. *Int. J. Adv. Comput. Sci. Appl. (IJACSA)* 8 (6). doi:10.14569/ijacsa.2017.080610

Angheluta, A. M., and Lee, B. K. (2011). Art therapy for chronic pain: Applications and future directions. *Can. J. Couns. psychotherapy* 45 (2).

Arkhipova, S., and Lazutkina, O. (2022). Psychomotor development of preschoolers with speech pathologies by means of art therapy techniques. *Rev. Tempos Espaços em Educ.* 15 (34), e17214. doi:10.20952/revtee.v15i34.17214

Bailenson, J. (2018). *Experience on demand: What virtual reality is, how it works, and what it can do*. WW Norton & Company.

Baldursson, B. R., Peterson, D., and Gamboa, M. (2022). "Nebula: Artistic somaesthetic appreciation with biosignals in virtual reality," in Adjunct Proceedings of the 2022 Nordic Human-Computer Interaction Conference, 1–3.

Bamodu, O., and Ye, X. M. (2013). "Virtual reality and virtual reality system components," in *Advanced materials research* (Paris, France: Trans Tech Publications Ltd, Atlantis Press), 765, 1169–1172.

Banakou, D., Beacco, A., Neyret, S., Blasco-Oliver, M., Seinfeld, S., and Slater, M. (2020). Virtual body ownership and its consequences for implicit racial bias are dependent on social context. *R. Soc. open Sci.* 7 (12), 201848. doi:10.1098/rsos.201848

Banakou, D., Groten, R., and Slater, M. (2013). Illusory ownership of a virtual child body causes overestimation of object sizes and implicit attitude changes. *Proc. Natl. Acad. Sci.* 110 (31), 12846–12851. doi:10.1073/pnas.1306779110

Baron, L., Wang, Q., Segear, S., Cohn, B. A., Kim, K., and Barmaki, R. (2021). "Enjoyable physical therapy experience with interactive drawing games in immersive virtual reality," in Symposium on Spatial User Interaction, 1–8.

Berkman, M. I., and Akan, E. (2019). "Presence and immersion in virtual reality," in *Encyclopedia of computer graphics and games*. Editor N. Lee (Cham: Springer). doi:10.1007/978-3-319-08234-9_162-1

Bertrand, S., Vassiliadi, M., Zikas, P., Geronikolakis, E., and Papagiannakis, G. (2021). From readership to usership: Communicating heritage digitally through presence, embodiment and aesthetic experience. *Front. Commun.* 6, 676446. doi:10.3389/fcomm.2021.676446

Biocca, F., and Levy, M. R. (2013). *Communication in the age of virtual reality*. Routledge, New York: Routledge.

Brahnam, S., and Brooks, A. L. (2014). "Two innovative healthcare technologies at the intersection of serious games, alternative realities, and play therapy," in *Innovation in medicine and healthcare 2014* (Springer, Cham: IOS Press), 153–162.

Brahnam, S. (2014). "HCI prototyping and modeling of future psychotherapy technologies in second life," in International Conference on Human-Computer Interaction (Cham: Springer), 273–284.

Brien, S. E., Lorenzetti, D. L., Lewis, S., Kennedy, J., and Ghali, W. A. (2010). Overview of a formal scoping review on health system report cards. *Implement. Sci.* 5 (1), 2–12. doi:10.1186/1748-5908-5-2

Buk, A. (2009). The mirror neuron system and embodied simulation: Clinical implications for art therapists working with trauma survivors. *Arts Psychotherapy* 36 (2), 61–74. doi:10.1016/j.aip.2009.01.008

Carl, E., Stein, A. T., Levihn-Coon, A., Pogue, J. R., Rothbaum, B., Emmelkamp, P., et al. (2019). Virtual reality exposure therapy for anxiety and related disorders: A meta-analysis of randomized controlled trials. *J. anxiety Disord.* 61, 27–36. doi:10.1016/j.janxdis.2018.08.003

Carlton, N. R. (2014). Digital culture and art therapy. *Arts Psychotherapy* 41 (1), 41–45. doi:10.1016/j.aip.2013.11.006

Cheng, C., Elamin, M. E., May, H., and Kennedy, M. (2021). Drawing on emotions: The evolving role of art therapy. *Ir. J. Psychol. Med.*, 1–3. doi:10.1017/ipm.2021.20

Cheung, K. L., Tunik, E., Adamovich, S. V., and Boyd, L. A. (2014). "Neuroplasticity and virtual reality," in *Virtual reality for physical and motor rehabilitation* (New York, NY: Springer), 5–24.

Christofi, M., Michael-Grigoriou, D., and Kyrlitsias, C. (2020). A virtual reality simulation of drug users' everyday life: The effect of supported sensorimotor contingencies on empathy. *Front. Psychol.* 11, 1242. doi:10.3389/fpsyg.2020.01242

Coogan, C. G., and He, B. (2018). Brain-computer interface control in a virtual reality environment and applications for the internet of things. *IEEE Access* 6, 10840–10849. doi:10.1109/access.2018.2809453

Czamanski-Cohen, J., Sarid, O., Huss, E., Ifergane, A., Niego, L., and Cwikel, J. (2014). CB-ART—the use of a hybrid cognitive behavioral and art based protocol for treating pain and symptoms accompanying coping with chronic illness. *Arts Psychotherapy* 41 (4), 320–328. doi:10.1016/j.aip.2014.05.002

Deng, W., Hu, D., Xu, S., Liu, X., Zhao, J., Chen, Q., et al. (2019). The efficacy of virtual reality exposure therapy for ptsd symptoms: A systematic review and meta-analysis. *J. Affect. Disord.* 257, 698–709. doi:10.1016/j.jad.2019.07.086

Farmer, H., Ciaunica, A., and Hamilton, A. F. D. C. (2018). The functions of imitative behaviour in humans. *Mind Lang.* 33 (4), 378–396. doi:10.1111/mila.12189

Farokhi, M. (2011). Art therapy in humanistic psychiatry. *Procedia-Social Behav. Sci.* 30, 2088–2092. doi:10.1016/j.sbspro.2011.10.406

Feniger-Schaal, R., Orkibi, H., Keisari, S., Sajnani, N. L., and Butler, J. D. (2022). Shifting to tele-creative arts therapies during the COVID-19 pandemic: An international study on helpful and challenging factors. *Arts psychotherapy* 78, 101898. doi:10.1016/j.aip.2022.101898

Finisguerra, A., Ticini, L. F., Kirsch, L. P., Cross, E. S., Kotz, S. A., and Urgesi, C. (2021). Dissociating embodiment and emotional reactivity in motor responses to artworks. *Cognition* 212, 104663. doi:10.1016/j.cognition.2021.104663

Frewen, P., Mistry, D., Zhu, J., Kielt, T., Wekerle, C., Lanius, R. A., et al. (2020). Proof of concept of an eclectic, integrative therapeutic approach to mental health and well-being through virtual reality technology. *Front. Psychol.* 11, 858. doi:10.3389/fpsyg.2020.00858

Gatto, C., D'Errico, G., Nuccetelli, F., De Luca, V., Paladini, G. I., and De Paolis, L. T. (2020). "Xr-based mindfulness and art therapy: Facing the psychological impact of Covid-19 emergency," in International Conference on Augmented Reality, Virtual Reality and Computer Graphics (Cham: Springer), 147–155.

Geraets, C. N., Van der Stouwe, E. C., Pot-Kolder, R., and Veling, W. (2021). Advances in immersive virtual reality interventions for mental disorders: A new reality? *Curr. Opin. Psychol.* 41, 40–45. doi:10.1016/j.copsyc.2021.02.004

Hacmun, I., Regev, D., and Salomon, R. (2021). Artistic creation in virtual reality for art therapy: A qualitative study with expert art therapists. *Arts Psychotherapy* 72, 101745. doi:10.1016/j.aip.2020.101745

Hacmun, I., Regev, D., and Salomon, R. (2018). The principles of art therapy in virtual reality. *Front. Psychol.* 9, 2082. doi:10.3389/fpsyg.2018.02082

Hadjipanayi, C., and Michael-Grigoriou, D. (2022). Arousing a wide range of emotions within educational virtual reality simulation about major depressive disorder affects knowledge retention. *Virtual Real.* 26 (1), 343–359. doi:10.1007/s10055-021-00568-5

Haeyen, S., Jans, N., Glas, M., and Kolijn, J. (2021a). VR health experience: A virtual space for arts and psychomotor therapy. *Front. Psychol.* 12, 704613. doi:10.3389/fpsyg.2021.704613

Haeyen, S., Jans, N., and Heijman, J. (2021b). The use of VR tilt brush in art and psychomotor therapy: An innovative perspective. *Arts Psychotherapy* 76, 101855. doi:10.1016/j.aip.2021.101855

Harley, D. (2020). Palmer Luckey and the rise of contemporary virtual reality. *Convergence* 26 (5-6), 1144–1158. doi:10.1177/1354856519860237

Hodges, L. F., Kooper, R., Meyer, T. C., Rothbaum, B. O., Opdyke, D., Graaff, J. J. D., et al. (1995). Virtual environments for treating the fear of heights. *IEEE Comput.* 28 (7), 27–34. doi:10.1109/2.391038

Iosa, M., Aydin, M., Candelise, C., Coda, N., Morone, G., Antonucci, G., et al. (2021). The Michelangelo effect: Art improves the performance in a virtual reality task developed for upper limb neurorehabilitation. *Front. Psychol.* 11, 611956. doi:10.3389/fpsyg.2020.611956

Jin, R., Pilozzi, A., and Huang, X. (2020). Current cognition tests, potential virtual reality applications, and serious games in cognitive assessment and non-pharmacological therapy for neurocognitive disorders. *J. Clin. Med.* 9 (10), 3287. doi:10.3390/jcm9103287

Joachimczak, M., Liu, J., and Ando, H. (2022). "Creating 3D personal avatars with high quality facial expressions for telecommunication and telepresence," in 2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW) (IEEE), 856–857.

Kaimal, G., Carroll-Haskins, K., Berberian, M., Dougherty, A., Carlton, N., and Ramakrishnan, A. (2020a). Virtual reality in art therapy: A pilot qualitative study of the novel medium and implications for practice. *Art. Ther.* 37 (1), 16–24. doi:10.1080/07421656.2019.1659662

Kaimal, G., Carroll-Haskins, K., Ramakrishnan, A., Magsamen, S., Arslanbek, A., and Herres, J. (2020b). Outcomes of visual self-expression in virtual reality on psychosocial well-being with the inclusion of a fragrance stimulus: A pilot mixed-methods study. *Front. Psychol.* 11, 589461. doi:10.3389/fpsyg.2020.589461

Kaimal, G., Carroll-Haskins, K., Topoglu, Y., Ramakrishnan, A., Arslanbek, A., and Ayaz, H. (2022). Exploratory fNIRS assessment of differences in activation in virtual reality visual self-expression including with a fragrance stimulus. *Art. Ther.* 39 (3), 128–137. doi:10.1080/07421656.2021.1957341

Kim, G., and Biocca, F. (20182018). "Immersion in virtual reality can increase exercise motivation and physical performance," in Virtual, Augmented and Mixed Reality: Applications in Health, Cultural Heritage, and Industry: 10th International Conference, VAMR 2018, Held as Part of HCI International 2018, Las Vegas (NV, USA: Springer International Publishing), 10, 94–102. Proceedings, Part II.

Kim, Y., and Lee, H. (2021). Falling in love with virtual reality art: A new perspective on 3D immersive virtual reality for future sustaining art consumption. *Int. J. Human–Computer Interact.* 38, 371–382. doi:10.1080/10447318.2021.1944534

Kumarapeli, D., Jung, S., and Lindeman, R. W. (2022). "Emotional avatars: Effect of uncanniness in identifying emotions using avatar expressions," in 2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW) (IEEE), 650–651.

Kyrlitsias, C., and Michael-Grigoriou, D., (2022). Social Interaction With Agents and Avatars in Immersive Virtual Environments: A Survey. *Front. Virtual Real.*, 2, p.786665. doi:10.3389/frvir.2021.786665

Li, B., and Shen, M. (2022). The psychological recovery of patients in the context of virtual reality application by a complementary medicine scheme based on visual art. *Evidence-based Complementary Altern. Med.* 2022, 7358597. eCAM. doi:10.1155/2022/7358597

Li, G., McGill, M., Brewster, S., and Pollick, F. (2020). "A review of electrostimulation-based cybersickness mitigations," in 2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR) (IEEE), 151–157.

Liu, Z., Yang, Z., Xiao, C., Zhang, K., and Osmani, M. (2021). An investigation into art therapy aided health and well-being research: A 75-year bibliometric analysis. *Int. J. Environ. Res. Public Health* 19 (1), 232. doi:10.3390/ijerph19010232

Maggio, M. G., De Luca, R., Manuli, A., Buda, A., Foti Cuzzola, M., Leonardi, S., et al. (2020). Do patients with multiple sclerosis benefit from semi-immersive virtual reality? A randomized clinical trial on cognitive and motor outcomes. *Appl. Neuropsychol. Adult* 29, 59–65. doi:10.1080/23279095.2019.1708364

Maister, L., Slater, M., Sanchez-Vives, M. V., and Tsakiris, M. (2015). Changing bodies changes minds: Owning another body affects social cognition. *Trends cognitive Sci.* 19 (1), 6–12. doi:10.1016/j.tics.2014.11.001

C. A. Malchiodi (Editor) (2011). *Handbook of art therapy* (New York, NY: Guilford Press).

Malchiodi, C. A. (2002). *The soul's palette: Drawing on art's transformative powers*. Boston: Shambhala Publications.

Malchiodi, C. A. (2020). *Trauma and expressive arts therapy: Brain, body, and imagination in the healing process*. New York, NY: Guilford Publications.

Marks, K., Marks, S., and Brown, A. (2017). Step into my (virtual) world: An exploration of virtual reality drawing applications for arts therapy. *Aust. N. Z. J. Arts Ther.* 12 (1), 99–111.

Maselli, A., and Slater, M. (2013). The building blocks of the full body ownership illusion. *Front. Hum. Neurosci.* 7, 83. doi:10.3389/fnhum.2013.00083

Mishina, A. V., Blinova, J. L., and Belomoyeva, O. G. (2017). Multimodal art therapy for overcoming negative emotional states among adolescents. *Helix* 8, 2307–2311.

Moller, H. J., Waterworth, J. A., and Chignell, M. (2020). "Returning to nature: VR mediated states of enhanced wellness," in International Conference on Human-Computer Interaction (Cham: Springer), 593–609.

Montana, J. I., Matamala-Gomez, M., Maisto, M., Mavrodiev, P. A., Cavalera, C. M., Diana, B., et al. (2020). The benefits of emotion regulation interventions in virtual reality for the improvement of wellbeing in adults and older adults: A systematic review. *J. Clin. Med.* 9 (2), 500. doi:10.3390/jcm9020500

Moon, C. H. (2011). *Materials & media in art therapy: Critical understandings of diverse artistic vocabularies*. New York, NY: Routledge.

Mozgai, S., Hartholt, A., Leeds, A., and Rizzo, A. S. (2020). "Iterative participatory design for VRET domain transfer: From combat exposure to military sexual trauma," in Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, 1–8.

Orr, P. (2012). Technology use in art therapy practice: 2004 and 2011 comparison. *Arts Psychotherapy* 39 (4), 234–238. doi:10.1016/j.aip.2012.03.010

Osimo, S. A., Pizarro, R., Spanlang, B., and Slater, M. (2015). Conversations between self and self as Sigmund Freud—a virtual body ownership paradigm for self counselling. *Sci. Rep.* 5 (1), 13899–13914. doi:10.1038/srep13899

Ouzzani, M., Hammady, H., Fedorowicz, Z., and Elmagarmid, A. (2016). Rayyan—A web and mobile app for systematic reviews. *Syst. Rev.* 5 (1), 210–10. doi:10.1186/s13643-016-0384-4

Pendse, A., Gravier, N., Deedwania, D., Gotsis, M., Patterson, M., and Summers, C. (2016). "Inner activity," in ACM SIGGRAPH 2016 VR Village, New York NY, United States (New York, NY: Association for Computing Machinery), 1–2.

Pifalo, T. (2007). Jogging the cogs: Trauma-focused art therapy and cognitive behavioral therapy with sexually abused children. *Art. Ther.* 24 (4), 170–175. doi:10.1080/07421656.2007.10129471

Pissini, J. (2020). *Embodied by design: The presence of creativity, art-making, and self in virtual reality*. PhD thesis. Columbus (Ohio): The Ohio State University.

Richesin, M. T., Baldwin, D. R., and Wicks, L. A. (2021). Art making and virtual reality: A comparison study of physiological and psychological outcomes. *Arts Psychotherapy* 75, 101823. doi:10.1016/j.aip.2021.101823

Riva, G. (2005). Virtual reality in psychotherapy: Review. *Cyberpsychology Behav.* 8 (3), 220–230. doi:10.1089/cpb.2005.8.220

Rosal, M. L. (2018). *Cognitive-behavioral art therapy: From behaviorism to the third wave*. New York, NY: Routledge.

Rubio-Tamayo, J. L., Gertrudix Barrio, M., and García García, F. (2017). Immersive environments and virtual reality: Systematic review and advances in communication, interaction and simulation. *Multimodal Technol. Interact.* 1 (4), 21. doi:10.3390/mti1040021

Rzeszewski, M., and Evans, L. (2020). Virtual place during quarantine–a curious case of VRChat. *Rozw. Reg. i Polityka Reg.* (51), 57–75. doi:10.14746/rrpr.2020.51.06

Saffo, D., Di Bartolomeo, S., Yildirim, C., and Dunne, C. (2021). "Remote and collaborative virtual reality experiments via social vr platforms," in Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 1–15.

Salles, J., Charras, M., and Schmitt, L. (2020). "Digital tools in psychiatry and art-therapy, which possible meeting points?," in *Annales medico-psychologiques* (Masson), 178, 43–47. 21 STREET CAMILLE DESMOULINS, ISSY, 92789 MOULINEAUX CEDEX 9, FRANCE: MASSON EDITEUR.

Sandmire, D. A., Gorham, S. R., Rankin, N. E., and Grimm, D. R. (2012). The influence of art making on anxiety: A pilot study. *Art. Ther.* 29 (2), 68–73. doi:10.1080/07421656.2012.683748

Sandmire, D. A., Rankin, N. E., Gorham, S. R., Eggleston, D. T., French, C. A., Lodge, E. E., et al. (2016). Psychological and autonomic effects of art making in college-aged students. *Anxiety, Stress, & Coping* 29 (5), 561–569. doi:10.1080/10615806.2015.1076798

Sarid, O., and Huss, E. (2010). Trauma and acute stress disorder: A comparison between cognitive behavioral intervention and art therapy. *arts psychotherapy* 37 (1), 8–12. doi:10.1016/j.aip.2009.11.004

Schaverien, J., 2000. The triangular relationship and the aesthetic countertransference in analytical art psychotherapy. The changing shape of art therapy: New developments in theory and practice, pp.55–83.

Shamri Zeevi, L. (2021). Making art therapy virtual: Integrating virtual reality into art therapy with adolescents. *Front. Psychol.* 12, 584943. doi:10.3389/fpsyg.2021.584943

Skinner, M. K., and Nagel, P. J. (1996). Painting a mural and writing an article: Creative rehabilitation strategies. *Rehabil. Nurs.* 21 (2), 63–66. doi:10.1002/j.2048-7940.1996.tb01678.x

Slater, M. (2018). Immersion and the illusion of presence in virtual reality. *Br. J. Psychol.* 109 (3), 431–433. doi:10.1111/bjop.12305

Slater, M., Neyret, S., Johnston, T., Iruretagoyena, G., Crespo, M. Á. D. L. C., Alabèrnia-Segura, M., et al. (2019). An experimental study of a virtual reality counselling paradigm using embodied self-dialogue. *Sci. Rep.* 9 (1), 10903–10913. doi:10.1038/s41598-019-46877-3

Slater, M., Pérez Marcos, D., Ehrsson, H., and Sanchez-Vives, M. V. (2009). Inducing illusory ownership of a virtual body. *Front. Neurosci.* 3, 214–220. doi:10.3389/neuro.01.029.2009

Slater, M., and Sanchez-Vives, M. V. (2016). Enhancing our lives with immersive virtual reality. *Front. Robotics AI* 3, 74. doi:10.3389/frobt.2016.00074

Slobounov, S. M., Ray, W., Johnson, B., Slobounov, E., and Newell, K. M. (2015). Modulation of cortical activity in 2D versus 3D virtual reality environments: An EEG study. *Int. J. Psychophysiol.* 95 (3), 254–260. doi:10.1016/j.ijpsycho.2014.11.003

Song, M., Tadeo, T., Sandor, I., Ulas, S., and DiPaola, S. (2019). "BioFlockVR: Exploring visual entrainment through amorphous nature phenomena in bio-responsive multi-immersant VR interactives," in Proceedings of the 2nd International Conference on Image and Graphics Processing, 150–154.

Tricco, A. C., Lillie, E., Zarin, W., O'Brien, K. K., Colquhoun, H., Levac, D., et al. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Ann. Intern. Med.* 169 (7), 467–473. doi:10.7326/m18-0850

Vallance, M., and Towndrow, P. A. (2022). Perspective: Narrative storyliving in virtual reality design. *Front. Virtual Real* 3, 779148. doi:10.3389/frvir.2022.779148

Vizcay, S., Kourtesis, P., Argelaguet, F., Pacchierotti, C., and Marchal, M., 2021. Electrotactile feedback for enhancing contact information in virtual reality. *arXiv preprint arXiv:2102.00259*.

Wagener, N., Duong, T. D., Schöning, J., Rogers, Y., and Niess, J. (2021). "The role of mobile and virtual reality applications to support well-being: An expert view and systematic app review," in IFIP Conference on Human-Computer Interaction (Cham: Springer), 262–283.

Walker, M. S., Kaimal, G., Koffman, R., and DeGraba, T. J. (2016). Art therapy for ptsd and tbi: A senior active duty military service member's therapeutic journey. *Arts Psychotherapy* 49, 10–18. doi:10.1016/j.aip.2016.05.015

Worden, M. (2020). 17 june *the difference between art Therapy and therapeutic art-making*. Art from the streets. Available at: https://artfromthestreets.org/blogs/news/art-therapy-near-me (Accessed: October 20, 2021).

Xiong, Z., Weng, X., and Wei, Y. (2022). SandplayAR: Evaluation of diagnosis game for people with generalized anxiety disorder. *Arts Psychotherapy* 80, 101934. doi:10.1016/j.aip.2022.101934

Yee, N., and Bailenson, J. (2007). The Proteus effect: The effect of transformed self-representation on behavior. *Hum. Commun. Res.* 33 (3), 271–290. doi:10.1111/j.1468-2958.2007.00299.x

Ying-Chun, L., and Chwen-Liang, C. (2018). "The application of virtual reality technology in art therapy: A case of tilt brush," in 2018 1st IEEE International Conference on Knowledge Innovation and Invention (ICKII) (IEEE), 47–50.

Zhang, Y., Wang, H., and Shi, B. E. (2021). "Gaze-controlled robot-assisted painting in virtual reality for upper-limb rehabilitation," in 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (IEEE), 4513–4517.

frontiers | Frontiers in Artificial Intelligence

TYPE Review
PUBLISHED 13 June 2024
DOI 10.3389/frai.2024.1394386

OPEN ACCESS

EDITED BY
Kezhi Li,
University College London, United Kingdom

REVIEWED BY
Laurent Moccozet,
University of Geneva, Switzerland
Antonio Sarasa-Cabezuelo,
Complutense University of Madrid, Spain

*CORRESPONDENCE
Gabriella Spinelli
✉ gabriella.spinelli@brunel.ac.uk

RECEIVED 01 March 2024
ACCEPTED 27 May 2024
PUBLISHED 13 June 2024

CITATION
Khatun N, Spinelli G and Colecchia F (2024) Technology innovation to reduce health inequality in skin diagnosis and to improve patient outcomes for people of color: a thematic literature review and future research agenda.
Front. Artif. Intell. 7:1394386.
doi: 10.3389/frai.2024.1394386

Technology innovation to reduce health inequality in skin diagnosis and to improve patient outcomes for people of color: a thematic literature review and future research agenda

Nazma Khatun, Gabriella Spinelli* and Federico Colecchia

College of Engineering, Design and Physical Science, Brunel Design School, Brunel University London, Uxbridge, United Kingdom

The health inequalities experienced by ethnic minorities have been a persistent and global phenomenon. The diagnosis of different types of skin conditions, e.g., melanoma, among people of color is one of such health domains where misdiagnosis can take place, potentially leading to life-threatening consequences. Although Caucasians are more likely to be diagnosed with melanoma, African Americans are four times more likely to present stage IV melanoma due to delayed diagnosis. It is essential to recognize that additional factors such as socioeconomic status and limited access to healthcare services can be contributing factors. African Americans are also 1.5 times more likely to die from melanoma than Caucasians, with 5-year survival rates for African Americans significantly lower than for Caucasians (72.2% vs. 89.6%). This is a complex problem compounded by several factors: ill-prepared medical practitioners, lack of awareness of melanoma and other skin conditions among people of color, lack of information and medical resources for practitioners' continuous development, under-representation of people of color in research, people of color being a notoriously hard to reach group, and 'whitewashed' medical school curricula. Whilst digital technology can bring new hope for the reduction of health inequality, the deployment of artificial intelligence in healthcare carries risks that may amplify the health disparities experienced by people of color, whilst digital technology may provide a false sense of participation. For instance, Derm Assist, a skin diagnosis phone application which is under development, has already been criticized for relying on data from a limited number of people of color. This paper focuses on understanding the problem of misdiagnosing skin conditions in people of color and exploring the progress and innovations that have been experimented with, to pave the way to the possible application of big data analytics, artificial intelligence, and user-centred technology to reduce health inequalities among people of color.

KEYWORDS

dermatology, artificial intelligence, skin of color, people of color, ethnic minorities, data augmentation, health inequalities

1 Introduction

Healthcare inequalities have been persistent throughout healthcare globally (Stuart and Soulsby, 2011). These imbalances are present in healthcare access, treatments, and outcomes among minority communities (WHO, 2018) and can lead to detrimental health consequences. Disparity in health outcomes can be based on several factors such as gender, age, ethnicity, access to support and care services, and familiarity with digital technology. Digital technology, including artificial intelligence (AI), has been implemented into several areas of healthcare to combat inequalities. Despite targeted approaches, challenges associated with resource constraints and unintentional biases pose threats to successful execution and development, predominantly for people of color (POC).

Studies have illustrated the use of AI within dermatological settings for skin diagnostics of lesions including melanoma. Melanoma is a common type of skin cancer that originates from melanocyte skin cells (Cancer Council, 2023). Recognising signs of melanoma is crucial for early detection: lesions often appear as moles undergoing changes in color, growth patterns, shape irregularities, or being elevated and itchy (Cancer Council, 2023). Unfortunately, POC are at a greater disadvantage in melanoma mortality rates for reasons including late diagnosis or incorrect treatment (Mahendraraj et al., 2017), the integration of AI could address these issues by benefiting both healthcare workers and POC, considering internal medicine and physician trainees were less likely to refer POC to specialists for further management, with only 25% of trainees referring a drug rash for POC compared to 40% for Caucasian patients (Hutchison et al., 2023).

This paper explores the role of digital technology and AI to reduce health inequality, while also evaluating the benefits and challenges of AI adoption. The use of AI in diagnosing skin conditions, especially among POC, has the potential to magnify existing health inequalities for POC. This paper is concerned with diagnostic accuracy, equity in healthcare, potential biases in the technology, and the use of appropriate terminology to enable a more considerable adoption of digital health technologies.

2 Methodology

For this literature review, an opportunistic search was carried out through Google and Google Scholar. The interconnection of health inequality, dermatology, and AI was investigated in several fields of research including engineering, computing, medicine, and healthcare by selecting relevant keywords. Only published academic literature and grey literature from reputable sources (e.g., American Journal of Clinical Dermatology and International Journal of Equity in Health) were selected. A total of 94 relevant papers were shortlisted based on the matching between keywords and the papers' title. A further selection took place following the review of the abstracts. This led to 45 publications (42 academic papers and 3 conferences) that were determined appropriate and relevant for this review. Other research databases, including PubMed, Science Direct and IEEE Xplore, have also been searched using the same selection criteria to ensure all recent, and key literature has been identified and included. From this cross-check, no new additional papers have been identified. Geographical locations or date of publication were not restricting factors. This was to ensure that all potential AI advancements in skin lesion recognition and approaches to mitigating health inequality were explored. Papers were selected regardless of whether the studies had POC representation; if a paper had information on skin tone or ethnicity, it was considered. This was to ensure a comprehensive understanding of the problem was identified, to remove chances of biases, and for a clear and transparent comparative analysis of skin color representation within AI. Literature not written in English was not considered to avoid the chances of misinterpreting any findings. Biases have also been mitigated by defining and using consistently certain keywords, which collectively establish the objective criteria for papers' selection at the title screening level (see Figure 1). This method ensured that the selection process was based on specific, predefined criteria rather than subjective judgment, resulting in reduced chances of potential bias. Papers were screened by all three authors, and any discrepancies were resolved through discussions. Taking this approach allowed for a transparent review process. Figure 1 shows a flowchart of the selection process to identify target papers.

The following combination of keywords was used to identify relevant papers: 'artificial intelligence within dermatology," 'people of color and skin diagnosis accuracy in artificial intelligence', 'clinical pathway and artificial intelligence use', 'skin diagnosis tools for people of color', 'AI skin diagnosis in people of color', 'Artificial intelligence use within healthcare', 'digital technology to reduce healthcare inequalities', 'artificial intelligence', 'overfitting in artificial intelligence and skin diagnosis', 'data augmentation in artificial intelligence and skin diagnosis', 'image selection for artificial intelligence and skin diagnosis', 'people of color representation within artificial intelligence', 'artificial intelligence vs. experts diagnosis accuracy of skin disease', 'health inequality', and 'language barriers'. The search for relevant literature stopped upon reaching saturation, where no additional literature matching the keywords could be found. The search end date was March 2024, to ensure the most recent publications were considered.

The initial search on Google Scholar was undertaken using the keywords previously listed. This search identified 94 papers. A comparison with searches on other scientific databases did not identify additional papers. The initial search identified 6 duplicate papers that were removed from the set prior to reaching title screening. 88 papers reached the title screening level and 15 were excluded at this stage. 73 papers reached the abstract screening level and 28 were excluded. 45 were considered eligible for full review. At this stage, no papers were excluded. The final set of papers considered in this review was 45.

3 Results

3.1 Digital technologies to reduce health inequality

Digital technology plays an important role in addressing and presenting opportunities to overcome several barriers within health inequality. Deployment of technology can be through virtual health services, telemedicine consultations, or educational initiatives. Technologies as such benefit marginalized communities that may be constrained by geographical locations, financial situations, or inadequacies in equal access to healthcare resources and services for all (Table 1).

The Core20PLUS5 is a national NHS strategy to reduce health inequalities on a system and national level. The approach identifies

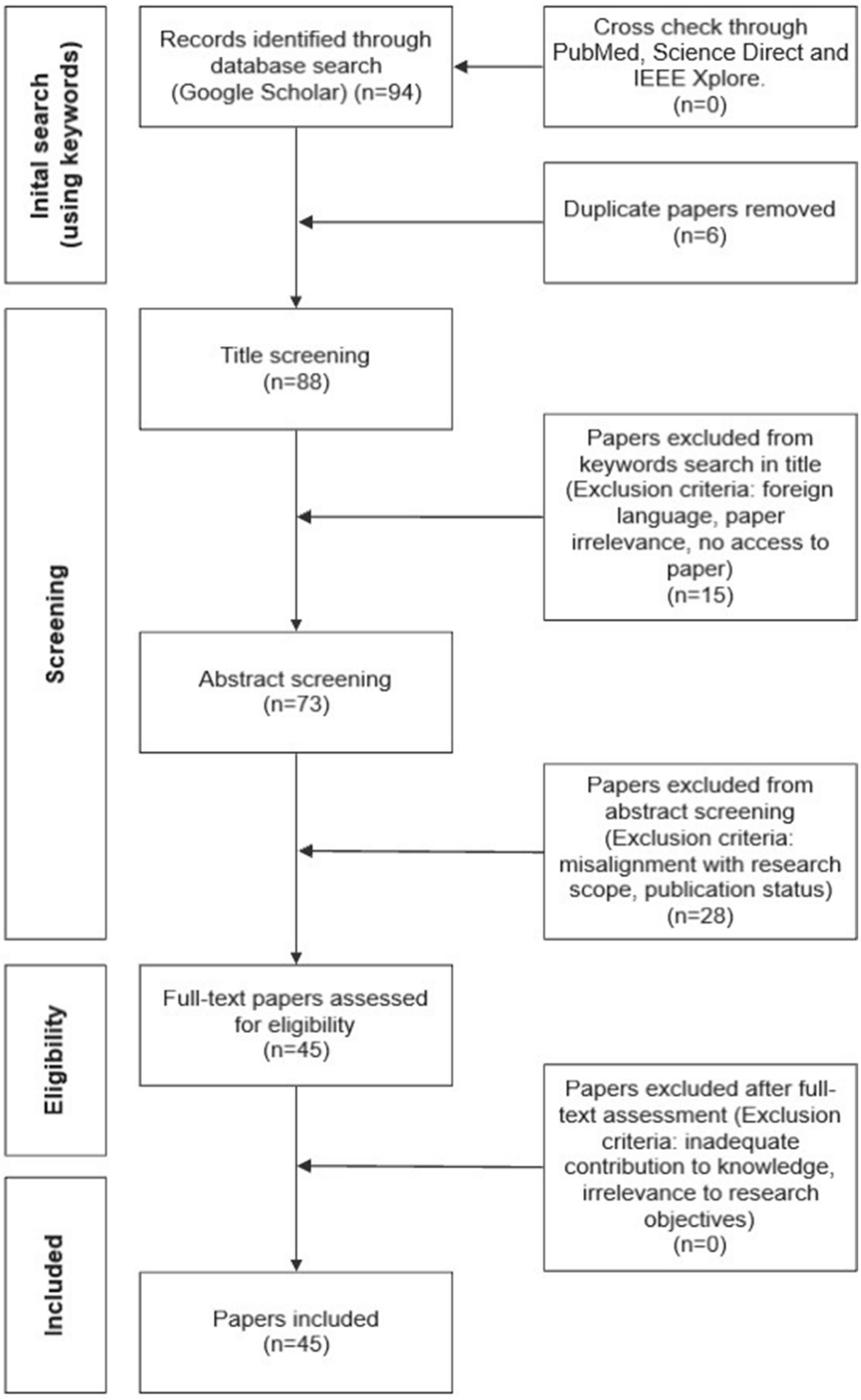

FIGURE 1
Flowchart of the selection process to identify target papers.

target populations among adults, young people and children, and clinical areas that need improvement (NHS, 2021a). Core20PLUS5 has three components: Core20 refers to the 20% of the most deprived national population, identified by the Index of Multiple Deprivation (IMD), PLUS relates to individuals including ethnic minorities or groups defined by the Equality Act 2010, and 5 stands for the five clinical areas which need improvement including severe mental illness or early diagnosis of cancer (NHS, 2021a). The strategy provides platforms, builds networks, and creates opportunities for sharing best practices. The targeted approach of Core20PLUS5 demonstrates clinical priority areas being addressed to attain health equality and inclusivity. However, the success of the recently developed approach relies on robust monitoring and evaluation to ensure the program is continuously relevant and appropriate.

TABLE 1 Literature sourced organized by theme. Some references fit into multiple categories due to their overlapping relevance.

Category	Reference	No. of papers
Understanding of the healthcare system, dermatology, and skin conditions	DermNet (2012), Eedy (2015), Mahendraraj et al. (2017), Chuchu et al., (2018), Goddu et al. (2018), Johnson et al. (2022), Al-Janabi et al. (2023), Cancer Council (2023), Heldreth et al. (2024), and Department of Health and Social Care, (2024)	8
Exploration of health inequality faced by POC	Hutchison et al. (2023), Kelly and Haidet (2007), Kenison et al. (2017), Stuart and Soulsby (2011), WHO (2018), Lester et al. (2019), Chauhan et al. (2020) and Raney et al. (2021)	8
Solutions to address skin diagnosis inequality faced by POC	Mukwende (2020), St. George's University (2020), Smith (2021), NHS (2021a), and NHS (2022)	5
Understanding of AI	Mitrani (2019) and Du-Harpur et al. (2020)	2
Current uses of AI in healthcare and skin diagnosis	Hakim (2023), Healthy.io (2024), Lacobucci, 2023, NHS (2021a), Obermeyer et al. (2019), Schakermann et al. (2024), Shore et al. (2019) and While (2023)	8
Understanding of AI in Skin diagnosis	Nasr-Esfahani et al. (2016), Aggarwal (2019), Brinker et al. (2019), Mitrani (2019), Khosla and Saini (2020), and Nahm (2022)	6
Understanding of data augmentation in AI	Perez et al. (2018), Aggarwal (2019), Chlap et al. (2021), Wen et al., 2022, Abayomi-Alli et al. (2021), Rezk et al. (2022) and Saeed et al. (2023)	7
Understanding of image selection in AI	Ribeiro et al. (2016), Koziarski and Cyganek, 2018, Aggarwal (2019), Brinker et al. (2019), Hogarty et al. (2019), Winkler et al., 2019 and Liopyris et al. (2022)	7
AI performance compared to Dermatologists	Brinker et al. (2019), Han et al. (2020), Philips et al. (2020)	3
AI performance with POC representation	Chen et al. (2016), Jinnai et al. (2020), Liu et al. (2020), and Liu and Primiero (2023)	4
Data augmentation to increase POC data	Aggarwal (2019) and Abhari and Ashok (2023)	2
Assessment of Skin Image Search	Zaar et al. (2020), Kamulegeya et al. (2023), and (2021)	3
Google AI development	Bui and Liu (2021) and Liu et al. (2020)	2

The clinical pathway within the UK and globally has shown that a choice of language matters when describing medical conditions (Chauhan et al., 2020; NHS, 2022). This can be for reasons including the reoccurrence of negative biases (Goddu et al., 2018; Raney et al., 2021), difficulty in understanding the choice of terminology (Kelly and Haidet, 2007; Kenison et al., 2017) or irrelevancy for minority groups through descriptions of medical conditions and images (NHS, 2022). The issue of language is evident within the NHS, particularly in the implementation of the comprehensive digital tool, Health A-Z (NHS, 2022). Health A-Z is designed to provide information on conditions, symptoms, and treatments for the public; however, at times, it fails to provide relevant symptom descriptions for all groups of people. When addressing skin conditions, the language used tends to focus on physical appearances and is often tailored to Caucasian skin types. While beneficial for some, it often leads to confusion among minority groups including POC or the visually impaired. Descriptions like "becoming pale" or "lips turning blue" may be relevant for Caucasians but may be challenging for minority groups to interpret (St. George's University, 2020). Smith (2021), a content designer for the NHS website, revealed patients want inclusive language such as "there are approximately ten spots that vary in size from about 1 mm to 1 cm, some spots are close together" to describe chickenpox which offers a neutral description that is independent from color reference. The implementation of a more neutral and objective language is underway, but the lack of medical sources detailing symptoms on Brown and Black skin poses a challenge to accurately describe how symptoms appear on diverse skin tones, slowing down the creation of inclusive material and the adoption of a neutral language.

Inadequate resources and knowledge for skin lesion diagnosis in POC is a persistent issue. Malone Mukwende, a medical student, developed Mind the Gap (Mukwende, 2020) after identifying a gap in the representation of POC in medical textbooks. Mind the Gap is a free online photographic repository with and without supporting text descriptions of various skin conditions with Fitzpatrick scale (FST) V and VI (DermNet, 2012). This tool is used worldwide in educational and professional settings (St. George's University, 2020) and relies on the public information sharing of skin conditions. The initiative addresses the representation gap and enhances global accessibility to a valuable resource, but the reliance on external contribution can stagnate the growth of the digital tool. There is also a risk of individuals self-misdiagnosing skin conditions if there is a lack of professional follow-up.

3.2 Artificial intelligence to reduce health inequality

AI describes the ability of machines to learn, communicate, reason, conduct different tasks simultaneously, or operate independently in different scenarios similarly to humans (Hogarty et al., 2019; Du-Harpur et al., 2020). Within the realm of AI, machine

learning can be supervised, semi-supervised or unsupervised (Hogarty et al., 2019), depending on the level of human intervention in correcting and directing the machine learning process.

Numerous instances of AI implementations within the clinical process have demonstrated promising outcomes in addressing health inequalities but have drawn attention to underlying issues. Examples of AI integration are Healthy.io and mobile applications such as Mindful Kidney (Healthy.io, 2024). The self-testing urine kit produces real-time clinical results through colorimetric analysis, computer vision, AI, and a smartphone camera that transforms into a clinical-grade medical device (NHS, 2021b; Healthy.io, 2024). This AI-powered digital technology reduces health inequality through accessibility to remote testing which may be challenging for some due to cost, transportation, or geographical locations. Findings show that patients favor the use of Healthy.io over taking a urine sample at their GP, possibly due to the comfort of their home and the ability to conduct the test at a convenient time (Shore et al., 2019). Considering user requirements can contribute to the success of AI integration; however ethical concerns have risen from a pilot study at a GP based in Oxford, where patient data were shared with a third party. This consequently led to the GP withdrawing from the study (Lacobucci, 2023) because the study became perceived as one with high risks for patients' privacy.

The Virtual AI Ward treating remote patients hosted by the NHS Croydon Primary Care Trust demonstrated the potential of AI. All users reported positive outcomes, especially regarding the ease of learning and understanding of the provided medical kits; the overall experience led to an improvement in participants' quality of life (Hakim, 2023). Success of the Virtual AI Ward was attributed to being run by community services with pathways to emergency treatment, when needed, upskilled staff, knowing when to choose continuous monitoring over spot monitoring, and having access to a cross-system multi-disciplinary team (Hakim, 2023; While, 2023). Challenges within the NHS including underfunding, understaffing, and overworked staff (Johnson et al., 2022; Al-Janabi et al., 2023), could adversely impact the success rates of implementing Virtual Wards across the NHS.

The US-based study by Obermeyer et al. (2019) explores the integration of AI into a medical system used within hospitals that raised ethical concerns. The AI program aims to predict complex health needs for the purpose of developing an intervention that manages those in need (Obermeyer et al., 2019). Patients are enrolled in the AI system through their insurance program if their risk score falls above the 97th percentile. The metadata gathered for the AI program includes demographic, insurance type, diagnoses, medications, and detailed costs, but specifically excludes race. Obermeyer et al. (2019) suggest that the algorithm's prediction on health needs is based on costing. Black and Caucasian patients have roughly the same costs per year, with Black patients generating an average of $1,801 less than Caucasians annually, despite having 26.3% more ongoing health issues. This suggests that the AI program failed to highlight health needs by predicting an equal level of risk for both groups. Identifying this, Obermeyer et al. (2019) adjusted the labels used within the algorithm, inevitably showing an increase in the percentage of additional help received by Black patients from 17.7 to 46.5%. This study is a distinct example of biases and ethical concerns that arise inversely through label choices, affecting predictive performance and creating racial biases, and exhibits why AI needs close monitoring.

3.3 Artificial intelligence in skin diagnosis

The integration of AI in dermatological settings has been investigated on multiple occasions and has proven to achieve the desired results in identifying skin conditions at varying levels (Nasr-Esfahani et al., 2016; Brinker et al., 2019). Considering the limited number of Dermatologists available, within the UK and globally (Eedy, 2015), it would benefit patients, GPs, and Dermatologists for AI to be successfully integrated into the clinical pathway. The current clinical pathway of checking the health of the skin and diagnosing possible conditions, within the UK, is shown in Figure 2. This flowchart has been adapted from the figure presented by Chuchu et al. (2018), illustrating the clinical pathway for skin lesions. The revised version incorporates the UK Government's guidelines on promoting the Pharmacy First Scheme (Department of Health and Social Care, 2024), which aims to alleviate the burden on GPs by encouraging patients to seek advice or treatment at a pharmacy as an initial step first, or they may choose to consult a GP directly. At the primary care level, skin concerns are categorized as melanoma, high risk, low risk, or benign. High-risk cases and melanoma are referred to Dermatologists, while low or benign cases are treated by GPs, and if no concern is confirmed, patients are discharged. AI integration can occur at various points in the clinical process (Points A, B, C, and D in Figure 2). An AI skin recognition tool at these decision points may assist in diagnosing skin concerns, collecting relevant images and descriptions, and expanding data sets that serve to improve future diagnostic accuracy. Implementing AI at these points could potentially alleviate the workload for primary care providers, whilst providing better outcomes for patients.

AI success consists of these factors:

- Sensitivity: This assesses a model's ability to predict true positive values of each available category (Mitrani, 2019).
- Specificity: This evaluates a model's ability to predict true negative values (Mitrani, 2019).
- Area under the receiver operating characteristic curve (AUROC): This is used to measure accuracy on classification tasks, the closer the receiver operating characteristic curve is to the upper left corner of the graph, the higher the accuracy of the test as the upper left corner is where the sensitivity = 1 and the false positive rate = 0 (specificity = 1).
- Receiver operating characteristic (ROC): This is used to evaluate the overall diagnostic performance of a test and to compare the performance of two or more tests (Nahm, 2022). The ideal ROC curve has an AUC = 1.0. However, when the coordinates of the x-axis (1 – specificity) and the y-axis correspond to 1: 1, a graph is drawn on the 45° diagonal (y = x) of the ROC curve (AUC = 0.5). An AUC greater than 0.5 is essential for any diagnostic technique to be meaningful, and it is often required to exceed 0.8 to be considered acceptable (Nahm, 2022).

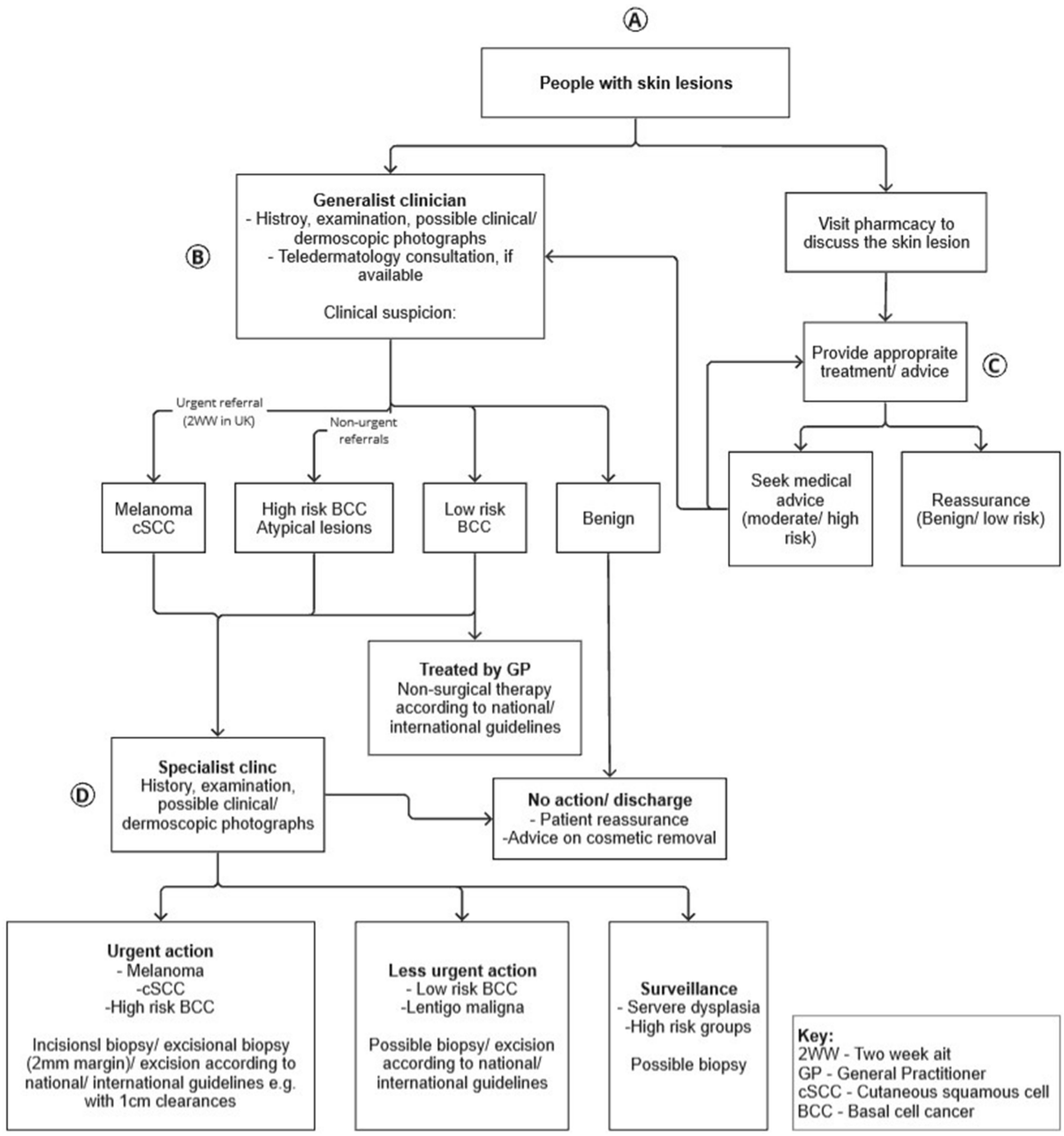

FIGURE 2
Clinical pathway of skin diagnosis within the UK (Adapted from Chuchu et al., 2018).

There are several factors to take into consideration during the development of AI for dermatological use and the impact they can have on its outcome. Overfitting is a significant challenge in supervised machine learning, where models exhibit high accuracy on training data but perform poorly on new data (Aggarwal, 2019). This can be problematic in skin lesion diagnostics due to the variability in data such as, size of skin lesions and variation in the angle images are taken (Aggarwal, 2019). To mitigate overfitting, steps such as data augmentation which help increase diversity and number of images, are taken (Khosla and Saini, 2020).

Data augmentation is the practice of artificially modifying images to account for a variability that exists in image taking (Aggarwal, 2019) and helps to expand training sets. This may be beneficial when limited images of skin conditions are available (Aggarwal, 2019; Chlap et al., 2021; Wen et al., 2022). Supervised machine learning typically relies on substantial amounts of training data to reduce the risk of overfitting; however obtaining well-annotated medical data is challenging, expensive and time-consuming, making data augmentation valuable in such situations. Chlap et al. (2021) categorize data augmentation into three main types:

- Basic augmentation (involving geometric transformations, cropping, occlusion, intensity operations, noise injection, filtering, and combinations)

- Deformable augmentation (utilising random displacement, spline interpolation, deformable image registration, and statistical shape models)
- Deep learning augmentation techniques (including Generative Adversarial Networks (GAN)-based augmentation methods).

Studies (Perez et al., 2018; Abayomi-Alli et al., 2021; Rezk et al., 2022; Saeed et al., 2023) highlight the positive impact of using data augmentation techniques to expand training sets on skin conditions and classification models, including increasing the number of images for POC, which is already very sparse. Although augmentation enhances data diversity, it introduces the risk of generating synthetic patterns that may not accurately represent real data, potentially affecting the model's performance.

Image selection is a fundamental aspect of AI development for skin diagnosis (Aggarwal, 2019; Brinker et al., 2019; Hogarty et al., 2019). Excluding inadequate low-quality images is essential to maintain a high level of sensitivity and specificity, consequently limiting the amount of usable training data. Low image quality refers to images affected by low resolution, presence of noise or small dynamic range where detail in an image may be lost due to dark or bright areas (Koziarski and Cyganek, 2018). Factors including hair, background skin issues, sun damage, rulers, blurry images, or dark corners of lenses contribute to poor image quality, causing confusion and miscalculation in results (Winkler et al., 2019; Liopyris et al., 2022). Ribeiro et al. (2016) conducted a study looking at AI distinguishing between photos of wolves and huskies. Results indicated that the AI predominantly relied on the entire image to differentiate between a wolve and huskie. Images which contained a light background or snow at the bottom were identified as wolves, if not they were identified as huskies, this is mainly due to images of wolves being taken in the snow. This is an example of overestimating the validity of AI models accuracy and would be problematic, especially for use within healthcare. In the application of AI to skin diagnosis, if a program is familiar with seeing melanoma on Caucasian skin, it may struggle considerably to identify the same on POC.

The study of Nasr-Esfahani et al. (2016) was one of the first to introduce AI into Dermatology; it was used to detect melanoma and benign cases using convolutional neural networks (CNN). CNN refers to a type of neural network where layers apply filters for specific features to areas within an image (Du-Harpur et al., 2020). The dataset for this study comprised of original images and augmented images subjected to cropping, scaled, and rotated and produced promising specificity and sensitivity results (Nasr-Esfahani et al., 2016). The success of the AI being able to distinguish melanoma from benign cases heavily relied on dataset illumination corrections which increased its ability to differentiate between the two conditions.

Brinker et al. (2019) investigated the performance of CNN-based classification of clinical images compared to dermatologists in sensitivity, specificity, and ROC. Dermatologists collectively achieved a mean sensitivity and specificity of 89% and 64%, respectively. In comparison, the CNN demonstrated a mean specificity of 68% and achieved the same sensitivity levels as the dermatologists (Brinker et al., 2019). Similar results are reported in a study by Han et al. (2020): clinicians' results indicated a sensitivity and specificity of 70% and 96%, respectively, while the CNN achieved 63% and 90%, respectively. Comparable outcomes are presented in Philips et al. (2020) with the AI program achieving 85% for both sensitivity and specificity and dermatologists achieving 87%, and 81%, respectively. The studies highlight promising AI performance and show good prospects of AI integration within dermatological workflows for skin diagnostics. Despite this, each study's drawback consists of the underrepresentation of POC in its dataset affecting the generalisability of results.

There is a growing body of literature that acknowledges the gravity of POC underrepresentation in AI training datasets. Jinnai et al. (2020) used images of only Black and Brown pigmented skin lesions on a faster region-based convolutional neural network (FRCNN) program. This produced a specificity and sensitivity of 94% and 83%, while board certified Dermatologists produced results of 86% for both sensitivity and specificity (Jinnai et al., 2020). Similar results are seen in Chen et al. (2016) study using images of different ethnicities to assess AI performance in identifying melanoma; sensitivity, and specificity results of 90% and 91% were reported. Liu et al. (2020) study for Google Health produced results of 'top-1 accuracy' of 71% and 'top-1 sensitivity' of 58% when diagnosing a range of contrasting skin conditions across different skin tones varying from FST I – V. Furthermore, Liu and Primiero (2023) systematic review presented evidence of accurate AI programs for POC within multiple studies showing accuracy levels from 70% to almost 100%.

Despite the observed high levels of accuracy reported in these studies, a comprehensive analysis of the dataset used shows little to no representation of POC data. Liu et al. (2020) study had 2.7% of participants with FST V and 0% of participants with FST VI. Chen et al. (2016) study had a range of ethnic participants but were not in a balanced ratio to Caucasian participants (American Indian or Alaska Native 2%, Asian or Pacific Islander 13.9%, Black or African American 4.3%, White, or Caucasian 30%). Jinnai et al. (2020) study did not provide a breakdown in the number of Brown and Black participants from each FST group, which is key as a limited number of FST VI and a higher number of IV will affect its validity. Additionally, Liu and Primiero (2023) systematic review predominantly consisted of papers with participants of East Asian origin with some studies containing only 10% of participants with FST type IV–VI. Schakermann et al. (2024) study developed the Health Equity Assessment of machine Learning (HEAL) framework to assess the performance of health AI in a case study. While Schakermann et al. (2024) case was carefully sampled to create a balance in demographics, there was still a poor representation of FST V-VI and American Indian/Alaska Natives. These studies' results are skewed due to poor representation of POC affecting the results generalisability or show the struggle in trying to work with balanced data sets due to limited resources.

Aggarwal (2019) study proves AIs ability to correctly diagnose melanoma through CNN programs. Augmentation of data was carried out by artificially darkening light skin toned images to input into the program. Results produced higher sensitivity (0.82) and specificity (0.76) rates for darker skin images compared to lighter skin tones (0.63 and 0.60). However, the 'darkening' of the images was only able to create data belonging to FST II, still excluding FST III – VI groups. This is a result of wanting to preserve the characteristics of the skin lesion on the original light skin toned images. Despite the potential misinterpretation of the study, it still

shows the capability of AI accuracy in melanoma diagnosis when training with minimal inclusive data sets. Similarly, Abhari and Ashok (2023) investigation used data augmentation techniques to increase the POC data set to improve the studies accuracy. However, the study generalized darker skin tones and failed to present information on skin tone categories (such as FST), making it difficult to comprehend the breadth of skin tones explored.

AI powered digital tools for skin diagnosis's have been made publicly accessible. Skin Image Search, developed by First Derm, was established to increase the availability of expert skin information. The application works by uploading two pictures of a skin lesion (an overview and close-up) to produce a diagnosis. The app has been used globally, in countries such as Sweden, Chile, China, Australia, and Ghana. Zaar et al. (2020) assessed the diagnostic accuracy of Skin Image Search developing interesting insights. The dataset consisted of all skin phototypes but low levels of FST type IV (4.2%), V (0.9%) and VI (1.4%) (type I 16.7%, II 59.5%, III 17.2%) were included. Evaluation results also indicated high and low levels of accuracy across varying skin conditions; and a top-5 accuracy rate of 56.4, and 22.8% accuracy for the most probable diagnosis. The poor accuracy rates, with a high FST I, II, and III and low FST IV, V, and VI test images, suggest that the program needs further refinement and development. Kamulegeya et al. (2023) tested Skin Image Search's diagnostic performance using predominantly FST VI images extracted from The Medical Concierge Group in Uganda. Data sets were anonymised and filtered to ensure a quality dataset was used. Skin Image Search was able to correctly diagnose 17% of images compared to the 69.9% performance reported from the AI training results. The subpar results could indicate that First Derm was heavily trained on images with FST I and II. FirstDerm has stated in a blog that Skin Image Search has an accuracy rate of 80% (Börve, 2021) with no supporting data for the claim. Such disinformation can increase the problems already caused by the underrepresentation of POC by creating a false sense of security among those who take information at face value, further increasing the health inequality gap.

Some AI tools are under development for skin diagnostics. Google has recognized that consumers conduct 10 billion searches annually related to skin, nail, and hair conditions and is now developing Derm Assis (Bui and Liu, 2021). This program operates by users capturing three images of the skin condition, answering questions about their skin type and the duration of symptoms, and then presenting possible diagnosis to the users. Google emphasises that this tool serves as an ancillary support, providing users with information before deciding on their next steps. Google's Health study for the development of the deep learning system, revealed a top differential diagnosis in validation with an acceptable accuracy and sensitivity rate when given the option to provide one diagnosis (Liu et al., 2020). When given the chance to provide three diagnoses, accuracy and sensitivity levels were significantly better across all 26 skin conditions (Liu et al., 2020). While there are promising results, Google's identification of consumer need with the current response of a dermatological level tool, fails in its generalization ability. This is a consequence of using a dataset that is not representative of all ethnic groups; groups with skin tones in categories FST V were represented by 2.7% of participants and 0% for FST VI. This action formulates potential misdiagnoses and biases, especially among ethnic groups.

4 Discussion

Health inequalities have been tackled in multiple ways through strategies and digital technological approaches. The NHS Core20PLUS5 strategy presents a targeted approach to reducing health inequalities with a focus on specific communities and groups. The future success of this strategy could also serve as a foundation for tackling inequalities in health globally, considering the impact on population composition that economic and political migration are generating. Other approaches including Healthy.io, NHS Croydon Primary Care Trust Virtual AI Ward, and the USA medical system present the case of successful AI capabilities in addressing health inequalities through ease and appropriate access to medical care, treatment, and results with the condition that it is supervised correctly suggesting that unsupervised AI would not be appropriate, and possibly detrimental, in medical settings.

Achieving success in tackling health inequalities through AI usage in complex areas such as dermatological settings is possible. However, for such success to occur some foundational issues must be resolved first to create the conditions for an effective and rigorous application of AI. The NHS (2022) approach to the expansion of the Health A-Z free public website and Malone Mukwandes' Mind the Gap initiative (https://www.blackandbrownskin.co.uk/mindthegap) emphasise the limited representation of POC in current data sets, and the possibilities of false-positive reassurance in self-diagnoses when primary care follow ups are not carried out. The inadequate representation of skin tones is commonly seen within research and educational settings as a reoccurring issue (Lester et al., 2019). This is a barrier faced by many researchers and has consistently been a failure in AI development, despite the attempts made through data augmentation. Whilst data augmentation creates the potential to expand the dataset of POC through various techniques, it creates the possibility of generating synthetic patterns that are unrepresentative of the real population. This could be detrimental not only to a particular study's reliability, but generally to public trust in AI usage in healthcare.

Within dermatology, it is evident that the capability of AI to match or surpass dermatologists' performance is achievable. Addressing challenges such as overfitting and implementing effective data augmentation is important for the development and accuracy of AI in the diagnosis of skin lesions. Ensuring diversity in image datasets is equally crucial to prevent biases, as highlighted by multiple studies that demonstrated poor diagnostic performance when AI was predominantly trained on lighter skin tones. Some studies claim to include POC in training datasets or in the testing of AI programs, suggesting insightful findings; however, looking specifically at the number of POC data used, it is clear that statistical representation has yet to be achieved. Not only are more patients of color needed within studies, but transparency and clarity from researchers on participant skin tones need to be shared to avoid misleading interpretations. The consistent use of the FST scale throughout clinical studies could be considered a contributing factor to the lack of POC representation. The scale is currently inclusive of non-marginalized and ethnoracial minorities alike (Heldreth et al., 2024), compressing under type IV-VI a plethora of diverse skin tones that are therefore unfairly represented in the scale. This creates poor dermatological learning resources and, consequently, AI studies in dermatology.

Before AI can be used, within clinical studies, for skin diagnostic purposes several interventions need to take place to reduce biases and to show the potential and reliability of AI. This can be achieved in many ways, including:

- An increased database of expert confirmed diagnoses across a variety of skin tones.
- Targeted campaigns for hard-to-reach groups. This will result in higher participation of POC in clinical studies.
- An improvement in learning resources providing accurate and diverse clinical representation of POC through detailed supportive text and images.
- Continuous professional development (CPD) for GPs to create a better understanding of unintentional biases and awareness of skin lesions among POC.
- An appropriate skin color categorization technique, which encapsulates different skin color variations, and can also be used within clinical and educational settings.

Without these interventions in place, the systemic issue of the under representation of POC in AI cannot be solved and will only continue to amplify the disparities and exclusion POC face.

The limitation of this study includes the lack of full details in the reviewed literature about skin tones used for training data, making it difficult to understand if the findings are generalisable. Additionally, it is unclear whether the literature on AI being reviewed used the same AI programming system. For instance, Brinker et al. (2019) and Han et al. (2020) highlight, in their methodology, the use of CNN, while Jinnai et al. (2020) study uses FRCNN, but Abhari and Ashok (2023), Liu et al. (2020) and Philips et al. (2020) AI programming systems are not clarified. The lack of clear parameters can make it harder to compare the performance of different AI approaches. A clinical validation of the findings highlighted in this review could have also been beneficial.

5 Conclusion

Evidence demonstrates a notable disadvantage for POC in various aspects of healthcare. This is seen for skin diagnostics within clinical studies at both primary and secondary care levels. These situations result in lower survival rates, poorer quality of life for POC in comparison to Caucasians, and a disproportionate underrepresentation of POC in medical advancements.

Digital technologies, including the integration of AI, in dermatology have shown promise within healthcare, particularly in addressing the scarcity of dermatologists globally and in providing accurate diagnoses of skin conditions when executed efficiently, as shown through the NHS Croydon Primary Care Trust Virtual AI ward (Hakim, 2023). However, challenges have unexpectedly emerged in AI development that require attention and upstream interventions to improve the lack of diverse representation impacting the reliability and generalisability of AI models. This has also inadvertently highlighted ongoing issues faced by POC within healthcare, such as unintentional biases made by healthcare professionals or incorrect diagnoses of skin conditions. While interesting techniques, such as data augmentation, show potential in overcoming problems, such as the number of limited imagery available on POC, they do not address the unintentional biases shown within healthcare and show the need for more care to be placed in ensuring POC are being cared for at the same pace and level as Caucasians.

To ensure technology advancements continue and to prevent the widening of pre-existing racial disparities, the inclusion of POC in studies needs to be a priority and can be achieved through targeted campaigns to include hard to reach participants. A more effective approach to categorising POC to ensure a comprehensive representation of skin tones is also needed. The current use of the FST scale to represent POC fails to encompass the full diversity of human skin tones. Relevant participant data, such as ethnicity and skin tone, also needs to be transparently shared within clinical studies for a clearer understanding on whether studies are truly generalisable.

Digital tools including Healthy.io and the NHS Croydon Primary Care Trust Virtual AI Ward are successful in their execution, which could be due to the user-centred approach applied. Many studies have taken a technical approach to address skin diagnosis among POC through AI. Comparatively fewer studies have adopted a user-centred approach throughout their development process. Whilst AI-augmented skin diagnosis is technically promising, caution, additional research, measures and regulations are needed. The fundamental issue of the lack of balanced data set representation of all skin types and transparency in research is a gap that needs addressing for both traditional clinical diagnosis and AI-assisted diagnostic pathways.

Author contributions

NK: Writing – original draft, Writing – review & editing. GS: Supervision, Writing – review & editing. FC: Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. Brunel University London funded the publication of this study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Abayomi-Alli, O. O., Damaševičius, R., Misra, S., Maskeliūnas, R., Abayomi-Alli, A., Damasevicius, R., et al. (2021). Malignant skin melanoma detection using image augmentation by oversampling in nonlinear lower-dimensional embedding by oversampling in nonlinear lower-dimensional embedding manifold. *Turk. J. Electr. Eng. Comput. Sci.* 29, 2600–2614. doi: 10.3906/elk-2101-133

Abhari, J., and Ashok, A. (2023). *Mitigating racial biases for machine learning based skin Cancer detection', MobiHoc '23: Proceedings of the twenty-fourth international symposium on theory, algorithmic foundations, and protocol Design for Mobile Networks and Mobile Computing*. New York, NY: MobiHoc. 556–561.

Aggarwal, P. (2019). Data augmentation in dermatology image recognition using machine learning. *Skin Res. Technol.* 25, 815–820. doi: 10.1111/srt.12726

Al-Janabi, H., Williams, I., and Powell, M. (2023). Is the NHS underfunded? Three approaches to answering the question. *Royal Soci. Med. J.* 116, 409–412. doi: 10.1177/01410768231214340

Brinker, T. J., Hekler, A., Enk, A. H., Klode, J., Hauschild, A., Berking, C., et al. (2019). A convolutional neural network trained with Dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. *Eur. J. Cancer* 111, 148–154. doi: 10.1016/j.ejca.2019.02.005

Börve, A. (2021). Artificial intelligence and the future of skin Care. First Derm, 20 March. Available at: https://www.firstderm.com/artificial-intelligence-and-the-future-of-skin-care/#:~:text=In%20many%20cases%20we%20are,in%20detecting%20a%20skin%20disease

Bui, P., and Liu, Y. (2021). Using AI to help find answers to common skin conditions. Google Health. Available at: https://blog.google/technology/health/ai-dermatology-preview-io-2021/ (Accessed June 12, 2023).

Cancer Council (2023). Types of Cancer melanoma. Cancer council. Available at: https://www.cancer.org.au/cancer-information/types-of-cancer/melanoma (Accessed June 15, 2023).

Chuchu, N., Takwoingi, Y., Dinnes, J., Matin, R. N., Bassett, O., Moreau, J. F., et al. (2018). "Smartphone applications for triaging adults with skin lesions that are suspicious for melanoma (Review)," in *Research Gate.* 4, 1–68. doi: 10.1002/14651858.CD013192

Chauhan, A., Walton, M., Manias, E., Walpola, R. L., Seale, H., Latanik, M., et al. (2020). The safety of health care for ethnic minority patients: A systematic review. *Int. J. Equity Health* 19, 1–25. doi: 10.1186/s12939-020-01223-2

Chen, R., Snorrason, M., Enger, S. M., Mostafa, E., Ko, J. M., Aoki, V., et al. (2016). Validation of a skin-lesion image-matching algorithm based on computer vision technology. *Telemed. e-Health* 22, 45–50. doi: 10.1089/tmj.2014.0249H., Snorrason, M., Enger, S, M., Mostafa, E., Ko, J, M., Aoki, V., Bowling, J

Chlap, P., Min, H., Vandenberg, N., Dowling, J., Holloway, L., and Haworth, A. (2021). A review of medical image data augmentation techniques for deep learning applications. *J. Med. Imaging Radiat. Oncol.* 65, 545–563. doi: 10.1111/1754-9485.13261

Department of Health and Social Care (2024). Pharmacy First: what you need to know. Available at: https://healthmedia.blog.gov.uk/2024/02/01/pharmacy-first-what-you-need-to-know/ (Accessed February 15, 2024).

DermNet (2012). Fitzpatrick skin Phototype. Derm Net. Available at: https://dermnetnz.org/topics/skin-phototype (Accessed September 16, 2023)

du-Harpur, X., Watt, F., Luscombe, N. M., and Lynch, M. D. (2020). What is AI? Applications of artificial intelligence to dermatology. *Br. J. Dermatol.* 183, 423–430. doi: 10.1111/bjd.18880M., Luscombe, N, M., Lynch, M, D

Eedy, D. (2015). Dermatology: a specialty in crisis. *Clin. Med.* 15, 509–510. doi: 10.7861/clinmedicine.15-6-509

Goddu, A. P., O'Conor, K. J., Lanzkron, S., Saheed, M. O., Saha, S., Peek, M. E., et al. (2018). Do words matter? Stigmatizing language and the transmission of Bias in the medical record. *J. Gen. Intern. Med.* 33, 685–691. doi: 10.1007/s11606-017-4289-2P., O'Conor, K, J., Lanzkron, S., Saheed, M, O., Saha, S., Peek, M, E., Haywood, C., Beach, M, C

Hakim, R (2023). *Realising the potential of virtual wards*. NHS Confediration. Available at: https://www.nhsconfed.org/publications/realising-potential-virtual-wards (Accessed October 18, 2023)

Han, S., Moon, I. J., Kim, S. H., Na, J. I., Kim, M. S., Park, G. H., et al. (2020). Assessment of deep neural networks for the diagnosis of benign and malignant skin neoplasms in comparison with dermatologists: A retrospective validation study. *PLoS Med.* 17:e1003381. doi: 10.1371/journal.pmed.1003381

Healthy.io. (2024). Improve adherence with home kidney testing. Healthy.io. Available at: https://healthy.io/eu/services/kidney/ (Accessed May 04, 2023).

Heldreth, C., Monk, E. P., Clark, A. T., Schumann, C., Eyee, X., and Ricco, S. (2024). Which skin tone measures are the Most inclusive? An investigation of skin tone measures for artificial intelligence. *ACM J. Responsib. Comput.* 1, 1–21. doi: 10.1145/3632120M., Monk, E, P., Clark, A, T., Schumann, C., Eyee, X., Ricco, S.

Hogarty, D. T., Su, J. C., Phan, K., Attia, M., Hossny, M., Nahavandi, S., et al. (2019). Artificial intelligence in dermatology-where we are and the way to the future: A review. *Am. J. Clin. Dermatol.* 21, 41–47. doi: 10.1007/s40257-019-00462-6

Hutchison, E., Yoseph, R., and Wainman, H. (2023). Skin of colour: essential for the non-dermatologist. *Clin. Med. J.* 23, 2–8. doi: 10.7861/clinmed.2022-0335

Jinnai, S., Yamazaki, N., Hirano, Y., Sugawara, Y., Ohe, Y., and Hamamoto, R. (2020). The development of a skin Cancer classification system for pigmented skin lesions using deep learning. *MDPI* 10, 1–13. doi: 10.3390/biom10081123

Johnson, A., Conroy, S., Thompson, D., Hassett, G., Clayton, A., and Backhouse, E. (2022). Staff experience in the NHS: A National Study—an Experience-Based Design Approach. *J. Pat. Exp.* 9:237437352211439. doi: 10.1177/23743735221143921

Kamulegeya, L., Bwanika, J., Okello, M., Rusoke, D., Nassiwa, F., Lubega, W., et al. (2023). Using artificial intelligence on dermatology conditions in Uganda: a case for diversity in training data sets for machine learning. *Science* 23, 753–763. doi: 10.4314/ahs.v23i2.86

Kelly, P., and Haidet, P. (2007). Physician overestimation of patient literacy: A potential source of health care disparities. *Patient Educ. Couns.* 66, 119–122. doi: 10.1016/j.pec.2006.10.007

Kenison, T., Madu, A., Krupat, E., Ticona, L., Vargas, I. M., and Green, A. R. (2017). Through the veil of language: exploring the hidden curriculum for the Care of Patients with limited English proficiency. *J. Assoc. Am. Med. Coll.* 92, 92–100. doi: 10.1097/ACM.0000000000001211

Khosla, C., and Saini, B. S. (2020). "Enhancing performance of deep learning models with different data augmentation techniques: A survey" in *2020 international conference on intelligent engineering and management (ICIEM)*, vol. *2020* (London, UK: IEE Xplore), 79–85.

Koziarski, M., and Cyganek, B. (2018). Impact of low resolution on image recognition with deep neural networks: an experimental study. *Int. J. Appl. Math. Comput. Sci.* 28, 735–744. doi: 10.2478/amcs-2018-0056

Lacobucci, G. (2023). Data privacy: GP surgery withdraws from kidney screening pilot after patients voice concerns. *Br. Med. J.* 380:157. doi: 10.1136/bmj.p157

Lester, J. C., Taylor, S. C., and Chren, M.-M. (2019). Under-representation of skin of colour in dermatology images: not just an educational issue. *Br. J. Dermatol.* 180, 1521–1522. doi: 10.1111/bjd.17608

Liopyris, K., Gregorious, S., Dias, J., and Stratigos, A. J. (2022). Artificial intelligence in dermatology: challenges and perspectives. *Dermatol. Ther.* 12, 2637–2651. doi: 10.1007/s13555-022-00833-8

Liu, Y., Jain, A., Eng, C., Way, D., Lee, K., Bui, P., et al. (2020). A deep learning system for differential diagnosis of skin diseases. *Nat. Med.* 26, 900–908. doi: 10.1038/s41591-020-0842-3

Liu, Y., and Primiero, C. (2023). Artificial intelligence for the classification of pigmented skin lesions in populations with skin of color: a systematic review. *Dermatology* 239, 499–513. doi: 10.1159/000530225A., Kulkarni, V., Soyer, H, P., Stablein, B, B

Mahendraraj, K., Sidu, K., Lau, C. S. M., McRoy, G. J., Chamberlain, R. S., and Smith, F. O. (2017). Malignant melanoma in African-Americans. *Medicine* 96:e6258. doi: 10.1097/MD.0000000000006258

Mitrani, A, (2019). Evaluating categorical models II: Sensitivity and specificity, medium, 6 December. Towards data science. Available at: https://towardsdatascience.com/evaluating-categorical-models-ii-sensitivity-and-specificity-e181e573cff8 (Accessed September 25, 2023).

Mukwende, Malone (2020) *Black and Brown skin*. Available at: https://www.blackandbrownskin.co.uk/ (Accessed July 04, 2023).

Nahm, F. S. (2022). Receiver operating characteristic curve: overview and practical use for clinicians. *Kor. J. Anaesthesiol.* 75, 25–36. doi: 10.4097/kja.21209

Nasr-Esfahani, E., Samavi, S., Karimi, N., Soroushmehr, SMR, Jafari, MH., Ward, K., et al., (2016). 'Melanoma detection by analysis of clinical images using convolutional neural Network', 016 38th annual international conference of the IEEE engineering in medicine and biology society (EMBC), Orlando, FL, USA, 2016, 1373–1376.

NHS (2021a) Core20PLUS5 an approach to reducing health inequalities: Supporting information. Available at: https://www.england.nhs.uk/publication/core20plus5-an-approach-to-reducing-health-inequalities-supporting-information/ (Accessed September 18, 2023).

NHS (2021b) Healthy.io: Smartphone albuminuria urine self-testing. Available at: https://transform.england.nhs.uk/ai-lab/explore-all-resources/understand-ai/healthyio-smartphone-albuminuria-urine-self-testing/ (Accessed May 04, 2023).

NHS (2022) Inclusive content skin symptoms. Available at: https://service-manual.nhs.uk/content/inclusive-content/skin-symptoms (Accessed October 15, 2023).

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainthan, S. (2019). Dissecting racial Bias in an algorithm used to manage the health of populations. *Science* 366, 447–453. doi: 10.1126/science.aax2342

Perez, F., Vasconcelos, C., Avila, S., and Valle, E. (2018) *'Data augmentation for skin lesion Analysis', context-aware operating theaters, computer assisted robotic endoscopy, clinical image-based procedures, and skin image analysis*. Springer Link. Available at: https://link.springer.com/chapter/10.1007/978-3-030-01201-4_33#chapter-info

Philips, M., Greenhalgh, J., Marsden, H., and Palamaras, I. (2020). Detection of malignant melanoma using artificial intelligence: an observational study of diagnostic accuracy. *Dermatology Practical and Conceptual* 10:e2020011. doi: 10.5826/dpc.1001a11

Raney, J., Pal, R., Lee, T., and Saenz, S. (2021). Words matter: an Antibias workshop for health care professionals to reduce stigmatizing language. *J. Teach. Learn. Res.* 17, 1–6. doi: 10.15766/mep_2374-8265.11115

Rezk, E., Eltorki, M., and El-Dakhankhni, W. (2022). 'Improving skin color diversity in Cancer detection: deep learning Approach', JMIR. *Dermatology* 5, 1–14. doi: 10.2196/39143

Ribeiro, E. T., Singh, S., and Guestrin, C. (2016). Why should I trust you?' expelling the predictions of any classifier. *Arxiv* 3, 1135–1144. doi: 10.48550/arXiv.1602.04938

Saeed, M., Asma, N., Masood, H., Rehman, S. U., and Gruhn, A. V. (2023). The power of generative AI to augment for enhanced skin Cancer classification: A deep learning approach. *IEEE Access* 11, 130330–130344. doi: 10.1109/ACCESS.2023.3332628

Schakermann, M., Spitz, T., Pyles, M., Lewis, H., Wulczyn, E., Pfohl, S. R., et al. (2024). Health equity assessment of machine learning performance (HEAL): A framework and dermatology AI model case study. *eClinicalMedicine* 70:102479. doi: 10.1016/j.eclinm.2024.102479

Shore, J., Green, M., Hardy, A., and Livesey, D. (2019). The compliance and cost-effectiveness of smartphone urinalysis albumin screening for people with diabetes in England. *Expert Rev. Pharmacoecon. Outcomes Res.* 20, 387–395. doi: 10.1080/14737167.2019.1650024

Smith, R. (2021). Making content about skin symptoms more inclusive', NHS England Digital, 30th July. NHS. Available at: https://digital.nhs.uk/blog/design-matters/2021/making-content-about-skin-symptoms-more-inclusive (Accessed October 15, 2023).

St. George's University (2020) *Mind the gap: A handbook of clinical signs on black and Brown skin*. London: St. George's University London. Available at: https://www.sgul.ac.uk/news/mind-the-gap-a-handbook-of-clinical-signs-on-black-and-brown-skin-on#:~:text=I%20know%20the%20handbook%20is,handbook%20on%20the%20nurses'%20desk. (Accessed at: July 04, 2023).

Stuart, K., and Soulsby, E. J. L. (2011). Reducing Global Health inequalities. Part 1. *J. R. Soc. Med.* 104, 321–326. doi: 10.1258/jrsm.2011.100396

Wen, D., Khan, S. M., Xu, A. J., Ibrahim, H., Smith, L., Caballero, J., et al. (2022). Characteristics of Publicly Available Skin Cancer Image Datasets: A Systematic Review. *Lancet Digital Health.* 4, 64–74. doi: 10.1016/S2589-7500(21)00252-1

While, A. (2023). Digital health and technologies. *Br. J. Community Nurs.* 28, 120–126. doi: 10.12968/bjcn.2023.28.3.120

WHO (2018) *Health inequalities and their causes*. WHO. Available at: https://www.who.int/news-room/facts-in-pictures/detail/health-inequities-and-their-causes (Accessed August 18, 2023)

Winkler, J. K., Fink, C., and Toberer, F. (2019). Association between surgical skin markings in Dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA Dermatol.* 155, 1135–1141. doi: 10.1001/jamadermatol.2019.1735

Zaar, O., Larson, A., Polesie, S., Saleh, K., Tarstedt, M., Olives, A., et al. (2020). Evaluation of the diagnostic accuracy of an online artificial intelligence application for skin disease diagnosis. *Acta Dermato Venereol.* 100:260. doi: 10.2340/00015555-3624

frontiers | Frontiers in Big Data

TYPE Original Research
PUBLISHED 17 June 2024
DOI 10.3389/fdata.2024.1359906

OPEN ACCESS

EDITED BY
Rui Qin,
Manchester Metropolitan University,
United Kingdom

REVIEWED BY
Iroju Olaronke,
Adeyemi College of Education, Nigeria
Sakib Jalil,
James Cook University, Australia

*CORRESPONDENCE
Grace Ataguba
✉ grace.ataguba@dal.ca

RECEIVED 22 December 2023
ACCEPTED 10 May 2024
PUBLISHED 17 June 2024

CITATION
Ataguba G and Orji R (2024) Toward the design of persuasive systems for a healthy workplace: a real-time posture detection. *Front. Big Data* 7:1359906. doi: 10.3389/fdata.2024.1359906

Toward the design of persuasive systems for a healthy workplace: a real-time posture detection

Grace Ataguba* and Rita Orji

Department of Computer Science, Dalhousie University, Halifax, NS, Canada

Persuasive technologies, in connection with human factor engineering requirements for healthy workplaces, have played a significant role in ensuring a change in human behavior. Healthy workplaces suggest different best practices applicable to body posture, proximity to the computer system, movement, lighting conditions, computer system layout, and other significant psychological and cognitive aspects. Most importantly, body posture suggests how users should sit or stand in workplaces in line with best and healthy practices. In this study, we developed two study phases (pilot and main) using two deep learning models: convolutional neural networks (CNN) and Yolo-V3. To train the two models, we collected posture datasets from creative common license YouTube videos and Kaggle. We classified the dataset into comfortable and uncomfortable postures. Results show that our YOLO-V3 model outperformed CNN model with a mean average precision of 92%. Based on this finding, we recommend that YOLO-V3 model be integrated in the design of persuasive technologies for a healthy workplace. Additionally, we provide future implications for integrating proximity detection taking into consideration the ideal number of centimeters users should maintain in a healthy workplace.

KEYWORDS

persuasive technology, healthy workplace, posture, machine learning, YOLO-V3, convolutional neural networks

1 Introduction

The importance of persuasive technologies in influencing changes in human behavior is significant and cannot be overemphasized. Persuasive technologies have an impact on users' behavior and the choices they make (Rapoport, 2017; Orji et al., 2018; Darioshi and Lahav, 2021; Wang et al., 2023). As a result, persuasive technologies prioritize user-centered design, and they can assist users in leading a healthy lifestyle. Considering this, research has demonstrated the valuable roles these technologies play in preventing and aiding the management of illnesses (Schnall et al., 2015; Karppinen et al., 2016; Sonntag, 2016; Bartlett et al., 2017; Faddoul and Chatterjee, 2019; Fukuoka et al., 2019; Kim M. T. et al., 2019; Oyibo and Morita, 2021), promoting fitness and exercise (Bartlett et al., 2017; Schooley et al., 2021), and other significant ones (Jafarinaimi et al., 2005; Anagnostopoulou et al., 2019; Beheshtian et al., 2020).

The workplace, a location, setting, or environment where people engage in work, have recorded significant unhealthy practices, including bad posture, over the years (Nanthavanij et al., 2008; Ko Ko et al., 2020; Roy, 2020; van de Wijdeven et al., 2023). In the context of this study, we consider work-from-home (WFH) contexts, offices, and other spaces where computers are employed to be workplaces. Best workplace practices are significant for a healthy working style. These practices cover the need to ensure that computer users maintain the right posture, follow the right movement practices, take

regular breaks from computer systems, ensure they have proper lighting conditions, adhere to computer system layout, and other significant psychological and cognitive aspects. Poor workplace practices can lead to various health issues, such as repetitive strain injuries, eyestrain, and postural problems (Ofori-Manteaw et al., 2015; Workineh and Yamaura, 2016; Alaydrus and Nusraningrum, 2019). Research has shown that over 70% of stress, neck injuries, other types of sprains and pains (for example, arm sprains and back pain), and stress are work-related (Tang, 2022). This study presents the design of a persuasive system based on the best posture practices. In addition, this study presents implications for designing persuasive systems based on their proximity to computer system requirements.

Machine learning, a subfield of artificial intelligence (AI), deals with developing models. These models assist computers in learning and detecting patterns of objects in the real world (Mahesh, 2020; Sarker, 2021). Hence, machine learning has contributed to several studies that have significantly detected patterns in human behaviors (Cheng et al., 2017; Krishna et al., 2018; Xu et al., 2019; Chandra et al., 2021; Jupalle et al., 2022; Cob-Parro et al., 2023), human emotions (Jaiswal and Nandi, 2020; Gill and Singh, 2021), and health-related behaviors (Reddy et al., 2018; Mujumdar and Vaidehi, 2019; Ahmad et al., 2021). In this study, we leverage the opportunity of machine learning algorithms to design a persuasive system for detecting patterns of unhealthy postures and proximity to computers in workplaces.

As part of persuasive technology's goal to provide users with real-time feedback on their actions (which, in turn, influences their behavior), we report on our experiment comparing the convolutional neural networks (CNN) and Yolo-V3 models. Research has shown the success of these models in real-time object detection (Tan et al., 2021; Alsanad et al., 2022). One of the significant drawbacks of CNN compared with Yolo-V3 from research is its requirement for a large number of training sets (Han et al., 2018). On the other hand, the Yolo-V3 model generates regions or boxes around objects and returns its accuracy values within these boxes. This implies that several boxes are marked within an object, and its performance can be implied from the confidence of predictions (Figure 1). For example, in Figure 1, the YOLO-V3 model predicted the hardhat with 95% confidence. Yolo-V3 and CNN work in real time by analyzing images extracted from frames per second and providing a consistent update as these images change.

Though we found significant studies in the application of persuasive systems to encourage computer users to take regular breaks from workplaces (Jafarinaimi et al., 2005; Reeder et al., 2010; Ludden and Meekhof, 2016; Ren et al., 2019), little is yet known about how they maintain the right posture before these regular breaks. Based on this limitation, the overarching goal of our study is to explore how people can be conscious of their unhealthy posture practices in workplaces (while sitting or standing). This connects with the main research question we seek to answer (RQ): RQ: Can we design persuasive computers to detect unhealthy posture practices (such as sitting and standing) in workplaces?

People in workplaces have two types of posture positions: sitting and standing (Botter et al., 2016). The sitting position affords the computer user space to relax the back correctly on a chair (Figure 2, L). This, compared with the standing position,

FIGURE 1
A YOLO-V3 detection on a sample image. Reproduced from "YOLOv3 on custom dataset," YouTube, uploaded by "Aman Jain," 22 July 2021, https://www.youtube.com/watch?v=D4RQ7Rkrass, Permissions: YouTube Terms of Service.

allows computer users to stand while using the computer system (Figure 3). It is significant to recall that before COVID-19, these workplaces were office spaces. However, most recently, after COVID-19, workplaces have extended to home spaces (Abdullah et al., 2020; Javad Koohsari et al., 2021). People now work from home, and the posture practices in these spaces have not been evaluated.

The scientific contributions of this research are in 4-folds:

1. Provision of ground truth posture datasets:

We are contributing ground-truth posture datasets for the research community to explore related concepts in the future. These datasets can be increased in future work to enhance the accuracy and effectiveness of future technological interventions. Hence, this contribution will support researchers and designers in developing more robust and context-aware persuasive technologies.

2. Implementation of deep learning models for posture detection:

We present the development and implementation of deep learning models for detecting the posture practices of computer users. These models leverage advanced techniques to interpret and classify diverse body positions, contributing to the evolving landscape of human–computer interaction. The models offer a technological solution to the challenge of real-time posture detection in the workplace. This contribution aligns with the forefront of research in machine learning and computer vision.

3. Real-time persuasive design for healthy workplace behavior:

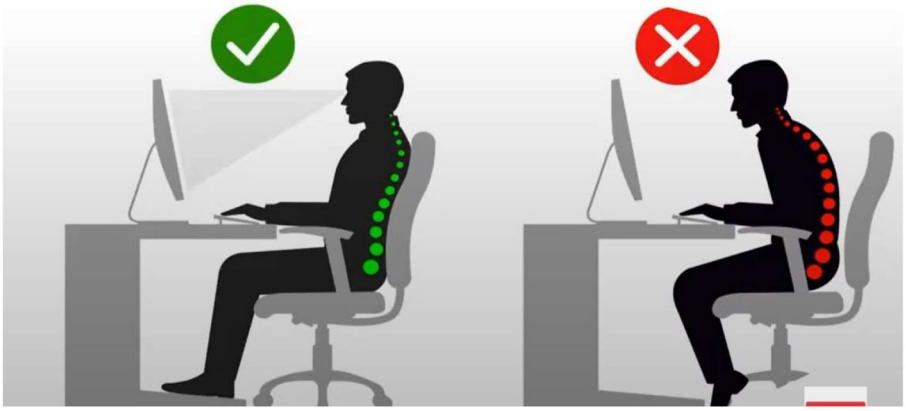

FIGURE 2
Correct ergonomics (L) and incorrect ergonomics (R) in a sitting workstation. Reproduced from "Computer Ergonomics," YouTube, uploaded by "Pearls Classroom," 5 October 2021, https://www.youtube.com/watch?v=XQTQ578wLzo, Permissions: YouTube Terms of Service.

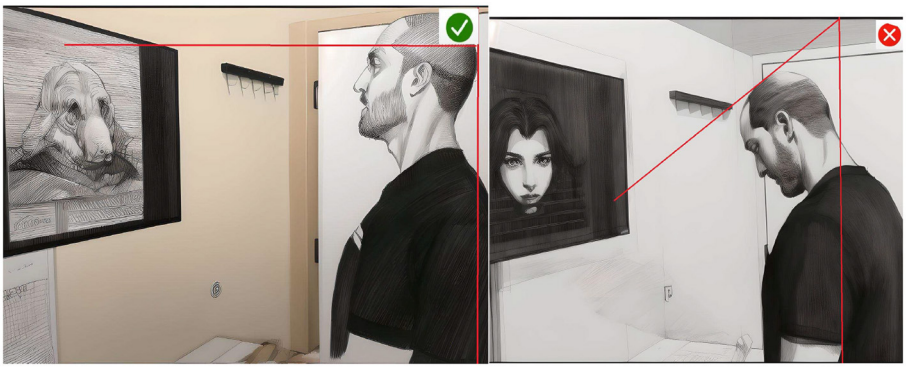

FIGURE 3
Edited scenes. Reproduced from "Libertyville IL neck pain—prevent bad posture with the right workstation," YouTube, uploaded by "Functional Pain Relief," 22 August 2018, https://www.youtube.com/watch?v=0M5C1BJdVsA, Permissions: YouTube Terms of Service.

We present a real-time persuasive design based on posture practices, thereby introducing a novel approach to promoting healthy workplace behavior. This contribution has practical implications for addressing issues related to sedentary work habits, discomfort, and potential health impacts associated with poor posture.

4. Integrating real-time feedback and persuasive elements:

Our design presents the potential and feasibility of persuasive technology to positively influence user behavior, fostering increased awareness and conscious efforts toward maintaining proper posture. This interdisciplinary contribution merges insights from computer science, psychology, and workplace health.

Collectively, these scientific contributions play a significant role in the advancement of knowledge in the fields of human–computer interaction, machine learning, and persuasive technology, with direct applications for improving workplace wellbeing and behavior. The rest of the study is structured as follows: First, we reviewed significant scholarly works on workplace practices, user health, and productivity; persuasive technologies and the workplace; machine learning and workplace practices; and accessibility technologies and healthy practices. Second, we present the methodology based on data collection and deep learning model deployment for the pilot study and the main study. Third, we report on the results of the pilot and main studies. In addition, we compare outcomes for deploying CNN and Yolo-V3 models toward persuasive, healthy workplace designs. Fourth, we present a discussion on the results from the pilot and main studies. Fifth, we report on the limitations of the study and present design recommendations to guide future research. Sixth, we conclude by summarizing the study and drawing an inference based on the results, limitations, and recommendations for future studies.

2 Related work

This section provides an in-depth exploration of related work comparing the relationship between workplace practices, user health and productivity, and other significant ones such as persuasive technologies and workplace practices, machine

TABLE 1 Relationship between the human head anatomy and exerted force leading to spine damage.[a]

S/N	Degrees	Force (lb)	Spine damage risk level
1.	0	10–12	Low or no risk
2.	15	27	Medium
3.	30	40	High
4.	60	50	Very high

[a]https://www.youtube.com/watch?v=0M5C1BJdVsA.

learning and workplace practices, and accessibility technologies and healthy practices.

2.1 Workplace practices, user health, and productivity

Workplace practices cover significant areas such as the proper chair and desk height, appropriate monitor placement, ergonomic keyboard and mouse usage, reduction of glare and reflection, importance of regular breaks, and promoting movement through sit-stand workstations (Dainoff et al., 2012; , 2023). Research has established a relationship between failing to adhere to good workplace practices and the consequences for computer users' health. These include the potential for musculoskeletal disorders, eye strain, and other common health issues related to prolonged computer use (Dainoff et al., 2012; Woo et al., 2016; Boadi-Kusi et al., 2022). According to Nimbarte et al. (2013), Shahidi et al. (2015), and Barrett et al. (2020), the force on the neck increases proportionately as the head angle tilts at a higher degree. The long-term impact of this, as shown in Table 1, is a spine damage risk.

In addition, computer users' health is typically at risk due to repetitive stress injuries (Borhany et al., 2018; Mowatt et al., 2018; Iyengar et al., 2020; Roy, 2020; Steiger et al., 2021). Repetitive strain injury (RSI) is defined as "a chronic condition that develops because of repetitive, forceful, or awkward hand movements for prolonged periods leading to damage to muscles, tendons, and nerves of the neck, shoulder, forearm, and hand, which can cause pain, weakness, numbness, or impairment of motor control" (Sarla, 2019). This implies that computer use involving extended periods of typing and mouse use without proper ergonomics can increase the risk of RSIs. In addition, maintaining poor posture and not adhering to ergonomic requirements when setting up workstations can contribute to this risk. For example, Borhany et al. (2018) carried out a study to examine common musculoskeletal problems arising from the repetitive use of computers. They conducted a survey with 150 office workers and found that 67 of these workers suffer from repetitive stress injuries on the low back, neck, shoulder, and wrist/hand. In addition, they found that these injuries were caused by continuous use of computers without breaks, bad lighting, bad posture, and poorly designed ergonomics in offices. While it is typical that workplace tasks are characterized by repetitive tasks and actions, it has become imperative to design workplace technologies to support users in carrying out repetitive tasks without straining any part of the body (Moore, 2019; Johnson et al., 2020).

It is important to state that research has found the impact of computer users' health due to repetitive stress injuries and other related health issues on the productivity of users in workplaces. In other words, a well-designed workplace not only improves the user's comfort but also enhances work efficiency and overall job satisfaction (Pereira et al., 2019; Baba et al., 2021; Franke and Nadler, 2021). Pereira et al. (2019) examined 763 office workers in a 12-week study. They interpreted office productivity to be relative to absenteeism from work due to neck pain. The results from this study show that those exposed to healthy workplace practices and neck-specific exercise training had limited records of absenteeism. Pereira at al. reported that individuals with unhealthy workplace practices and limited access to health promotion information were more likely to be less productive, i.e., absent from work. Baba et al. (2021) conducted a study involving 50 newly employed staff in an organization. The staff was divided into experimental groups (with healthy workplace practices, e.g., comfortable computer desks) and control groups (with unhealthy workplace practices, such as less comfortable furniture). The study revealed a significant impact on the work productivity of the experimental group compared with the control groups (based on a t-test showing that t.cal = 0.08; t.tab = 1.71, where t.cal is the calculated t-test value and t.tab is the value of t in the distribution table).

While many organizations focus on employee training and sensitization programs for healthy workplace practices, limited research has been reported on workplace culture, employee training, computer workstation assessment, and the benefits of posture assessment tools. This study explores the potential of persuasive technologies for enhancing effective workplace posture practices. These technologies can serve as posture assessment tools, providing valuable feedback to organizations on the best ways to support their employees.

2.2 Persuasive technologies and the workplace

Persuasive technologies and workplace practices are two distinct areas of study and practice, but they intersect in designing user interfaces and technology systems that promote healthy workplace practices for technology users. Overall, this will enhance technology users' wellbeing and productivity. Research has explored persuasive technologies in relation to best workplace practices. This includes taking regular breaks (Jafarinaimi et al., 2005; Ludden and Meekhof, 2016; Ren et al., 2019), fitness apps (Mohadis et al., 2016; Ahtinen et al., 2017; Paay et al., 2022), feedback systems and wearable devices (Bootsman et al., 2019; Jiang et al., 2021), workstation movement (Min et al., 2015; Damen et al., 2020a,b), chair, desk, and monitor height adjustments (Kronenberg and Kuflik, 2019; Kronenberg et al., 2022), posture correction (Min et al., 2015; Bootsman et al., 2019; Kim M. T. et al., 2019), mouse/keyboard use and reduction of glare and reflection (Bailly et al., 2016), and other healthy work behaviors (Berque et al., 2011; Mateevitsi et al., 2014; Gomez-Carmona and Casado-Mansilla, 2017; Jiang et al., 2021; Brombacher et al., 2023; Haliburton et al., 2023; Robledo Yamamoto et al., 2023).

Table 2 summarizes closely related work on persuasive technologies with respect to workplace practices. We present discussions based on instances of workplace practices we listed previously. This includes taking regular breaks, fitness apps, feedback systems, workstation movement, chair, desk, monitor height adjustments, posture correction, mouse/keyboard use, reduction of glare and reflection, and other healthy practices. Jafarinaimi et al. (2005) developed sensor-based office chairs that encourage users to break away from their computers. Every 2 min, the chair slouches its position from upright to backward bend, signifying the need for computer users to take a break. In view of this, they experimented with a single user (55-year-old university staff). The results from the study showed how the sensor-based office chair greatly influenced the user's attitude to break away from their computer.

Mohadis et al. (2016) developed a low-fidelity web-based prototype to encourage physical activity among older office workers. They considered 23 persuasive principles as they relate to physical activity. These include reduction, tunneling, tailoring, personalization, self-monitoring, simulation, rehearsal, dialogue support, praise, rewards, reminders, suggestions, similarity, social role, credibility support expertise, real-world feel, third-party endorsements verifiability, social support/social learning, social comparison, normative influence, social facilitation, competition, and recognition. Reduction was targeted at making complex tasks simple to complete. Tunneling was driven by using the system to guide users while persuading them to change their behavior. Self-monitoring ensures that users can keep track of their behavior. Simulation covers demonstrating aspects of behaviors to interpret cause-and-effect relationships. Rehearsal provides an opportunity to continue to practice behavior toward change. In addition, the other persuasive principles (dialogue support praise, rewards, reminders, suggestions, similarity, social role, credibility support expertise, real-world feel, third-party endorsements verifiability, social support/social learning, social comparison, normative influence, social facilitation, competition, and recognition) were driven toward enhancing a change in the user's physical activity behaviors. The authors experimented with 10 participants and found that only two (2) persuasive principles were perceived positively. This includes dialogue support and credibility support.

Bootsman et al. (2019) explored wearable posture monitoring systems for nurses in workplaces. Nurses were considered to carry out repetitive bending throughout their work shifts. The system was designed to track their lower back posture. The system is connected to a mobile application that provides feedback on the different posture positions of users and tips for changing bad postures. The system was evaluated with six (6) nurses (aged between 20 and 65 years) for 4 days during work hours. Based on the intrinsic motivation inventory, the results show interest, perceived competence, usefulness, relatedness, and effort/importance scored more points. In addition, the results from the qualitative analysis show that participants appreciated the comfortability of the wearable system, though they were not in support of the frequency of beeps as it caused some distractions.

Haque et al. (2020) explored computer workstation movements similar to regular breaks. Unlike the regular break, computer users are encouraged to walk around and keep track of their physical activity level. The authors conducted an experiment with 220 office workers from the United Kingdom, Ireland, Finland, and Bangladesh for 4 weeks while evaluating their "IGO mHealth app." The app monitors office workers' meal intake and work periods to send a 10-min interval walk-around reminder. The app tracks this movement while setting a target limit of 1,000 steps every 10 min. The app incorporates the leaderboard gaming element, encouraging competition through persuasion. The results from this study show a trend in weight loss, and a follow-up interview revealed three (3) persuasive principles that were perceived positively: (1) autonomy, (2) competence, and (3) relatedness. Autonomy shows how the app helped them achieve their set goals. Competence reflects how confident they were about their capability to use the app to perform different tasks. Relatedness shows how they were able to use the app to establish social connections.

Kronenberg et al. (2022) developed robotic arms that can be used to automatically adjust computer system screens. The robot detects the distance between the screen and the user's seating position. Then, the robot calculates the new screen orientation and adjusts to keep a healthy distance between the users and their computer screens. The authors conducted an experiment with 35 participants (25–68 years old) in their workspaces. The results of one-sample Wilcoxon Signed Rank Test show that participants could effectively complete the tasks and scenarios using this system at ($p < 0.001$), the screen did not move at the right pace when it moved (given that $p = 0.189$ was not significant that it moved at the right pace), the screen did not move at the appropriate moment (given that $p = 0.904$ was not significant that it moved at the appropriate moment), the screen was not well-adjusted to users' pose (given that $p = 0.163$ was not significant that it was well-adjusted to users pose), and the users felt distracted by the movement of the screen (given that $p = 0.028$ was not significant that users felt less distracted by the movement of the screen).

Kim M. T. et al. (2019) conducted experiments with a robot to support posture corrections during object lifting with 10 adults (30–34 years old). They considered five (5) different joints in the human body: (1) hips, (2) knees, (3) ankles, (4) shoulders, and (5) elbows. The results of their *t*-test analysis showed that the robot significantly lowered the overloading effect in all joints: shoulder ($p < 0.001$), elbow ($p < 0.001$), hip ($p < 0.001$), knee ($p < 0.001$), and ankle ($p < 0.001$). This implies that the robot can promote better posture practices in workplaces.

Bailly et al. (2016) developed a *"LivingDesktop"* that supports users to reduce reflection from the monitor screen. In addition, the system allows users to adjust the mouse and keyboard positions to improve ergonomics. The authors evaluated the system with 36 desktop users (22–40 years old). The results from this study show that users liked adjustable features because they fit their needs for video conferencing, tidying their workspace, and maintaining the right posture. On the other hand, some users criticized the system for its distractions in workspaces.

Jiang et al. (2021) developed a smart t-shirt wearable application for depression management in workplaces. They considered emotion regulation for depression management based on the movement of the shoulders and arms. The smart t-shirt

TABLE 2 Summary of research on persuasive technologies and workplace practices.

S/N	References	Technology	Workplace practices covered							
			Chair and desk height	Monitor placement	Keyboard and mouse use	Reduction of glare and reflection	Regular breaks	Workstation movement	Posture correction	Other healthy practices
1.	Haque et al. (2020)	Mobile App						☐		
2.	Damen et al. (2020a)	Tangible						☐		
3.	Damen et al. (2020b)	Phones, Tablets and Notebooks						☐		
4.	Min et al. (2015)	Sensors						☐	☐	
5.	Ludden and Meekhof (2016)	Tangible					☐			
6.	Jafarinaimi et al. (2005)	Tangible					☐			
7	Kronenberg and Kuflik (2019)	Robot	☐							
8.	Jiang et al. (2021)	Tangible								☐
9.	Mohadis et al. (2016)	Web App						☐		
10.	Gomez-Carmona and Casado-Mansilla (2017)	Tangible								☐
11.	Bootsman et al. (2019)	Tangible and Mobile App							☐	
12.	Kronenberg et al. (2022)	Robot		☐						
13.	Kim W. et al. (2019)	Robot							☐	
14.	Bailly et al. (2016)	Actuators			☐	☐				

10.3389/fdata.2024.1359906

changes resistance based on users' emotions. The fabric maintains a resistance of 180 kΩ while relaxed (positive emotion) and 400 kΩ when stretched (negative emotion). In view of this, they tested the smart t-shirt with six (6) healthcare workers for 5 days and found that the smart t-shirts regulated healthcare workers' emotions positively at work.

While most of these persuasive technologies have explored user interface design and user experience evaluation, we found other state-of-the-art practices employing machine learning techniques. Machine learning designs present more intelligent and data-oriented systems. This makes them more flexible to learn new patterns while users continue to interact with them. We present the extent to which machine learning has been tailored to enhance workplace practices in the next Section 2.3.

2.3 Machine learning and workplace practices

Machine learning can significantly impact the design of products for healthy workplaces. It interprets a wide range of data types, including sensor data, motion, eye movements, and human body movement. Machine learning models can be embedded into wearable devices, phones, and computers, enabling the detection of patterns in data and the optimization of communication with humans based on the diverse data they were trained on. For instance, facial recognition models, as supported in self-service photo booths (Kember, 2014), can detect specified height, width, and head position orientations (Chen et al., 2016).

Some significant research studies have delved into the application of machine learning in the realm of workplace practices. These studies have particularly focused on classifying healthy and active work styles (Rabbi et al., 2015) and automatic adjustments of chair and desk heights (Kronenberg and Kuflik, 2019). In their study, Kronenberg and Kuflik (2019) proposed a deep learning design for robotic arms that are capable of adjusting chair and desk heights based on body positions. Although the system was still in the implementation stage, initial results demonstrated the potential of embedding a camera in a robotic arm. This camera would interact with their proposed deep learning model.

Despite extensive research within this domain, limited study has been conducted on camera posture positions on the face, head, neck, and arms. While Min et al. (2015) explored body positions such as the back and spine using sensors, there is still a need to explore additional body positions captured by cameras. In a related study, Mudiyanselage et al. (2021) evaluated a workplace that involved lifting work-related materials using wearable sensors and various machine learning models (Decision Tree, Support Vector Machine, K-Nearest Neighbor, and Random Forest). The results indicated that the decision tree models outperformed others with a precision accuracy of 99.35%. Although these results were significant and focused on back body positions, there are still gaps within the context of computer workstations.

In another relevant study by Nath et al. (2018), significant work on lifting arm and wrist positions was considered using wearable sensors and the support vector machine (SVM) model. The study results demonstrated that SVM recognized over 80% of the risky positioning of the arm and wrist.

Hence, based on the persuasive and machine learning perspectives of workplace system design, different body positions are captured, and feedback is provided to support users. Nevertheless, there is a need to understand the extent to which research has supported making these technologies more accessible to diverse users. In the next section, we covered related work done with respect to making workplace posture technologies more accessible.

2.4 Accessibility technologies and healthy practices

Most accessibility technologies focus on providing feedback based on machine learning detection to address the needs of disabled individuals (Kulyukin and Gharpure, 2006). Brik et al. (2021) developed an IoT-machine learning system designed to detect the thermal comfort of a room for disabled persons, offering feedback on the room's thermal condition. The machine learning system was trained on artificial neural networks (ANNs). The performance of ANNs was compared with other algorithms such as logistic regression classifiers (LRC), decision tree classifiers (DTC), and gaussian naïve bayes classifiers (NBC). ANN performed better, achieving 94% accuracy compared with the other algorithms.

In a related study, Ahmetovic et al. (2019) investigated navigation-based assistive technologies for the blind and visually impaired. They identified rotation errors and utilized a multi-layer perceptron machine learning model to correct rotation angles, providing positive feedback. The multi-layer perceptron achieved lower rotation errors (18.8° on average) when tested with 11 blind and visually impaired individuals in real-world settings.

Overall, we found that though related studies have explored healthy practices in workplace settings based on different persuasive technologies ranging from mobile to tangible, little work has covered real-time posture detection for important areas of the body such as the back, neck, hands, and head. These parts of the body have been associated with a lot of repetitive workplace stress injuries based on bad postures (Anderson and Oakman, 2016; Catanzarite et al., 2018; Krajnak, 2018). The study by Min et al. (2015) and Mudiyanselage et al. (2021) presents closely related concepts. Though these studies explored parts of the body such as the back, spine, arm, and wrists, they used sensors, which might not be comfortable for users of systems. Considering that laptop cameras can detect these parts of the body in an unobstructive way, we explored this in our current study.

3 Materials and methods

We outline the materials and methods employed in the study. This aligns with the overarching goal of our research to investigate how individuals can become aware of their unhealthy posture practices in workplaces (both while sitting and standing) and the main research question (RQ: Can persuasive computers be designed to detect unhealthy posture practices in workplaces?). We provide details on the experimental materials used for developing

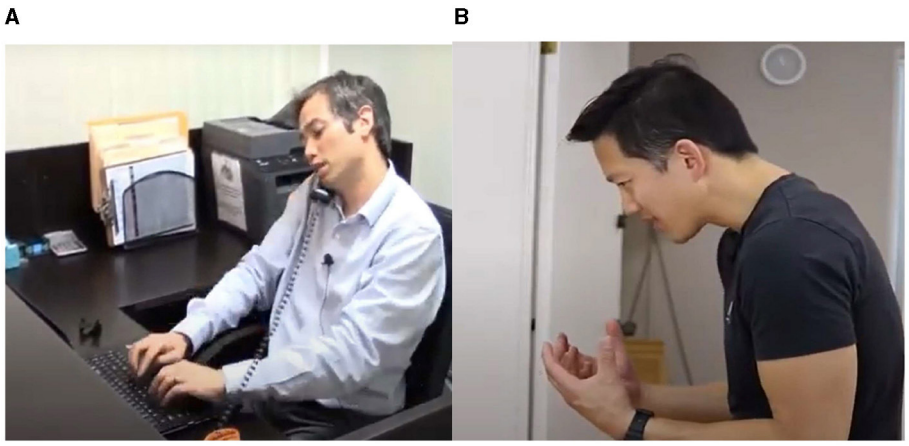

FIGURE 4
Samples of bad practices. **(A)** Reproduced from "Center for Musculoskeletal Function: Workspace Ergonomics and MicroBreak Exercises," YouTube, uploaded by "Dr. Daniel Yinh DC MS," 10 Apr 2017, https://www.youtube.com/watch?v=HS2KrPmKySc, Permissions: YouTube Terms of Service. **(B)** Reproduced from "Correct Ergonomic Workstation Set-up | Daily Rehab #23 | Feat. Tim Keeley | No.112 | Physio REHAB," YouTube, uploaded by "Physio REHAB," 13 December 2017, https://www.youtube.com/watch?v=FgW-9_28N8E&t=314s, Permissions: YouTube Terms of Service.

FIGURE 5
Samples of the good practices. **(A)** Reproduced from "Working from home—how to set up your laptop (correctly!) | Tim Keeley | Physio REHAB," YouTube, uploaded by "Physio REHAB," 19 March 2020, https://www.youtube.com/watch?v=6GlkoFnZpFk, Permissions: YouTube Terms of Service. **(B)** Reproduced from "How to set up workstation at home," YouTube, uploaded by "Sundial Clinics," 12 April 2021, https://www.youtube.com/watch?v=wN-Ww1sCWNY, Permissions: YouTube Terms of Service.

deep learning models, specifically convolutional neural networks and Yolo-V3.

3.1 Data collection and preprocessing

We conducted data collection in three phases (phase 1, phase 2, and phase 3). In the first phase, we gathered data by extracting Creative Commons image datasets from YouTube using the search terms ({bad} OR {good} AND {ergonomic posture}). Utilizing the Snip and Sketch tools, we extracted key frames depicting instances of both good and bad ergonomics. In total, we amassed 269 image datasets, comprising 157 examples of bad practices and 112 examples of good practices. The datasets from this initial phase were utilized for the pilot study, which aimed to assess the feasibility of employing machine learning for the detection of posture practices. Figures 4, 5 provide a cross-section of the datasets collected from YouTube.

In addition, we gathered more image datasets from Pexels using the Snip and Sketch tools. Pexels offers royalty-free images that match both the good and bad workplace practices of computer users. Utilizing related search terms such as "people AND {using the computer}" OR "{looking head straight}" OR "{sitting in the office}," we extracted key frames, resulting in 618 instances of bad

FIGURE 6
Samples of bad posture. Reproduced from Pexels.

FIGURE 7
Samples of good posture. Reproduced from Pexels.

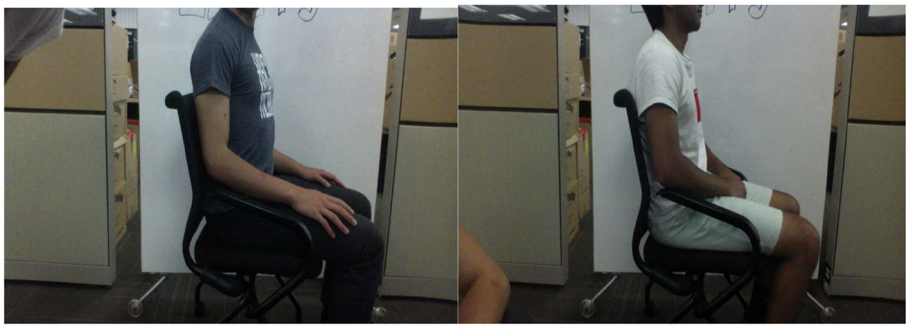

FIGURE 8
Samples of the good practices. Reproduced from Kaggle.

practices and 90 instances of good practices. These datasets were combined with those from Phase 1 to conduct the main study for YOLO-V3.

Recognizing the limitations of convolutional neural networks (CNN) with small datasets (Han et al., 2018), we addressed this concern in Phase 3 by collecting additional datasets. To enhance the dataset, we collected both zoomed-in and zoomed-out resolution images from Pexels. Research has shown that zooming, as one of the techniques of data augmentation, increases the number of datasets (Shorten and Khoshgoftaar, 2019). Figures 6, 7 offer a cross-section of the datasets collected from Pexels.

For the Phase 3 data collection task, we explored the posture dataset available on Kaggle. Kaggle, known for its extensive repository of public datasets for machine learning (Tauchert et al., 2020), provided a valuable resource. We added 311 images depicting good practices to the datasets from Phases 1 and 2.

TABLE 3 Summary of datasets distribution by source.

S/N	Source	Comfortable	Uncomfortable
1.	YouTube	112	157
2.	Pexels	90	618
3.	Kaggle	311	-
Total		513	775

The combined datasets from this phase were used to conduct the main study experiment for convolutional neural networks (CNN). Figure 8 showcases a cross-section of sample images collected from Kaggle9. Though Kaggle had a couple of images for bad postures, we considered using the good ones to balance our datasets (we initially had more bad postures compared with good postures).

Additionally, we defined the two classes as "comfortable" and "uncomfortable." All the image datasets depicting good practices were assigned to the "comfortable" class, while those depicting bad practices were assigned to the "uncomfortable" class. Table 3 offers a summary of all the datasets collected for the study. We employed static image datasets as they are applicable to existing real-time detection studies (Huang et al., 2019; Lu et al., 2019), and a video is a sequence of moving images in frames (Lienhart et al., 1997; Perazzi et al., 2017). Hence, the computer vision library provides functionality to help capture this image frame per second and parse them to the machine learning model to quickly predict the class in real time.

3.2 Study description

We covered two significant steps, namely, the pilot and main studies. We explored the feasibility of designing with a few datasets in a pilot study. We present this pilot study to guide the research community on the impact of dataset size in this area. In the main study, we extended the number of datasets to show improvements in the accuracy of models. The datasets collected from YouTube during Phase 1 data collection were pre-processed and used to train the two models for the pilot study (CNN-pilot and Yolo-V3-pilot). We evaluated their performance through loss graphs and in real-time (mean average precision). The mean average precision is a metric for evaluating the accuracy of object detection, especially in real time (Padilla et al., 2021). Furthermore, we combined datasets from YouTube and Pexels to train the YOLO-V3-main model. Additionally, we combined datasets from YouTube, Pexels, and Kaggle to train the CNN-main model. Both the YOLO-V3-main and CNN-main models were developed for the main study.

3.2.1 Pilot study

We conducted two experiments for the pilot study. The first experiment involved the development of the Yolo-V3 model (Yolo-V3-pilot). We performed an automatic data annotation task[1] on the entire datasets collected from YouTube. Subsequently, we trained our datasets on the Yolo-V3 model implementation of keras-yolo3[2] on the CPU and we tested this implementation on Google Colab. The second experiment was implemented on the CNN model of Abhishekjl.[3] Our selection of Abhishekjl's framework was based on its relevance in the application of the cv2 python library which is applicable in the recent study by Singh and Agarwal (2022). In addition, the keras-yolo3 implementation has been recently applied to the current state-of-the-art pedestrian detection system by Jin et al. (2021) and other systems (Chen and Yeo, 2019; Silva and Jung, 2021). Hence, datasets collected from YouTube were trained on the CNN model (CNN-pilot). The CNN-pilot model was trained and tested on Google Colab.

3.2.2 Main study

We conducted two experiments for the main study. In the first experiment, we combined datasets from YouTube and Pexels (from phases 1 and 2 of data collection). We performed automatic data annotation exclusively for datasets from Pexels. The annotation data were then added to pre-existing annotations from the pilot study to train a new Yolo-V3 model (Yolo-V3-main) for the main study, utilizing CPU resources. In the second experiment, we combined datasets from YouTube, Pexels, and Kaggle (from phases 1–3) and trained them using Google Colab on the CNN model (CNN-main). Like the pilot study, both Yolo-V3 and CNN models were implemented based on the architectures of Keras-Yolo3 and Abhishekjl. In addition, we tested Yolo-V3-main and CNN-main in Google Colab.

3.3 Overview of the CNN model

The CNN model (Figure 9) consists of 2 convolutional 2D layers, 2 max_pooling 2D layers, one flatten, and twi dense layers. Furthermore, the hyperparameters for the model include 3 activation functions (rectified linear unit, RELU) for the convolutional 2D layers and one of the dense layers, one sigmoid activation function added to the last dense layer, Adam optimizer, a learning rate of 1e-3, a batch size of 5, and 10 epochs. The loss of the CNN-pilot model was set to binary_crossentropy. The convolutional 2D layers combine the 2D input after filtering, computing the weights, and adding a bias term (Li et al., 2019). The max_pooling2d layers reduce the input dimensions, leading to a reduction in outputs (Keras[4]). The flatten layer combines all the layers into a flattened 2-D array that fits into the neural network classifier (Christa et al., 2021). The dense layers are regular, deeply connected neural network layers that are used to return outputs from the model (Keras[5]). We employed the rectified linear unit (RELU) activation function as it is one of the most widely used functions because of its improved performance (Dubey et al., 2022). The sigmoid function was selected because it is suitable for

1 https://github.com/iwinardhyas/auto_annotation/tree/master/auto_annotatation

2 https://github.com/qqwweee/keras-yolo3

3 https://github.com/Abhishekjl/Facial-Emotion-detection-webcam-

4 https://keras.io/api/layers/pooling_layers/max_pooling2d/

5 https://keras.io/api/layers/core_layers/dense/

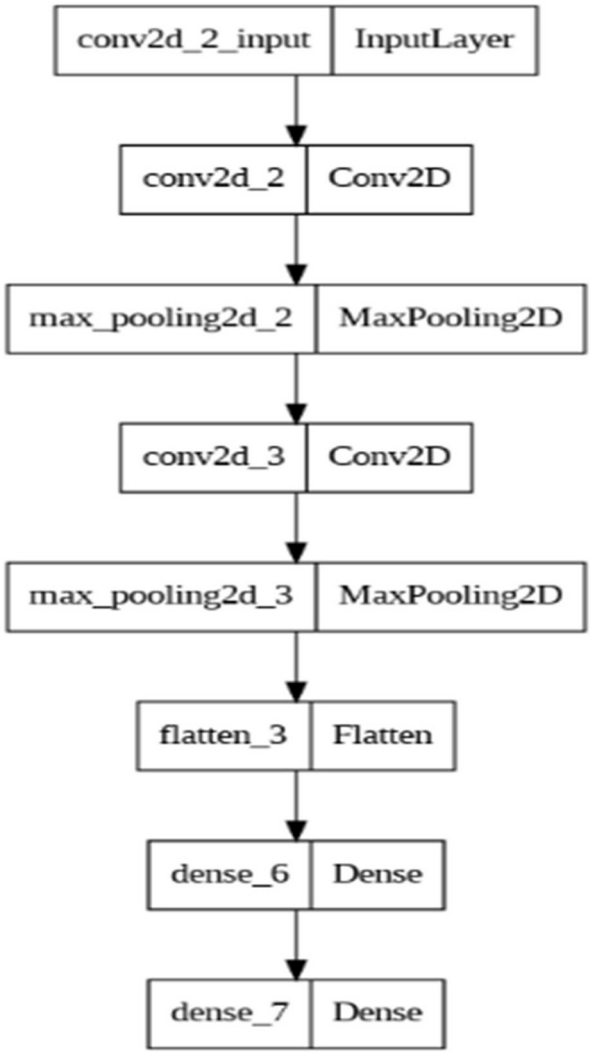

FIGURE 9
CNN model architecture.

binary classification tasks (Keras[6]) as we employed in our study. We employed the Adam optimizer because it is memory efficient and requires limited processing resources (Ogundokun et al., 2022). We set the learning rate of 1 e-3 and batch size 5 as we considered the sensitivity of CNN models to small datasets (Brigato and Iocchi, 2021).

3.4 Overview of the Yolo-V3 model

The Yolo-V3 model (Figure 10) consists of 74 convolutional 2D layers, 71 batch normalization layers, 70 leaky rectified linear unit (RELU) activation layers, two UpSampling2D layers, and one ZeroPadding2D layer. We set the hyperparameters for the model as follows: Adam optimizer, learning rate of 1e-4, and batch size of 16. We consider Adam Optimizer to be appropriate as it is memory efficient and requires limited processing resources (Ogundokun et al., 2022). In addition, we considered a reduced learning rate and batch size because of the number of datasets we have. This will help the model learn efficiently. Unlike CNN, YOLO-V3 yielded more annotated datasets with different dimensions. This is typical with YOLO-V3 data annotations (Diwate et al., 2022). Furthermore, we varied the number of epochs for both the pilot and main studies. We used four epochs for the pilot study (Section 4.12) and a maximum of 40 epochs for the main study (Section 4.2.2). We used the default loss function (binary_crossentropy) for the YOLO model. The convolutional 2D layers combine the 2D input after filtering, computing the weights, and adding a bias term (Li et al., 2019). The batch normalization layer normalizes inputs to ensure that they fit the model as their weights continue to change with each batch that the model processes (Arani et al., 2022; Keras[7]). The leaky RELU activation layer is a leaky version of a rectified linear unit activation layer (Keras[8]). It introduces non-linearity among the outputs between layers of a neural network (Xu et al., 2020). The UpSampling2D layer is used to repeat the dimensions of the input to improve its quality (Liu et al., 2022; Keras[9]). The ZeroPadding2D layer adds extra rows and columns of zeros around images to preserve their aspect ratio while being processed by the model (Dang et al., 2020; Keras[10]).

4 Results

In this section, we present our findings from the pilot and main studies. This section covers reports from our experiments with Yolo-V3 and CNN models using datasets collected from YouTube, Pexels, and Kaggle.

4.1 The pilot study

To visualize the feasibility of the study, we developed two models for detecting workplace practices in real time: CNN and Yolo-V3. We chose these models based on their proven capabilities for supporting real-time object detection in previous research (Tan et al., 2021; Alsanad et al., 2022). For the CNN model, we divided the datasets into 75% training and 25% validation datasets (refer to Table 4). We used 75% training to 25% validation set split for the CNN model considering how similar tasks employed this ratio (Azimjonov and Özmen, 2021; Bavankumar et al., 2021; Akter et al., 2022). Programmatically, we split the datasets into 90% training and 10% validation datasets for the Yolo-V3 model. The reason for the difference in this split ratio was based on previous studies employing similar ratios, especially for Yolo models (Akut, 2019; Setyadi et al., 2023; Wong et al., 2023).

4.1.1 CNN pilot study posture detection

We trained the CNN-pilot model for 10 epochs, employing hyperparameter tuning variables such as the stochastic gradient

6 https://keras.io/api/layers/activations/

7 https://keras.io/api/layers/normalization_layers/batch_normalization/

8 https://keras.io/api/layers/activation_layers/leaky_relu/

9 https://keras.io/api/layers/reshaping_layers/up_sampling2d/

10 https://keras.io/api/layers/reshaping_layers/zero_padding2d/

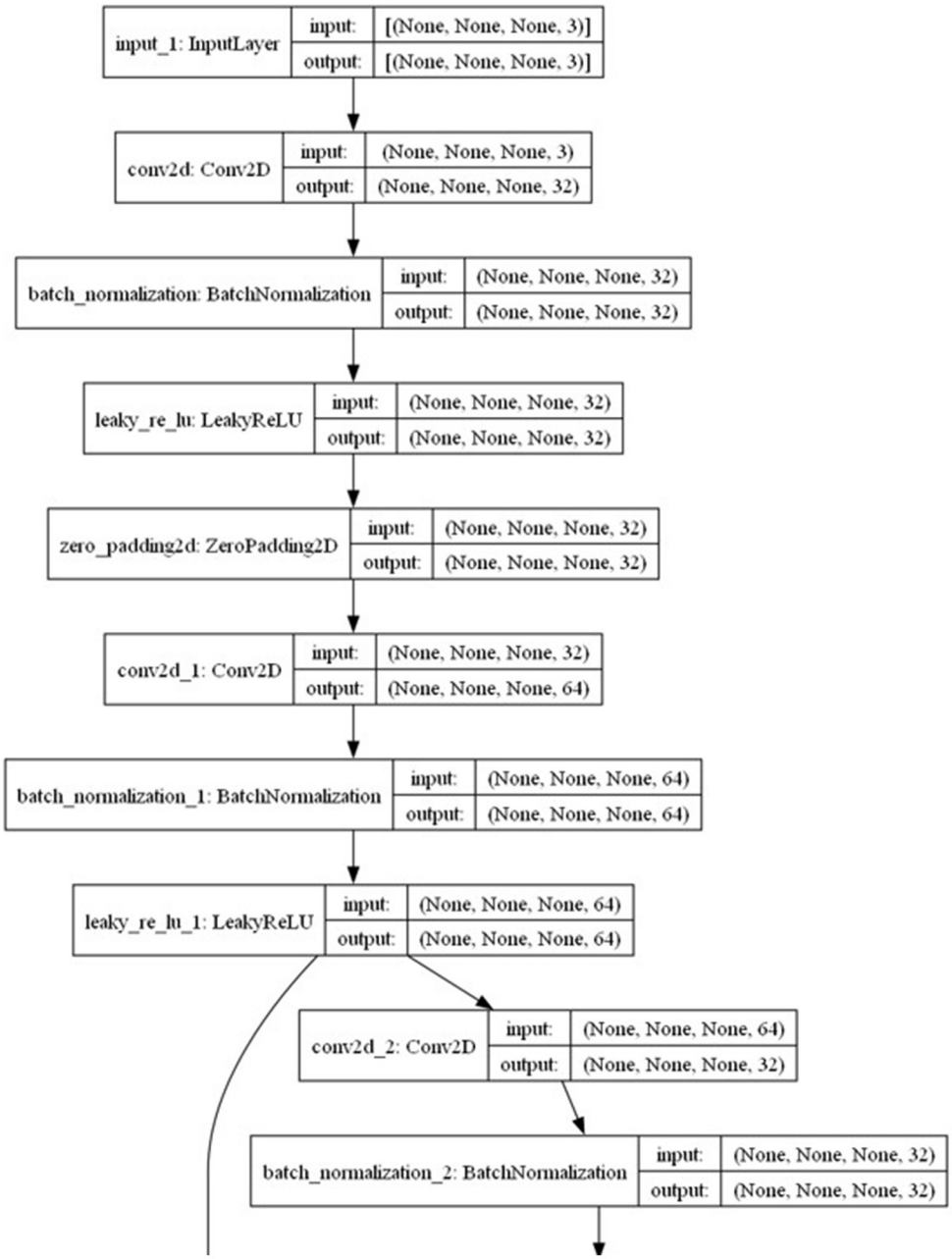

FIGURE 10
Cross-section of the YOLO-V3 model architecture (full architecture is available at Appendix A1).

TABLE 4 Summary of dataset distribution for the pilot study.

S/N	Model	Comfortable		Uncomfortable		Total	
		Training	Validation	Training	Validation	Training	Validation
1.	CNN	84	28	118	39	202	67
2.	Yolo-V3	101	11	141	16	242	27
Total		185	39	259	55	444	94

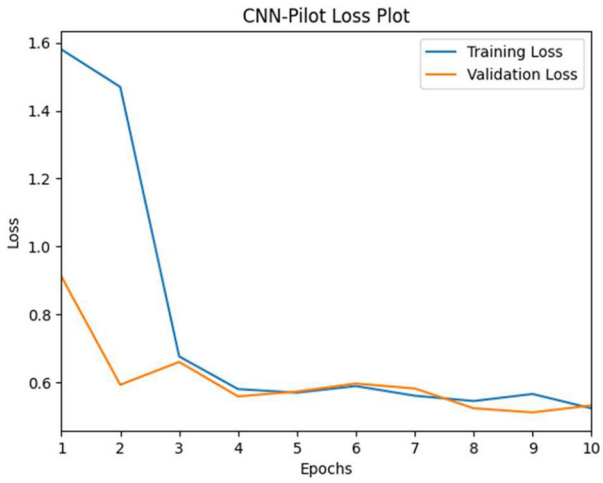

FIGURE 11
CNN-pilot model's training vs. validation loss.

descent optimizer with a learning rate of 1e-3. The results of our CNN training indicate a significant decrease in both training and validation loss values, approaching the 10th epoch (see Figure 11). The validation loss was minimal at epoch 10 compared with the training loss, suggesting a slight underfitting of the model.

We deployed the model in real-time using the computer vision Python library. Running the model on six real-time test instances, it achieved a mean average precision of 52%. In most instances, better precision values were observed for "comfortable" compared with "uncomfortable" (see Figure 12).

4.1.2 Yolo-V3 pilot study posture detection

The Yolo-V3-pilot model was trained with two layers, employing a strategy of frozen layers to stabilize the loss and unfrozen layers to further reduce the loss, over four epochs. These layers were configured to train with hyper-tuning parameters, including the Adam optimizer with a learning rate of 1e-4 and a batch size of 16. The results of our YOLO-V3 layers 1 and 2 training reveal a decrease in the training loss toward epoch 4 compared with the validation loss (refer to Figure 13). However, it is typical for YOLO-V3 to return a high level of loss values below epoch 10 (Li et al., 2020).

We deployed the Yolo-V3-pilot model in real time for the classes "comfortable" and "uncomfortable." For exceptional cases, we included a "neutral" class. This addition allows Yolo-V3 to handle instances where the detections do not match the expected classes. Figures 14, 15 showcase instances where the Yolo-V3-pilot model segmented areas of comfort compared with discomfort. In other cases, the model returned "neutral" while one of the researchers tested it in real time using the computer vision Python library. The model achieved a mean average precision of 64% across six real-time test instances.

From the results of both models (CNN-pilot and Yolo-V3-pilot), the Yolo-V3-pilot model's boxes extended beyond the face, capturing other significant areas of comfort or discomfort such as the eyes, neck, and back (see Figures 14, 15).

4.2 The main study

To enhance the performance of both models (CNN-main and Yolo-V3-main) in the main study, we trained these models on additional datasets collected from Pexels and Kaggle. For the Yolo-V3-main model, we combined YouTube datasets with those from Pexels, while the CNN-main model was trained on a combination of datasets from YouTube, Pexels, and Kaggle. In the case of the CNN-main model, we split the datasets into 75% training and 25% validation sets (refer to Table 5). We maintained the 90% training and 10% validation set split for the Yolo-V3-main model.

4.2.1 CNN main study posture detection

We maintained the hyper-tuning parameters from the pilot study for CNN, and the model was trained for 10 epochs. The results of our CNN training indicate a significant decrease in both training and validation loss values, approaching the 10th epoch (see Figure 16). The training loss was minimal at epoch 10 compared with the validation loss, indicating better convergence of the training and validation losses compared with those reported earlier in the pilot study (see Figure 11).

In real time, the CNN-main model predicts uncomfortable classes better (Figure 17: 89.6, 98.7, 93.5, and 93.0%). The CNN-main model attained a mean average precision of 91% on 19 real-time test data points.

4.2.2 Yolo-V3 main study posture detection

Like the pilot study, the Yolo-V3-main model was trained with two layers, incorporating frozen layers for a stable loss and unfrozen layers to further reduce the loss. The first layer was set to train for 10 epochs, and the second layer started at the 11th epoch (continuing from the first layer) and concluded at the 39th epoch. These layers were trained with hyper-tuning parameters, including the Adam optimizer with a learning rate of 1e-4 and a batch size of 16. The results for both layers 1 and 2 of the Yolo-V3-main model show that the training and validation loss curves converged at epoch 10 for the first layer and diverged slightly upward at epoch 39 for the second layer (see Figure 18). This implies slight overfitting of our Yolo-V3-main model.

We deployed the Yolo-V3-main model in real time, and the results indicate that the model performed significantly better in detecting both classes, "comfortable" and "uncomfortable" (refer to Figure 19). The Yolo-V3-main model achieved a mean average precision of 92% across 11 real-time test instances.

5 Discussion

The study explored design opportunities for persuasive systems based on real-time posture detection. We conducted two experiments, namely, the pilot and main studies, utilizing two

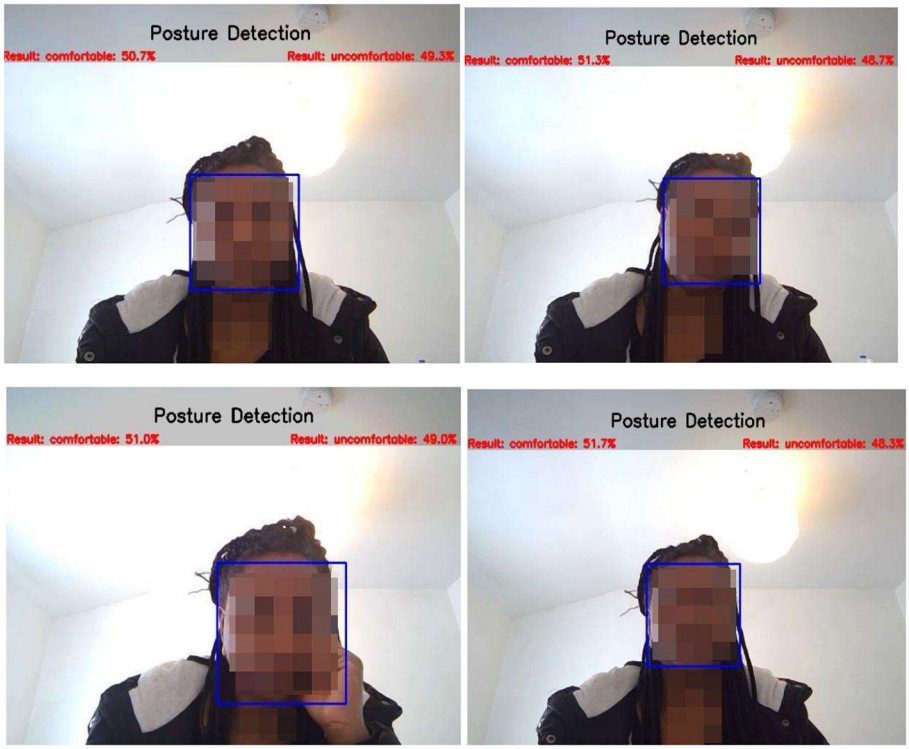

FIGURE 12
CNN-pilot model's detection of posture.

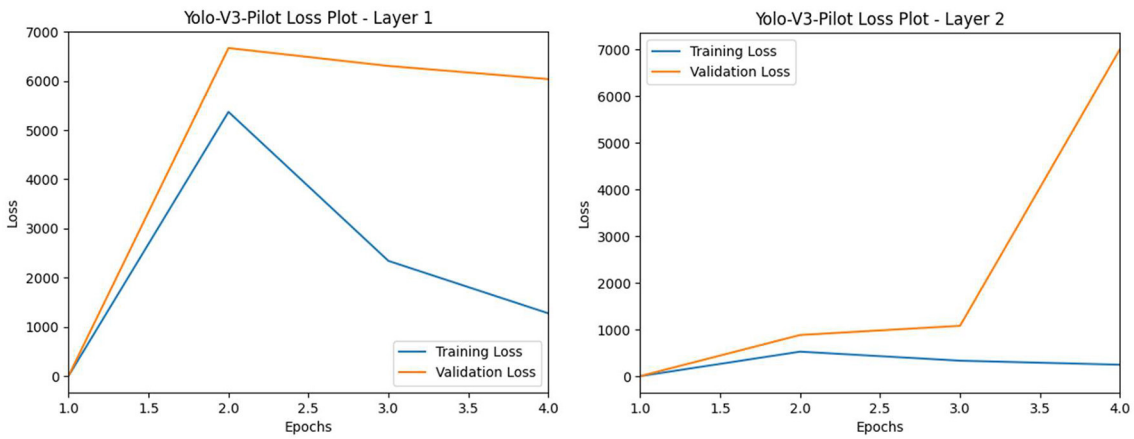

FIGURE 13
L-R: Yolo-V3-pilot model's training vs. validation loss (L: Layer 1 and R: Layer 2).

deep learning algorithms: CNN and Yolo-V3. In this section, we discuss the results and propose design recommendations aligned with the overarching goal of the study, addressing how people can become conscious of their unhealthy posture practices in workplaces, whether sitting or standing. Furthermore, we relate these findings to answering the main research question: RQ: Can we design persuasive computers to detect unhealthy posture practices, such as sitting and standing, in workplaces?

From the pilot study, we observed that the CNN-pilot model tends to generalize its detection based on facial regions, occasionally extending to the neck regions. Additionally, for the CNN-pilot model, we reported on the detection of comfortable and uncomfortable postures with similar precision accuracy values. The lack of generalizability in the model raises concerns, particularly given our overarching goal of ensuring that persuasive technologies encourage people to maintain the right posture practices. It would

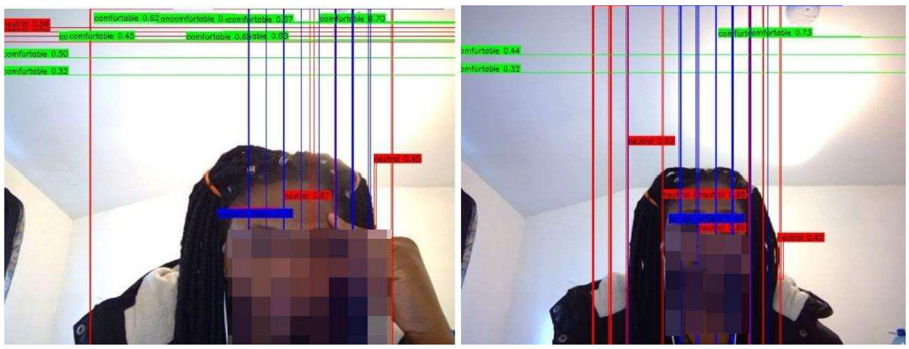

FIGURE 14
L-R: Yolo-V3-pilot model's posture detection: comfortable; uncomfortable; neutral. L: showing areas of discomfort around the eyes and where the hand intercepts the eyes. R: showing discomfort from the eye to the neck regions.

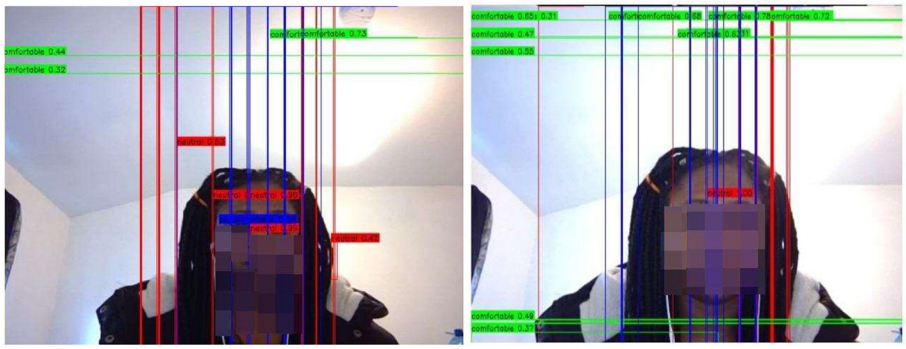

FIGURE 15
L-R: Yolo-V3-pilot model's posture detection: comfortable; uncomfortable; neutral. L: showing areas of discomfort around the eyes, neck, and back regions. R: showing discomfort from the eye to the neck regions.

be more suitable for individuals to be prompted to change their uncomfortable postures more frequently.

In contrast, the Yolo-V3-pilot model, with its anchor boxes, provided more comprehensive coverage and detection of postures. While it is common for Yolo models to generate multiple anchor boxes when detecting objects (Zhang et al., 2022), we observed trends of it detecting various body positions and regions associated with the required postures.

The main study results demonstrated a significant improvement in the CNN-main model compared with the CNN-pilot model. The convergence and drop of the loss values toward epoch 10 were notably pronounced, and the achieved mean average precision of 91% aligns well with the overarching goal of the study. The enhanced recognition of uncomfortable posture positions by the CNN-main model suggests that users of persuasive technologies would be more conscious.

Furthermore, there was a substantial improvement in the performance of the Yolo-V3-main model compared with the Yolo-V3-pilot model. The increased precision around both comfortable and uncomfortable body positions resulted in a mean average precision of 92%. Considering these results, we address the main research question by recommending the following.

D1. Persuasive systems can be customized to detect the posture positions of users. While there are promising prospects with the CNN model, particularly with additional training datasets, the Yolo-V3 model stands out in addressing crucial body positions such as the eyes, face, head, neck, and arms. The successes of Yolo-V3 models have been reported in real-time workplace monitoring, showcasing its capability to report multiple and significant positions (Saumya et al., 2020).

D2. Persuasive systems based on the Yolo-V3 model can be trained to recognize various environmental conditions, such as the lighting conditions of the room, desk height, and leg position of users. While previous study by Min et al. (2015) demonstrated the potential of using sensor reading based on back and arm movements, expanding to recognize more positions would necessitate multimodal datasets, sensors, and strategically positioned cameras to provide users with comprehensive feedback. It is important to note that this approach may require privacy permissions. The importance of aligning such feedback with users' privacy expectations, both in private and social spaces, has been emphasized in the study by Brombacher et al. (2023). Additionally, a study by Bootsman et al. (2019) was limited to reading lumbar (back) posture data, overlooking other key postures that directly impact the back, as we have reported (eyes, head, neck, and arms).

TABLE 5 Summary of dataset distribution for the main study.

S/N	Model	Comfortable		Uncomfortable		Total	
		Training	Validation	Training	Validation	Training	Validation
1.	CNN	384	129	581	194	965	323
2	Yolo-V3	182	20	698	77	880	97
Total		**566**	**149**	**1,279**	**271**	**1,845**	**420**

D3. Persuasive systems based on the Yolo-V3 model can be trained to provide auditory feedback to users, particularly benefiting individuals with visual impairments. This customization could involve real-world feedback systems, such as a single beep sound for correct posture positions and a buzzer sound for incorrect posture positions. To enhance usability, additional concepts may be implemented, such as helping users locate body positions through a screen reader. Feedback systems, as reported in the study by Brombacher et al. (2023), have been recognized as effective in capturing users' attention, especially when working behind a desk and receiving posture-related feedback.

5.1 The present study vs. related studies

We present our methodology and results compared with existing studies. Deep learning models, compared with SVM and other algorithms used in existing studies (Tang et al., 2015; Nath et al., 2018; Mudiyanselage et al., 2021; Zhang and Callaghan, 2021), capture the variability of highly complex patterns in datasets. Hence, while SVM performs significantly better with small datasets, deep learning models require a substantial number of datasets. In a related study (Mudiyanselage et al., 2021), SVM yielded 99.5% with 54 datasets for five weightlifting classes (10, 15, 20, 30, and 35 lbs.). The results from this study showed significant overfitting of the SVM model. In addition, in a related study conducted by Nath et al. (2018) with 9,069 datasets for three classes of ergonomic weightlifting risks (low, moderate, and high), SVM achieved ~80% accuracy.

We employed deep learning models (CNN and Yolo-v3) in this study, considering the variability of good and bad posture patterns that SVM and other non-deep learning models might not significantly capture. While deep learning requires large datasets, we report on our findings (Yolo-v3: 92% and CNN: 91% accuracy values using 2,265 posture images for two classes, good and bad) to propose future work with additional datasets. In another related real-time study by Zhang and Callaghan (2021) with different human postures (sitting, walking, standing, running, and lying) using deep learning multi-layer perceptron (MLP), the authors reported accuracy up to 82% with few datasets (30 training and 19 testing samples). Nevertheless, results from the study by Tang et al. (2015) revealed a significant number of misclassifications. Deep neural networks (DNN) in a similar task of human gesture recognition achieved an accuracy of 98.12%. This level of accuracy was attained using a dataset comprising 21,600 images across 10 distinct classes of hand gestures. While Yolo-v3 compared with CNN has not been explored in previous study, our results present the baseline performance of both models to guide future work.

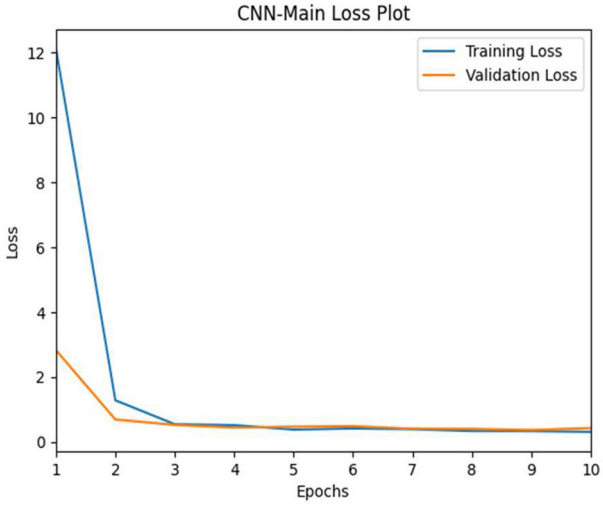

FIGURE 16
CNN-main model's training vs. validation loss.

5.2 Limitation of the study

While we report these significant findings of our study, we present the following limitations to improve future work. Though we found significant posture practices such as leg position and lying position, our findings are limited to the areas captured by the camera for sitting and standing body postures. Exploring these contexts further in future studies could inform the design of more wearable persuasive devices. In addition, our datasets are limited in size because there are a few instances of them publicly available. In the future, we will explore running experiments to collect additional ground truth datasets to enhance our model. In addition, to comprehensively assess the effectiveness of this technology in different workplaces (work-from-home, offices, and other spaces), a future study should include an evaluation of users' perceptions, considering both the advantages and disadvantages. We propose this framework as a valuable posture assessment tool which is applicable to any workplace setting, whether at home or in an office. Evaluating both contexts in future studies would contribute to a more comprehensive understanding of the applicability of technology. Finally, we had variations in the design of both models (YOLO-V3 and CNN); our comparisons might have favored YOLO-V3, especially with the dataset split ratio of 90% training and 10% validation sets. This is inconclusive at this point. We recommend that future studies explore setting the same standards for testing both models.

FIGURE 17
CNN-main model's posture detection.

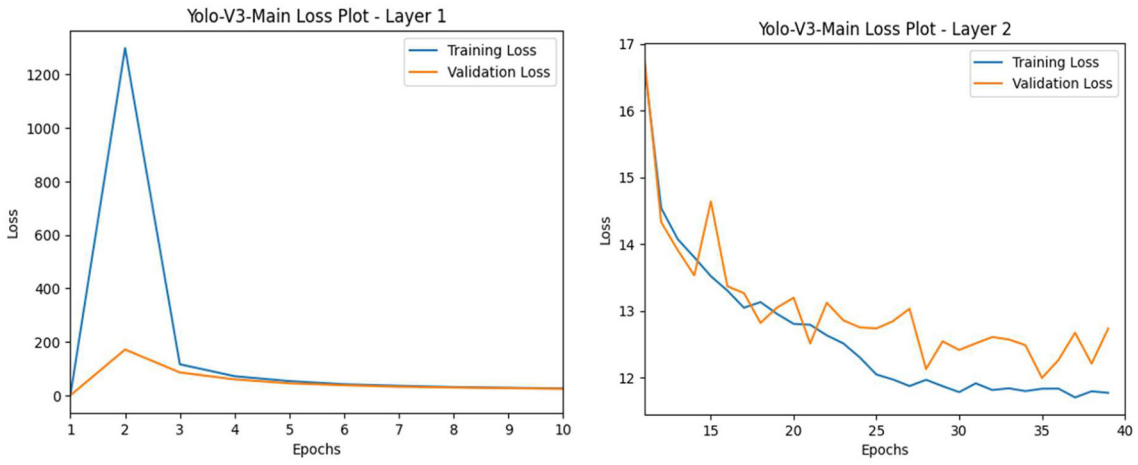

FIGURE 18
L-R: Yolo-V3-main model's training vs. validation loss (L: Layer 1 and R: Layer 2).

5.3 Implication of future design on system proximity detection and posture

Considering the prospects of posture evaluation based on proximity detection, we designed a system to integrate with our proposed Yolo-V3 and CNN models in the future. It is recommended that a computer user maintain 40 cm from the computer (Woo et al., 2016). To meet this requirement, we modified the proximity detection program by Harsh Jaggi[11] and presented the preliminary results, as shown in Figure 20.

11 https://www.linkedin.com/pulse/face-distance-measurement-python-haar-cascade-unlocking-harsh-jaggi

FIGURE 19
L-R: Yolo-V3-main model's posture detection: comfortable; uncomfortable; neutral.

FIGURE 20
Proximity detection of uncomfortable and comfortable posture.

6 Conclusion and future work

We explored potential designs for persuasive systems based on real-time posture detection. Given how significant persuasive systems and human factor engineering contribute to changing human behavior in workplaces, we conducted experiments using two deep learning models: convolutional neural networks (CNN) and Yolo-V3. These models have proven valuable in real-time detection of emotions, human activities, and behavior in previous research (Tan et al., 2021; Alsanad et al., 2022). Despite their effectiveness in various domains, little attention has been given to designing persuasive systems specifically for promoting proper postures in workplaces. Our overarching goal was to investigate how individuals can become more conscious of their posture practices while sitting and standing with a computer system. Additionally, we aimed to address the main research question: RQ: Can we design persuasive computers to detect unhealthy posture practices (such as sitting and standing) in workplaces?

Hence, based on the results of this study, we conclude with the following key insights:

1. Posture detection based on deep learning models would require a lot of datasets to implement.
2. Persuasive systems based on real-time posture detection should be tailored to capture more body positions. Overall, this helps to address more workplace requirements for behavioral changes.
3. There are prospects around eye strains, pupil datasets, and other contexts linked with stress. Hence, the framework of this study can be extended in the future.

In conclusion, our study highlights the potential for developing persuasive technologies that are specifically designed to support users in adhering to proper posture practices. The significance of this study prompts consideration for future exploration into themes such as more in-depth studies with large datasets, proximity detection, support for individuals with visual impairments in adopting optimal posture practices, eye strain detection, addressing various workplace requirements, and comparing outcomes of user studies with our technology from different workplaces such as work-from-home contexts, offices, and other ones.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary material, further inquiries can be directed to the corresponding author.

Ethics statement

Ethical approval was not required for the study involving human data in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required in accordance with the national legislation and the institutional requirements.

Author contributions

GA: Conceptualization, Data curation, Methodology, Writing – original draft, Writing – review & editing. RO: Supervision, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

The authors wish to acknowledge the efforts of colleagues who critically reviewed and provided insightful feedback on our study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdata.2024.1359906/full#supplementary-material

References

Abdullah, N. A. A., Rahmat, N. H., Zawawi, F. Z., Khamsah, M. A. N., and Anuarsham, A. H. (2020). Coping with post COVID-19: Can work from home be a new norm? *Eur. J. Soc. Sci. Stud.* 5:933. doi: 10.46827/ejsss.v5i6.933

Ahmad, H. F., Mukhtar, H., Alaqail, H., Seliaman, M., and Alhumam, A. (2021). Investigating health-related features and their impact on the prediction of diabetes using machine learning. *Appl. Sci.* 11:1173. doi: 10.3390/app11031173

Ahmetovic, D., Mascetti, S., Bernareggi, C., Guerreiro, J., Oh, U., and Asakawa, C. (2019). Deep learning compensation of rotation errors during navigation assistance for people with visual impairments or blindness. *ACM Trans. Access. Comput.* 12, 1–19. doi: 10.1145/3349264

Ahtinen, A., Andrejeff, E., Harris, C., and Väänänen, K. (2017). "Let's walk at work: persuasion through the brainwolk walking meeting app," in *Proceedings of the 21st International Academic Mindtrek Conference* (Tampere: ACM), 73–82.

Akter, S., Prodhan, R. A., Pias, T. S., Eisenberg, D., and Fresneda Fernandez, J. (2022). M1M2: deep-learning-based real-time emotion recognition from neural activity. *Sensors* 22:8467. doi: 10.3390/s22218467

Akut, R. R. (2019). FILM: finding the location of microaneurysms on the retina. *Biomed. Eng. Lett.* 9, 497–506. doi: 10.1007/s13534-019-00136-6

Alaydrus, L. L., and Nusraningrum, D. (2019). Awareness of workstation ergonomics and occurrence of computer-related injuries. *Ind. J. Publ. Health Res. Dev.* 10:9. doi: 10.5958/0976-5506.2019.04091.9

Alsanad, H. R., Sadik, A. Z., Ucan, O. N., Ilyas, M., and Bayat, O. (2022). YOLO-V3 based real-time drone detection algorithm. *Multimed. Tools Appl.* 81, 26185–26198. doi: 10.1007/s11042-022-12939-4

Anagnostopoulou, E., Magoutas, B., Bothos, E., and Mentzas, G. (2019). "Persuasive technologies for sustainable smart cities: the case of urban mobility," in *Companion Proceedings of The 2019 World Wide Web Conference* (San Francisco, CA: ACM), 73–82.

Anderson, S. P., and Oakman, J. (2016). Allied health professionals and work-related musculoskeletal disorders: a systematic review. *Saf. Health Work* 7, 259–267. doi: 10.1016/j.shaw.2016.04.001

Arani, E., Gowda, S., Mukherjee, R., Magdy, O., Kathiresan, S., and Zonooz, B. (2022). A comprehensive study of real-time object detection networks across multiple domains: a survey. *arXiv preprint arXiv:2208.10895*. doi: 10.48550/arXiv.2208.10895

Azimjonov, J., and Özmen, A. (2021). A real-time vehicle detection and a novel vehicle tracking systems for estimating and monitoring traffic flow on highways. *Adv. Eng. Informat.* 50:101393. doi: 10.1016/j.aei.2021.101393

Baba, E. I., Baba, D. D., and Oborah, J. O. (2021). Effect of office ergonomics on office workers' productivity in the polytechnics, Nigeria. *J. Educ. Pract.* 12, 67–75. doi: 10.7176/JEP/12-3-10

Bailly, G., Sahdev, S., Malacria, S., and Pietrzak, T. (2016). "LivingDesktop: augmenting desktop workstation with actuated devices," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (ACM), 5298–5310.

Barrett, J. M., McKinnon, C., and Callaghan, J. P. (2020). Cervical spine joint loading with neck flexion. *Ergonomics* 63, 101–108. doi: 10.1080/00140139.2019.1677944

Bartlett, Y. K., Webb, T. L., and Hawley, M. S. (2017). Using persuasive technology to increase physical activity in people with chronic obstructive pulmonary disease by encouraging regular walking: a mixed-methods study exploring opinions and preferences. *J. Med. Internet Res.* 19:e124. doi: 10.2196/jmir.6616

Bavankumar, S., Rajalingam, B., Santhoshkumar, R., JawaherlalNehru, G., Deepan, P., Balaraman, N., et al. (2021). A real time prediction and classification of face mask detection using CNN model. *Turk. Online J. Qual. Inquiry* 12, 7282–7292.

Beheshtian, N., Moradi, S., Ahtinen, A., Väänanen, K., Kähkonen, K., and Laine, M. (2020). "Greenlife: a persuasive social robot to enhance the sustainable behavior in shared living spaces," in *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society* (ACM), 1–12.

Berque, D., Burgess, J., Billingsley, A., Johnson, S., Bonebright, T. L., and Wethington, B. (2011). "Design and evaluation of persuasive technology to encourage healthier typing behaviors," in *Proceedings of the 6th International Conference on Persuasive Technology: Persuasive Technology and Design: Enhancing Sustainability and Health*, 1–10.

Boadi-Kusi, S. B., Adueming, P. O. W., Hammond, F. A., and Antiri, E. O. (2022). Computer vision syndrome and its associated ergonomic factors among bank workers. *Int. J. Occup. Saf. Ergon.* 28, 1219–1226. doi: 10.1080/10803548.2021.1897260

Bootsman, R., Markopoulos, P., Qi, Q., Wang, Q., and Timmermans, A. A. (2019). Wearable technology for posture monitoring at the workplace. *Int. J. Hum. Comput. Stud.* 132, 99–111. doi: 10.1016/j.ijhcs.2019.08.003

Borhany, T., Shahid, E., Siddique, W. A., and Ali, H. (2018). Musculoskeletal problems in frequent computer and internet users. *J. Fam. Med. Prim. Care* 7, 337–339. doi: 10.4103/jfmpc.jfmpc_326_17

Botter, J., Ellegast, R. P., Burford, E. M., Weber, B., Könemann, R., and Commissaris, D. A. (2016). Comparison of the postural and physiological effects of two dynamic workstations to conventional sitting and standing workstations. *Ergonomics* 59, 449–463. doi: 10.1080/00140139.2015.1080861

Brigato, L., and Iocchi, L. (2021). "A close look at deep learning with small data," in *2020 25th International Conference on Pattern Recognition (ICPR)* (Milan: IEEE), 2490–2497.

Brik, B., Esseghir, M., Merghem-Boulahia, L., and Snoussi, H. (2021). An IoT-based deep learning approach to analyse indoor thermal comfort of disabled people. *Build. Environ.* 203:108056. doi: 10.1016/j.buildenv.2021.108056

Brombacher, H., Houben, S., and Vos, S. (2023). Tangible interventions for office work well-being: approaches, classification, and design considerations. *Behav. Inform. Technol.* 2023, 1–25. doi: 10.1080/0144929X.2023.2241561

Catanzarite, T., Tan-Kim, J., Whitcomb, E. L., and Menefee, S. (2018). Ergonomics in surgery: a review. *Urogynecology* 24, 1–12. doi: 10.1097/SPV.0000000000000456

Chandra, R., Bera, A., and Manocha, D. (2021). Using graph-theoretic machine learning to predict human driver behavior. *IEEE Trans. Intell. Transport. Syst.* 23, 2572–2585. doi: 10.1109/TITS.2021.3130218

Chen, J., Wu, J., Richter, K., Konrad, J., and Ishwar, P. (2016). "Estimating head pose orientation using extremely low resolution images," in *2016 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI)* (Santa Fe, NM: IEEE), 65–68.

Chen, W., and Yeo, C. K. (2019). "Unauthorized parking detection using deep networks at real time," in *2019 IEEE International Conference on Smart Computing (SMARTCOMP)* (Washington, DC: IEEE), 459–463.

Cheng, L., Guan, Y., Zhu, K., and Li, Y. (2017). "Recognition of human activities using machine learning methods with wearable sensors," in *2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC)* (Las Vegas, NV: IEEE), 1–7. doi: 10.1109/CCWC.2017.7868369

Christa, G. H., Jesica, J., Anisha, K., and Sagayam, K. M. (2021). "CNN-based mask detection system using openCV and MobileNetV2," in *2021 3rd International Conference on Signal Processing and Communication (ICPSC)* (Coimbatore: IEEE), 115–119.

Cob-Parro, A. C., Losada-Gutiérrez, C., Marrón-Romera, M., Gardel-Vicente, A., and Bravo-Muñoz, I. (2023). A new framework for deep learning video based Human Action Recognition on the edge. *Expert Syst. Appl.* 2023:122220. doi: 10.1016/j.eswa.2023.122220

Dainoff, M., Maynard, W., Robertson, M., and Andersen, J. H. (2012). Office ergonomics. *Handb. Hum. Fact. Ergon.* 56, 1550–1573. doi: 10.1002/9781118131350.ch56

Damen, I., Heerkens, L., Van Den Broek, A., Drabbels, K., Cherepennikova, O., Brombacher, H., et al. (2020a). "PositionPeak: stimulating position changes during meetings," in *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (ACM), 1–8.

Damen, I., Kok, A., Vink, B., Brombacher, H., Vos, S., and Lallemand, C. (2020b). "The hub: facilitating walking meetings through a network of interactive devices," in *Companion Publication of the 2020 ACM Designing Interactive Systems Conference* (ACM), 19–24.

Dang, K. B., Nguyen, M. H., Nguyen, D. A., Phan, T. T. H., Giang, T. L., Pham, H. H., et al. (2020). Coastal wetland classification with deep u-net convolutional networks and sentinel-2 imagery: a case study at the tien yen estuary of vietnam. *Remote Sens.* 12:3270. doi: 10.3390/rs12193270

Darioshi, R., and Lahav, E. (2021). The impact of technology on the human decision-making process. *Hum. Behav. Emerg. Technol.* 3, 391–400. doi: 10.1002/hbe2.257

Diwate, R. B., Zagade, A., Khodaskar, M. R., and Dange, V. R. (2022). "Optimization in object detection model using YOLO.v3," in *2022 International Conference on Emerging Smart Computing and Informatics (ESCI)* (Pune: IEEE), 1–4.

Dubey, S. R., Singh, S. K., and Chaudhuri, B. B. (2022). Activation functions in deep learning: a comprehensive survey and benchmark. *Neurocomputing* 503, 92–108. doi: 10.1016/j.neucom.2022.06.111

Ergonomics (2023). *Ergonomics in the Work Environment. Available online at: https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=dc7357a6b312d394785c2f6beb0fcef29fd9e584 (accessed November* 5, 2023).

Faddoul, G., and Chatterjee, S. (2019). "The virtual diabetician: a prototype for a virtual avatar for diabetes treatment using persuasion through storytelling," in *Proceedings of the 25th Americas Conference on Information Systems* (Cancún), 1–10.

Franke, M., and Nadler, C. (2021). Towards a holistic approach for assessing the impact of IEQ on satisfaction, health, and productivity. *Build. Res. Inform.* 49, 417–444. doi: 10.1080/09613218.2020.1788917

Fukuoka, Y., Haskell, W., Lin, F., and Vittinghoff, E. (2019). Short-and long-term effects of a mobile phone app in conjunction with brief in-person counseling on physical activity among physically inactive women: the mPED randomized clinical trial. *J. Am. Med. Assoc. Netw. Open* 2:e194281. doi: 10.1001/jamanetworkopen.2019.4281

Gill, R., and Singh, J. (2021). "A deep learning approach for real time facial emotion recognition," in *2021 10th International Conference on System Modeling & Advancement in Research Trends (SMART)* (Moradabad: IEEE), 497–501.

Gomez-Carmona, O., and Casado-Mansilla, D. (2017). "SmiWork: an interactive smart mirror platform for workplace health promotion," in *2017 2nd International Multidisciplinary Conference on Computer and Energy Science (SpliTech)* (Split: IEEE), 1–6.

Haliburton, L., Kheirinejad, S., Schmidt, A., and Mayer, S. (2023). Exploring smart standing desks to foster a healthier workplace. *Proc. ACM Interact. Mob. Wear. Ubiquit. Technol.* 7, 1–22. doi: 10.1145/3596260

Han, D., Liu, Q., and Fan, W. (2018). A new image classification method using CNN transfer learning and web data augmentation. *Expert Syst. Appl.* 95, 43–56. doi: 10.1016/j.eswa.2017.11.028

Haque, M. S., Kangas, M., and Jämsä, T. (2020). A persuasive mHealth behavioral change intervention for promoting physical activity in the workplace: feasibility randomized controlled trial. *JMIR Form. Res.* 4:e15083. doi: 10.2196/preprints.15083

Huang, R., Gu, J., Sun, X., Hou, Y., and Uddin, S. (2019). A rapid recognition method for electronic components based on the improved YOLO-V3 network. *Electronics* 8:825. doi: 10.3390/electronics8080825

Iyengar, K., Upadhyaya, G. K., Vaishya, R., and Jain, V. (2020). COVID-19 and applications of smartphone technology in the current pandemic. *Diabet. Metabol. Syndr.* 14, 733–737. doi: 10.1016/j.dsx.2020.05.033

Jafarinaimi, N., Forlizzi, J., Hurst, A., and Zimmerman, J. (2005). "Breakaway: an ambient display designed to change human behavior," in *CHI'05 Extended Abstracts on Human Factors in Computing Systems* (ACM), 1945–1948.

Jaiswal, S., and Nandi, G. C. (2020). Robust real-time emotion detection system using CNN architecture. *Neural Comput. Appl.* 32, 11253–11262. doi: 10.1007/s00521-019-04564-4

Javad Koohsari, M., Nakaya, T., Shibata, A., Ishii, K., and Oka, K. (2021). Working from home after the COVID-19 pandemic: do company employees sit more and move less? *Sustainability* 13:939. doi: 10.3390/su13020939

Jiang, M., Nanjappan, V., Liang, H. N., and ten Bhömer, M. (2021). *In-situ* exploration of emotion regulation via smart clothing: an empirical study of healthcare workers in their work environment. *Behav. Inform. Technol.* 2021, 1–14. doi: 10.1080/0144929X.2021.1975821

Jin, C. J., Shi, X., Hui, T., Li, D., and Ma, K. (2021). The automatic detection of pedestrians under the high-density conditions by deep learning techniques. *J. Adv. Transport.* 2021, 1–11. doi: 10.1155/2021/1396326

Johnson, A., Dey, S., Nguyen, H., Groth, M., Joyce, S., Tan, L., et al. (2020). A review and agenda for examining how technology-driven changes at work will impact workplace mental health and employee well-being. *Austr. J. Manag.* 45, 402–424. doi: 10.1177/0312896220922292

Jupalle, H., Kouser, S., Bhatia, A. B., Alam, N., Nadikattu, R. R., and Whig, P. (2022). Automation of human behaviors and its prediction using machine learning. *Microsyst. Working Technol.* 28, 1879–1887. doi: 10.1007/s00542-022-05326-4

Karppinen, P., Oinas-Kukkonen, H., Alahäivälä, T., Jokelainen, T., Keränen, A. M., Salonurmi, T., et al. (2016). Persuasive user experiences of a health Behavior Change Support System: a 12-month study for prevention of metabolic syndrome. *Int. J. Med. Informat.* 96, 51–61. doi: 10.1016/j.ijmedinf.2016.02.005

Kember, S. (2014). Face recognition and the emergence of smart photography. *J. Vis. Cult.* 13, 182–199. doi: 10.1177/1470412914541767

Kim, M. T., Kim, K. B., Nguyen, T. H., Ko, J., Zabora, J., Jacobs, E., et al. (2019). Motivating people to sustain healthy lifestyles using persuasive technology: a pilot study of Korean Americans with prediabetes and type 2 diabetes. *Pat. Educ. Counsel.* 102, 709–717. doi: 10.1016/j.pec.2018.10.021

Kim, W., Lorenzini, M., Balatti, P., Nguyen, P. D., Pattacini, U., Tikhanoff, V., et al. (2019). Adaptable workstations for human-robot collaboration: a reconfigurable framework for improving worker ergonomics and productivity. *IEEE Robot. Automat. Mag.* 26, 14–26. doi: 10.1109/MRA.2018.2890460

Ko Ko, T., Dickson-Gomez, J., Yasmeen, G., Han, W. W., Quinn, K., Beyer, K., et al. (2020). Informal workplaces and their comparative effects on the health of street vendors and home-based garment workers in Yangon, Myanmar: a qualitative study. *BMC Publ. Health* 20, 1–14. doi: 10.1186/s12889-020-08624-6

Krajnak, K. (2018). Health effects associated with occupational exposure to hand-arm or whole body vibration. *J. Toxicol. Environ. Health B* 21, 320–334. doi: 10.1080/10937404.2018.1557576

Krishna, K., Jain, D., Mehta, S. V., and Choudhary, S. (2018). An lstm based system for prediction of human activities with durations. *Proc. ACM Interact. Mob. Wear. Ubiquit. Technol.* 1, 1–31. doi: 10.1145/3161201

Kronenberg, R., and Kuflik, T. (2019). "Automatically adjusting computer screen," in *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, 51–56.

Kronenberg, R., Kuflik, T., and Shimshoni, I. (2022). Improving office workers' workspace using a self-adjusting computer screen. *ACM Transactions on Interactive Intelligent Systems (TiiS).* 12, 1–32. doi: 10.1145/3545993

Kulyukin, V. A., and Gharpure, C. (2006). "Ergonomics-for-one in a robotic shopping cart for the blind," in *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction* (ACM), 142–149.

Li, J., Li, X., and He, D. (2019). A directed acyclic graph network combined with CNN and LSTM for remaining useful life prediction. *IEEE Access* 7, 75464–75475. doi: 10.1109/ACCESS.2019.2919566

Li, Y., Zhao, Z., Luo, Y., and Qiu, Z. (2020). Real-time pattern-recognition of GPR images with YOLO v3 implemented by tensorflow. *Sensors* 20:6476. doi: 10.3390/s20226476

Lienhart, R., Pfeiffer, S., and Effelsberg, W. (1997). Video abstracting. *Commun. ACM* 40, 54–62. doi: 10.1145/265563.265572

Liu, B., Su, S., and Wei, J. (2022). The effect of data augmentation methods on pedestrian object detection. *Electronics* 11:3185. doi: 10.3390/electronics11193185

Lu, S., Wang, B., Wang, H., Chen, L., Linjian, M., and Zhang, X. (2019). A real-time object detection algorithm for video. *Comput. Electr. Eng.* 77, 398–408. doi: 10.1016/j.compeleceng.2019.05.009

Ludden, G. D., and Meekhof, L. (2016). "Slowing down: introducing calm persuasive technology to increase wellbeing at work," in *Proceedings of the 28th Australian Conference on Computer-Human Interaction* (ACM), 435–441.

Mahesh, B. (2020). Machine learning algorithms-a review. *Int. J. Sci. Res.* 9, 381–386. doi: 10.21275/ART20203995

Mateevitsi, V., Reda, K., Leigh, J., and Johnson, A. (2014). "The health bar: a persuasive ambient display to improve the office worker's well being," in *Proceedings of the 5th Augmented Human International Conference* (ACM), 1–2.

Min, D. A., Kim, Y., Jang, S. A., Kim, K. Y., Jung, S. E., and Lee, J. H. (2015). "Pretty pelvis: a virtual pet application that breaks sedentary time by promoting gestural interaction," in *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems* (ACM), 1259–1264.

Mohadis, H. M., Mohamad Ali, N., and Smeaton, A. F. (2016). Designing a persuasive physical activity application for older workers: understanding end-user perceptions. *Behav. Technol.* 35, 1102–1114. doi: 10.1080/0144929X.2016.1211737

Moore, P. V. (2019). "OSH and the future of work: benefits and risks of artificial intelligence tools in workplaces," in *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Human Body and Motion: 10th International Conference, DHM 2019, Held as Part of the 21st HCI International Conference, HCII 2019, Orlando, FL, USA, July 26-31. 2019, Proceedings, Part I 21* (Berlin: Springer International Publishing), 292–315.

Mowatt, L., Gordon, C., Santosh, A. B. R., and Jones, T. (2018). Computer vision syndrome and ergonomic practices among undergraduate university students. *Int. J. Clin. Practice* 72:e13035. doi: 10.1111/ijcp.13035

Mudiyanselage, S. E., Nguyen, P. H. D., Rajabi, M. S., and Akhavian, R. (2021). Automated workers' ergonomic risk assessment in manual material handling using sEMG wearable sensors and machine learning. *Electronics* 10:2558. doi: 10.3390/electronics10202558

Mujumdar, A., and Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms. *Proc. Comput. Sci.* 165, 292–299. doi: 10.1016/j.procs.2020.01.047

Nanthavanij, S., Jalil, S., and Ammarapala, V. (2008). Effects of body height, notebook computer size, and workstation height on recommended adjustments for proper work posture when operating a notebook computer. *J. Hum. Ergol.* 37, 67–81. doi: 10.11183/jhe1972.37.67

Nath, N. D., Chaspari, T., and Behzadan, A. H. (2018). Automated ergonomic risk monitoring using body-mounted sensors and machine learning. *Adv. Eng. Informat.* 38, 514–526. doi: 10.1016/j.aei.2018.08.020

Nimbarte, A. D., Sivak-Callcott, J. A., Zreiqat, M., and Chapman, M. (2013). Neck postures and cervical spine loading among microsurgeons operating with loupes and headlamp. *IIE Trans. Occup. Erg. Hum. Fact.* 1, 215–223. doi: 10.1080/21577323.2013.840342

Ofori-Manteaw, B. B., Antwi, W. K., and Arthur, L. (2015). Ergonomics and occupational health issues in diagnostic imaging: a survey of the situation at the Korle-Bu Teaching Hospital. *Ergonomics* 19, 93–101.

Ogundokun, R. O., Maskeliunas, R., Misra, S., and Damaševičius, R. (2022). "Improved CNN based on batch normalization and adam optimizer," in *International Conference on Computational Science and Its Applications* (Cham: Springer International Publishing), 593–604.

Orji, R., Tondello, G. F., and Nacke, L. E. (2018). "Personalizing persuasive strategies in gameful systems to gamification user types," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (ACM), 1–14.

Oyibo, K., and Morita, P. P. (2021). Designing better exposure notification apps: the role of persuasive design. *JMIR Publ. Health Surveill.* 7:e28956. doi: 10.2196/28956

Paay, J., Kjeldskov, J., Papachristos, E., Hansen, K. M., Jørgensen, T., and Overgaard, K. L. (2022). Can digital personal assistants persuade people to exercise? *Behav. Inform. Technol.* 41, 416–432. doi: 10.1080/0144929X.2020.1814412

Padilla, R., Passos, W. L., Dias, T. L., Netto, S. L., and Da Silva, E. A. (2021). A comparative analysis of object detection metrics with a companion open-source toolkit. *Electronics* 10:279. doi: 10.3390/electronics10030279

Perazzi, F., Khoreva, A., Benenson, R., Schiele, B., and Sorkine-Hornung, A. (2017). "Learning video object segmentation from static images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE), 2663–2672.

Pereira, M., Comans, T., Sjøgaard, G., Straker, L., Melloh, M., O'Leary, S., et al. (2019). The impact of workplace ergonomics and neck-specific exercise versus ergonomics and health promotion interventions on office worker productivity: a cluster-randomized trial. *Scand. J. Work Environ. Health* 45, 42–52. doi: 10.5271/sjweh.3760

Rabbi, M., Pfammatter, A., Zhang, M., Spring, B., and Choudhury, T. (2015). Automated personalized feedback for physical activity and dietary behavior change with mobile phones: a randomized controlled trial on adults. *JMIR mHealth uHealth* 3:e4160. doi: 10.2196/mhealth.4160

Rapoport, M. (2017). Persuasive robotic technologies and the freedom of choice and action. *Soc. Robot.* 12, 219–238. doi: 10.4324/9781315563084-12

Reddy, U. S., Thota, A. V., and Dharun, A. (2018). "Machine learning techniques for stress prediction in working employees," in *2018 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)* (Madurai: IEEE), 1–4.

Reeder, S., Kelly, L., Kechavarzi, B., and Sabanovic, S. (2010). "Breakbot: a social motivator for the workplace," in *Proceedings of the 8th ACM Conference on Designing Interactive Systems* (ACM), 61–64.

Ren, X., Yu, B., Lu, Y., Zhang, B., Hu, J., and Brombacher, A. (2019). LightSit: an unobtrusive health-promoting system for relaxation and fitness microbreaks at work. *Sensors* 19:2162. doi: 10.3390/s19092162

Robledo Yamamoto, F., Cho, J., Voida, A., and Voida, S. (2023). "We are researchers, but we are also humans": creating a design space for managing graduate student stress. *ACM Trans. Comput. Hum. Interact.* 30, 1–33. doi: 10.1145/3589956

Roy, D. (2022). "Occupational health services and prevention of work-related musculoskeletal problems," in *Handbook on Management and Employment Practices* (Cham: Springer International Publishing), 547–571.

Sarker, I. H. (2021). Machine learning: algorithms, real-world applications and research directions. *SN Comput. Sci.* 2:160. doi: 10.1007/s42979-021-00592-x

Sarla, G. S. (2019). Excessive use of electronic gadgets: health effects. *Egypt. J. Intern. Med.* 31, 408–411. doi: 10.4103/ejim.ejim_56_19

Saumya, A., Gayathri, V., Venkateswaran, K., Kale, S., and Sridhar, N. (2020). "Machine learning based surveillance system for detection of bike riders without helmet and triple rides," in *2020 International Conference on Smart Electronics and Communication (ICOSEC)* (Trichy: IEEE), 347–352.

Schnall, R., Bakken, S., Rojas, M., Travers, J., and Carballo-Dieguez, A. (2015). mHealth technology as a persuasive tool for treatment, care and management of persons living with HIV. *AIDS Behav.* 19, 81–89. doi: 10.1007/s10461-014-0984-8

Schooley, B., Akgun, D., Duhoon, P., and Hikmet, N. (2021). "Persuasive AI voice-assisted technologies to motivate and encourage physical activity," in *Advances in Computer Vision and Computational Biology: Proceedings from IPCV'20, HIMS'20, BIOCOMP'20, and BIOENG'20* (Cham: Springer International Publishing), 363–384.

Setyadi, A., Kallista, M., Setianingsih, C., and Araffathia, R. (2023). "Deep learning approaches to social distancing compliance and mask detection in dining environment," in *2023 IEEE Asia Pacific Conference on Wireless and Mobile (APWiMob)* (IEEE), 188–194.

Shahidi, B., Curran-Everett, D., and Maluf, K. S. (2015). Psychosocial, physical, and neurophysiological risk factors for chronic neck pain: a prospective inception cohort study. *J. Pain* 16, 1288–1299. doi: 10.1016/j.jpain.2015.09.002

Shorten, C., and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *J. Big Data* 6, 1–48. doi: 10.1186/s40537-019-0197-0

Silva, S. M., and Jung, C. R. (2021). A flexible approach for automatic license plate recognition in unconstrained scenarios. *IEEE Trans. Intell. Transport. Syst.* 23, 5693–5703. doi: 10.1109/TITS.2021.3055946

Singh, A. P., and Agarwal, D. (2022). "Webcam motion detection in real-time using Python," in *2022 International Mobile and Embedded Technology Conference (MECON)* (Noida: IEEE), 1–4.

Sonntag, D. (2016). "Persuasive AI technologies for healthcare systems," in *2016 AAAI Fall Symposium Series.* (Stanford, CA; Washington, DC: AAAI Press).

Steiger, M., Bharucha, T. J., Venkatagiri, S., Riedl, M. J., and Lease, M. (2021). "The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (ACM), 1–14.

Tan, L., Huangfu, T., Wu, L., and Chen, W. (2021). Comparison of YOLO v3, faster R-CNN, and SSD for real-time pill identification. *Res. Square.* doi: 10.21203/rs.3.rs-668895/v1

Tang, A., Lu, K., Wang, Y., Huang, J., and Li, H. (2015). A real-time hand posture recognition system using deep neural networks. *ACM Trans. Intell. Syst. Technol.* 6, 1–23. doi: 10.1145/2735952

Tang, K. H. D. (2022). The prevalence, causes and prevention of occupational musculoskeletal disorders. *Glob. Acad. J. Med. Sci.* 4, 56–68. doi: 10.36348/gajms.2022.v04i02.004

Tauchert, C., Buxmann, P., and Lambinus, J. (2020). "Crowdsourcing data science: a qualitative analysis of organizations' usage of kaggle competitions," in *Proceedings of the 53rd Hawaii International Conference on System Sciences*, 229–238.

van de Wijdeven, B., Visser, B., Daams, J., and Kuijer, P. P. (2023). A first step towards a framework for interventions for individual working practice to prevent work-related musculoskeletal disorders: a scoping review. *BMC Musculoskelet. Disord.* 24:87. doi: 10.1186/s12891-023-06155-w

Wang, R., Bush-Evans, R., Arden-Close, E., Bolat, E., McAlaney, J., Hodge, S., et al. (2023). Transparency in persuasive technology, immersive technology, and online marketing: facilitating users' informed decision making and practical implications. *Comput. Hum. Behav.* 139:107545. doi: 10.1016/j.chb.2022.107545

Wong, T. L., Chou, K. S., Wong, K. L., and Tang, S. K. (2023). Dataset of public objects in uncontrolled environment for navigation aiding. *Data* 8:42. doi: 10.3390/data8020042

Woo, E. H. C., White, P., and Lai, C. W. K. (2016). Ergonomics standards and guidelines for computer workstation design and the impact on users' health-a review. *Ergonomics* 59, 464–475. doi: 10.1080/00140139.2015.1076528

Workineh, S. A., and Yamaura, H. (2016). Multi-position ergonomic computer workstation design to increase comfort of computer work. *Int. J. Indus. Erg.* 53, 1–9. doi: 10.1016/j.ergon.2015.10.005

Xu, J., Li, Z., Du, B., Zhang, M., and Liu, J. (2020). "Reluplex made more practical: leaky ReLU," in *2020 IEEE Symposium on Computers and Communications (ISCC)* (Rennes: IEEE), 1–7.

Xu, X., Wang, J., Peng, H., and Wu, R. (2019). Prediction of academic performance associated with internet usage behaviors using machine learning algorithms. *Comput. Hum. Behav.* 98, 166–173. doi: 10.1016/j.chb.2019.04.015

Zhang, S., and Callaghan, V. (2021). Real-time human posture recognition using an adaptive hybrid classifier. *Int. J. Machine Learn. Cybernet.* 12, 489–499. doi: 10.1007/s13042-020-01182-8

Zhang, Y., Ma, B., Hu, Y., Li, C., and Li, Y. (2022). Accurate cotton diseases and pests detection in complex background based on an improved YOLOX model. *Comput. Electr. Agri.* 203:107484. doi: 10.1016/j.compag.2022.107484

frontiers | Frontiers in Virtual Reality

TYPE Original Research
PUBLISHED 17 June 2024
DOI 10.3389/frvir.2024.1364207

OPEN ACCESS

EDITED BY
Daniele Giunchi,
University College London, United Kingdom

REVIEWED BY
Diego Vilela Monteiro,
ESIEA University, France
Nitesh Bhatia,
Imperial College London, United Kingdom

*CORRESPONDENCE
Murat Yalcin,
murat.yalcin@uni-wuerzburg.de

RECEIVED 01 January 2024
ACCEPTED 15 April 2024
PUBLISHED 17 June 2024

CITATION
Yalcin M, Halbig A, Fischbach M and Latoschik ME (2024), Automatic cybersickness detection by deep learning of augmented physiological data from off-the-shelf consumer-grade sensors.
Front. Virtual Real. 5:1364207.
doi: 10.3389/frvir.2024.1364207

Automatic cybersickness detection by deep learning of augmented physiological data from off-the-shelf consumer-grade sensors

Murat Yalcin*, Andreas Halbig, Martin Fischbach and Marc Erich Latoschik

Human-Computer Interaction (HCI) Group, University of Würzburg, Würzburg, Germany

Cybersickness is still a prominent risk factor potentially affecting the usability of virtual reality applications. Automated real-time detection of cybersickness promises to support a better general understanding of the phenomena and to avoid and counteract its occurrence. It could be used to facilitate application optimization, that is, to systematically link potential causes (technical development and conceptual design decisions) to cybersickness in closed-loop user-centered development cycles. In addition, it could be used to monitor, warn, and hence safeguard users against any onset of cybersickness during a virtual reality exposure, especially in healthcare applications. This article presents a novel real-time-capable cybersickness detection method by deep learning of augmented physiological data. In contrast to related preliminary work, we are exploring a unique combination of mid-immersion ground truth elicitation, an unobtrusive wireless setup, and moderate training performance requirements. We developed a proof-of-concept prototype to compare (combinations of) convolutional neural networks, long short-term memory, and support vector machines with respect to detection performance. We demonstrate that the use of a conditional generative adversarial network-based data augmentation technique increases detection performance significantly and showcase the feasibility of real-time cybersickness detection in a genuine application example. Finally, a comprehensive performance analysis demonstrates that a four-layered bidirectional long short-term memory network with the developed data augmentation delivers superior performance (91.1% F1-score) for real-time cybersickness detection. To encourage replicability and reuse in future cybersickness studies, we released the code and the dataset as publicly available.

KEYWORDS

virtual reality, cybersickness detection, deep learning, data augmentation, CGAN, physiological signals, data processing, sensors

1 Introduction

Today, virtual reality (VR) is used in many different application areas. VR has shown its potential for gaming (Pallavicini et al., 2019), teaching and learning (Oberdörfer et al., 2017; Checa and Bustillo, 2020), tourism and hospitality (Huang et al., 2016), and marketing and advertising (Alcañiz et al., 2019; Loureiro et al., 2019). The power and benefits of VR are particularly prominent in the field of therapy. For example, VR can be used in psychology to treat fear of heights (Abdullah and Shaikh, 2018; Bălan et al., 2020), of spiders Hildebrandt et al. (2016); Miloff et al. (2016); Lindner et al. (2020), of speaking in front of an audience (Barreda-Ángeles et al., 2020; Glémarec et al., 2022), or of disorders of body perception by leveraging personalized photorealistic avatars (Wolf et al., 2021; 2020). It is also used to treat neurological disorders, for example, gait impairments as a result of Parkinson's disease or strokes (Hamzeheinejad et al., 2019; Kern et al., 2019), as well as in orthopedics for the physical recovery after surgery (Gianola et al., 2020; Bartl et al., 2022; Gazendam et al., 2022).

While the areas of application for the utilization of VR technology constantly increase, immersive VR applications, in particular, still face the risk of potentially inducing cybersickness (CS). CS is a prominent risk factor potentially affecting the usability of VR applications (Chang et al., 2020; Stauffert et al., 2020), which is exceptionally critical for medical applications. Hence, to avoid and/or counteract potential occurrences of CS, we first need reliable methods to measure and detect CS. Measuring the occurrence and severity of CS is often done with subjective self-reports (Kennedy et al., 1993; Keshavarz and Hecht, 2011). Using such questionnaire tools, however, has notable drawbacks. Most prominently, it requires active user feedback, potentially inducing distraction and additional workload or breaking the current immersion and flow.

Here, alternative approaches to measuring CS use physiological and behavioral data, for example, using heart rate, skin conductance, electroencephalography (EEG), or eye-tracking data (Nakagawa, 2015; Dennison et al., 2016; Garcia-Agundez et al., 2019; Kim et al., 2019; Islam et al., 2020b; Tauscher et al., 2020). However, many of the existing solutions need an extensive setup (Jeong et al., 2018; Garcia-Agundez et al., 2019; Kim et al., 2019; Lee et al., 2019; Tauscher et al., 2020). Such elaborated setups and expensive devices render a widespread adaptation of objective CS detection unlikely for many use-cases. With this work, we address these problems and show how a CS detection that is based on a very simple setup can be realized. In our approach, we use wearable of-the-shelf sensors and the data provided by a standalone VR headset to achieve a reliable detection of CS. We apply a deep-learning-based data augmentation technique to achieve a significant improvement in CS detection even for smaller and imbalanced datasets.

1.1 Contribution

We first conducted a data-collection process with 20 participants who rode a VR rollercoaster while giving feedback about possible onsets of CS using a controller. We collected several types of physiological data using three different easy-to-use wearable sensors. We analyzed and compared the resulting data with different deep learning algorithms, aiming for automatic real-time detection of CS. Specifically, we used standard and bidirectional long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997; Schuster and Paliwal, 1997), a combination of convolutional neural networks (CNN) and LSTM, and a support vector machine (SVM) (Cortes and Vapnik, 1995) model for CS detection. Comprehensive performance analysis showed the highest accuracy for a four-layered bidirectional LSTM model, achieving 84.2% accuracy for our original dataset. To enhance detection performance, we pioneered the application of conditional generative adversarial networks (cGAN) to augment physiological time-series data in CS detection. The results increased to 91.7% accuracy and show that it is possible to detect the onset of CS with a fairly simple, unobtrusive setup based on wearable devices without the need for more complex electrode-based sensors and without a large dataset. The detection quality is higher than that in the previous works (Martin et al., 2020; Islam et al., 2021). However, we also propose that a mere accuracy metric is insufficient to evaluate a model's robustness and feasibility. Accordingly, we computed more detailed metrics that further confirmed the excellent performance of our developed method for detecting CS.

2 Related work

2.1 Phenomenology, causes, theories, and prevention of cybersickness

Cybersickness refers to symptoms accompanying VR applications, ranging from headache, dizziness, eyestrain, and blurred vision to nausea and vomiting (LaViola Jr, 2000; Sharples et al., 2008). CS is closely related to simulator sickness as they share many symptoms (Rebenitsch and Owen, 2016). However, Stanney et al. (1997) argue that the tow conditions have different profiles. While the sickness that occurs in simulators is mainly determined by oculomotor symptoms, the main symptom of CS is disorientation. Additionally, the symptoms of CS are approximately three times more severe than those of simulator sickness (Stanney et al., 1997).

CS and simulator sickness share not only a set of symptoms but also common origin theories because many of the theories that apply to simulators could be transferred relatively easily to head-mounted displays (HMDs) (see (Rebenitsch and Owen, 2016) for an overview). The sensory mismatch theory suggests that people experiencing VR receive input on different modalities that might be incongruent or conflicting, for example, visual and vestibular input (Oman, 1990). Because such incongruencies could have been triggered by toxins in the evolutionary history of humans, CS and simulator sickness could also be protective survival mechanisms of the body, deployed in the wrong context (Treisman, 1977). Another common theory references postural instability. It is similar to the sensory conflict theory and suggests that sickness symptoms occur in situations where humans do not have an effective strategy to maintain postural stability (Riccio and Stoffregen, 1991). When a person is using immersive technology, they may not receive the usual sensory input that helps them maintain their balance and posture (Chen Y.-C. et al., 2011). Possible triggers and causes for CS are also very diverse.

On the content level, one of the biggest factors is the optical flow. It is more likely for people to show sickness symptoms when they see

moving visual content instead of static content (Chen W. et al., 2011; Lubeck et al., 2015). As the movement becomes faster, the severity of symptoms can increase (Chardonnet et al., 2015; Liu and Uang, 2012). Human factors such as age (Saredakis et al., 2020), gender Freitag et al., 2016), or motion sickness susceptibility (Llorach et al., 2014) can also play a role.

Moreover, some hardware-specific factors can increase the probability of the occurrence of CS. Decisive factors include tracking accuracy (Chang et al., 2016), motion-to-photon latency (the time that elapses between the movement of a tracked object and the graphical representation of the associated movement in the virtual environment) (Stauffert et al., 2020), or latency jitter (Stauffert et al., 2018). Too-high latency or too-inaccurate tracking also causes a mismatch between input modalities.

Through continuous advances in hardware manufacturing and tailored software solutions, for example, asynchronous timewarp (Oculus) or asynchronous reprojection (Valve), modern HMDs significantly reduce the risk for CS. Nevertheless, some symptoms occur regularly and as intensely in contemporary applications Caserman et al., 2021; Cobb et al., 1999). CS must be given particular importance in healthcare applications. Supervisors leading a therapy session, for example, have a special duty of care toward the health of their patients. People working in the healthcare sector who want to integrate VR into their work routines need support in averting potential hazards to their patients Halbig et al., 2022). One possible solution to assist supervisors in protecting their clients from negative effects would be to use a warning system that detects possible signs of CS and warns the supervisor.

Over the years, different techniques that prevent CS were developed and tested, for example, having a virtual nose as a rest frame (Wienrich et al., 2018) or a dynamic restriction of the field of view (Groth et al., 2021). Nevertheless, CS symptoms are still widespread when it comes to the usage of HMDs, as it was shown by a survey among gamers (Rangelova et al., 2020).

2.2 Cybersickness measurement and detection

There are several options for measuring CS. The most widely used technique is the self-report questionnaire (Davis et al., 2014; Chang et al., 2016). Typical examples are the Simulator Sickness Questionnaire (SSQ) (Kennedy et al., 1993) and the Fast Motion Sickness Scale (FMS) (Keshavarz and Hecht, 2011). In addition to the advantages, such as the easy implementation and simple evaluation, these subjective methods also have drawbacks. For example, they only allow a discrete evaluation of the user state. In addition, longer self-reports usually take place after exposure to the VR stimulus and are, therefore, based on the active recapitulation of the experience by the user. Shorter mid-immersion assessments avoid these problems and closely link feedback to experience. However, they require active participation, potentially inducing unwanted breaks (especially immersion) and additional work load.

Alternative approaches to subjective self-reports measure CS via (objective) physiological and behavioral data, for example, using heart rate, skin conductance, electroencephalography (EEG), or eye-tracking data (Nakagawa, 2015; Dennison et al., 2016; Garcia-Agundez et al., 2019; Kim et al., 2019; Islam et al., 2020b; Tauscher et al., 2020). The analysis of the physiological data usually happens with the help of machine learning (ML), deep learning, or similar techniques (Halbig and Latoschik, 2021; Yang et al., 2022). These techniques can overcome many of the drawbacks of subjective methods. They could be used in a continuous online monitoring system that can warn the user or a supervisor in case the user/client felt sick or could even apply automatic counter-measures.

Many existing solutions for classifying CS based on physiological and behavioral measures need an extensive setup. For example, many setups are based on EEG, which often requires the application and preparation of many (up to 128) individual electrodes (Jeong et al., 2018; Garcia-Agundez et al., 2019; Kim et al., 2019; Lee et al., 2019; Tauscher et al., 2020). Even the examples without EEG data are often based on elaborate setups with different single electrodes (Islam et al., 2020a). It is hard to imagine that physical therapists, psychologists, or physicians would be willing to integrate such setups in their daily working routines. In contrast to EEG systems, the sensors used in this study are easy to attach to a person's body and non-disruptive to their behavior in the VR environment.

Several prominent ML algorithms have been applied to the CS detection task in the past (Yang et al., 2022), including the multilayer perceptron (MLP), SVM, linear discriminant analysis (LDA), and k-nearest neighbors (kNN) methods. However, these algorithms are not tailored to interpret time-series data and did not lead to satisfying results (Garcia-Agundez et al., 2019; Recenti et al., 2021). In recent years, deep learning has shown great performance for many classification and detection tasks. However, a limited number of works used deep learning for CS detection. Because deep learning models need very large amounts of data to train the models, they cannot be implemented if only a limited number of participants are available.

Some studies used wearable sensors and deep learning together. Islam et al. (2020a) used changes in physiological signals (heart rate, heart rate variability, galvanic skin response, and breathing rate) as CS predictors. They used an LSTM deep learning model with complicated electrode-based skin conductance and heart rate sensors. The hands were not moving freely, and the subjective feedback from SSQ was not consistently correlated with the physiological output. One of the recent works from Islam et al. (2021) used CNN + LSTM models and stereoscopic video data combined with eye-tracking (ET) and movement data. They achieved 52% accuracy using only video data, which is far from practical to be used as a CS detector. The same study used a physiological sensory setup with PPG EDA data and achieved 87% accuracy. Although they had an imbalanced dataset, they did not attempt to augment and balance it to get better results. Garcia-Agundez et al. (2019) proposed an electrode-based setup with ECG, EOG (electrooculographic), skin conductance, and respiratory data. They used SVM, kNN, and neural networks for binary CS detection and acquired 82% accuracy. Another interesting study Wang et al. (2023) used in-game characters' movement and users' eye motion data during gameplay in VR games. They trained an LSTM model to predict CS in real-time and acquired 83.4% accuracy.

2.3 Data augmentation

Collecting a huge amount of data for studies is often time-consuming, costly, and difficult. This becomes even harder if deep learning algorithms are used for classification or detection tasks. Because deep learning algorithms are data-hungry models, the size of the data should increase drastically to enhance the generalization capability of the models and to hinder overfitting issues. In some VR scenarios, physiological events that correspond to specific stimuli like CS, fear, or anxiety rarely occur, and this leads to imbalanced and skewed datasets. Recently, machine-learning approaches have been used for data augmentation, specifically for image classification tasks where images can be rotated, flipped, cropped, sheared, *etc.* (Shorten and Khoshgoftaar, 2019). However, unlike image data, physiological signals have a complex structure and dynamics that can be easily disrupted by transformations such as rotation or warping.

Especially in the medical and healthcare domains, when classifying time series physiological data, we often encounter imbalanced, skewed datasets in the literature. Some data augmentation techniques have already been proposed to tackle this problem (Iwana and Uchida, 2020; Wen et al., 2021). For example, Um et al. (2017) propose cropping, rotating, and wrapping the sensory data as a solution for this problem, but it also includes the risk of changing the respective data labels. In recent years, it can be seen that deep learning methods have increasingly been used for data augmentation on small and skewed datasets, and GAN, especially, increases classification performance. Harada et al. (2018) showed that using GAN to augment physiological data can improve the performance of the data classifier on imbalanced datasets. Specifically, conditioning GAN by target class labels offers two key advantages: it enhances GAN performance and facilitates the generation of samples belonging to a specific target class. Ehrhart et al. (2022) leveraged a cGAN to detect moments of stress. Nikolaidis et al. (2019) used cGAN for apnea detection tasks.

We address these limitations by using unobtrusive wearable devices with mid-immersion ground truth elicitation and proven deep learning models with the help of the cGAN data augmentation. Furthermore, to promote replicability and facilitate future research in cybersickness detection, we made our code and dataset publicly available.[1]

3 System description

Our end-to-end system mainly consists of sensory devices, virtual environment data acquisition, and data processing.

3.1 Sensory devices

We used three different devices to measure the participants' physiological signals during their VR experience. Because wearable sensors offer superior practicability with respect to cost, ease of use, and portability, we selected a Polar H10 (Polar Electro Oy, Finland) sensory device, which is an electrode-based chest strap, and an Empatica E4 (Empatica Inc., United States) device, which is a medical-grade wristband. Both of these devices transmit the data to the computer via Bluetooth communication. The Pico Neo 2 Eye VR headset (Pico Interactive, China) HMD, with a resolution of 3,840 × 2,160 px per eye and a total field of view of 101° running at a refresh rate of 75 Hz, was provided to participants. The eye movements were captured by the HMD's built-in eye tracker running at 90 Hz with a 0.5° accuracy. These sensors are easy to deploy and can, therefore, be used in a wide variety of scenarios without requiring too much effort. Figure 1 illustrates these sensory devices.

3.2 Virtual environment

The rollercoaster experience in VR is a well-known experiment when investigating CS in VR due to abundant motion that can elicit certain related symptoms of CS (Cebeci et al., 2019; Islam et al., 2020a). We implemented such a virtual environment for our study by adapting a rollercoaster that has many up-and-down bends, loopings, and sharp turns. It was initially obtained from the Unity Asset Store (2023) as a development environment. Then, we made adjustments to the rollercoaster to have a slightly lower speed and acceleration in the first 30 s. To get the exact time interval when a participant felt cybersick during the rollercoaster ride, we added functionality to collect the timestamps when the participant pressed the trigger button of the right controller and while hold the trigger during the CS symptoms continued. The system was implemented using Unity 2020.3.11f1 LTS (Unity Technologies, 2020). A screenshot of the scenario and the participant with sensory devices is shown in Figure 1.

3.3 Study and data acquisition

We conducted an experiment to acquire physiological data for the development of our CS detection approach. The experiments were completed with 20 participants aged between 18 and 57 years. Twelve participants were men, and eight were women. All participants provided their written informed consent to participate in this study. Before the study, the participants were debriefed about the study's purpose and noticeable effects of CS. In addition, they were informed and agreed to continue the study if the effects occurred during the study in terms of ethical considerations. During the study, no serious effects were observed or reported by the participants. At the start of the procedure, the participant filled out the pre-SSQ questionnaire to assess the level of CS before the VR exposure. Then, the participant put on the sensor devices and the connections between the sensors and the measurement engine (Viavr_Measurement_Engine, 2022) of the VIA-VR project (Viavr_Project, 2019) were established. Data streaming started simultaneously for every sensor. The participants started to have the rollercoaster experience. Whenever they felt symptoms, they reported CS occurrences by pressing the trigger button of the right controller and holding it as long as the symptoms were noticeable. Respective timesteps were stored in a *.CSV file, and all sensory data were stored in *.JSON files at the end of the experiments. The experiment and the data collection were stopped after one

1 https://github.com/m1237/automatic-cybersickness-detection

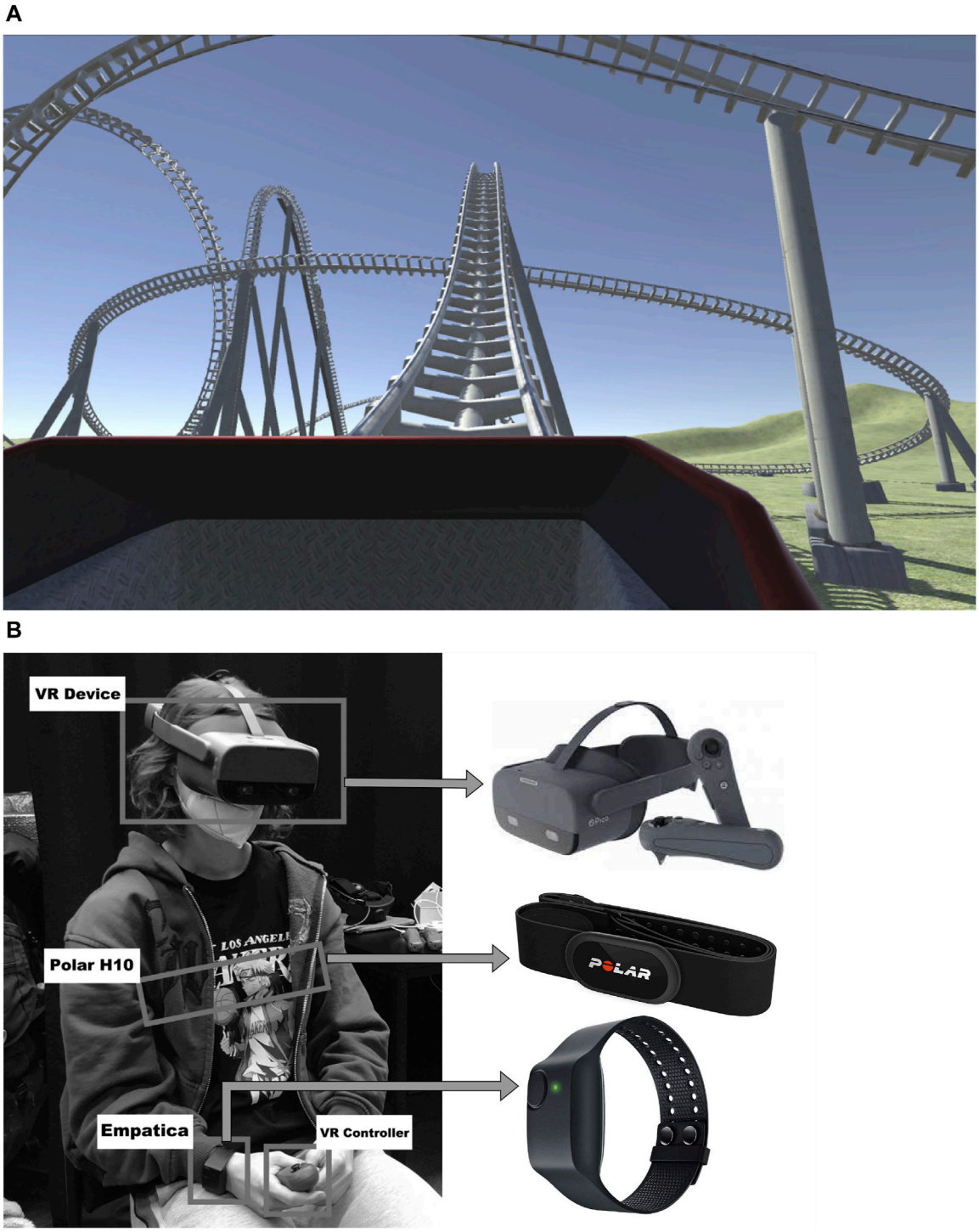

FIGURE 1
The overview of the CS detection setup. **(A)** Screenshot of the virtual roller coaster environment used to intentionally induce CS. **(B)** An equipped participant (center) and the respective sensors used during the experiments in detail.

rollercoaster cycle that lasted 80 s. The participant filled out a post-SSQ questionnaire to assess the level of CS after the experiment.

The recorded data types are summed-up in Table 1: electrocardiography (ECG) and acceleration (ACC) data collected using the Polar H10 chest strap; photoplethysmography (PPG), ACC, electrodermal activity (EDA), inter-beat interval (IBI) and peripheral body temperature (TEMP) data collected using the Empatica E4 wristband; and eye-tracking (ET) data collected using the Pico Neo 2 Eye HMD. Table 1 shows the overview of the data types, sampling rates, and number of features that we extracted from the physiological data.

TABLE 1 The features extracted and preprocessed from the raw sensor data to train the cybersickness classifier.

Device type	S. Rate (Hz)	Data type	Features
Eye tracker	90 Hz	Pupil diameter (left, right eye)	2
	90 Hz	Gaze direction (left, right eye; x, y, z values)	6
Empatica	32	ACC (x, y, z values, rma, pc, max, min)	12
	64	BVP (rma, pc, max, min)	4
	4	EDA (rma, pc, max, min)	4
	64	IBI (rma, pc, max, min)	4
	4	TEMP (rma, pc, max, min)	4
Polar H10	200 Hz	ACC (x, y, z values; rma, pc, max, min)	12
	130 Hz	ECG (rma, pc, max, min)	4
	1	HR (rma, pc, max, min)	4

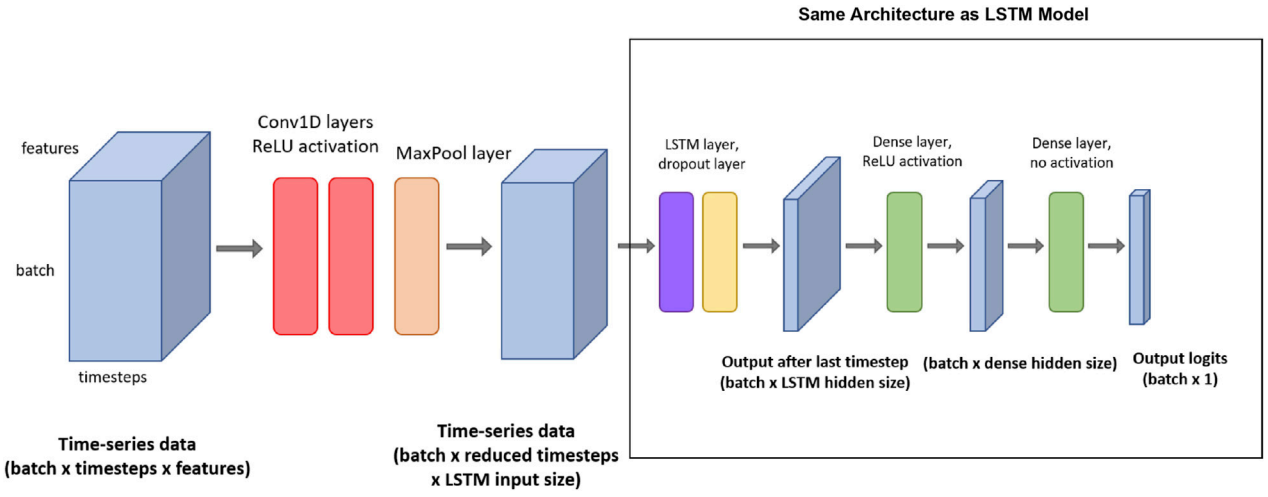

FIGURE 2
The architecture of the CNN + LSTM model with the respective input/output tensor shapes. (The red rectangle shows the whole CNN + LSTM model, and the black-colored rectangle shows the LSTM model used by itself as LSTM and bidirectional-LSTM models).

3.4 Data processing and feature extraction

After raw data acquisition, the collected data were preprocessed to apply deep learning algorithms. Instead of using only raw data, we calculated pre-features as input to the models. First, the data were normalized. Z-score normalization (D_z) was used for the ET (pupil diameters and gaze directions) and ACC (x, y, and z-axes) data. Here, for each data sample D_i, the Z-score normalization can be calculated with the mean of the training samples D_u and the standard deviation of the training samples D_s in Eq. 1:

$$D_z = \frac{D_i - D_u}{D_s} \quad (1)$$

The new data sample value D_z replaces the old sample i. This standard technique guarantees that the model will learn more smoothly on the data because it is standard and normally distributed (Islam et al., 2020b). The normalization of the BVP, EDA, IBI, TEMP, and ECG data was done using a min–max scaler (D_{minmax}) as stated in Eq. 2 that sets the values in relation to their max and min values, which can vary for each person. Here, D_{min} and D_{max} refer to the minimum and maximum of the data samples per person, respectively:

$$D_{minmax} = \frac{D_i - D_{min}}{D_{max} - D_{min}} \quad (2)$$

The features were extracted from the normalized data collected by Empatica and H10 sensors by using a rolling moving average (D_{rma}) in Eq. 3. This technique smooths the data by taking the average value over the last n samples instead of the raw signal. Given a fixed sampling rate, n corresponds to a time window. According to Courtney et al. (2010), an appropriate time for recognizing a change in physiological signals is 3 s. Depending on the sampling rate, this

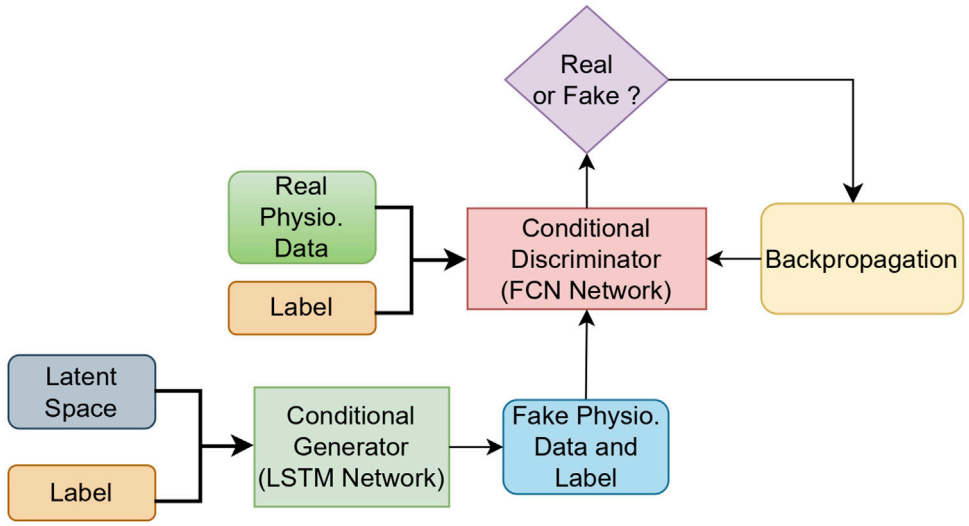

FIGURE 3
The overall cGAN data augmentation model with physiological data.

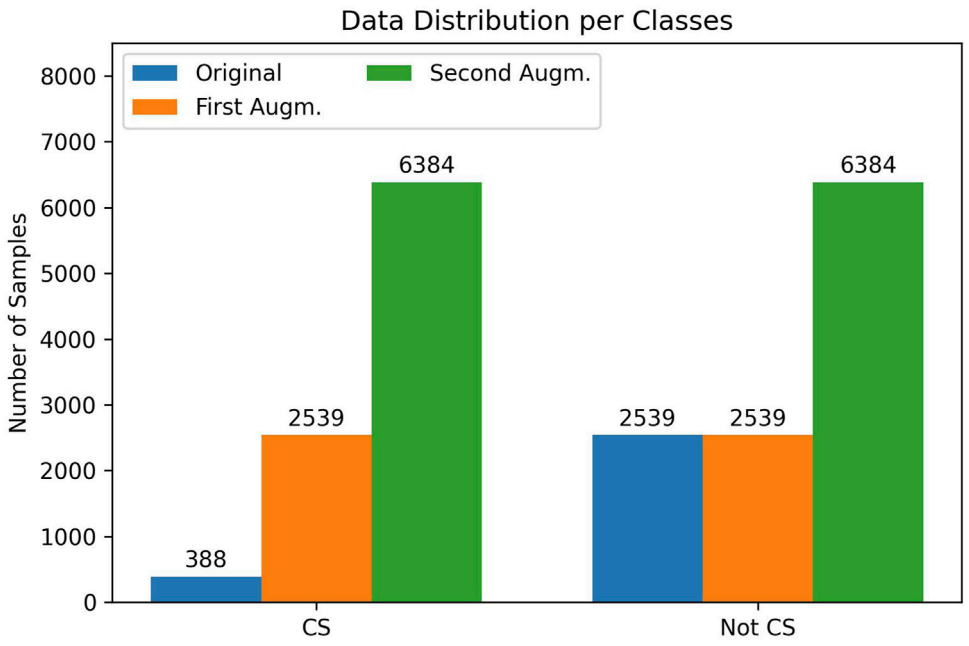

FIGURE 4
Data sample (timestep) distribution per class for the original dataset (blue), after the first data augmentation (orange) and after the second data augmentation (green).

value defines the final choice for the parameter n. Hence, $n = 3 \times$ *sampling rate*.

$$D_{rma} = \frac{1}{n}\sum_{z=0}^{n} D_z \tag{3}$$

For another feature class, the percentage of change (D_{pc}) was calculated from the normalized values using Eq. 4. It indicates how much the value has proportionally changed in the number of n timesteps rather than computing a nominal difference:

$$D_{pc} = \frac{D_z - D_{z-n}}{D_{z-n}} \tag{4}$$

The last two features that were constructed from the normalized data are the maximum (max) and the minimum (min) in the last n

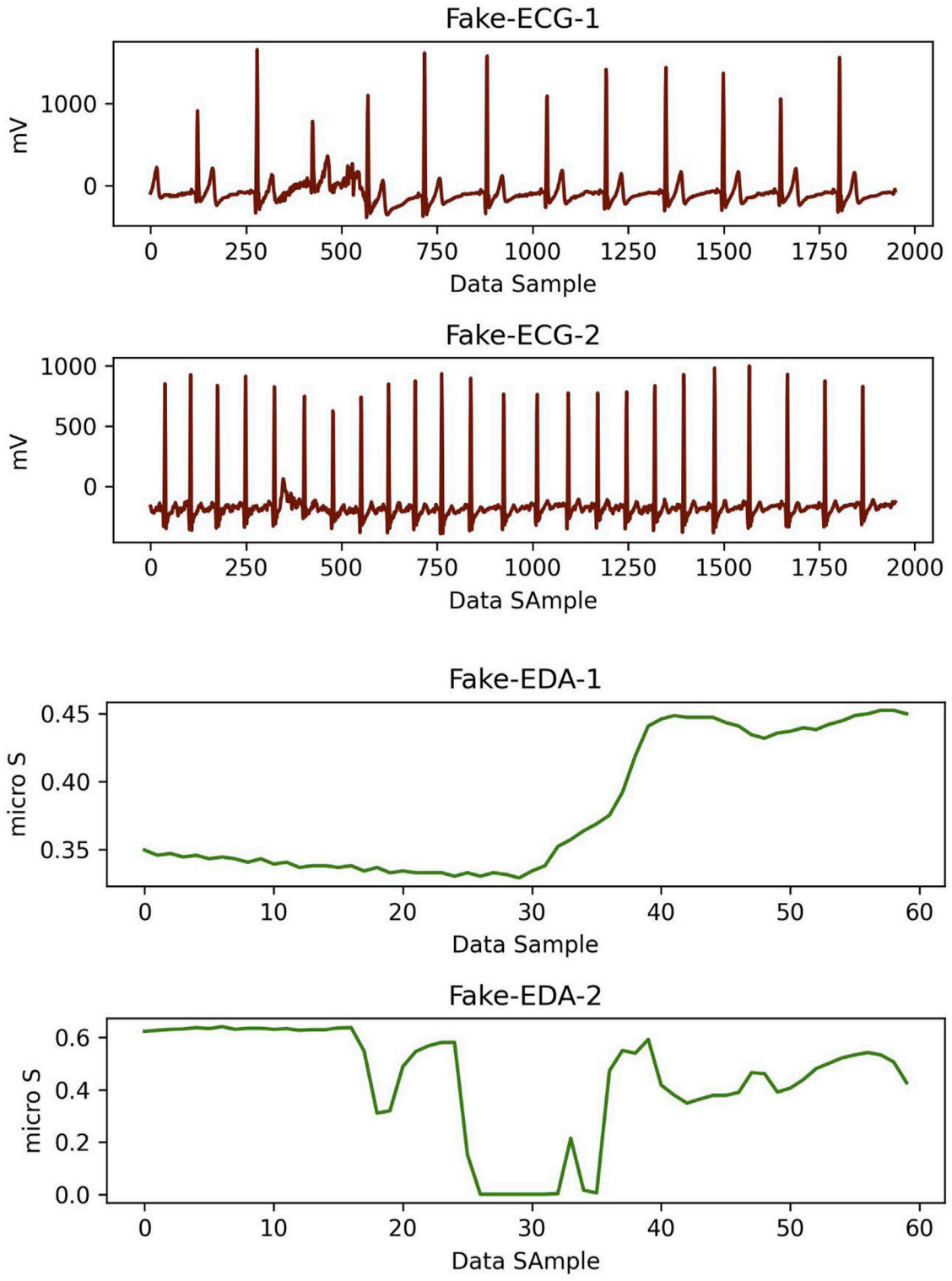

FIGURE 5
Two pairs of synthetic ECG (red) and EDA (green) data samples which were created using cGAN model for CS label. Here, time window for each data is 15 s.

timesteps. As an additional data source, the heart rate data were calculated from ECG data by using the algorithm in Christov (2004). This technique detects the current beat by leveraging specified thresholds and R-R interval analysis. The aforementioned four features (rolling moving average (rma), percent of change (pc), min, and max) were also calculated for the HR data.

Instead of utilizing HR data obtained from Empatica, we deliberately derived it from ECG data collected by the Polar

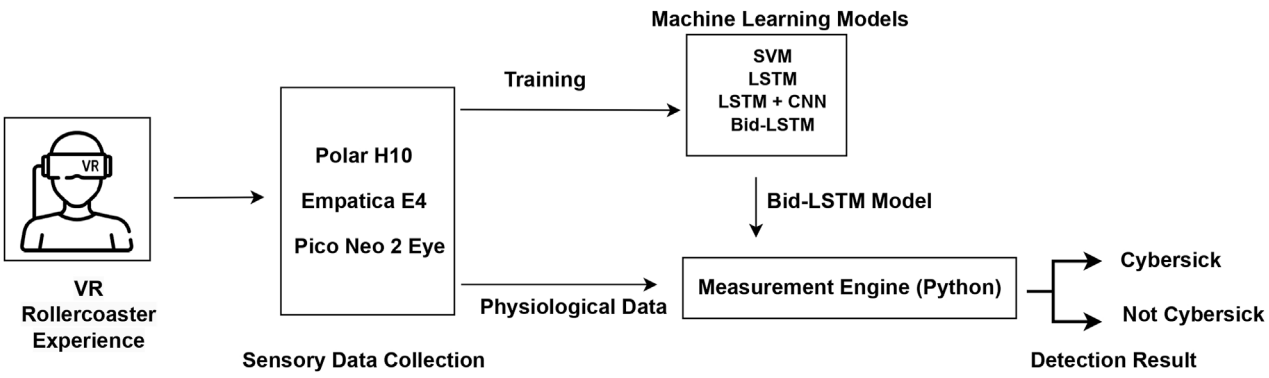

FIGURE 6
The final workflow scheme of the model for real-time CS detection.

H10 device due to its superior data quality. The Empatica wristband may be susceptible to motion artifacts, potentially leading to inaccuracies in heart rate readings compared to the Polar H10 chest strap, which is situated on a less-mobile body part. For instance, although Hadadi et al. (2022) gathered HR data using Empatica, they excluded it from their analysis due to its lower precision, reduced stability, and a notable increase in standard deviation.

4 Deep learning models for detection

After the feature extraction steps, the processed data contains 56 features (see Table 1) from the three different sensor devices for each sample to train the SVM and deep learning algorithms.

4.1 SVM

For the SVM model, we used a linear kernel and a class weight ratio of 1:8. Here, the class weight ratio automatically compensates for the data imbalance by increasing the weights of the minority class.

4.2 LSTM

To implement the LSTM model, we used the LSTM architecture described in Islam et al. (2020b). The model consists of four layers: an LSTM layer, a dropout layer, and two dense layers. The input for the LSTM layer is a tensor of shape (batch size, timesteps, and features). The LSTM module produces a tensor of shape (batch size and LSTM hidden size) as output, which contains the final hidden states of the input sequence after the last timesteps. After applying dropout, this output tensor is fed into the two dense layers, which both reduce the feature dimension. A ReLU activation function (Nair and Hinton, 2010) was used for the first dense layer, and no activation function was used for the second dense layer.

In addition to the four-layered LSTM network, we also used a bidirectional LSTM network (Schuster and Paliwal, 1997) with the same LSTM architecture for the detection task. Standard LSTM networks have restrictions as future input information cannot be reached from the current state. In contrast, bidirectional LSTM networks do not require input data to be in the same dimension. Moreover, their future input information can be reached from the current state. The main idea of bidirectional LSTM is to connect two hidden layers of opposite directions to the same output. By this structure, the output layer can access information from past and future states and interpret them better. The model can be seen in Figure 2.

4.3 CNN + LSTM

To improve the classification performance, we also deployed from-scratch CNN + LSTM architecture to acquire the spatial features and time-invariant patterns. Figure 2 shows a visual representation of the CNN + LSTM model architecture. This model consists of seven layers: two 1D convolution layers (Conv1D), and a pooling layer, followed by the four layers that were also present in the LSTM model (an LSTM layer, a dropout layer, and two dense layers). The input tensor for the first Conv1D layer is of shape (batch size, timestep, and features). Then, two 1D convolutions are applied. For both Conv1D layers, the number of filters is equal to the input size for the LSTM layer. The kernel size is 4, and ReLU is applied as an activation function. After the Conv1D layers, max pooling is used in the pooling layer, with a pool size of 2 and a stride of 2. After the max pooling function, the output tensor is of shape (batch size, reduced timesteps, and LSTM input size) and can be used as an input for the LSTM layer. The following LSTM and dense layers are set up in a similar way to the LSTM model described previously.

4.4 Hyperparameter optimization and model training

After preprocessing and merging, the dataset was divided into a training set and a testing set in a ratio of approximately 0.80/0.20, resulting in the training set containing data from 16 participants and the testing set containing data from the other four participants. We consciously selected different persons for the testing set to

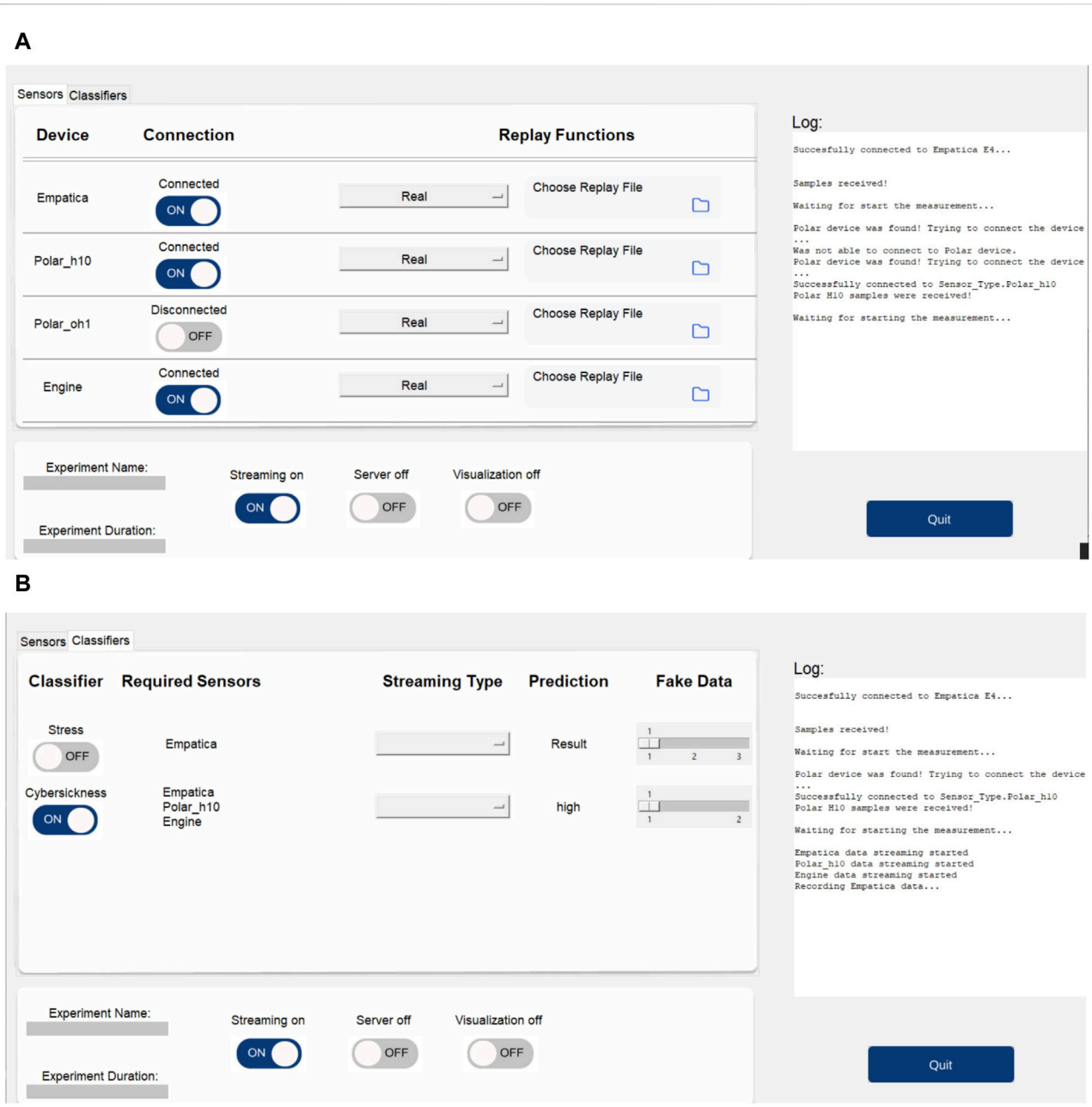

FIGURE 7
The measurement engine (viavr_measurement_engine) used in the study for data collection and real-time CS detection. **(A)** The GUI of the measurement engine shows the three connected sensors (Empatica, Polar H10, and Engine, which is the Unity application). **(B)** The GUI of the measurement engine shows the CS classifier and the detection result.

investigate the generalization capability of each model on never-seen participants. Afterward, we randomly divided the training set into 10-fold subsets and separated one set as a validation set to check the optimization performance of the training model. This technique is known as k-fold cross-validation in the literature (Hastie et al., 2001), and it minimizes the bias effect of one validation set. 10-fold cross-validation then iterates through the folds and uses one of the 10 folds as the validation set while using all remaining folds as the training set at each iteration. This process is repeated until every fold has been used as a validation set.

We investigated the best hyperparameters by deploying the grid-search technique throughout the implementation of all methods. We specified the deep learning model dependent variables are hidden layer size, dense layer size, timesteps, dropout, and learning rate.

4.5 Merging data

To merge the data from different sensors that have different sampling rates, we specified a different variable as a hyperparameter called sensor buffer with 0.1 s and 0.5 s time windows to have a mean value for each buffer size of data from different sensors. As a result, each data source is prepared as input for the models without depending on sampling rates.

TABLE 2 The variables and their values that were used in the grid search to optimize the models' hyperparameters and best-performing values for the respective model type.

H.Params	Values	SVM	LSTM	Bid-LSTM	LSTM + CNN
Timesteps	30, 50	30	30	30	50
CS buffer (seconds)	0, 1, 2	1.0	1.0	1.0	1.0
Sensor buffer (seconds)	0.1, 0.5	0.5	0.5	0.5	0.1
LSTM hidden	32, 64	-	64	64	64
Dense hidden	8, 16	-	8	8	16
Learning rate	0.001, 0.005	-	0.005	0.005	0.001
Dropout	0.5, 0.7	-	0.5	0.5	0.7

TABLE 3 Results of the CTST evaluation method using CNN + LSTM and bidirectional LSTM models.

Accuracy		
Method	CNN + LSTM	Bid-LSTM
CTST (cGAN)	0.604	0.573

Additionally, we used a timespan of 0 s, 1 s, and 2 s around a CS occurrence as a CS buffer. The aim is to include the before and after effects of physiological responses that participants felt. We combined these parameters with the hyperparameters of the learning algorithm and did a grid search to determine the best hyperparameters for training the data. During the SVM training, the binary cross-entropy (BCE) (Good, 1952) loss is calculated. After that, predictions and loss calculations are repeated using the testing set. The LSTM and CNN + LSTM models are trained with a 256 batch size for 30 epochs on the training set. For LSTM and CNN + LSTM training, the training loss is calculated using binary cross-entropy in each batch. We used Adam (Kingma and Ba, 2014) as the optimization algorithm, with a learning rate of 0.001 or 0.005, respectively. Every five epochs, the model's current performance is evaluated on the validation set by calculating the validation loss. Each model was trained on a machine with an Intel Core i7 9700K CPU and 32 GB of memory with NVIDIA RTX 2070 Super GPU. All models were trained by using the PyTorch 1.10 deep learning library.

5 Data augmentation using cGAN

To tackle the problems of small and imbalanced datasets, we deployed cGAN (Mirza and Osindero, 2014) to augment the original dataset. cGAN is the conditionally extended version of the GAN model (Goodfellow et al., 2014).

A GAN model architecture consists of two networks. One network generates candidate data (generator), and the other evaluates them (discriminator). Typically, the generative network learns to map from a latent space (sampled from Gaussian distribution) to a particular data distribution of interest, in our case, physiological data, while the discriminative network discriminates between instances from the true data distribution and candidates produced by the generator. The objective of the generator G) is to fool the discriminator D) such that it classifies generated data as real. Through the training, the generator learns to produce realistic-looking synthetic data. Consequently, the generated data distribution converges to the real data distribution. The generator G_{θ_g} is a directed latent variable model that deterministically generates samples x from latent space z. Because discriminator D) wants to classify real or fake samples, V (D, G) is considered an objective function as an aspect of the classification problem. The general form of the objective function can be written as Eq. 5 follows:

$$\min_{\theta_g}\max_{\theta_d} V(D,G) = \left[E_{x\sim p_{data}} \log D_{\theta_d}(x) + E_{z\sim p(z)} \log\left(1 - D_{\theta_d}\left(G_{\theta_g}(z)\right)\right)\right] \tag{5}$$

Here, the main difference between the cGAN and the two player minimax game objective function of the GAN is that cGAN includes labels as auxiliary information indicated as y. Hence, the objective function can be written as Eq. (6)

$$\min_{\theta_g}\max_{\theta_d} V(D,G) = \left[E_{x\sim p_{data}} \log D_{\theta_d}(x|y) + E_{z\sim p(z)} \log\left(1 - D_{\theta_d}\left(G_{\theta_g}(z|y)\right)\right)\right] \tag{6}$$

During the training process, Eq. (6) often results in mode collapse, which means that many samples out of the latent space map to the same generated sample. This results in a dataset with less diversity. To counteract this problem, the diversity term was introduced by Yang et al. (2019) to simply regularize and penalize the generator for producing the same samples. The diversity term is defined as Eq. 7

$$\max_{\theta_d} f(G) = E_{z_1,z_1}\left[\frac{\|G(z_1,y) - G(z_2,y)\|}{\|z_1 - z_2\|}\right] \tag{7}$$

The logic in this approach is if two samples are different, but the generated sequences are the same, the term is 0. This results in the following new objective function in Eq. 8

$$\min_{\theta_g}\max_{\theta_d} f(D,G) - \lambda f(G) \tag{8}$$

where λ is a hyperparameter that describes the importance of the term in Eq. (8), and $\|$ denotes a norm.

5.1 The cGAN architecture

5.1.1 Generator

The generator takes latent space and class labels as input. Sixteen hidden units per layer of stacked LSTM are used to generate the

TABLE 4 First table shows the detection evaluation results (accuracy, precision, recall, and F1-score) for the best-performing models in each model type on the original, first augmented, and second augmented training sets (with 10-fold cross-validation). The second table shows the detection evaluation results for the testing set. The model name and the numbers in bold indicate the highest value of the experimental results.

10-Fold cross-validation												
Model type	Original dataset				First data augmentation				Second data augmentation			
	acc	Pr	Rec	F1	acc	Pr	Rec	F1	acc	Pr	Rec	F1
(Naive Model)	0.133	0.133	1	0.235	0.5	0.5	1	0.667	0.5	0.5	1	0.667
SVM	0.66	0.310	0.911	0.463	0.843	0.675	0.883	0.766	0.874	0.752	0.902	0.820
CNN + LSTM	0.88	0.575	0.919	0.707	0.88	0.915	0.919	0.917	0.905	0.889	0.920	0.905
LSTM	0.86	0.555	0.871	0.678	0.912	0.914	0.882	0.898	0.925	0.932	0.906	0.919
Bid-LSTM	**0.87**	**0.568**	**0.882**	**0.691**	**0.919**	**0.921**	**0.909**	**0.915**	**0.939**	**0.945**	**0.912**	**0.928**
Testing Set												
Model Type	acc	pr	rec	F1	acc	pr	rec	F1	acc	pr	rec	F1
(Naive Model)	0.156	0.156	1	0.27	0.156	0.156	1	0.27	0.156	0.156	1	0.27
SVM	0.652	0.207	0.660	0.316	0.825	0.650	0.854	0.738	0.857	0.869	0.841	0.855
CNN + LSTM	0.773	0.361	0.641	0.465	0.871	0.892	0.911	0.901	0.893	0.902	0.881	0.891
LSTM	0.836	0.421	0.673	0.473	0.901	0.886	0.921	0.903	0.910	0.929	0.889	0.908
Bid-LSTM	**0.842**	**0.452**	**0.742**	**0.471**	**0.907**	**0.913**	**0.902**	**0.908**	**0.917**	**0.936**	**0.885**	**0.911**

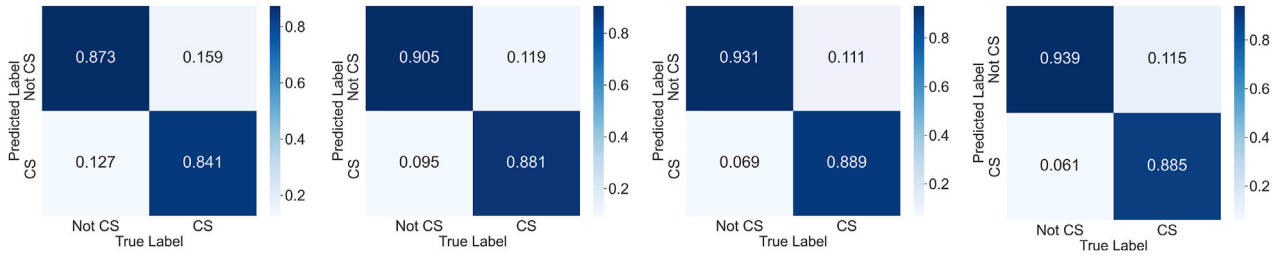

FIGURE 8
The confusion matrices which show the test results of the second augmented dataset. The matrices belong to the SVM, CNN + LSTM, LSTM, and bidirectional LSTM models, respectively.

physiological signals. The mapping from the random space is performed via a dense layer using a Leaky ReLU (Xu et al., 2015) activation function. Then, the LSTM layer group was applied. The output was fed through a linear activation. The final output of the generator has the shape of the matrix, which is batch size times time window. Here, the time window for created data is 15 s. After training, we can apply random Gaussian noise N (0, 1) and labels to create the physiological data.

5.1.2 Discriminator

In our architecture, the temporal convolutional layers are used to extract features from the time series signal. The convolutional layer for the discriminator is chosen because in our experiments, we saw that the fully convolutional network (FCN) discriminator outperformed the recurrent discriminator. This indicates that the convolutional network, especially the FCN, provides the generator with better gradients during training. Therefore, 1D filters were applied to capture the changes in the signal according to the different classes of physiological signals. The filters per layer are 32, 64, and 32, and the kernel size per layer was set to 8, 5, and 3, respectively. After the three convolutional blocks, the resulting feature maps are followed by a pooling layer and a sigmoid activation function, which outputs a scalar value in the range of 0 to 1 for the sequence, indicating whether it is real or fake. For the optimization process, the Adam optimizer (Kingma and Ba, 2014), with a learning rate of 0.0002 and a beta value of 0.5 (Christopoulos et al., 2019), was used and trained for 1,650 epochs. A batch size of 32 was used to ensure stable

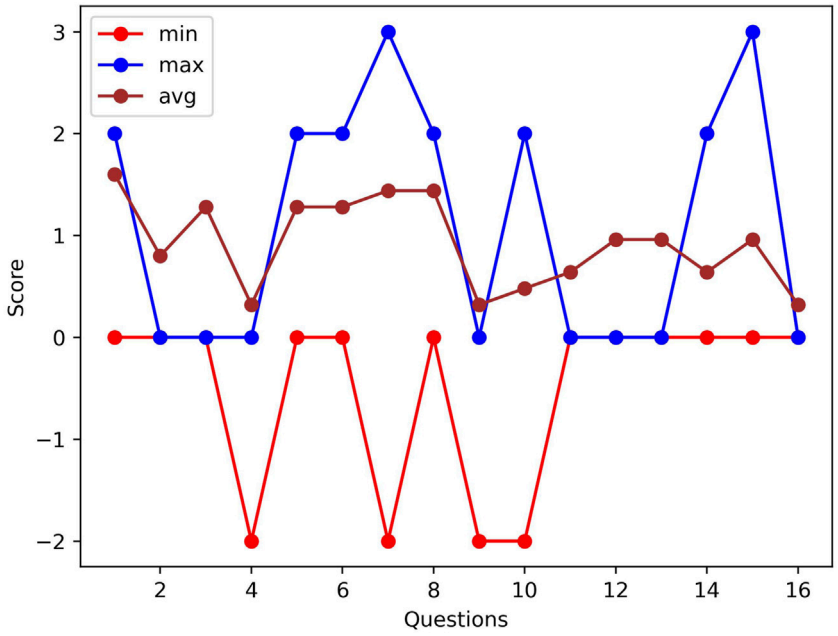

FIGURE 9
A comparison of the pre- and post-SSQ difference results of two participants who have min and max scores and the average differences of all participants per question.

TABLE 5 Cybersickness detection accuracy results from the literature and comparison with our result.

Work	Setup	Physio. Data	Methods	Best Acc. (%)
Hadadi et al. (2022)	HMD, Empatica	EDA,TEMP,BVP,ACC	SVM + TDA	71
Garcia-Agundez et al. (2019)	HMD, Electrode-based	ECG, EOG, EDA, RESP	SVM, KNN, NN	82
Islam et al. (2021)	HMD, Electrode-based	Video, Eye-Track., Head-Track	LSTM + CNN	87
Pane et al. (2018)	EEG Setup	EEG	SVM, CNN	88.9
Kim et al. (2019)	EEG Setup	Video, EEG	CNN-RNN	90.4
Our work	HMD, Polar H10, Empatica	ECG, ACC, EDA, BVP, TEMP	Bid-LSTM, CNN	91.7

training. Figure 3 depicts the overall cGAN algorithm with physiological data.

5.2 cGAN implementation

The original dataset that we collected during experiments is quite skewed and unbalanced. The data samples consist of 2539 CS labeled and 388 not-CS labeled timesteps (with a 0.5 s CS buffer). Because we have already split the dataset into testing and training sets, only the training set was used for the data augmentation process. After training our cGAN, we created 2,151 synthetic CS timesteps data as the first data augmentation and made the data equally distributed. In the second data augmentation, we wanted to investigate the result with the equally enriched synthetic data for both classes. After this process, we augmented the data, which includes 6,384 timesteps for each class. The data distribution per class can be seen in Figure 4 for the original dataset and the first and second augmented datasets.

5.2.1 Evaluation of the cGAN model

Synthetic data samples produced by the cGAN model are of good quality if real data and synthetic data are indistinguishable from each other. To measure the similarity, we used the classifier two-sample test (CTST) proposed by Lopez-Paz and Oquab (2017). In this approach, a binary classifier is trained to distinguish data samples belonging to the synthetic dataset from the real (original) dataset. For the training set, we randomly selected 214 synthetic and real timesteps data samples for CS and 528 synthetic and real timesteps data samples for non-CS sequences. For the training set, we randomly selected 87 synthetic and real timesteps data elements for CS and 161 synthetic and real timesteps data samples for non-CS sequences. We trained our CNN + LSTM and bidirectional LSTM model with the best hyperparameters (see Section 5.1). As can be observed

in Table 3 the accuracy result is close to the chance level. Figure 5 depicts two synthetic ECG and EDA data samples for the CS label.

5.3 Real-time CS detection

5.3.1 Data Capture

To start data streaming, all sensors must establish a connection to the measurement engine (viavr_measurement_engine). Here, socket programming (Socket, 2022) was used for the Empatica connection, and the Bleak library (Bleak, 2022) was used for the Polar H10 connection. The Pico Neo 2 includes a built-in eye tracker (Tobii Ocumen AB, 2021) that can collect raw eye-tracking data (binocular gaze, pupil size, and blink status) using the advanced API. The measurement engine starts recording the data when the "Streaming On" button is clicked. We implemented the data streaming via the threading method. The streaming of each data source is independent of each other and could be started or stopped separately. In the case of a connection breakdown related to the sensors, the engine log screen informs the user about the current state. To prevent the data drifting, we used the same data acquisition architecture for the real-time classification. Additionally, the sensor buffer was used to prevent a lack of data in the streaming. The sensor buffer gets the mean of the data for a specified period of time, and then one value for each data element can be calculated. Hence, we prevent potential missing data and system performance decrease.

5.3.2 Real-time data processing

After finding the best model for the detection task, we also implemented the whole procedure as a real-time CS detection system. All sensory devices are connected to the measurement engine, which was written in Python. Data are feed-forwarded to the four-layered bidirectional LSTM model. Each feed-forward data processing time is around 60 m. In every 5 s period of time, the measurement engine produces detection results by using already trained model parameters. Because the sensor buffer is 0.5 s for the best model, depending on the output of the last layer's sigmoid function, the engine produces 10 different results in 5 s. If the mean value of the results is higher than 0.5, the engine detects CS; otherwise, it detects not CS. This period of time can easily be selected to be higher or lower because we selected 5 s as an example. Figure 6 shows the overview of the real-time detection system. Figure 7 shows the real-time Python GUI implementation of the measurement engine. The result of the classification is shown using the labels "high" and "low" on the GUI.

The system is ready to use in real-time VR applications. It demonstrates an average latency of 60 ms between classifying the data and providing feedback to the user, ensuring a seamless and responsive experience. The prototype achieves a high accuracy (91.7% with the testing set), indicating a high level of accuracy in detecting CS symptoms. Although we used a Pico Neo 2 in our study, any VR headset that included eye-tracking could be used in future studies.

6 Results

To find the best hyperparameters for the respective model type, we conducted a grid search covering 584 different model configurations with 10-fold cross-validation, resulting in 5,840 total model trainings. The best hyperparameters that led to the best classification results are shown in Table 2.

After the training, we assessed each model's performance based on the performance metrics accuracy, precision, recall, and F1-score on the testing set, and we also calculated these metrics for the validation set (10-fold cross-validation) to get a better insight into the model's learning behavior. We also calculated a naive classifier to compare the results of the given classifier model with a baseline (for example, accuracy is the random occurrence of the CS label in this case). In addition to these metrics, the confusion matrix was also calculated to assess the ratio between true/false positives/negatives for a second augmented dataset.

All performance metrics are based on the true and false results and their real values. They are called true negative (TN), true positive (P), false negative (FN), and false positive (FP). TP is an outcome where the model correctly predicts the positive class (in our case, CS), while TN is an outcome where the model correctly predicts the negative class (in our case, not CS). FP is an outcome where the model incorrectly predicts the positive class, and FN is an outcome where the model incorrectly predicts the negative class. We can formulate these metrics as Eqs. 9–12, follows:

$$Accuracy = \frac{TP + \mathrm{TN}}{TP + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{9}$$

$$Precision = \frac{TP}{TP + \mathrm{FP}} \tag{10}$$

$$Recall = \frac{TP}{TP + \mathrm{FN}} \tag{11}$$

$$f_1 = 2 \times \frac{Precision \times Recall}{Precision + \mathrm{Recall}} \tag{12}$$

The results of the best-performing models are shown in Table 4 on the original, first, and second augmented training datasets. All model types achieved higher F1-scores on the training datasets with 10-fold cross-validation than on the testing set. To tackle the imbalanced dataset problem and increase the classifier performance, we augmented the dataset with the previously explained methods (see Section 5.2). After data augmentation and training with the new dataset, the results for all models increased significantly. The confusion matrices of the second data augmentation test results can be seen in Figure 8. We acquired the best result with a four-layered bidirectional LSTM model with 91.7% accuracy and a 91.1% F1-score. We also evaluated the cGAN performance using the CTST (see Section 5.2) method. As we can see in Table 3, after testing, accuracy is close to the chance level, which means that our cGAN model created synthetic data that are almost similar to real data.

Because SSQ results cannot provide actual data labels during the experiments, we only used the SSQ results for validation. Each participant answered 16 different questions with four options standing for 1–4 score scales as pre- and post-SSQ: "None," "Rather not applicable," "Rather applicable," or "Often or a lot." We calculated the SSQ scores for each participant and evaluated the SSQ results. We showed the average difference of pre- and post-SSQ scores per question in Figure 9. In SSQ results, the participant scores for questions 1 (general discomfort), 3 (headache), 5 (difficulty focusing), 6 (salivation increase), 7 (sweating), and 8 (nausea) were slightly higher than other questions. The average score of the difference of all the symptoms was 1.1, which indicates that the

participants felt a bit worse after the experiment than before. This validates that the experiment resulted in a cybersick feeling for most participants, although it might be rather small. Hence, it justifies the correlation with physiological data.

7 Discussion

We have demonstrated that the utilization of unobtrusive wearable devices in a simple setup, combined with appropriate deep learning algorithms and a supportive data augmentation technique, yields excellent results in detecting CS. Our proposed approach involves the use of a bidirectional LSTM model in conjunction with conditional GAN data augmentation, achieving an accuracy of 91.7% and an F1-score of 91.1%. This outperforms previous works employing similar physiological sensory setups, including more complex ones such as EEG.

A comparison with recent literature is presented in Table 5 Hadadi et al. (2022) incorporated physiological data from an Empatica wristband and topological point cloud data from HMD. This combination was not sufficient to capture CS responses properly, using their (TDA + SVM) model. Garcia-Agundez et al. (2019) additionally used game parameters with electrode-based data using machine learning models (SVM, KNN, and NN) but could not reach satisfying classification performance. (Pane et al. (2018) and Kim et al. (2019) used EEG setups for their studies. However, EEG setups are not easy to deploy for studies because they have complex, error-prone, and time-consuming features. Although some studies worked on CS severity classification (Islam et al., 2021), the F1-scores of these works are not as high as the accuracy results because they also have imbalanced datasets. Furthermore, none of these works attempted:

- to implement data augmentation to overcome lower generalization capability issues for imbalanced datasets.
- to implement a real-time mid-immersion ground truth elicitation method.

In our work, we mainly pioneered to address these issues, hence improving the detection performance.

Upon evaluating the test results, we observed that the four-layered bidirectional LSTM model outperformed the CNN + LSTM and SVM models and slightly outperformed the standard LSTM model. Incorporating hidden layers in opposite directions, enabling access to past and future states, played a significant role in capturing sequential data patterns through the bidirectional LSTM. Notably, the recall scores for all models surpassed the precision scores in the original dataset, mainly due to the class imbalance issue. While the models correctly classified a substantial quantity of CS labels, they exhibited a high number of FPs, indicating a compromise in the quality of the classification. Additionally, both the training and testing sets on the original dataset showed a higher number of FPs than FNs.

One significant finding that we wish to emphasize is that our models with data augmentation exhibit remarkable generalization capability on a testing set comprising participants who differ from those used in the training set. Unlike previous research in the literature (see Table 5), our models effectively generalize their learning to new participants. With data augmentation, precision scores increased significantly by decreasing FPs, which is strong proof that the models gained enhanced detection capabilities for CS labels. Additionally, we successfully implemented a real-time CS detection system using our best model, which is a four-layered bidirectional LSTM. This system can be readily deployed in various VR scenarios, including medical and therapy applications.

During our experiments, we found that instructing participants to press the controller button when experiencing the rollercoaster simulation provided reliable ground truth data. However, this procedure resulted in an imbalanced dataset, as there were fewer instances of participants experiencing CS during the rollercoaster scenario than instances when they did not experience CS. This was particularly the case during the first 40 s of the experiment because it took time to elicit the CS effects. To capture the before and after effects on participants, we deployed a CS buffer as a hyperparameter during the optimization process. By labeling the data one second before and after CS occurrences, we observed an improvement in the classification performance. This can be attributed to the time required for participants to make decisions, such as pressing and releasing the button, and the continuation of physiological responses during the label transition phase. This hyperparameter also increased the number of CS-labeled data by approximately 5%.

Our data augmentation technique generates synthetic data that closely align with the data distribution of the original dataset. We evaluated the similarity between the synthetic and real data, and our results indicate that the bidirectional LSTM model achieves classification performance close to the chance level with 57.3% accuracy (see Table 3), which is evidence of an indistinguishable synthetic dataset. Moreover, we successfully addressed the issue of data imbalance through the implementation of the cGAN data augmentation model. Following the first round of data augmentation, the dataset achieved equal distribution per class, and the testing results revealed significant improvements not only in accuracy but also in other evaluation metrics. The recall and precision scores approached each other, indicating robust and accurate detection of both classes by the models. Performance evaluation of the second augmented dataset also indicated similar results across different metrics. Notably, precision and F1 scores surpassed those obtained from the original dataset, signifying improved accuracy in classifying instances of CS.

7.1 Limitations

We showed that the augmented physiological data can increase classifier performance significantly. However, the cGAN model is difficult to train in a stable way. We tried to overcome this problem with a diversity term. This could also improve the generalization capability of the learning models. In addition, the choice of the virtual scenario highly influences the responses from the participants. Even though many past experiments, for example, by Islam et al. (2020a) or Nalivaiko et al. (2015), chose to expose participants to a rollercoaster ride in VR, it might have influenced the resulting data negatively. A person might feel sick or nauseous during the experiment, not due to CS, but because a rollercoaster ride might have made them feel exactly the same way in real life. Because these borders are hard to define, another choice of virtual scenario might be an improvement. The rollercoaster ride might not be the perfect virtual scenario, but it can efficiently provoke CS symptoms. On the other hand, the measurement engine that we will provide can be

used for data collection as well as a real-time detection system with the same sensory device setup. Hence, the system can be used by researchers in validation studies.

We used a relatively small data set in our study and enhanced the result with data augmentation to acquire generalization capability. However, more data can be collected in the future to acquire more robust results in different studies. A wider range of experimental scenarios would provide more variability and enable better model training and validation. Larger datasets that can be collected from a more extensive and diverse user population can help improve the model's performance by reducing bias and overfitting and help ensure that the model is robust across different contexts.

8 Conclusion

In this work, we used a VR environment that includes a rollercoaster to elicit cybersickness and used a simple setup with sensory devices to collect physiological responses. We deployed three different deep learning models and one classical machine learning model to detect CS. In addition, we realized a completely real-time system using our best model. We demonstrated that a four-layered bidirectional LSTM with data augmentation gives superior results (91.7% accuracy; 91.1% F1-score), and this combination is the best solution for sensor-based CS detection in real-time applications, particularly for wearable devices. Furthermore, we showed that small, skewed, and imbalanced datasets can be augmented with our pioneered cGAN approach to increase the classifier performance significantly. In future works, we plan to investigate different VR scenarios for cybersickness elicitation and state-of-the-art models to enable multi-level CS classification.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

MY: methodology, software, visualization, writing–original draft, writing–review and editing, conceptualization, data curation, formal analysis, validation. AH: writing–review and editing, and software. MF: supervision, writing–review and editing, and investigation. ML: supervision, writing–review and editing, funding acquisition, and investigation.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This research has been funded by the German Federal Ministry of Education and Research (BMBF) in the project VIA-VR (project number: 16SV8444) and has been funded/was supported by the Bavarian State Ministry for Digital Affairs in the project XR Hub (Grant A5-3822-2-16). This publication was supported by the Open-Access Publication Fund of the University of Würzburg.

Acknowledgments

Special thanks to Jennifer Häfner, Felix Achter, Mohammad Farrahi and Marja Wahl for their help to this work and Florian Heinrich for proofreading and feedback.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

References

Abdullah, M., and Shaikh, Z. A. (2018). An effective virtual reality based remedy for acrophobia. *Int. J. Adv. Comput. Sci. Appl.* 9. doi:10.14569/ijacsa.2018.090623

AB, T (2021). Tobii pro lab. *Comput. Softw.*

Alcañiz, M., Bigné, E., and Guixeres, J. (2019). Virtual reality in marketing: a framework, review, and research agenda. *Front. Psychol.* 10, 1530. doi:10.3389/fpsyg.2019.01530

Bălan, O., Moise, G., Moldoveanu, A., Leordeanu, M., and Moldoveanu, F. (2020). An investigation of various machine and deep learning techniques applied in automatic fear level detection and acrophobia virtual therapy. *Sensors* 20, 496. doi:10.3390/s20020496

Barreda-Ángeles, M., Aleix-Guillaume, S., and Pereda-Baños, A. (2020). Users' psychophysiological, vocal, and self-reported responses to the apparent attitude of a virtual audience in stereoscopic 360°-video. *Virtual Real.* 24, 289–302. doi:10.1007/s10055-019-00400-1

Bartl, A., Merz, C., Roth, D., and Latoschik, M. E. (2022). "The effects of avatar and environment design on embodiment, presence, activation, and task load in a virtual reality exercise application," in *IEEE international symposium on mixed and augmented reality (ISMAR)*.

Bleak (2022). Bluetooth low energy platform agnostic klient. *Comput. Softw.*

Caserman, P., Garcia-Agundez, A., Gámez Zerban, A., and Göbel, S. (2021). Cybersickness in current-generation virtual reality head-mounted displays: systematic review and outlook. *Virtual Real.* 25, 1153–1170. doi:10.1007/s10055-021-00513-6

Cebeci, B., Celikcan, U., and Capin, T. K. (2019). A comprehensive study of the affective and physiological responses induced by dynamic virtual reality environments. *Comput. Animat. Virtual Worlds* 30, e1893. doi:10.1002/cav.1893

Chang, C.-M., Hsu, C.-H., Hsu, C.-F., and Chen, K.-T. (2016). "Performance measurements of virtual reality systems: quantifying the timing and positioning accuracy," in *Proceedings of the 24th ACM international conference on Multimedia*, 655–659.

Chang, E., Kim, H. T., and Yoo, B. (2020). Virtual reality sickness: a review of causes and measurements. *Int. J. Human–Computer Interact.* 36, 1658–1682. doi:10.1080/10447318.2020.1778351

Chardonnet, J.-R., Mirzaei, M. A., and Mérienne, F. (2015). "Visually induced motion sickness estimation and prediction in virtual reality using frequency components analysis of postural sway signal," in *International conference on artificial reality and telexistence eurographics symposium on virtual environments*, 9–16.

Checa, D., and Bustillo, A. (2020). A review of immersive virtual reality serious games to enhance learning and training. *Multimedia Tools Appl.* 79, 5501–5527. doi:10.1007/s11042-019-08348-9

Chen, W., Chen, J. Z., and Richard, S. (2011a). Visually induced motion sickness: effects of translational visual motion along different axes. *Contemp. ergonomics Hum. factors*, 281–287. doi:10.1201/b11337-47

Chen, Y.-C., Dong, X., Hagstrom, J., and Stoffregen, T. (2011b). Control of a virtual ambulation influences body movement and motion sickness. *BIO Web Conf.* 1, 00016. doi:10.1051/bioconf/20110100016

Christopoulos, G. I., Uy, M. A., and Yap, W. J. (2019). The body and the brain: measuring skin conductance responses to understand the emotional experience. *Organ. Res. Methods* 22, 394–420. doi:10.1177/1094428116681073

Christov, I. (2004). Real time electrocardiogram QRS detection using combined adaptive threshold. *Biomed. Eng. online* 3, 28. doi:10.1186/1475-925X-3-28

Cobb, S. V., Nichols, S., Ramsey, A., and Wilson, J. R. (1999). Virtual reality-induced symptoms and effects (vrise). *Presence Teleoperators Virtual Environ.* 8, 169–186. doi:10.1162/105474699566152

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297. doi:10.1023/A:1022627411411

Courtney, C., Dawson, M., Schell, A., Iyer, A., and Parsons, T. (2010). Better than the real thing: eliciting fear with moving and static computer-generated stimuli. *Int. J. Psychophysiol. official J. Int. Organ. Psychophysiol.* 78, 107–114. doi:10.1016/j.ijpsycho.2010.06.028

Davis, S., Nesbitt, K., and Nalivaiko, E. (2014). "A systematic review of cybersickness," in *Proceedings of the 2014 conference on interactive entertainment*, 1–9.

Dennison, M. S., Wisti, A. Z., and D'Zmura, M. (2016). Use of physiological signals to predict cybersickness. *Displays* 44, 42–52. doi:10.1016/j.displa.2016.07.002

Ehrhart, M., Resch, B., Havas, C., and Niederseer, D. (2022). A conditional gan for generating time series data for stress detection in wearable physiological sensor data. *Sensors* 22, 5969. doi:10.3390/s22165969

Freitag, S., Weyers, B., and Kuhlen, T. W. (2016). Examining rotation gain in cave-like virtual environments. *IEEE Trans. Vis. Comput. Graph.* 22, 1462–1471. doi:10.1109/tvcg.2016.2518298

Garcia-Agundez, A., Reuter, C., Becker, H., Konrad, R. A., Caserman, P., Miede, A., et al. (2019). Development of a classifier to determine factors causing cybersickness in virtual reality environments. *Games health J.* 8, 439–444. doi:10.1089/g4h.2019.0045

Gazendam, A., Zhu, M., Chang, Y., Phillips, S., and Bhandari, M. (2022). Virtual reality rehabilitation following total knee arthroplasty: a systematic review and meta-analysis of randomized controlled trials. *Knee Surg. Sports Traumatol. Arthrosc.* 30, 2548–2555. doi:10.1007/s00167-022-06910-x

Gianola, S., Stucovitz, E., Castellini, G., Mascali, M., Vanni, F., Tramacere, I., et al. (2020). Effects of early virtual reality-based rehabilitation in patients with total knee arthroplasty: a randomized controlled trial. *Medicine* 99, e19136. doi:10.1097/md.0000000000019136

Glémarec, Y., Lugrin, J.-L., Bosser, A.-G., Buche, C., and Latoschik, M. E. (2022). Controlling the stage: a high-level control system for virtual audiences in virtual reality. *Front. Virtual Real. – Virtual Real. Hum. Behav.* 3. doi:10.3389/frvir.2022.876433

Good, I. J. (1952). Rational decisions. *J. R. Stat. Soc. Ser. B Methodol.* 14, 107–114. doi:10.1111/j.2517-6161.1952.tb00104.x

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., et al. (2014). *Generative adversarial networks*. doi:10.48550/ARXIV.1406

Groth, C., Tauscher, J.-P., Heesen, N., Castillo, S., and Magnor, M. (2021). "Visual techniques to reduce cybersickness in virtual reality," in *2021 IEEE conference on virtual reality and 3D user interfaces abstracts and workshops (VRW)* (IEEE), 486–487.

Hadadi, A., Guillet, C., Chardonnet, J.-R., Langovoy, M., Wang, Y., and Ovtcharova, J. (2022). Prediction of cybersickness in virtual environments using topological data analysis and machine learning. *Front. Virtual Real.* 3. doi:10.3389/frvir.2022.973236

Halbig, A., Babu, S., Gatter, S., Latoschik, M., Brukamp, K., and von Mammen, S. (2022). Opportunities and challenges of virtual reality in healthcare – a domain experts inquiry. *Virtual Real.* 3, 837616. doi:10.3389/frvir.2022.837616

Halbig, A., and Latoschik, M. E. (2021). A systematic review of physiological measurements, factors, methods, and applications in virtual reality. *Front. Virtual Real.* 2, 89. doi:10.3389/frvir.2021.694567

Hamzeheinejad, N., Roth, D., Götz, D., Weilbach, F., and Latoschik, M. E. (2019). "Physiological effectivity and user experience of immersive gait rehabilitation," in *2019 IEEE conference on virtual reality and 3D user interfaces (VR)* (IEEE), 1421–1429.

Harada, S., Hayashi, H., and Uchida, S. (2018). Biosignal data augmentation based on generative adversarial networks. in *2018 40th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*, 368–371.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). The elements of statistical learning. in *Springer series in statistics*. New York, NY, USA: Springer New York Inc.

Hildebrandt, L. K., McCall, C., Engen, H. G., and Singer, T. (2016). Cognitive flexibility, heart rate variability, and resilience predict fine-grained regulation of arousal during prolonged threat. *Psychophysiology* 53, 880–890. doi:10.1111/psyp.12632

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi:10.1162/neco.1997.9.8.1735

Huang, Y. C., Backman, K. F., Backman, S. J., and Chang, L. L. (2016). Exploring the implications of virtual reality technology in tourism marketing: an integrated research framework. *Int. J. Tour. Res.* 18, 116–128. doi:10.1002/jtr.2038

Islam, R., Desai, K., and Quarles, J. (2021). "Cybersickness prediction from integrated hmd's sensors: a multimodal deep fusion approach using eye-tracking and head-tracking data," in *2021 IEEE international symposium on mixed and augmented reality (ISMAR)* (IEEE), 31–40.

Islam, R., Lee, Y., Jaloli, M., Muhammad, I., Zhu, D., and Quarles, J. (2020a). "Automatic detection of cybersickness from physiological signal in a virtual roller coaster simulation," in *2020 IEEE conference on virtual reality and 3D user interfaces abstracts and workshops (VRW)* (IEEE), 648–649.

Islam, R., Lee, Y., Jaloli, M., Muhammad, I., Zhu, D., Rad, P., et al. (2020b). "Automatic detection and prediction of cybersickness severity using deep neural networks from user's physiological signals," in *2020 IEEE international symposium on mixed and augmented reality (ISMAR)*, 400–411. doi:10.1109/ISMAR50242.2020.00066

Iwana, B. K., and Uchida, S. (2020). An empirical survey of data augmentation for time series classification with neural networks. *PLoS ONE* 16, e0254841. doi:10.1371/journal.pone.0254841

Jeong, D. K., Yoo, S., and Jang, Y. (2018). "Vr sickness measurement with eeg using dnn algorithm," in *Proceedings of the 24th ACM symposium on virtual reality software and technology*, 1–2.

Kennedy, R. S., Lane, N. E., Berbaum, K. S., and Mg, L. (1993). Simulator sickness questionnaire: an enhanced method for quantifying simulator sickness. *Int. J. Aviat. Psychol.* 3, 203–220. doi:10.1207/s15327108ijap0303_3

Kern, F., Winter, C., Gall, D., Käthner, I., Pauli, P., and Latoschik, M. E. (2019). "Immersive virtual reality and gamification within procedurally generated environments to increase motivation during gait rehabilitation," in *2019 IEEE conference on virtual reality and 3D user interfaces (VR)* (IEEE), 500–509.

Keshavarz, B., and Hecht, H. (2011). Validating an efficient method to quantify motion sickness. *Hum. factors* 53, 415–426. doi:10.1177/0018720811403736

Kim, J., Kim, W., Oh, H., Lee, S., and Lee, S. (2019). "A deep cybersickness predictor based on brain signal analysis for virtual reality contents," in *Proceedings of the IEEE/CVF international conference on computer vision*, 10580–10589.

Kingma, D. P., and Ba, J. (2014). *Adam: a method for stochastic optimization*. doi:10.48550/ARXIV.1412.6980

LaViola Jr, J. J. (2000). A discussion of cybersickness in virtual environments. *ACM Sigchi Bull.* 32, 47–56. doi:10.1145/333329.333344

Lee, S., Kim, S., Kim, H. G., Kim, M. S., Yun, S., Jeong, B., et al. (2019). Physiological fusion net: quantifying individual vr sickness with content stimulus and physiological response. In *2019 IEEE international conference on image processing (ICIP)* (IEEE), 440–444.

Lindner, P., Rozental, A., Jurell, A., Reuterskiöld, L., Andersson, G., Hamilton, W., et al. (2020). Experiences of gamified and automated virtual reality exposure therapy for spider phobia: qualitative study. *JMIR serious games* 8, e17807. doi:10.2196/17807

Liu, C.-L., and Uang, S.-T. (2012). "A study of sickness induced within a 3d virtual store and combated with fuzzy control in the elderly," in *2012 9th international conference on fuzzy systems and knowledge discovery* (IEEE), 334–338.

Llorach, G., Evans, A., and Blat, J. (2014). "Simulator sickness and presence using hmds: comparing use of a game controller and a position estimation system," in *Proceedings of the 20th ACM symposium on virtual reality software and technology*, 137–140.

Lopez-Paz, D., and Oquab, M. (2017). "Revisiting classifier two-sample tests," in *International conference on learning representations*.

Loureiro, S. M. C., Guerreiro, J., Eloy, S., Langaro, D., and Panchapakesan, P. (2019). Understanding the use of virtual reality in marketing: a text mining-based review. *J. Bus. Res.* 100, 514–530. doi:10.1016/j.jbusres.2018.10.055

Lubeck, A. J., Bos, J. E., and Stins, J. F. (2015). Motion in images is essential to cause motion sickness symptoms, but not to increase postural sway. *Displays* 38, 55–61. doi:10.1016/j.displa.2015.03.001

Martin, N., Mathieu, N., Pallamin, N., Ragot, M., and Diverrez, J.-M. (2020). "Virtual reality sickness detection: an approach based on physiological signals and machine learning," in *2020 IEEE international symposium on mixed and augmented reality (ISMAR)*, 387–399. doi:10.1109/ISMAR50242.2020.00065

Miloff, A., Lindner, P., Hamilton, W., Reuterskiöld, L., Andersson, G., and Carlbring, P. (2016). Single-session gamified virtual reality exposure therapy for spider phobia vs. traditional exposure therapy: study protocol for a randomized controlled non-inferiority trial. *Trials* 17, 60–68. doi:10.1186/s13063-016-1171-1

Mirza, M., and Osindero, S. (2014). *Conditional generative adversarial nets*. doi:10.48550/ARXIV.1411.1784

Nair, V., and Hinton, G. E. (2010). "Rectified linear units improve restricted Boltzmann machines," in Proceedings of the 27th international Conference on international Conference on machine learning *(madison, WI, USA: omnipress), ICML'10*, 807–814.

Nakagawa, C. (2015). "Toward the detection of the onset of virtual reality sickness by autonomic indices," in *2015 IEEE 4th global conference on consumer electronics (GCCE)* (IEEE), 662–663.

Nalivaiko, E., Davis, S., Blackmore, K., Vakulin, A., and Nesbitt, K. (2015). Cybersickness provoked by head-mounted display affects cutaneous vascular tone, heart rate and reaction time. *Aut. Neurosci.* 192, 63. doi:10.1016/j.autneu.2015.07.032

Nikolaidis, K., Kristiansen, S., Goebel, V., Plagemann, T., Liestøl, K., and Kankanhalli, M. (2019). "Augmenting physiological time series data: a case study for sleep apnea detection," in *Machine learning and knowledge discovery in databases: European conference, ECML PKDD 2019, würzburg, Germany, september 16–20, 2019, proceedings, Part III* (Berlin, Heidelberg: Springer-Verlag), 376–399. doi:10.1007/978-3-030-46133-1_23

Oberdörfer, S., Heidrich, D., and Latoschik, M. E. (2017). "Interactive gamified virtual reality training for affine transformations," in *DeLFI/GMW workshops*.

Oman, C. M. (1990). Motion sickness: a synthesis and evaluation of the sensory conflict theory. *Can. J. physiology Pharmacol.* 68, 294–303. doi:10.1139/y90-044

Pallavicini, F., Pepe, A., and Minissi, M. E. (2019). Gaming in virtual reality: what changes in terms of usability, emotional response and sense of presence compared to non-immersive video games? *Simul. Gaming* 50, 136–159. doi:10.1177/1046878119831420

Pane, E. S., Khoirunnisaa, A. Z., Wibawa, A. D., and Purnomo, M. H. (2018). Identifying severity level of cybersickness from eeg signals using cn2 rule induction algorithm. *2018 Int. Conf. Intelligent Inf. Biomed. Sci. (ICIIBMS)* 3, 170–176. doi:10.1109/iciibms.2018.8549968

Rangelova, S., Motus, D., and André, E. (2020). "Cybersickness among gamers: an online survey," in *Advances in human factors in wearable Technologies and game design*. Editor T. Ahram (Cham: Springer International Publishing), 192–201.

Rebenitsch, L., and Owen, C. (2016). Review on cybersickness in applications and visual displays. *Virtual Real.* 20, 101–125. doi:10.1007/s10055-016-0285-9

Recenti, M., Ricciardi, C., Aubonnet, R., Picone, I., Jacob, D., Svansson, H. Á., et al. (2021). Toward predicting motion sickness using virtual reality and a moving platform assessing brain, muscles, and heart signals. *Front. Bioeng. Biotechnol.* 9, 635661. doi:10.3389/fbioe.2021.635661

Riccio, G. E., and Stoffregen, T. A. (1991). An ecological theory of motion sickness and postural instability. *Ecol. Psychol.* 3, 195–240. doi:10.1207/s15326969eco0303_2

Saredakis, D., Szpak, A., Birckhead, B., Keage, H. A., Rizzo, A., and Loetscher, T. (2020). Factors associated with virtual reality sickness in head-mounted displays: a systematic review and meta-analysis. *Front. Hum. Neurosci.* 14, 96. doi:10.3389/fnhum.2020.00096

Schuster, M., and Paliwal, K. (1997). Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* 45, 2673–2681. doi:10.1109/78.650093

Sharples, S., Cobb, S., Moody, A., and Wilson, J. R. (2008). Virtual reality induced symptoms and effects (vrise): comparison of head mounted display (hmd), desktop and projection display systems. *Displays* 29, 58–69. doi:10.1016/j.displa.2007.09.005

Shorten, C., and Khoshgoftaar, T. (2019). A survey on image data augmentation for deep learning. *J. Big Data* 6, 60. doi:10.1186/s40537-019-0197-0

Socket (2022). Socket — low-level networking interface. *Comput. Softw.*

Stanney, K. M., Kennedy, R. S., and Drexler, J. M. (1997). Cybersickness is not simulator sickness. in *Proceedings of the human factors and ergonomics society annual meeting*. CA: Los Angeles, CA: SAGE Publications Sage, 1138–1142. doi:10.1177/107118139704100292

Stauffert, J.-P., Niebling, F., and Latoschik, M. E. (2018). "Effects of latency jitter on simulator sickness in a search task," in *2018 IEEE conference on virtual reality and 3D user interfaces (VR)* (IEEE), 121–127.

Stauffert, J.-P., Niebling, F., and Latoschik, M. E. (2020). Latency and cybersickness: impact, causes, and measures. a review. *Front. Virtual Real.* 1, 582204. doi:10.3389/frvir.2020.582204

Tauscher, J.-P., Witt, A., Bosse, S., Schottky, F. W., Grogorick, S., Castillo, S., et al. (2020). Exploring neural and peripheral physiological correlates of simulator sickness. *Comput. Animat. Virtual Worlds* 31, e1953. doi:10.1002/cav.1953

Treisman, M. (1977). Motion sickness: an evolutionary hypothesis. *Science* 197, 493–495. doi:10.1126/science.301659

Um, T. T., Pfister, F. M. J., Pichler, D., Endo, S., Lang, M., Hirche, S., et al. (2017). "Data augmentation of wearable sensor data for Parkinson's disease monitoring using convolutional neural networks," in *Proceedings of the 19th ACM international conference on multimodal interaction* (New York, NY, USA: Association for Computing Machinery), 216–220. doi:10.1145/3136755.3136817

Unity Asset Store (2023). *Unity asset store*.

Unity Technologies (2020). *Unity real-time development platform*.

Viavr_Measurement_Engine, Yalcin, M.H., Latoschik, A., and Erich, M. (2022). Measurement Engine–Technology Platform for Virtual Adventures (VIA-VR). Available at: https://hci.uni-wuerzburg.de/projects/via-vr/ (Accessed May 10, 2024).

Viavr_Project, von Mammen, S., Latoschik, A., Botsch, M.E., and Brukamp, M. (2019). "VIA VR: A technology platform for virtual adventures for healthcare and well-being," in 2019 11th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games), 1–2. doi:10.1109/VS-Games.2019.8864580

Wang, J., Liang, H.-N., Monteiro, D. V., Xu, W., and Xiao, J. (2023). Real-time prediction of simulator sickness in virtual reality games. *IEEE Trans. Games* 15, 252–261. doi:10.1109/tg.2022.3178539

Wen, Q., Sun, L., Yang, F., Song, X., Gao, J., Wang, X., et al. (2021). *Time series data augmentation for deep learning: a survey*, 4653–4660. doi:10.24963/ijcai.2021/631

Wienrich, C., Weidner, C. K., Schatto, C., Obremski, D., and Israel, J. H. (2018). "A virtual nose as a rest-frame-the impact on simulator sickness and game experience," in *2018 10th international conference on virtual worlds and games for serious applications (VS-Games)* (IEEE), 1–8.

Wolf, E., Döllinger, N., Mal, D., Wienrich, C., Botsch, M., and Latoschik, M. E. (2020). "Body weight perception of females using photorealistic avatars in virtual and augmented reality," in *2020 IEEE international symposium on mixed and augmented reality (ISMAR)*, 462–473.

Wolf, E., Merdan, N., Döllinger, N., Mal, D., Wienrich, C., Botsch, M., et al. (2021). "The embodiment of photorealistic avatars influences female body weight perception in virtual reality," in *Proceedings of the 28th IEEE virtual reality conference* (VR 21), 65–74. doi:10.1109/VR50410.2021.00027

Xu, B., Wang, N., Chen, T., and Li, M. (2015). Empirical evaluation of rectified activations in convolutional network

Yang, A., Kasabov, N., and Cakmak, Y. (2022). Machine learning methods for the study of cybersickness: a systematic review. *Brain Inf.* 9, 24. doi:10.1186/s40708-022-00172-6

Yang, D., Hong, S., Jang, Y., Zhao, T., and Lee, H. (2019). "Diversity-sensitive conditional generative adversarial networks," in *International conference on learning representations*.

frontiers | Frontiers in Big Data

TYPE Original Research
PUBLISHED 26 June 2024
DOI 10.3389/fdata.2024.1371680

OPEN ACCESS

EDITED BY
Rui Qin,
Manchester Metropolitan University,
United Kingdom

REVIEWED BY
Hanqing Zhao,
Hebei University, China
Patrick Mikalef,
NTNU, Norway
Li Xia,
South China University of Technology, China

*CORRESPONDENCE
Alaa Marshan
✉ a.marshan@surrey.ac.uk

RECEIVED 16 January 2024
ACCEPTED 10 June 2024
PUBLISHED 26 June 2024

CITATION
Marshan A, Almutairi AN, Ioannou A, Bell D, Monaghan A and Arzoky M (2024) MedT5SQL: a transformers-based large language model for text-to-SQL conversion in the healthcare domain. *Front. Big Data* 7:1371680. doi: 10.3389/fdata.2024.1371680

MedT5SQL: a transformers-based large language model for text-to-SQL conversion in the healthcare domain

Alaa Marshan[1]*, Anwar Nais Almutairi[2], Athina Ioannou[3], David Bell[4], Asmat Monaghan[5] and Mahir Arzoky[6]

[1]School of Computer Science and Electronic Engineering, University of Surrey, Guildford, United Kingdom, [2]College of Business Studies, PAAET, Kuwait City, Kuwait, [3]Surrey Business School, University of Surrey, Guildford, United Kingdom, [4]Department of Computer Science, Brunel University London, London, United Kingdom, [5]School of Business and Management, Royal Holloway, University of London, London, United Kingdom, [6]Department of Computer Science, Brunel University London, London, United Kingdom

Introduction: In response to the increasing prevalence of electronic medical records (EMRs) stored in databases, healthcare staff are encountering difficulties retrieving these records due to their limited technical expertise in database operations. As these records are crucial for delivering appropriate medical care, there is a need for an accessible method for healthcare staff to access EMRs.

Methods: To address this, natural language processing (NLP) for Text-to-SQL has emerged as a solution, enabling non-technical users to generate SQL queries using natural language text. This research assesses existing work on Text-to-SQL conversion and proposes the MedT5SQL model specifically designed for EMR retrieval. The proposed model utilizes the Text-to-Text Transfer Transformer (T5) model, a Large Language Model (LLM) commonly used in various text-based NLP tasks. The model is fine-tuned on the MIMICSQL dataset, the first Text-to-SQL dataset for the healthcare domain. Performance evaluation involves benchmarking the MedT5SQL model on two optimizers, varying numbers of training epochs, and using two datasets, MIMICSQL and WikiSQL.

Results: For MIMICSQL dataset, the model demonstrates considerable effectiveness in generating question-SQL pairs achieving accuracy of 80.63%, 98.937%, and 90% for exact match accuracy matrix, approximate string-matching, and manual evaluation, respectively. When testing the performance of the model on WikiSQL dataset, the model demonstrates efficiency in generating SQL queries, with an accuracy of 44.2% on WikiSQL and 94.26% for approximate string-matching.

Discussion: Results indicate improved performance with increased training epochs. This work highlights the potential of fine-tuned T5 model to convert medical-related questions written in natural language to Structured Query Language (SQL) in healthcare domain, providing a foundation for future research in this area.

KEYWORDS

text-to-SQL conversion, large language model, transformers, T5 model, NLP, MIMICSQL dataset, healthcare domain

1 Introduction

Large businesses, government departments, healthcare providers, financial services and many others store their vast amounts of data in large relational databases or [data centers. To handle, manage and retrieve information from these databases, it is required to know the necessary technical background which non-technical people lack. For example, Structured Query Language (SQL), a standardized programming language that performs a variety of data operations to manage databases, provides special communication with databases typically required for efficient data management, including retrieval, deletion and updating records (Groff et al., 2002). One prominent use of relational databases is in today's healthcare domain, where patients' health information is stored in databases as electronic medical records (EMRs), designed to ensure that every patient receives the correct medical care, based on their entire health history. EMRs also help researchers gather the statistics required for clinical trials, in turn helping the study of diseases and the provision of suitable cures. To carry out their duties, healthcare professionals must be able to access EMRs, however, while they are considered experts in their medical fields, they often lack formal training in database query languages like SQL. This can result in significant inefficiencies when attempting to extract relevant patient information from Electronic Medical Records (EMRs). Studies have shown that clinicians spend a considerable amount of their time on documentation and data entry tasks, often leading to frustration and burnout (Shanafelt et al., 2012; Sinsky et al., 2016). A survey of over 4,000 physicians revealed that 49% reported spending more than half their workday interacting with EHRs (American Medical Association, 2018). Moreover, the complexity of EMR databases, with their intricate schemas and vast amounts of data, can further exacerbate these challenges. This difficulty in accessing data can hinder clinical decision-making, delay patient care, and impede research efforts. For instance, a study found that difficulties in retrieving relevant information from EMRs contributed to diagnostic errors in 25% of cases (Singh et al., 2013). Therefore, an intermediate system is therefore needed that can assist end-users, such as the healthcare staff, to handle database records smoothly without needing to learn SQL.

Responding to this need, researchers started to explore the possibility of employing automated Text-to-SQL conversion, using machine learning (ML) and natural language processing (NLP) to convert questions written in natural language to SQL queries; the principle is shown in Figure 1 (Iyer et al., 2017; Kate et al., 2018; Kim et al., 2020). NLP is a pervasive artificial intelligence (AI) technology in which computers simulate human intelligence through machine learning. Without explicit programming, machine learning automates the learning of computers using a collected data based on the required task. In this way, computers are given the ability to understand human language and turn it into machine language to perform required tasks, such as text summarization and translation. Text-to-SQL conversion facilitates the development of flexible, highly interactive communication with databases to handle the records without the need for end-users to know SQL.

Previous research papers have analyzed the creation of SQL through NLP and proposed Text-to-SQL conversion models such as SQLNet, proposed by Xu et al. (2017), Seq2SQL, developed by Zhong et al. (2017) and MedTS, created by Pan et al. (2021). Recently, the NLP technology has progressed with the development of Transformer, a deep neural network architecture capable of multiple NLP tasks, such as automatic summarization and translation (Vaswani et al., 2017). This architecture became the baseline for various language models trained on large data to perform NLP tasks, such as Bidirectional Encoder Representations from Transformers (BERT), proposed by Devlin et al. (2019) and Multi-Task Deep Neural Networks (MT-DNN) for Natural Language, proposed by Liu X. et al. (2019). Transfer learning these pre-trained models, in which they are fine-tuned on a downstream task such as translation, has become an effective approach in NLP research. In Text-to-SQL conversion, fine-tunning pre-trained models has raised the performance of Text-to-SQL models to near human performance levels (Guo et al., 2019; Wang et al., 2019; Pan et al., 2021). Subsequently, Raffel et al. (2020) proposed their model, namely Text-to-Text Transfer Transformer (T5) as a unified model for various NLP tasks and is considered one of the first Large language Models (LLMs). The T5 model transforms text-based language problems, such as translation, into a text-to-text format and has become the state-of-the-art for various NLP tasks, such as summarization, question answering and text classification (Raffel et al., 2020; Xie et al., 2022). Using the T5 model for Text-to-SQL conversion resulted in a significant improvement in the performance of such task (Scholak et al., 2021; Xie et al., 2022).

While many researchers have proposed Text-to-SQL conversion models, few have focused explicitly on the healthcare domain to assist healthcare staff in managing and retrieving information from EMRs (Wang et al., 2020; Pan et al., 2021). This relative scarcity can be attributed to several factors. First, healthcare data presents unique challenges, including complex medical terminologies, diverse data formats across different EMR systems, and stringent privacy and security requirements. These challenges necessitate the development of specialized Text-to-SQL models that can accurately understand medical language and comply with healthcare-specific regulations. Second, the integration of Text-to-SQL systems with existing EMR systems can be complex and time-consuming. The heterogeneity of EMR systems across different healthcare institutions, with varying data structures and terminologies, poses a significant barrier to generalizability. Developing a Text-to-SQL model that seamlessly integrates with diverse EMR systems requires extensive customization and validation, which may deter researchers and practitioners from focusing on this domain.

Despite these challenges, the need for efficient and user-friendly access to EMR data remains critical for healthcare professionals. Therefore, this work aims to develop a T5-based model, namely MedT5SQL, which is a transformers-based fine-tuned large language model to perform Text (questions)-to-SQL conversion specifically within the healthcare domain. The objective of the MedT5SQL model is to empower medical staff by enabling them to express their data requests in natural language, thereby overcoming the barriers associated with traditional SQL query formulation.

In achieving the above, this paper is structured as follows. First, a theoretical background of the work related to text-to-SQL conversion is discussed. The following section clarifies the research

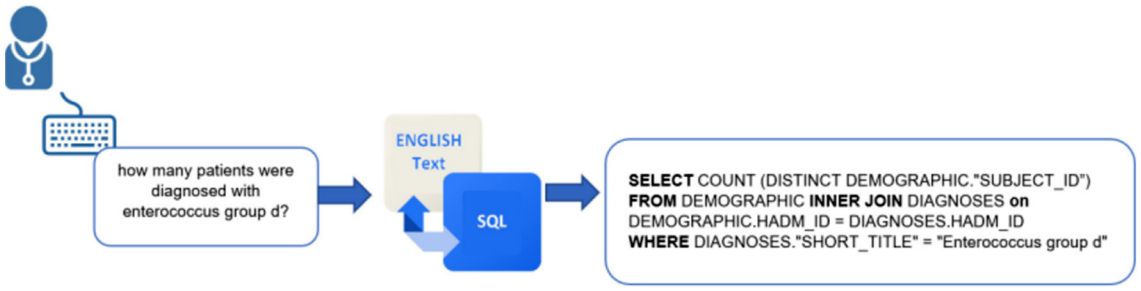

FIGURE 1
Text-to-SQL conversion model.

methodology, namely CRoss Industry Standard Process for Data mining (CRISP-DM), that is followed to pre-process the data and develop and validate MedT5SQL model. Third, the evaluation results are discussed in detail and compared to past research. Finally, the conclusion section concludes this work and offers some suggestions for future research.

2 Theoretical background

Nowadays, patients' health information is stored in a digital format in electronic medical records (EMRs) that are used by healthcare staff to retrieve patients' historical health details or to use for clinical trials. At the beginning of 2020, the world experienced a global pandemic of coronavirus known as COVID-19. This pandemic has left hospitals overloaded with patients, causing enormous stress on healthcare workers due to shortages of medical staff in relation to the number of patients (Birkmeyer et al., 2020; Kruizinga et al., 2021; Iness et al., 2022). The pandemic highlighted the importance of EMRs and revealed the need for a faster communication method to handle it (Dagliati et al., 2021). It is essential to have an interface that provides easy user-to-database interactions; in particular, a system that generates an SQL query in response to a question in human language. This section reviews the state-of-the-art in natural language processing for Text-to-SQL conversion to facilitate interactions between users and databases.

2.1 Rule-based systems

Converting natural language to SQL is a subtask of semantic parsing, in which natural language is converted into a machine-understandable logical form (Zettlemoyer and Collins, 2005). Semantic parsing seeks to understand the meaning of natural language and map it to logical forms such as SQL. Rule-based systems were used to support non-technical users in communicating with databases through a set of predefined rules mapping natural language words with SQL keywords and database schemas (Androutsopoulos et al., 1995; Popescu et al., 2004; Li and Jagadish, 2014; Saha et al., 2016). An expert programmer constructs these rules to translate users' requirements into SQL queries (Masri et al., 2019). However, it is required for non-technical users to train before using them and are domain-specific, since each system is built for a specific schema. These systems have limited intelligence, as they only operate based on the rules created by humans and do not learn, change or update on their own (Kamath and Das, 2018). This limits the ability of non-technical users to manage their data without relying on expert programmers.

2.2 Deep learning models for text-to-SQL

To increase usability and generalize Text-to-SQL conversion, researchers began using deep learning (DL) by training neural networks to generate executable SQL queries. Training neural networks means performing supervised learning, in which the network is provided with natural language questions and their corresponding SQL queries so it can learn the conversion. The trained neural networks is called a DL model that generates a query from a given question. This has led to the release of several Text-to-SQL datasets that boost the accuracy of the models by delivering sufficient data for supervised learning: GeoQuery, created by Zelle (1996) for US geography and updated later by Iyer et al. (2017) to include SQL; ATIS, created by Price (1990) for flight bookings and updated by Iyer et al. (2017) to include SQL; Scolar, created by Iyer et al. (2017) for academic publications; WikiSQL, created by Zhong et al. (2017) from Wikipedia; and Spider, created by Yu et al. (2018a) and representing a cross-domain dataset.

Due to their large sizes and multiple domain coverage, Spider and WikiSQL are the most used datasets among researchers. WikiSQL is a corpus of 80,654 hand-annotated pairs of questions and corresponding SQL queries for 24,241 tables covering multiple domains. However, each question-SQL pair is related to a single table in which the SQL only has SELECT and WHERE clauses, as presented in Figure 2. The Spider dataset was introduced to overcome WikiSQL's simple SQL structure and to present the first cross-domain dataset. It includes 200 complex databases with multiple tables, 10,181 questions, and 5,693 corresponding complex SQL queries with nested queries. Table 1 presents a comparison of existing datasets for text-to-SQL translation.

2.2.1 Deep learning models architecture

Deep learning models for Text-to-SQL conversion are built as neural networks in an encoder-decoder architecture that was initially embraced by Sutskever et al. (2014) for

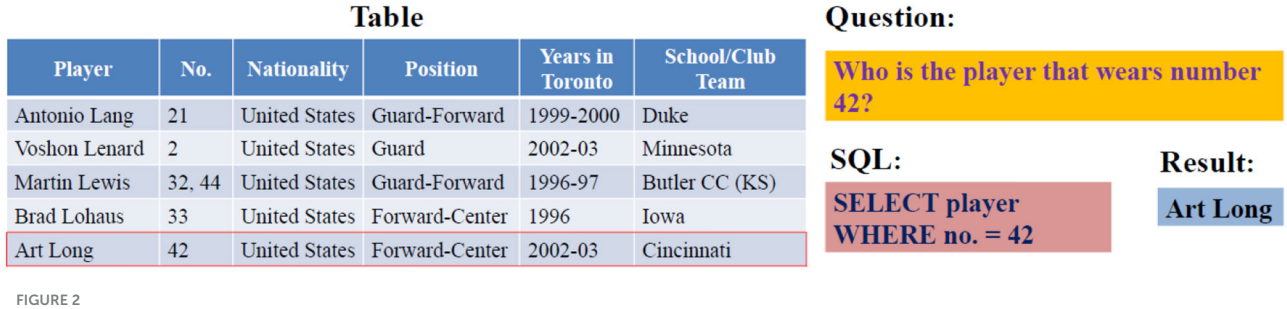

FIGURE 2
Sample of WikiSQL question-SQL pairs (Xu et al., 2017).

TABLE 1 Comparison of existing text-to-SQL benchmarking databases.

Dataset	#Databases	#Tables per database	#Question-SQL pairs	SQL query level
ATIS	1	32	5,280	Complex (no HAVING and ORDER BY)
GeoQuery	1	6	877	Complex
Scolar	1	7	817	Simple
Spider	200	On average 5	10,181	Complex
WikiSQL	24241	1	80,654	Simple

translation purposes. Given a natural language question (NLQ) and its corresponding SQL as source sequences, models operate as follows:

- The source sequences are always tokenized into tokens before encoding, and each token represents a word in the sequence (Webster and Kit, 1992).
- As deep learning models only take numbers as inputs, each token is embedded into a vector representation, called word embedding, using embedding algorithms such as Glove or Word2vec (Mikolov et al., 2013; Pennington et al., 2014). This process reveals the relationship between tokens and reduces input dimensionality as tokens with similar meanings have similar vector representation.
- The encoder takes the NLQ tokens' embeddings and encodes their information/features into a vector named "hidden states."
- For training purposes, the decoder takes the encoder's hidden states and the word embedding of the SQL tokens for the supervised training. The decoder is built and trained as a classifier to decode the hidden states into a target SQL query.
- For generalization to an unseen schema, the database schema is usually considered as an input to the models.

To provide an accurate conversion, models must develop an understanding of source sequences by understanding words' dependencies and memorizing previously gathered information. To meet this need, researchers have built encoders and decoders with recurrent neural networks (RNNs), particularly long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997). LSTM can remember long-term information and capture dependencies between sequence tokens. The understanding and encoding of each token depends on the previously seen token. Therefore, it can improve natural language understanding and help with translation tasks (Graves, 2013; Yin et al., 2017).

Encoders built using LSTM take input tokens sequentially and produce their hidden states one at a time. At the end, it outputs a single hidden states vector compressing all the tokens' hidden states. The decoder alone needs to interpret the information compressed in this vector into a complex target sequence, leading to the risk of information loss. To circumvent this risk, an attention mechanism was proposed to allow the decoder to look at all tokens' hidden states when predicting the final output (Bahdanau et al., 2014; Galassi et al., 2021). This is accomplished by passing the weighted sum of the hidden states to the decoder, allowing it to focus on the required information to generate the next target token. This simplifies the encoder task by avoiding encoding the entire source sequence into a single vector. The architecture of the encoder-decoder with and without attention mechanism can be seen in Figure 3 where 'h' corresponding to hidden states vectors.

2.2.2 Deep learning approaches

In Text-to-SQL tasks, this sub-section outlines the approaches used as (1) sequence-to-sequence (2) sequence-to-set (3) fine-tuning a pre-trained language model (transfer learning).

Sequence-to-sequence (Seq2Seq), introduced by Sutskever et al. (2014), is an LSTM-based machine translation that operates by sequentially taking source tokens and translating them into sequence target tokens. Seq2Seq relies on a single ground truth query as the optimal correct query. This raises the issue of "order matter" because in SQL, the order in the WHERE clause does not matter, making it a challenge when using this approach. Seq2Seq

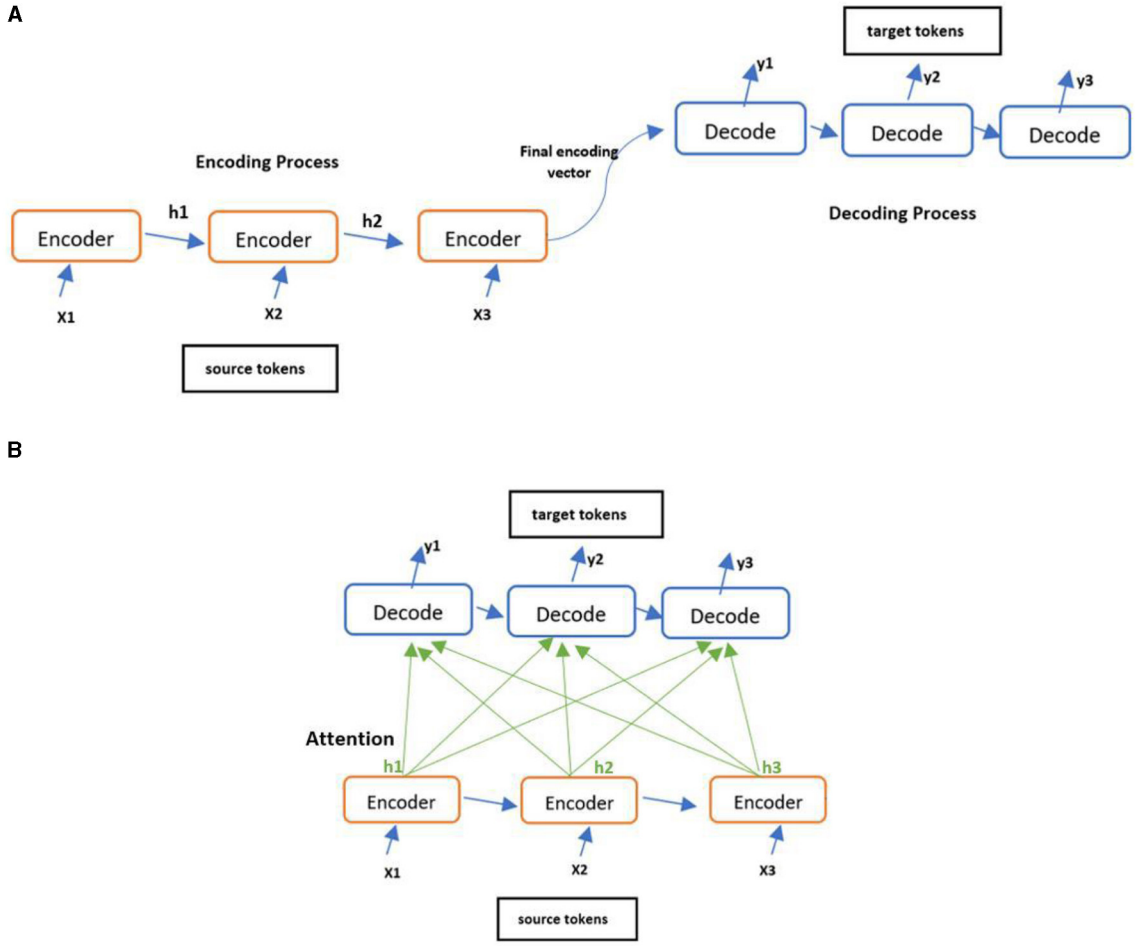

FIGURE 3
Various encode-decoder architectures. **(A)** Encoder-Decoder architecture and **(B)** Encoder-Decoder architecture with attention.

does not require the attention mechanism; however, it is possible to combine the two for better results.

Sequence-to-set was first introduced by Xu et al. (2017). It is similar to Seq2Seq, apart from its ability to overcome the order matter by producing an unordered set of sequences after dividing the prediction into sections. The decoder prediction is performed based on the dependency between the predicted tokens, which is captured using the attention mechanism. Sequence-to-set usually uses an approach of sketch matching and slot filling, where each slot has its own decoder. The slots present parts of the SQL, such as the column name or the aggregation operator, in the SELECT clause. Using a sketch structure presenting the dependencies of the query slots, the decoding of each slot in the query is based only on the decoding of other slots it depends on. For example, decoding the aggregation operator in the SELECT clause depends on the decoding of the column name and is independent of the WHERE clause.

Pre-trained language models are transformer-based neural networks for word embedding that learn contextual relations between tokens without recurrent connections (Peters et al., 2018; Yang et al., 2019). The Transformer is an encoder-decoder-based neural network proposed by Vaswani et al. (2017). It is built and trained to work on multiple NLP tasks, such as summarization and translation. The transformer has three main functioning concepts. The first is positional encoding, in which transformers are fed with all the tokens at once, with each token appended with its order, unlike the recurrent neural network of sequential input of token. Second, through learning from training data, transformers use the attention mechanism and consider each input token in the source before any translation prediction is generated. Third, both the encoder and decoder use a self-attention mechanism in which a word is understood based on the context of the words around it (Vaswani et al., 2017).

Although transformers are encoder-decoder neural networks, pre-trained language models only use the encoding mechanism, as they aim to learn representations of a language. The most commonly used language model in text-to-SQL conversion is bidirectional encoder representation from transformers (BERT), introduced by Devlin et al. (2019). The term "bidirectional" means positional encoding and the term "representation" refers to the attention mechanisms. BERT is a multi-layer bidirectional transformer encoder for contextual-bidirectional embeddings that can be finetuned for specific NLP tasks. It was trained by two learning mechanisms—masked learning mechanism (MLM) and

next sentence prediction (NSP)—to increase its accuracy and minimize the loss values. In MLM, 15% of input tokens are placed with masked tokens (MASK) before being given to BERT. Therefore, through contextual relations between tokens, BERT learns to predict the original token. In NSP, BERT is given pairs of sentences and trained to predict whether the pairs are subsequent to each other in the source text. It is fed by 50% subsequent pairs during training, where sentences are separated by special tokens at the start of the first sentence in each pair and at the end of each sentence. Most pre-trained models were built later, based on BERT (Sanh et al., 2019; Liu X. et al., 2019).

To apply transfer learning with pre-trained models, researchers must perform fine-tuning by re-training the model using one of the Text-to-SQL datasets. In GloVe and Word2Vec, each token is embedded into one static vector representation. However, as a result of the attention mechanism in BERT, a token appearing in multiple locations in the source is treated as different tokens, thus embedded into multiple word embeddings/vectors based on its context.

Most of the text-to-SQL models were evaluated using:

Execution accuracy: this metric compares the results of executing the ground truth query (gold standard) with the results of executing the model-generated query. While intuitive, it can be misleading, especially in situations where multiple queries produce the same result. For instance, consider a query to find the average age of patients. Both SELECT AVG(age) FROM patients and SELECT AVG(age), COUNT(*) FROM patients would yield the same average age, but only the first query accurately captures the intent of the natural language question.

Logical form accuracy (exact match): This metric compares the structure of the ground truth query with the generated query using an accuracy matrix. It addresses the limitation of execution accuracy by focusing on structural correctness. However, it can be overly strict, as minor variations in query formulation (e.g., different ordering of clauses) can lead to incorrect results even if the queries are functionally equivalent.

Manual matching: In this approach, human evaluators manually compare the structure of the ground truth query with the generated query, often using a set of predefined criteria. Manual matching offers a nuanced assessment of query correctness, but it can be time-consuming and subjective.

Combination of metrics: Given the limitations of individual metrics, using a combination of execution accuracy, logical form accuracy, and manual matching provides a more comprehensive evaluation. Execution accuracy verifies the functional correctness of the query, while logical form accuracy and manual matching assess its structural correctness and alignment with the natural language question's intent.

2.2.3 Text-to-SQL in single domain dataset

WikiSQL is considered the biggest single dimension dataset used for Text-to-SQL, where each SQL is related to a single database table. Seq2SQL, created by Zhong et al. (2017), was the first model trained with WikiSQL. It uses a Seq2Seq approach designed to leverage the structure of SQL commands with three decoders for the SELECT column clause, aggregation operator and WHERE clause separately. It uses two encoders, one for the question tokens and another for the column name, to train the model to generate the SQL query given the question and column. The decoder was designed as an LSTM-augmented pointer network created by Vinyals et al. (2015). It augments the encoder's output along with an SQL vocabulary of required SQL operations to produce the SQL query with tokens taken exclusively from this augmentation. To minimize the effect of the order matter problem, Seq2SQL uses reinforcement learning with policy gradients presented by Sutton et al. (1999), allowing the decoder to evaluate the predicted query based on whether it is well formed or not. The model achieved an execution accuracy of 59.4% and a logical form of 48.3%. Even though it presented a state-of-the-art model for WikiSQL, an accuracy below 50% is considered insufficient.

SELECT <aggregator><Column>
WHERE <Column><Operator><Value>
(**AND** <Column><Operator><Value>) *

FIGURE 4
SQLNet SQL sketch.

SQLNet is a sequence-to-set sketch-based approach developed by Xu et al. (2017) to avoid the order matter. The dependency between the slots is based on the SQL sketch shown in Figure 4, where five decoders were used. Tokens between "< >" are the slots to be filled, while (*) indicates zero or more conditions. The aggregator options are NULL, MAX, MIN, COUNT, SUM and AVG, while the operator options are =, > and <. Additionally, SQLNet uses a column-attention mechanism in which one LSTM encoder is used over each column name and another is used to encode the NLQ conditional in each column. In this way, the model reflects the most relevant word in the question when predicting the column name. SQLNet structure allowed it to achieve around 10% improvement in the execution accuracy compared to Seq2SQL. TYPESQL, developed by Yu et al. (2018c), is an improved version of SQLNet with a 5.5% increase in accuracy. TYPESQL achieves 2% higher accuracy by concatenating each NLQ token with a type before encoding to assist the decoder in filling the slots. For example, the model uses INTEGER, FLOAT, DATE or YEAR for number tokens, COLUMN for column name tokens and PERSON, PLACE, COUNTRY, ORGANIZATION and SPORT for named entities. TYPESQL achieves the other 3.5% by grouping related slots together, resulting in three decoders. All models use GloVe word embedding for the encoder embedding layer.

Due to their functionality, pre-trained models are effective in revealing the connections between source sequences as well as portraying the meaning of the question. Therefore, researchers began using them to connect questions with table schema to produce accurate SQL queries. Hwang et al. (2019) developed SQLova, the first model to utilize BERT in text-to-SQL tasks for word embedding on WikiSQL. SQLova was created following a sequence-to-set approach with LSTM. It has two separate encoders: one for the question and one for the column names. BERT is used on top of the encoders to perform word embedding for the question and column names. This allows the model to capture a larger

context of the input with any possible different pronunciations of the question. Inspired by SQLNet, SQLova follows the same decoding process as well as a sixth decoder for "where-number," indicating the number of conditions. SQLova uses the execution guided (EG) decoding proposed by Wang et al. (2018) to exclude non-executable generated queries from the decoder output. By using BERT, SQLova achieves 80.7% logical form accuracy and 86.2% execution accuracy without EG and 83.6% and 89.6% with EG. Therefore, using a pre-trained model increased the accuracy of this task.

X-SQL, created by He et al. (2019), was built based on the SQLova structure. Similar to SQLova, X-SQL uses two encoders for the question and table schema. However, the question encoder is built using the multi-task deep neural network (MT-DNN), a pre-trained model proposed by Liu X. et al. (2019) based on BERT. With and without EG, X-SQL outperforms SQLova by 2–4%. This implies that using a pre-trained model as an encoder rather than a word embedder results in better performance. Lyu et al. (2020) argued that neither SQLova nor X-SQL benefit correctly from using a pre-trained language model and added complexity to the models. They proposed Hydranet, employing BERT alone as its encoder without using any other encoders. Instead of pairing the question with all table schemas, Hydranet pairs the question with each column one at a time before encoding. Hydranet was able to achieve state-of-the-art on WikiSQL by reaching 91.8% execution accuracy using EG and 92.2% when replacing BERT with RoBERTa (Liu Y. et al., 2019). Even though those models keep increasing accuracy for text-to-SQL, they are trained on WikiSQL, which means they can manage simple SQL structures.

2.2.4 Text-to-SQL in cross domain dataset

To develop text-to-SQL tasks for complex SQL queries, the Spider dataset was proposed, motivating researchers to develop models for more realistic SQL tasks. Yu et al. (2018c) evaluated SQLNet and TYPESQL on Spider to study their functionality for complex queries. It was found that both models failed to manage nested queries because they limited the query to a defined sketch structure. Motivated by SQLNet, they proposed SyntaxSQLNet (Yu et al., 2018b). As Spider question-SQL pairs can relate to multiple tables, SyntaxSQLNet encoding considers both tables and column names for column embeddings. They employed grammar-based decoding, in which a series of grammar rules are applied sequentially to generate the SQL query. By recursively calling nine independent sequence-to-set decoders, they obtained their SQL syntax tree to generate the SQL. In SyntaxSQLNet, decoders share their decoding history to facilitate the prediction of nested queries; thus, given the current training sample's SQL tokens and the history of previous decoded SQL, the relevant decoder is invoked. Even though its performance was better than SQLNet and TYPESQL on Spider, it achieved an accuracy below 30% due to the complexity of Spider's SQL.

Lee (2019) presented RCSQL, a clause-wise SQL decoding model, to predict syntactically correct SQL. Each clause decoder consists of sub-models matching its clause syntax and implied history sharing. For further improvement, they conducted a self-attention mechanism on database schema encoding. RCSQL's exact matching accuracy was 28.8%, indicating that improvement is still needed. IRNet, created by Guo et al. (2019), adopted the grammar-based model of SyntaxSQLNet. It focused on addressing the challenge of out-of-domain words affecting column prediction. They proposed the use of schema linking, where the model identifies the dataset's columns, tables and conditions appearances in the question. This enhanced the question and schema representations, aiding in their understanding. The model achieved around 20% improvement over SyntaxSQLNet. Inspired by SQLova, Guo et al. (2019) augmented BERT with both SyntaxSQLNet and IRNet. As a result, the performance of both models increased by around 5%. Choi et al. (2021) proposed a complete sketch to synthesize nested queries in the SELECT clause. They also proposed statement position code (SPC) to transform nested SQL queries into non-nested SELECT clauses and to apply sketch-based slot-filling decoding recursively on each statement. With BERT as an encoder, their model RYANSQL achieved 58.2% exact match accuracy on the Spider benchmark.

Unlike models built on WikiSQL, which deals with table schema, Spider models need to handle table schema relations or database schemas since the question-SQL pair represents multi-table relations. Accordingly, the researchers began contextualizing the dataset schema with the question to boost performance. As seen in IRNet, th performance improved with schema linking. RAT-SQL is a grammar-based model presented by Wang et al. (2019) with an encoder that contextualizes the schema and the question using a relation-aware self-attention mechanism. According to their alignment and schema relations, RAT-SQL explicitly links columns with corresponding question tokens, achieving logical form accuracy of 57.2% on Spider and 65.6% when augmenting with BERT. In BRIDGE, created by Lin et al. (2020), the relational DB schema is represented as a tagged sequence concatenated to the question. Using the database content, the model accesses the values of the columns identified in the question and appends them to their column names in the question. As a result, the input is a hybrid question-schema serialization containing the question, followed by the table name, column names, and column values. BRIDGE uses BERT to shape dependencies in the serialization and two single-layer LSTM encoders with a single LSTM-based pointer-generator with attention for decoding. This allowed the model to exceed the RAT-SQL by 1.9%. When applied to WikiSQL, BRIDGE was able to achieve 86.5% with EG.

2.2.5 Text-to-SQL in healthcare domain

Despite WikiSQL and Spider being multi-domain benchmarks, they lack sufficient suitable medical records. Therefore, Wang et al. (2020) proposed the first dataset for healthcare named MIMICSQL. It consists of five tables and 10,000 question-SQL pairs of real-world medical information. The syntax for SQL does not include nested queries, but includes multiple tables connected by the JOIN operation. The pairs are divided into template questions and natural language questions based on the collection method: machine-generated (template questions) or human-annotated. Along with the database, they released TREQS, a translate-edit model operating in two stages. Stage one involves translating a natural language question into SQL using a Seq2Seq

TABLE 2 Summary of T5 model sizes.

T5 Model	Model Size
Small	60 million parameters
Base	220 million parameters
Large	770 million parameters
3b	3 billion parameters
11b	11 billion parameters

model with attention, while stage two performs editing to the generated SQL using a look-up table. The look-up table contains the table's names, columns and keywords of each column to recover the exact information between the question and the schema. They also proposed a technique to ensure query execution by retrieving the condition values of the predicted SQL and matching them against the dataset. As they introduced the model with their dataset, their accuracy measurements were broken down based on the question-SQL pairs. They achieved 85.3% and 92.4% logical form accuracy and execution accuracy, respectively, for template questions and 55.6% and 65.4% for human-annotated questions. Pan et al. (2021) claimed that because TREQS is based on Seq2Seq, it did not consider SQL's intrinsic structure. To incorporate the results of IRNET, they proposed a model named MedTS, which applied schema linking and BERT as an encoder. MedTS adopts a grammar-based LSTM decoding strategy with designed grammar rules based on the MIMICSQL dataset. A logical form of 78.4% and execution accuracies of 89.9% were obtained by MedTS.

2.3 Text-to-text transfer transformer (T5)

Raffel et al. (2020) conducted a large-scale survey on existing transfer learning techniques in natural language processing, such as ELMO created by Peters et al. (2018) and BERT created by Devlin et al. (2019). After testing and refining several models in NLP, they created a Text-to-Text Transfer Transformer (T5) model built on insights from the survey. The T5 model is a pre-trained language model that uses the complete encoder-decoder architecture of the transformer (Vaswani et al., 2017). In addition, T5 uses layer normalization to stabilize the hidden state and reduce training time (Ba et al., 2016). It is a very large neural network that takes the source sequence tokens all at once and relies on self-attention alone to compute its source input and target output. The T5 model was created as a unified framework covering all NLP tasks, such as summarization and translation, by converting every language problem into a text-to-text format. Unlike other pre-trained models, this model takes the source sequence as input and produces a target text string rather than word embedding.

The T5 model has various sizes depending on the number of parameters used for building and training it, as summarized in Table 2. The model was trained with two learning methodologies, as follows:

- Unsupervised training, in which T5 was trained on the colossal clean crawled corpus (C4) created by Raffel et al. (2020). C4 is a huge clean dataset of English text collected from the web for pre-training the T5 model.
- Supervised training, in which T5 was fine-tuned for several NLP tasks by training it with labeled data for each task. T5 was pre-trained using the Adafactor optimizer created by Shazeer and Stern (2018) and cross-entropy loss function. The loss function is used to evaluate the model performance during training by comparing the generated result with the expected result to produce a loss value (Demirkaya et al., 2020). The optimizer is an algorithm used to update the model parameters to reduce the loss value, such as inputs' weight presenting the impact of an input on the model output.

In Raffel et al.'s (2020) evaluation, the T5 model achieved promising results on many NLP benchmarks and was shown to be flexible for fine-tuning a variety of NLP problems. Its development has shown that deep learning approaches are moving toward reaching human-level accuracy in performing NLP tasks. Xie et al. (2022) proposed a large-scale multi-task learning framework using T5 and studied its performance in 21 NLP tasks, including Text-to-SQL. On many SQL benchmarks, such as Spider and WikiSQL, their study showed that the T5 model achieved near and above the state-of-the-art performance of these benchmarks.

Inspired by Raffel et al. (2020), researchers have started considering the T5 model to directly convert NLQ into SQL with simpler architecture. Shaw et al. (2020) showed by experiment that the T5 model without modification achieved promising results compared to previous models on Spider. They proposed NQG-T5, a hybrid model combining a grammar-based approach with the T5 model, achieving competitive results with the state-of-the-art model on the Spider dataset with a 70% exact match accuracy using the T5-3b. In a study conducted by Scholak et al. (2021), the T5 model was fine-tuned on Spider and augmented with an additional method called PICARD at decoding. PICARD was implemented to guarantee semantically correct SQL by rejecting invalid tokens at each decoding step. To match the generated SQL with the question, PICARD uses the table schema when evaluating SQL tokens. They concluded that the conversion was accelerated, and performance was improved using the T5 model. Their T5+PICARD model became the state-of-the-art on Spider with 71% exact match and 75% execution accuracy.

2.4 Summary

In summary, for improved accuracy, Text-to-SQL conversion models are developed by deep learning with encoder-decoder architecture. As pre-trained models were introduced, researchers began focusing on employing transfer learning for Text-to-SQL conversion, which led to near-human performance level. Furthermore, using pre-trained models instead of building models from scratch simplified the process of model development. Upon its introduction, the Text-to-Text transfer transformer (T5) captured the attention of researchers due to its encoder-decoder transformer architecture and its multi-task training covering various NLP tasks, such as summarization and question answering. Researchers started fine-tuning the T5 model for Text-to-SQL conversion,

which significantly improved the performance, making it state-of-the-art. Table 3 presents a summary of the Text-to-SQL models discussed in this review where ACCLF, ACCEX and EG indicate logical form accuracy, execution accuracy, and execution-guided, respectively.

Existing text-to-SQL models have not been fully embraced in the healthcare domain. Wang et al. (2020) stated that Text-to-SQL for EMRs was still under-explored. Based on literature review, only MedTS and TRESQ were introduced to assist medical staff with databases. Encouraged by previous success in the improvements of Text-to-SQL with transfer learning of the T5 model, this research aims to utilize transfer learning by fine-tuning the T5 model to develop a Text-to-SQL conversion model on EMRs and evaluate its performance. To the best of our knowledge, no existing work has fine-tuned the T5 model in Text-to-SQL for the healthcare domain. Furthermore, this study uses the WikiSQL dataset to benchmark the intended model against other models, in which WikiSQL was used, for performance comparison.

3 Research methodology

This research leverages the MIMICSQL dataset (Wang et al., 2020), the first publicly available dataset designed for Text-to-SQL tasks in the healthcare sector, to train and evaluate the MedT5SQL model.

We adopted the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology (Ncr and Clinton, 1999) to guide our research process. CRISP-DM is a widely used, structured approach for data mining projects, encompassing six key stages: business understanding, data understanding, data preparation, modeling, evaluation, and deployment (Marbán et al., 2009). This methodology has been successfully applied in various domains, including healthcare (Martínez-Plumed et al., 2021; Marshan et al., 2021).

To implement the MedT5SQL model, we utilized the Python programming language along with the PyTorch and HuggingFace Transformers libraries (Paszke et al., 2019; Huggingface.co, 2022) on the Google Colaboratory platform. Google Colab's provided GPU resources accelerated the computationally intensive deep learning processes involved in model training and evaluation.

4 Data analysis and results

4.1 Clinical objectives definition (business understanding)

The primary aim of this work is to generate an SQL query from written questions in the healthcare domain by utilizing natural language processing (NLP) through deep learning. The review of the relevant literature has revealed that the current state-of-the-art for Text-to-SQL conversion is to employ deep learning approaches with encoder-decoder architecture to achieve the required conversion. With the rise of the transformer's encoder-decoder architecture, various language conversion models were developed to improve NLP tasks using the transformer's encoder. They present large neural networks operating under a pre-train-fine tune paradigm where they are pre-trained over a large text corpus for a generic task, such as understanding a language, and then fine-tune on specific downstream tasks, such as summarization. Pre-training and fine-tuning these models facilitate leveraging transfer learning to improve the accuracy of various NLP tasks, including Text-to-SQL.

Throughout the literature review, it was observed that with the growth of transfer learning through pre-trained language models, deep learning has achieved promising results in this field (Guo et al., 2019; Hwang et al., 2019; Lyu et al., 2020; Choi et al., 2021; Pan et al., 2021). To get the most out of the transformer's encoder-decoder architecture and explore the limits of transfer learning, Raffel et al. (2020) built the Text-to-Text transfer transformer (T5) model as a unified large language model for all NLP tasks. The T5 model operates as an encoder-decoder with position encoding, attention mechanism, and self-attention for modeling all source tokens at once while understanding each token based on the context of the words around it. In Text-to-SQL conversion, Shaw et al. (2020) showed that the T5 model is able to learn Text-to-SQL conversion and operate with promising results. Encouraged by this work, research has been conducted presenting significant improvements in both WikiSQL and Spider benchmarks (Scholak et al., 2021; Xie et al., 2022). Despite the wealth of research in the field of Text-to-SQL, however, only two studies have been conducted focusing on the healthcare domain, proposing TREQS and MedTS models (Wang et al., 2020; Pan et al., 2021). TREQS is an original model developed entirely by Wang, Shi and Reddy, and MedTS benefits from transfer learning using a pre-trained model encoder, allowing it to outperform TREQS.

Considering the findings from the literature review, this study utilizes deep learning for Text-to-SQL conversion in the healthcare domain to develop a Text-to-SQL model named MedT5SQL employing transfer learning of the T5 transformer model. To the best of our knowledge, this work establishes the first model employing T5 in the healthcare Text-to-SQL conversion. This work focuses on using supervised deep learning to train the model on a healthcare-related dataset to achieve high conversion accuracy. Furthermore, MedT5SQL is benchmarked on WikiSQL dataset to evaluate its performance between the two datasets.

4.2 EMR data exploration (data understanding)

In this research we use MIMICSQL dataset that is created by Wang et al. (2020), to train and evaluate the MedT5SQL model. MIMICSQL is the first dataset created for Text-to-SQL tasks in the healthcare field. It is a large-scale dataset with 10,000 question-SQL pairs collected based on the Medical Information Mart for Intensive Care III (MIMIC III) dataset (Johnson et al., 2016). The medical information from MIMIC III was grouped into five tables for MIMICSQL as: demographics (Demo), laboratory tests (Lab), diagnosis (Diag), procedures (Pro) and prescriptions (Pres) (See Table 4 for information regarding MIMICSQL dataset). The question-SQL pairs were carefully constructed based on these

TABLE 3 Summary of text-to-SQL model.

Research	Model specs				
	DL approach	Domain	Performance	Transfer learning (Yes/No)	Opportunity
Seq2SQL (Zhong et al., 2017)	Seq2Seq	Single	ACCLF: 48.3% ACCEX: 59.4%	No	The first model on WikiSQL
SQLNet (Xu et al., 2017)	Sequence-to-set sketch-based	Single	ACCEX: 68.0%	No	Avoid the "Order-Matter
TYPESQL (Yu et al., 2018c)	Sequence-to-set sketch-based	Single	ACCEX: 73.5%	No	Improving SQLNet
SQLova (Hwang et al., 2019)	Sequence-to-set	Single	ACCLF: 80.7% ACCEX: 86.2% –with EG— ACCLF: 83.6% ACCEX: 89.6%	Yes (BERT)	Utilize BERT in Text-to-SQL
X-SQL (He et al., 2019)	Sequence-to-set	Single	ACCLF: 83.3% ACCEX: 88.7% –with EG— ACCLF: 86.0% ACCEX: 91.8%	Yes (MT-DNN)	Utilize MT-DNN in Text-to-SQL
Hydranet (Lyu et al., 2020)	Pre-trained language model	Single	–with EG– ACCLF: 86.0% ACCEX: 91.8%	Yes (BERT)	BERT alone as encoder
			–with EG– ACCLF: 86.5% ACCEX: 92.2%	Yes (RoBERTa)	RoBERTa alone as encoder
SyntaxSQLNet (Yu et al., 2018b)	Sequence-to-set grammar-based	Cross-domain	ACCLF: 27.2%	No	First Model on Spider
RCSQL Lee (2019)	Sequence-to-set self-attention mechanism	Cross-domain	ACCLF: 28.8%	No	clause-wise SQL decoding with attention mechanism
IRNet (Guo et al., 2019)	Sequence-to-setgrammar-based	Cross-domain	–without BERT— ACCLF: 46.7% –with BERT— ACCLF: 54.7%	Yes (BERT)	Handle out-of-domain words in columns prediction + schema linking
RYANSQL (Choi et al., 2021)	Pre-trained language model	Cross-domain	ACCLF: 58.2%	Yes (BERT)	Handle nested SELECT clause +BERT as encoder
RAT-SQL (Wang et al., 2019)	Grammar-based with Pre-trained language model	Cross-domain	–without BERT— ACCLF: 57.2% –with BERT— ACCLF: 65.6%	Yes (BERT)	propose relation-aware self-attention mechanism
BRIDGE (Lin et al., 2020)	Pre-trained language model	Single + cross-domain	–on Spider– ACCLF: 67.5% –on WikiSQL– ACCLF: 91.9%	Yes (BERT)	hybrid question-schema serialization
TREQS (Wang et al., 2020)	Seq2Seq	healthcare	ACCLF: 55.6% ACCEX: 65.4%	No	First Healthcare Domain Text-to-SQL model
MedTS (Pan et al., 2021)	Grammar-based with pre-trained language model	healthcare	ACCLF: 78.4% ACCEX: 89.9%	Yes (BERT)	Introduce transfer learning to healthcare domain
NQG-T5 (Shaw et al., 2020)	Transformer	Cross-domain (In this work, the focus is on Spider)	On Spider development set: -Using T5-base- ACCLF: 57.1% -Using T5-3b- ACCLF: 70%	Yes (T5)	First grammar-based approach with T5 on Spider
PICARD (Scholak et al., 2021)	Transformer	Cross-domain	ACCLF: 71% ACCEX: 75%	Yes (T5)	Fine-tune T5 on Spider and introduce PICARD for semantically correct SQL
UnifiedSKG (Xie et al., 2022)	Transformer	Single (In this work, the focus is on WikiSQL)	-Using T5-base- ACCLF: 82.63% -Using T5-3b- ACCLF: 85.96%	Yes (T5)	Benchmarking T5 on Text-to-SQL

ACCLF, **ACCEX**, and **ACCASM** indicate the logical form accuracy, the execution accuracy, and approximate string-matching respectively.
EG stands for Execution guided.

TABLE 4 Statistical summary of MIMICSQL dataset.

Data	Stats
Number of patients	46,520
Number of tables	5
Number of columns per table	Demo: 23, Diag: 5, Pro: 5, Pres: 7, and Lab: 9
Number of question-SQL pairs	10,000
Average template question length (in words)	18.39
Average natural language question length (in words)	16.45
Average SQL query length	21.14

tables. The pairs include questions to retrieve patient information directly from the database and reasoning questions to collect patient information from multiple tables. The pairs are divided into template questions (machine-generated) and natural language questions (human-annotated).

The general structure of the SQL queries adopted in MIMICSQL is shown in Figure 5 and described as following:

- The SELECT clause allows multiple columns.
- The aggregation operators (AGG_OP) vary between NULL, MAX, MIN, COUNT and AVG.
- The column headers in the tables represent the question topic; therefore, AGG_COLUMN holds the question topic to retrieve the required information.
- The queries either retrieve the data from a single table or a new table generated from joining multiple tables through INNER JOIN by a condition.
- WHERE clause allows for one or multiple conditions.
- Only five condition operations (COND_OP) are considered in MIMICSQL, including =, >, <, >= and <=.

The WikiSQL dataset is used to benchmark MedT5SQL against other models that have used WikiSQL. This dataset contains 80,654 question-SQL pairs and it is larger than MIMICSQL with similar SQL structure.

4.3 Data acquisition and pre-processing

4.3.1 Data acquisition

MIMICSQL was downloaded from Wang and Shi's (2020) repository on GitHub. They uploaded MIMICSQL in three separate files as data partitioning of the dataset, in the ratio of 0.8:0.1:0.1 for training, validation and test sets, respectively. In this work, we adopt the same data partitioning, using 8,000 pairs for training, 1,000 pairs for validation, and 1,000 pairs for testing the MedT5SQL model. The sets were stored on GitHub in the form of JSON files, and we extracted them into Pandas dataframes for easier manipulation. Similarly, WikiSQL is partitioned into three sets collected from the Hugging Face dataset library.

4.3.2 Feature selection

The features relevant to this research in MIMICSQL dataset are (question_refine) and (sql), which represent the question-SQL pairs. Therefore, they were extracted for the training, validation and test datasets used in this research. The (question_refine) presents the (source_text) for the model, while (sql) presents the (target_text). In WikiSQL dataset, the question-SQL pairs are presented by (question) and (sql) features, renamed (source_text) and (target_text). However, this (target_text) was found to be a dictionary object where its entry (human_readable) presents the text form of the SQL, and thus, the SQL was extracted to form the (target_text).

4.3.3 Handling missing and duplicate records

The datasets are inspected for missing data or duplicate pairs. In addition, the structure of the question-SQL pairs was inspected by checking random records to detect irrelevant records. No issues were identified in MIMICSQL, while WikiSQL had 189 duplicate pairs in the training set, 42 in the test set, and 29 in the validation set. These pairs were deleted before feeding the model with the data for the purpose of maintaining accuracy and avoiding biased performance.

4.3.4 Tokenization

Prior to fine-tuning the T5 model, the source and target sequences were tokenized by splitting each text into its list of tokens (words) to understand the context. For the testing process, only the source text was tokenized before using it to generate the equivalent SQL query for model evaluation. A pre-trained T5Tokenizer from the T5ForConditionalGeneration module in the Hugging Face transformer package was used in this step. After number of experiments, the maximum number of tokens we were able to use for the source and target texts and train the MedT5SQL model are 150 tokens (original question) and 256 tokens (SQL Query), respectively. The pre-trained tokenizer not only splits the text into tokens but also converts the tokens into numeric representations to prepare the data before feeding it to the transformer-based deep neural model (Marshan et al., 2023). The tokenizer also adds padding tokens, which are used to fill the source and target text with extra tokens to standardize the number of tokens in each as required by deep neural models. Padding tokens also include guidance tokens that indicate the start and end of each text. As a result, the tokenizer results in "input_ids" and "attention_masks" fields, where the "input_ids" presents the list of tokens' IDs given to the model and "attention_masks" is a value of 0 or 1 mapped to each token, enabling the model to ignore padding tokens: 0 = masked/ignore and 1 = not masked.

4.3.5 Data loader

In order to accelerate the training, validation and testing processes, PyTorch DataLoader was used to create data loaders for the tokenized datasets. Data loaders make it easier to manage the data and simplify the deep learning pipeline. They navigate the dataset by synchronously loading multiple batches of data using background processes called workers. Batches present the number

```
SELECT $AGG_OP ($AGG_COLUMN) *
FROM $TABLE (INNER JOIN $TABLE on ($COND_COLUMN $COND_OP $COND_VAL)
WHERE ($COND_COLUMN $COND_OP $COND_VAL) *
```

FIGURE 5
MIMICSQL SQL query structure.

of data samples run by the model in each training, validation, or test epoch. An epoch presents a complete pass of the whole dataset through the model. Following the work done by Pan et al. (2021) on MIMICSQL dataset, we used eight data samples per batch. The number of workers is set to four to allow faster data loading. To make the model more robust and avoid overfitting, shuffling was enabled for the training data loader to shuffle the data in every training epoch.

4.4 Modeling: developing the MedT5SQL model

Shaw et al. (2020) and Scholak et al. (2021) have showed that a pre-trained T5 model, especially the T5-base and T5-3b models have shown promising results as the current state-of-the-art for Text-to-SQL conversion. Motivated by these papers, this study developed the MedT5SQL model as a fine-tuned T5 model for text-to-SQL conversion in the healthcare domain. The MedT5SQL model went through several iterations until the successful model was achieved, as explained below. However, they all used the same model configurations.

4.4.1 Model configuration and development environment

This study uses the T5ForConditionalGeneration module in the huggingface package to load the pre-trained T5-base model with its weights and operative configurations for fine-tuning. A list of the most important configurations related to the T5 architecture can be found in Table 5.

MedT5SQL is trained on the Tesla P100 GPU from NVIDIA Corp offered by Google Colab and we used the parameters settings presented in Table 6. MedT5SQL sets the training and validation batch sizes to 8, similar to Pan et al. (2021) on MIMICSQL and the learning rate to 1e-4 as used by Shaw et al. (2020) and Scholak et al. (2021) for fine-tuning T5. The MedT5SQL model was trained using three different Epochs numbers 10, 15, 20, 50 and 100 to study its effect on the model performance. The validation was done using 15 epochs. The remaining parameters follow Pytorch-lightning (2022).

TABLE 5 T5-base configurations from hugging face transformers package.

Configuration	Description	Value
vocab_size	The number of different tokens represented by 'inputs_ids' passed to the model	32,128
d_model	Encoder layers and the pooler layer size	768
num_layers	Number of encoder's hidden layers	12
feed_forward_proj	The activation function in the encoder	Relu
num_decoder_layers	Number of decoder's hidden layers	12
dropout_rate	Dropout rate for regularization	0.1
transformers_version	The version of transformers package	4.20.1
num_beams	Transformers use greedy decoding to select tokens with the highest probability	4
early_stopping	Use early stopping for regularization	True

4.4.2 MedT5SQL model

Initially, we attempted to fine-tune the Hugging Face T5-base model directly using PyTorch. However, despite successful training, the model failed to generalize to the Text-to-SQL task during testing, simply reproducing the input question instead of generating the corresponding SQL query. This indicated that the model had not adequately learned the translation task, likely due to insufficient task-specific guidance during fine-tuning.

To address this, we incorporated a task-specific prefix ("translate English to SQL") into the input sequence. This prefix acts as an explicit instruction to the model, prompting it to interpret the input as a Text-to-SQL translation problem. Additionally, we modified the T5-base model's configuration file to include parameters that reinforce the desired task (see Figure 6). These modifications guided the model's learning process and significantly improved its ability to generate correct SQL queries in response to natural language questions.

Nonetheless, training the T5 model with the new configuration file prevented it from using the weights of the pre-trained T5-base, and thus, the model functioned as a new model and not as a transfer learning of the T5 model, and with 8,000 training data, it achieved poor accuracy of <1% despite changing parameters, including training epochs, the optimizer, and the learning rate. Thus, as alluded to by Raffel et al. (2020) it is concluded that the T5 model should be able to understand the translation task without specifying the prefix.

To ensure that all the proper fine-tuning steps are performed, we developed the MedT5SQL model by utilizing the PyTorch

TABLE 6 MedT5SQL model configurations.

Parameter	Value
MODEL	t5-3b
TRAIN_BATCH_SIZE	8
VALID_BATCH_SIZE	8
TRAIN_EPOCHS	10, 15, 20, 50 and 100
VAL_EPOCHS	15
LEARNING_RATE	1e-4
MAX_SOURCE_TEXT_LENGTH	150
MAX_TARGET_TEXT_LENGTH	256
SEED	42
adam_epsilon	1e-8
weight_decay	0.0
n_gpu	1
gradient_accumulation_steps	16
warmup_steps	0
fp_16	False
output_dir	"/content/drive/MyDrive/MedT5SQL"
opt_level	apex
max_grad_norm	1.0

```
"task_specific_params": {
  "summarization": {
    "early_stopping": true,
    "length_penalty": 2.0,
    "max_length": 200,
    "min_length": 30,
    "no_repeat_ngram_size": 3,
    "no_beams": 4,
    "prefix": "summarize: "
  },
  "translate_en_to_sql: ": {
    "early_stopping": true,
    "max_length": 300,
    "num_beams": 4,
    "prefix": "translate English to SQL: "
  }
```

FIGURE 6
Modification on T5-base configuration model.

Lightning framework (Lightning, 2022), which organizes and facilitates the process of building a model by abstracting the details of the training. It has a good Graphics Processing Unit (GPU) utilization and makes deep learning models flexible and easier to reproduce (Sawarkar, 2022). Using the Lightning Framework to build and train deep learning models requires the configuration of a LightningModule and Trainer parameters. LightningModule is used to structure the intended module to specify its behavior with each batch of training and validation data. Trainer uses the LightningModule with a specified dataset to automate the training and validation processes for the intended module.

4.4.3 Lightning module configuration

In more details, to fine-tune the T5-base model using the Lightning Framework, MedT5SQL uses LightningModule to structure its implementation into four sections: initialization, training loop, validation loop and optimizer configuration (Pytorch-lightning, 2022). The LightningModule contains a function for each section to easily adopt any deep learning model to automate the training and validation loops with all the required components, such as epochs and optimizers. Overriding each of its functions allows MedT5SQL to specify its behavior in the training and validation to fine-tune the T5 model as required. The LightningModule for MedT5SQL was created and initialized given the model parameters listed in Table 6 and shown in Figure A1 in the Supplementary material.

4.4.4 Training loop configuration

To activate the training loop of Lightning Framework for the fine-tuned T5-base model, MedT5SQL overrides the training functions of the LightningModule: *training_step* and *training_epoch_end*, as shown in Figure A2 in the Supplementary material. This loop is performed on the training dataset, loaded as batches by the data loader, to fine-tune the T5 model and obtains the training loss value using the *training_step* and *_step* functions as displayed in Figures A2, A3 in the Supplementary material. In T5 training, the T5 model's encoder uses the source tokens' IDs and masked values as input, while the decoder takes the encoder's output along with the target tokens' IDs (labels) to compute the training loss. The T5 model uses the cross-entropy loss function to compute the loss value required to modify the model's parameters during training (Raffel et al., 2020). The function *training_epoch_end* returns the average loss value of each training epoch. Lower loss values indicate a well-trained model.

4.4.5 Validation loop configuration

To activate the validation loop of the Lightning Framework, MedT5SQL overrides the validation functions of LightningModule, as presented in Figure A4 in the Supplementary material. This loop uses the validation dataset, loaded as batches by the data loader, in the method *_step* to validate the model and obtain the validation loss (see Figure A3 in the Supplementary material). The *_step* method allows the model to generate the target text using the source text and evaluate the performance by comparing the generated query against the expected query and calculate the loss value. The function *validation_epoch_end* returns the average loss value of each validation epoch.

4.4.6 Optimizer configuration

MedT5SQL is trained using Adafactor optimizer, the same optimizer used to pre-train the T5 model by Raffel et al. (2020)

and other Text-to-SQL models built using the T5 model (Shaw et al., 2020; Scholak et al., 2021). The optimizer configuration is shown in Figure A5 in the Supplementary material. Moreover, we used AdamW optimizer, created by Loshchilov and Hutter (2017) to compare its performance against that of Adafactor (see Figure A6 in the Supplementary material).

4.4.7 Trainer configuration

To develop MedT5SQL, PyTorch Lightning Trainer we automate the training and validation loops as presented in Figure A7 in the Supplementary material. The trainer was first created with the required arguments for the training process, such as the number of epochs, and was then given an object of MedT5SQL LightningModule class that contained the training and validation DataLoaders and loops. The developed MedT5SQL model presents a structured version of the first model we have developed using the Lightning framework, yet, without the use of a task-specific prefix, since the fine-tuning process was performed and organized successfully by the Lightning Framework.

4.5 Evaluation

To evaluate MedT5SQL model performance, the test dataset was used to assess the model on unseen data. The source text was tokenized and loaded in a DataLoader to feed MedT5SQL with natural language questions to generate equivalent SQL queries, as seen in Figure A8 in the Supplementary material. To generate the target text given the source text, the function "generate()" from the module T5ForConditionalGeneration is used. At the end, the tokenizer decodes the generated tokens into string form to output the SQL query sequence. This generated query was evaluated against the test dataset's target text to measure MedT5SQL performance.

MedT5SQL's performance was evaluated using logical form accuracy, known as exact match, and manual evaluation, in line with previous papers (Hwang et al., 2019; Wang et al., 2020; Pan et al., 2021). Additionally, we used approximate string matching to evaluate how close the MedT5SQL predicted query is to the expected query. The performance evaluation for MedT5SQL is presented in Figure A9 in the Supplementary material.

The manual evaluation was conducted by an independent reviewer with expertise in the medical domain. The reviewer was presented with a random sample of generated SQL queries paired with their corresponding expected queries from the test dataset. They assessed each generated query's correctness based on the following criteria:

- **Correctness:** Does the generated query accurately reflect the intended meaning and structure of the expected query?
- **Completeness:** Does the generated query include all necessary clauses and conditions?
- **Syntax:** Is the generated query syntactically valid?
- **Functional Equivalence:** If there are minor differences, does the generated query produce the same result as the expected query when executed on the database?

The reviewer assigned a score of "correct," "partially correct," or "incorrect" to each query. The manual evaluation score reported in our results represents the percentage of queries deemed "correct."

A breakdown of logical form accuracy was performed on each SQL clause for further inspection. Figure A10 in the Supplementary material presents the evaluation process for the *SELECT* clause. MedT5SQL performance was evaluated in terms of the number of training epochs, as well as the optimizers, AdamW and Adafactor. It was also benchmarked against MIMICSQL and WikiSQL to examine its performance on different datasets.

5 Results and discussion

In this research, a Text-to-SQL conversion model named MedT5SQL was developed as the first fine-tuned T5-base model in the healthcare domain. The model was developed using MIMICSQL, a healthcare Text-to-SQL dataset. This section discusses the results of model evaluation and outlines its findings.

5.1 Performance evaluation on different training epochs and different optimizers

To understand the contribution of the number of training epochs to the performance of the model, MedT5SQL was trained on three different numbers of epochs: 10, 15, 20, 50 and 100. The performance was evaluated through accuracy measurement by comparing the generated SQL query against the expected SQL query using the test dataset (Hwang et al., 2019; Wang et al., 2020; Pan et al., 2021). The results are presented in Table 7, which shows that the accuracy measurements of MedT5SQL performance increased with the increasing number of training epochs.

To select the most efficient optimizer, the MedT5SQL model was developed using two different optimizers, Adafactor and AdamW, one at a time, and their performance was compared, as presented in Table 7. According to the analysis, Adafactor was more efficient for MedT5SQL, since it allowed the model to achieve higher accuracy compared to AdamW. Only when trained on 10 epochs did the model achieve 97.455% ACC_{ASM} with AdamW, compared to 97.369% with Adafactor. Nevertheless, it is worth noting that ACC_{LF} dropped by 0.2% when trained with AdamW on 20 epochs, compared to 15 epochs which could be a result of overfitting. ACC_{LF} rose by 0.5 under the same conditions using Adafactor. In general, Adafactor elevated MedT5SQL performance by 0.1–2% ACC_{LF}, 0.3% ACC_{ASM} and 5% ACC_{manual}, compared to AdamW. MedT5SQL achieved its highest accuracy of 80.1% ACC_{LF}, 98.937% ACC_{ASM}, and 90% ACC_{manual} when trained using Adafactor on 100 epochs. The values of ACC_{ASM} were extremely high, indicating the high similarities between the generated and the expected queries. Therefore, a breakdown evaluation was conducted on the SQL clauses to understand the reasons behind the differences between the ACC_{LF} and ACC_{ASM} values.

5.2 Performance on each SQL clause

To further analyse the generated SQL and investigate ACC_{ASM} values, we calculate the logical form accuracy (ACC_{LF}) for each

TABLE 7 MedT5SQL performance evaluation using different parameter.

# Training Epoc	AdamW Optimizer			Adafactor Optimizer		
	ACC_{LF}	ACC_{ASM}	ACC_{Manual}	ACC_{LF}	ACC_{ASM}	ACC_{Manual}
10 epochs	57.9 %	97.455%	60%, 12 out of 20	58%	97.369%	65%, 13 out of 20
15 epochs	61.3%	97.716%	65%, 13 out of 20	62.6%	98.054%	75%, 15 out of 20
20 epochs	61.1%	97.81%	75%, 15 out of 20	63.1%	98.1%	80%, 16 out of 20
50 epochs	63.2%	97.926%	80%, 16 out of 20	68.9%	98.572%	85%, 17 out of 20
100 epochs	66.7%	98.016%	90%, 18 out of 20	80.63%	98.937%	90%, 18 out of 20

ACC_{LF}, logical form accuracy by exact string matching; ACC_{ASM}, approximate string-matching accuracy. ACC_{manual}, manual evaluation derived by randomly examining 20 generated SQL queries.

clause and shows the results in Table 8. The results confirm that Adafactor is more efficient for the MedT5SQL model. On the best performance, the exact match between the generated and expected queries was 96.8% and 97.01% for the *SELECT* and *FROM* clauses, respectively, while achieving 68.6% on the *WHERE* clause, which indicates that the model suffers mostly when generating the *WHERE* clause.

As shown in Table 9, the reason for this is related to the condition's value and operator as found by the manual evaluation. This was also demonstrated by Pan et al.'s (2021) evaluation of the accuracy of each component of the SQL query using multiple models, which confirmed that the condition's operation and values had lower accuracy than other components.

Furthermore, the manual evaluation showed that the reasons behind the failed 3.2% ACC(SELECT) and 2.9% ACC(FROM) are related to column names and aggregators in the *SELECT* clause and the *INNER JOIN* or table names in the *FROM* clause as it can be noticed in Table 10. It was noted that a false *INNER JOIN* results in incorrect *WHERE* conditions.

5.3 Benchmarking MedT5SQL on two datasets

Based on findings from past research, MedT5SQL on MIMICSQL was developed using the **Adafactor** optimizer. MedT5SQL was benchmarked on the WikiSQL dataset, explained in Section 2.1, to compare the performance on different types of questions. Table 11 presents a performance comparison between MedT5SQL developed using WikiSQL and MedT5SQL developed using MIMICSQL. It was found that the MedT5SQL model performed better when fine-tuned on MIMICSQL. Using MIMICSQL, $\mathbf{ACC}_{LF}$ achieved 58% and 62.6% when trained on 10 and 15 epochs, respectively, compared to 43.63% and 44.2% when using WikiSQL on the same number of epochs. Similarly, the ACC_{ASM} values obtained using MIMICSQL were 3.2–3.7% higher than those attained using WikiSQL.

The difference in size between the datasets could be a contributing factor to this difference in performance. MIMICSQL has 8,000 question-SQL pairs for training and 1000 pairs for validation, while WikiSQL has 56,166 pairs for training and 8,392 for validation. Therefore, WikiSQL may need more training epochs to achieve better accuracies. However, benchmarking MedT5SQL on WikiSQL required longer execution time due to its enormous size, as shown in Table 11. For 20, 50 and 100 training epochs, the MedT5SQL did not run when it is trained on WikiSQL due to resource limitations, as Google Colab kept crashing when using the WikiSQL dataset on more than 15 epochs due to GPU memory shortage.

On WikiSQL, Xie et al.'s (2022) logical form accuracy evaluation of the UnifiedSKG model, baselined on T5-base, was shown to be 82.63% when trained on epochs between 50 and 200. In this work, WikiSQL achieved 43.63% on 10 epochs and 44.2% on 15 epochs. With sufficient resources, MedT5SQL may achieve equivalent results to UnifiedSKG. On MIMICSQL, Table 12 presents a comparison of the logical form accuracy between the developed model MedT5SQL and MedTS, the state-of-the-art model of MIMICSQL proposed by Pan et al. (2021). MedT5SQL outperforms MedTS knowing that it relies entirely on transfer learning, which offers a simpler architecture. According to Scholak et al. (2021) and Shaw et al. (2020), however, using T5-3b instead of T5-base, which we used in this research, can further improve the performance by around 13.5%. In this project, our attempt to create MedT5SQL by refining the T5-3b model was unsuccessful. The experiment faced challenges due to limitations in resources, specifically when the GPU exhausted its memory while processing the T5-3b model. This setback can be attributed to the substantial size of the T5-3b model, which comprises 3 billion parameters, in contrast to the 220 million parameters in the T5-base model.

Recent advancements in Text-to-SQL models have shown significant promise in improving the accuracy and efficiency of natural language interfaces for databases. In particular, models like ChatGPT (Liu et al., 2023), RASAT (Qi et al., 2022), and RESDSQL (Li et al., 2023) have reported impressive performance on various benchmark datasets. These models leverage large-scale pre-training and fine-tuning techniques, often employing transformers-based architectures, to achieve state-of-the-art results. However, their performance on healthcare-specific tasks and datasets remains less explored.

In the context of healthcare Text-to-SQL, the TREQS method proposed by Wang et al. (2020) stands out due to its reported 85% accuracy on the MIMICSQL dataset. While this accuracy is higher than that achieved by our MedT5SQL model, it is important to note that TREQS employs a rule-based approach with domain-specific templates, which may limit its generalizability to new datasets or query types. In contrast, our MedT5SQL model, based on the T5

TABLE 8 Break down logical form accuracy (ACCLF) of MedT5SQL.

# Training Epoc	AdamW Optimizer			Adafactor Optimizer		
	ACC (SELECT)	ACC (FROM)	ACC (WHERE)	ACC (SELECT)	ACC (FROM)	ACC (WHERE)
10 epochs	93.6 %	95.1%	63.2%	90.2%	95.9%	64.2%
15 epochs	95.4%	95.4%	64.9%	95%	96.2%	66.4%
20 epochs	93.1%	96.1%	65.5%	95.4%	96.6%	66.2%
50 epochs	93.9%	97.03%	67.9%	96.1%	96.8%	67.6%
100 epochs	94.8%	97.53%	72.1%	**96.8%**	**97.01%**	**68.6%**

Bold values highlight the best performance of the model.

TABLE 9 Manual evaluation of MedT5SQL with Adafactor on 20 Epoch.

Generated SQL query	Expected SQL query
SELECT COUNT (DISTINCT DEMOGRAPHIC.“SUBJECT_ID”) FROM DEMOGRAPHIC INNER JOIN LAB on DEMOGRAPHIC.HADM_ID = LAB.HADM_ID WHERE DEMOGRAPHIC.”AGE“ ”30“ AND LAB.”FLAG“ = ”abnormal“	SELECT COUNT (DISTINCT DEMOGRAPHIC.”SUBJECT_ID”) FROM DEMOGRAPHIC INNER JOIN LAB on DEMOGRAPHIC.HADM_ID = LAB.HADM_ID WHERE DEMOGRAPHIC.“AGE” < “30” AND LAB.“FLAG” = “abnormal”
SELECT COUNT (DISTINCT DEMOGRAPHIC.“SUBJECT_ID”) FROM DEMOGRAPHIC INNER JOIN PRESCRIPTIONS on DEMOGRAPHIC.HADM_ID = PRESCRIPTIONS.HADM_ID WHERE PRESCRIPTIONS.”DRUG“ = ”Capso Fungin“	SELECT COUNT (DISTINCT DEMOGRAPHIC.”SUBJECT_ID”) FROM DEMOGRAPHIC INNER JOIN PRESCRIPTIONS on DEMOGRAPHIC.HADM_ID = PRESCRIPTIONS.HADM_ID WHERE PRESCRIPTIONS.“DRUG” = “Caspofungin”
SELECT COUNT (DISTINCT DEMOGRAPHIC.“SUBJECT_ID”) FROM DEMOGRAPHIC INNER JOIN LAB on DEMOGRAPHIC.HADM_ID = LAB.HADM_ID WHERE DEMOGRAPHIC.”DOB_YEAR“ ”2170“ AND LAB.”LABEL“ = ”Other Cells“	SELECT COUNT (DISTINCT DEMOGRAPHIC.”SUBJECT_ID”) FROM DEMOGRAPHIC INNER JOIN LAB on DEMOGRAPHIC.HADM_ID = LAB.HADM_ID WHERE DEMOGRAPHIC.“DOB_YEAR” < “2170” AND LAB.“LABEL” = “Other Cells”

large language model, offers greater flexibility and potential for adaptation to different healthcare contexts.

5.4 Limitations and future work

An accurate Text-to-SQL conversion model (MedT5SQL) is successfully developed for the healthcare domain, with a promising performance of 80.63% using transfer learning of the T5-base model. We argue that employing a larger T5 variant such as T5-3B model may yield improved performance due to their increased capacity. Also, using higher number of epochs would result in superior performance compared to existing models. In this study, we opted for the T5-base model due to resource constraints. Also, our research aimed to establish the feasibility and effectiveness of fine-tuning the T5 architecture for the specific task of Text-to-SQL conversion in the healthcare domain. We viewed the T5-base model as a suitable starting point for this initial exploration, allowing us to assess the potential of this approach before committing to the resource-intensive fine-tuning of the T5-3B model. Additionally, leveraging transfer learning by pre-training the model on larger and more diverse datasets beyond MIMICSQL could further enhance its ability to generalize to a wider range of healthcare queries. In addition, incorporating domain-specific knowledge into the model's architecture or training process could be a promising direction. This could involve incorporating medical ontologies, semantic representations, or rules-based components to guide the model's understanding and generation of healthcare-related SQL queries. Furthermore, while we focused on question-SQL pairs in this study, future work could explore the model's ability to handle a wider range of SQL queries, including complex queries with multiple clauses and conditions. Expanding the scope of supported queries would make the MedT5SQL model even more versatile and valuable for real-world healthcare applications.

Our research acknowledges the dynamic nature of large language model (LLM) development. While the T5 model served as an effective foundation for our study, we recognize that its relative performance may have evolved since our initial experiments, potentially impacting its standing among other state-of-the-art models. In this work, our primary objective was to investigate the potential of fine-tuning the T5 model for the specific domain of healthcare. This targeted approach allowed us to thoroughly explore the unique challenges and opportunities presented by this domain, revealing insights that may not be as readily apparent in broader, comparative studies. We believe that this deep dive into domain-specific fine-tuning holds considerable value, regardless of the T5 model's shifting position in the broader LLM landscape.

While a direct comparison of our fine-tuned T5 model with other state-of-the-art, fine-tuned LLMs would undoubtedly offer valuable insights, such an undertaking was beyond the scope of this

TABLE 10 Evaluation of the SELECT and FROM clauses for MedT5SQL with Adafactor on 20 Epoch.

Generated SQL query	Expected SQL query	Argument
SELECT MAX (DEMOGRAPHIC."AGE") FROM DEMOGRAPHIC WHERE DEMOGRAPHIC."MARITAL_STATUS" = "MARRIED" AND DEMOGRAPHIC."DOB_YEAR" > "2064"	SELECT COUNT (DISTINCT DEMOGRAPHIC."SUBJECT_ID") FROM DEMOGRAPHIC WHERE DEMOGRAPHIC."MARITAL_STATUS" = "MARRIED" AND DEMOGRAPHIC."DOB_YEAR" < "2064"	Failed SELECT clause: Incorrect aggregator and column name Failed WHERE clause: Incorrect operator
SELECT AVG (DEMOGRAPHIC."AGE") FROM DEMOGRAPHIC WHERE DEMOGRAPHIC."ETHNICITY" = "WHITE" AND DEMOGRAPHIC."DIAGNOSIS" = "BRADYCARDIA"	SELECT COUNT (DISTINCT DEMOGRAPHIC."SUBJECT_ID") FROM DEMOGRAPHIC WHERE DEMOGRAPHIC."ETHNICITY" = "WHITE" AND DEMOGRAPHIC."DIAGNOSIS" = "BRADYCARDIA"	Failed SELECT clause: Incorrect aggregator and column name
SELECT COUNT (DISTINCT DEMOGRAPHIC."SUBJECT_ID") FROM DEMOGRAPHIC WHERE DEMOGRAPHIC."DIAGNOSIS" = "ACIDOSIS" AND DEMOGRAPHIC."DAYS_STAY" > "7"	SELECT COUNT (DISTINCT DEMOGRAPHIC."SUBJECT_ID") FROM DEMOGRAPHIC INNER JOIN DIAGNOSES on DEMOGRAPHIC.HADM_ID = DIAGNOSES.HADM_ID WHERE DEMOGRAPHIC."DAYS_STAY" > "7" AND DIAGNOSES."SHORT_TITLE" = "Acidosis"	Failed FROM clause: Unidentified INNER JOIN Failed WHERE clause: Incorrect WHERE condition
SELECT COUNT (DISTINCT DEMOGRAPHIC."SUBJECT_ID") FROM DEMOGRAPHIC WHERE DEMOGRAPHIC."DIAGNOSIS" = "SYNCOPE; COLLABORATION" AND DEMOGRAPHIC."ADMITYEAR" "2145"	SELECT COUNT (DISTINCT DEMOGRAPHIC."SUBJECT_ID") FROM DEMOGRAPHIC INNER JOIN DIAGNOSES on DEMOGRAPHIC.HADM_ID = DIAGNOSES.HADM_ID WHERE DEMOGRAPHIC."ADMITYEAR" < "2145" AND DIAGNOSES."SHORT_TITLE" = "Syncope and collapse"	Failed FROM clause: Unidentified INNER JOIN Failed WHERE clause: Incorrect WHERE condition and operator
SELECT DEMOGRAPHIC."DIAGNOSIS", PROCEDURES."SHORT_TITLE" FROM DEMOGRAPHIC INNER JOIN PROCEDURES on DEMOGRAPHIC.HADM_ID = PROCEDURES.HADM_ID WHERE DEMOGRAPHIC."NAME" = "Bruce Harris"	SELECT DEMOGRAPHIC."DIAGNOSIS", DIAGNOSES."ICD9_CODE" FROM DEMOGRAPHIC INNER JOIN DIAGNOSES on DEMOGRAPHIC.HADM_ID = DIAGNOSES.HADM_ID WHERE DEMOGRAPHIC."NAME" = "Bruce Harris"	Failed FROM clause: Incorrect table name resulting in incorrect INNER JOIN condition
SELECT COUNT (DISTINCT DEMOGRAPHIC."SUBJECT_ID") FROM DEMOGRAPHIC INNER JOIN PROCEDURES on DEMOGRAPHIC.HADM_ID = PROCEDURES.HADM_ID WHERE DEMOGRAPHIC."AGE" "54" AND PROCEDURES."LONG_TITLE" = "Squamous cell carcinoma of oral tongue/sda"	SELECT COUNT (DISTINCT DEMOGRAPHIC."SUBJECT_ID") FROM DEMOGRAPHIC WHERE DEMOGRAPHIC."DIAGNOSIS" = "SQUAMOUS CELL CARCINOMA ORAL TONGUE/SDA" AND DEMOGRAPHIC."AGE" < "54"	Failed FROM clause: Incorrectly generating INNER JOIN Failed WHERE clause: Incorrect WHERE condition

TABLE 11 Accuracy evaluation of benchmarking MedT5SQL on two datasets.

# Training Epoc	WikiSQL			MIMICSQL		
	ACC_{LF}	ACC_{ASM}	Time consumed	ACC_{LF}	ACC_{ASM}	Time consumed
10 epochs	43.63%	94.1%	13 h	58%	97.369%	1 h 20 min
15 epochs	44.2%	94.26%	17 h	62.6%	98.054%	2 h
20 epochs	Model did not run			63.1%	98.1%	3 h
50 epochs	Model did not run			68.9%	98.572%	7 h 25 min
100 epochs	Model did not run			80.63%	98.937%	13 h 40 min

initial study due to limitations on time and resources. However, we acknowledge the importance of such a comparison and consider it a crucial direction for future research. In our ongoing work, we aim to broaden our investigation by conducting comparative analyses that include other fine-tuned LLMs, further elucidating the strengths and weaknesses of various approaches in the context of healthcare.

We also acknowledge that the MIMICSQL dataset, while valuable, may not fully represent the diversity of EMR data and clinical queries encountered in real-world healthcare settings. This could lead to the model underperforming or exhibiting biases when applied to different patient populations or healthcare institutions. Additionally, the T5 model, like other large language models, can inadvertently learn and perpetuate biases present in its vast pre-training corpus. These biases could manifest as discriminatory or inequitable behavior in generated SQL queries. To deal with these biases, expanding and diversifying the training data to include a wider range of EMR types and clinical scenarios can help mitigate data bias. Model bias, on the other hand, can be addressed by developing evaluation metrics specifically

TABLE 12 MedT5SQL and MedTS performance comparison.

Model	ACC_{LF}
MedTS, trained on 100 epochs	78.4%
MedT5SQL, trained on 100 epochs	80.63%

for assessing bias in generated SQL queries and continuously monitoring the model's performance for potential biases. Finally, we argue that exploring techniques for fine-tuning the model to explicitly reduce biases, such as incorporating fairness constraints or re-weighting training examples should be an important direction of future research.

6 Conclusion

In recent times, patient health data is stored digitally in electronic medical records (EMRs), which healthcare professionals use to access patients' historical health information or for clinical trials. The onset of the global COVID-19 pandemic in early 2020 overwhelmed hospitals with patients, straining healthcare workers due to a shortage of medical staff relative to the patient load (Birkmeyer et al., 2020; Kruizinga et al., 2021; Iness et al., 2022). This crisis underscored the significance of EMRs and underscored the necessity for a more efficient communication method (Dagliati et al., 2021). The critical need is for an interface that facilitates seamless interactions between end users and databases, specifically a system capable of generating SQL queries in response to human language inquiries.

To meet this requirement, natural language processing (NLP) for Text-to-SQL, which allows non-technical users to generate SQL queries to communicate with databases using natural language text conversion has emerged as a suitable solution. This research reviews existing research on Text-to-SQL conversion and proposes a Text-to-SQL conversion model for EMRs retrieval. In this work we employ Large Language Model (LLM), namely Text-to-Text Transfer Transformer (T5) model, a transformer-based pre-trained model for all text-based NLP tasks, to develop the Text-to-SQL model.

The proposed model was developed by fine-tunning the T5 model on MIMICSQL dataset, the first Text-to-SQL dataset for healthcare domain. The model was benchmarked on two optimizers, different training epochs, and two datasets to compare the performance: WikiSQL and MIMICSQL datasets. The model's performance was evaluated by comparing the generated query, in which the model was given a text, against the expected query of the text. The experiments showed that the model was able to achieve high accuracy in generating SQL queries from natural language questions, particularly for medical question-SQL pairs. Further, evaluations of the performance on each SQL clause have shown the model's efficiency in generating these specific query types. This research demonstrates the potential of fine-tuning the T5 model to achieve state-of-the-art results for generating SQL queries from natural language questions in the healthcare domain. While the model's current scope is focused on question-SQL pairs, it provides a solid foundation for future research to expand into more comprehensive SQL generation tasks.

This research demonstrates the potential of fine-tuning the T5 model to achieve state-of-the-art results for generating SQL queries from natural language questions in the healthcare domain. While the model's current scope is focused on question-SQL pairs, it provides a solid foundation for future research to expand into more comprehensive SQL generation tasks. The MedT5SQL model, while promising, represents a significant step toward empowering healthcare professionals with efficient and intuitive access to EMR data. Its potential real-world deployment in clinical settings could revolutionize how medical staff interact with patient information, enabling them to quickly retrieve relevant data for informed decision-making. However, practical considerations such as seamless integration with existing Electronic Health Record (EHR) systems, development of user-friendly interfaces, and ensuring data security and privacy are crucial for successful implementation. Additionally, addressing potential limitations of the model, such as its current focus on question-SQL pairs and the need to adapt to varying EMR schemas, will be essential to maximize its impact.

Moving forward, further research should focus on expanding the model's capabilities to encompass a broader range of SQL queries, thoroughly evaluating its performance in real-world clinical environments, and exploring its potential applications in areas such as clinical decision support and medical research. By addressing these challenges and opportunities, MedT5SQL has the potential to transform the way healthcare professionals leverage EMR data, ultimately improving patient care and clinical outcomes.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://github.com/wangpinggl/TREQS/tree/master/mimicsql_data/mimicsql_natural_v2; https://huggingface.co/datasets/wikisql.

Author contributions

AMa: Writing – review & editing, Writing – original draft, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. AA: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. AI: Writing – review & editing, Validation, Methodology, Investigation, Conceptualization. DB: Writing – review & editing, Project administration, Investigation, Data curation, Conceptualization. AMo: Writing – review & editing, Validation, Methodology, Investigation, Formal analysis, Data curation. MA: Writing – review & editing, Methodology, Investigation, Formal analysis, Data curation.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdata.2024.1371680/full#supplementary-material

References

American Medical Association (2018). *2018 AMA STEPS Forward™ Practice Transformation Series: Overcoming EHR Challenges: Strategies for Improving Usability and Efficiency*. New York, NY: American Medical Association.

Androutsopoulos, I., Ritchie, G. D., and Thanisch, P. (1995). Natural language interfaces to databases – an introduction. *Nat. Lang. Eng.* 1, 29–81. doi: 10.1017/S135132490000005X

Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. *arXiv preprint* arXiv:1607.06450.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). "Neural machine translation by jointly learning to align and translate," in *Paper presented at the 3rd International Conference on Learning Representations, ICLR 2015*. San Diego, CA, United States.

Birkmeyer, J. D., Barnato, A., Birkmeyer, N., Bessler, R., and Skinner, J. (2020). The impact of the COVID-19 pandemic on hospital admissions in the United States. *Health Affairs* 39, 2010–2017. doi: 10.1377/hlthaff.2020.00980

Choi, D., Shin, M. C., Kim, E., and Shin, D. R. (2021). RYANSQL: recursively applying sketch-based slot fillings for complex text-to-SQL in cross-domain databases. *Comp. Ling.* 47, 309–332. doi: 10.1162/coli_a_00403

Dagliati, A., Malovini, A., Tibollo, V., and Bellazzi, R. (2021). Health informatics and EHR to support clinical research in the COVID-19 pandemic: an overview. *Brief. Bioinf.* 22, 812–822. doi: 10.1093/bib/bbaa418

Demirkaya, A., Chen, J., and Oymak, S. (2020). "Exploring the role of loss functions in multiclass classification," in *2020 54th Annual Conference on Information Sciences and Systems (CISS)*, 1–5.

Devlin, J., Chang, M. W., Lee, K., and Toutanova, K. (2019). "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 1, 4171–4186, Minneapolis, MN.

Galassi, A., Lippi, M., and Torroni, P. (2021). Attention in natural language processing. *IEEE Trans. Neur. Netw. Learn. Syst.* 32, 4291–4308. doi: 10.1109/TNNLS.2020.3019893

Graves, A. (2013). Generating sequences with recurrent neural networks. *arXiv preprint* arXiv:1308.0850.

Groff, J. R., Weinberg, P. N., and Oppel, A. J. (2002). *SQL: The Complete Reference, 3rd Edn.* London: McGraw-Hill/Osborne.

Guo, J., Zhan, Z., Gao, Y., Xiao, Y., Lou, J. G., Liu, T., et al. (2019). "Towards complex text-to-SQL in cross-domain database with intermediate representation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 4524–4535*. Florence, Italy.

He, P., Mao, Y., Chakrabarti, K., and Chen, W. (2019). X-SQL: reinforce schema representation with context. *arXiv preprint* arXiv:1908.08113.

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neur. Comput.* 9, 1735–1780. doi: doi: 10.1162/neco.1997.9.8.1735

Huggingface.co (2022). T5. Available online at: https://huggingface.co/docs/transformers/model_doc/t5 (accessed May 6, 2022).

Hwang, W., Yim, J., Park, S., and Seo, M. (2019). A comprehensive exploration on WikiSQL with table-aware word contextualization. *arXiv preprint* arXiv:1902.01069.

Iness, A. N., Abaricia, J. O., Sawadogo, W., Iness, C. M., Duesberg, M., Cyrus, J., et al. (2022). *The effect of hospital visitor policies on patients, their visitors, and health care providers during the COVID-19* pandemic: a systematic review. *The Am. J. Med.* 135, 1158–1167. doi: 10.1016/j.amjmed.2022.04.005

Iyer, S., Konstas, I., Cheung, A., Krishnamurthy, J., and Zettlemoyer, L. (2017). "Learning a neural semantic parser from user feedback," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics Conference*. Vancouver, Canada, 963–973.

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., et al. (2016). MIMIC-III, a freely accessible critical care database. *Sci. Data* 3, 1–9. doi: 10.1038/sdata.2016.35

Kamath, A., and Das, R. (2018). "A survey on semantic parsing," in *Proceedings of the 1st Conference on Automated Knowledge Base Construction (AKBC 2019)*. Amherst, MA: Association for Computational Linguistics.

Kate, A., Kamble, S., Bodkhe, A., and Joshi, M. (2018). Conversion of natural language query to SQL query," in *Proceedings of the 2nd International Conference on Electronics, Communication and Aerospace Technology, ICECA*. Coimbatore, India, 488–491.

Kim, H., So, B. H., Han, W. S., and Lee, H. (2020). Natural language to SQL: Where are we today?. *Proc. VLDB Endow.* 13, 1737–1750. doi: doi: 10.14778/3401960.3401970

Kruizinga, M. D., Peeters, D., van Veen, M., van Houten, M., Wieringa, J., Noordzij, J. G., et al. (2021). The impact of lockdown on pediatric ED visits and hospital admissions during the COVID19 pandemic: a multicenter analysis and review of the literature. *Eur. J. Pediatr.* 180, 2271–2279. doi: 10.1007/s00431-021-04015-0

Lee, D. (2019). "Clause-wise and recursive decoding for complex and cross-domain text-to-SQL generation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. Hong Kong, China, 6045–6051.

Li, F., and Jagadish, H. V. (2014). "Constructing an interactive natural language interface for relational databases," in *Proceedings of the 41st International Conference on Very Large Data Bases, Vol. 8*. Kohala Coast, Hawaii. doi: 10.14778/2735461.2735468

Li, H., Zhang, J., Li, C., and Chen, H. (2023). "RESDSQL: Decoupling schema linking and skeleton parsing for text-to-SQL," in *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI)*.

Lightning (2022). *PyTorch Lightning*. Available online at: https://www.pytorchlightning.ai/ (accessed July 4, 2022).

Lin, X. V., Socher, R., and Xiong, C. (2020). "Bridging textual and tabular data for cross-domain text-to-SQL semantic parsing," in *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, 4870–4888.

Liu, S., Wright, A. P., Patterson, B. L., Wanderer, J. P., Turer, R. W., Nelson, S. D., et al. (2023). Using AI-generated suggestions from ChatGPT to optimize clinical decision support. *J. Am. Med. Inform. Assoc.* 30, 1237–1245. doi: 10.1093/jamia/ocad072

Liu, X., He, P., Chen, W., and Gao, J. (2019). "Multi-task deep neural networks for natural language understanding," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence: Association for Computational Linguistics (ACL), 4487–4496.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., et al. (2019). RoBERTa: a robustly optimized BERT pretraining approach. *arXiv preprint* arXiv:1907.11692.

Loshchilov, I., and Hutter, F. (2017). "*Decoupled weight decay regularization*," in *Processing of Seventh International Conference on Learning Representations (ICLR 2019)*.

Lyu, Q., Chakrabarti, K., Hathi, S., Kundu, S., Zhang, J., and Chen, Z. (2020). Hybrid ranking network for text-to-SQL. *arXiv preprint* arXiv:2008.04759.

Marbán, Ó., Mariscal, G., and Segovia, J. (2009). "A data mining & knowledge discovery process model," in *Data Mining and Knowledge Discovery in Real Life Applications* (IntechOpen).

Marshan, A., Kansouzidou, G., and Ioannou, A. (2021). Sentiment analysis to support marketing decision making process: a hybrid model. *Adv. Int. Syst. Comput.* 1289, 614–626. doi: 10.1007/978-3-030-63089-8_40

Marshan, A., Nizar, F. N. M., Ioannou, A., and Spanaki, K. (2023). Comparing machine learning and deep learning techniques for text analytics: detecting the severity of hate comments online. *Inf. Syst. Front.* doi: 10.1007/s10796-023-10446-x

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Orallo, J. H., Kull, M., Lachiche, N., et al. (2021). CRISP-DM twenty years later: from data mining processes to data science trajectories. *IEEE Trans. Knowl. Data Eng.* 33, 3048–3061. doi: 10.1109/TKDE.2019.2962680

Masri, N., Sultan, Y. A., Akkila, A. N., Almasri, A., Ahmed, A., Mahmoud, A. Y., et al. (2019). Survey of rule-based systems. *IJAISR* 3, 1–22.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). "Efficient estimation of word representations in vector space," in *Proceedings of the 1st International Conference on Learning Representations, ICLR 2013*. Scottsdale, AZ: Workshop Track Proceedings.

Ncr and Clinton, J. (1999). *CRISP-DM 1.0 Step-by-Step Data Mining Guide*. CRISP-DM consortium: NCR Systems Engineering Copenhagen (USA and Denmark), DaimlerChrysler AG (Germany), SPSS Inc. (USA) and OHRA Verzekeringen en Bank Groep B.V (The Netherlands).

Pan, Y., Wang, C., Hu, B., Xiang, Y., Wang, X., Chen, Q., et al. (2021). A BERT-based generation model to transform medical texts to SQL queries for electronic medical records: model development and validation. *JMIR Med. Inf.* 9:698. doi: 10.2196/32698

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). Pytorch: an imperative style, high-performance deep learning library. *Adv. Neur. Inf. Proc. Syst.* 32, 1–14.

Pennington, J., Socher, R., and Manning, C. D. (2014). "GloVe: global vectors for word," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), Vol. 14*. Doha, Qatar, 1532–1543.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., et al. (2018). "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1*. New Orleans, LA, 2227–2237.

Popescu, A. M., Armanasu, A., Etzioni, O., Ko, D., and Yates, A. (2004). "Modern natural language interfaces to databases: composing statistical parsing with semantic tractability," in *Proceedings of the 20th International Conference on Computational Linguistics*. Geneva, Switzerland, 141–147.

Price, P. J. (1990). "Evaluation of spoken language systems: the ATIS domain," in *Proceedings Speech and Natural Language: a Workshop Held at Hidden Valley*. Valley, PA, 91–95.

Pytorch-lightning (2022). *LightningModule — PyTorch Lightning 1.6.5 Documentation*. Available online at: https://pytorch-lightning.readthedocs.io/en/stable/common/lightning_module.html (accessed July 4, 2022).

Qi, J., Tang, J., He, Z., Wan, X., Zhou, C., Wang, X., et al. (2022). Rasat: Integrating relational structures into pretrained seq2seq model for text-to-SQL. doi: 10.18653/v1/2022.emnlp-main.211

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21, 1–67.

Saha, D., Floratou, A., Sankaranarayanan, K., Minhas, U. F., and Mittal, A. R., Özcan, F., et al. (2016). ATHENA: an ontologydriven system for natural language querying over relational data stores. *Proc. VLDB Endowment* 9, 1209–1220. doi: 10.14778/2994509.2994536

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint* arXiv:1910.01108.

Sawarkar, K. (2022). *Deep Learning With PyTorch Lightning. Packt*. Available online at: https://www.packtpub.com/product/deep-learning-with-pytorch-lightning/9781800561618 (accessed September 23, 2022).

Scholak, T., Schucher, N., and Bahdanau, D. (2021). "PICARD: parsing incrementally for constrained auto-regressive decoding from language models," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic. Association for Computational Linguistics, 9895–9901.

Shanafelt, T. D., Hasan, O., Dyrbye, L. N., Sinsky, C. A., Satele, D., Sloan, J., et al. (2012). Changes in burnout and satisfaction with work-life balance in physicians and the general US working population between 2011 and 2014. *Mayo Clin. Proc.* 87, 431–440.

Shaw, P., Chang, M. W., Pasupat, P., and Toutanova, K. (2020). "Compositional generalization and natural language variation: Can a semantic parsing approach handle both?," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Vol. 1*. Association for Computational Linguistics, 922–938.

Shazeer, N., and Stern, M. (2018). "Adafactor: Adaptive learning rates with sublinear memory cost," *in Proceedings of the 35th International Conference on Machine Learning, Vol. 80*. Dy, Jennifer and Krause, Andreas, 4596–4604.

Singh, H., Giardina, T. D., Meyer, A. N., Forjuoh, S. N., Reis, M. D., Thomas, E. J., et al. (2013). Types and origins of diagnostic errors in primary care settings. *JAMA Int. Med.* 173, 418–425. doi: 10.1001/jamainternmed.2013.2777

Sinsky, C. A., Colligan, L., Li, L., Prgomet, M., Reynolds, S., Goeders, L., et al. (2016). Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. *Annal. Int. Med.* 165, 753–760. doi: 10.7326/M16-0961

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). "Sequence to sequence learning with neural networks," in *Proceedings of the 27th International Conference on Neural Information Processing Systems*. Montreal, Canada, 3104–3112.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. (1999). "Policy gradient methods for reinforcement learning with function approximation," in *Proceedings of the 12th International Conference on Neural Information Processing Systems (NIPS'99)*. Denver, CO, 1057–1063.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Long Beach, CA, 6000–6010.

Vinyals, O., Fortunato, M., and Jaitly, N. (2015). "Pointer networks," in *Proceedings of the 28th International Conference on Neural Information Processing Systems (NIPS'15), Vol. 2*. Montreal, Canada, 2692–2700.

Wang, B., Shin, R., Liu, X., Polozov, O., and Richardson, M. (2019). "RAT-SQL: relation-aware schema encoding and linking for text-to-SQL parsers," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 7567–7578.

Wang, C., Huang, P., Polozov, O., Brockschmidt, M., and Singh, R. (2018). "Execution-guided neural program decoding," in *ICML Work-Shop on Neural Abstract Machines and Program Induction v2 (NAMPI)*. Stockholm, Sweden.

Wang, P., and Shi, T. (2020). *mimicSQL_natural_v2*. Available online at: https://github.com/wangpinggl/TREQS/tree/master/mimicSQL_data/mimicSQL_natural_v2 (accessed April 5, 2022).

Wang, P., Shi, T., and Reddy, C. K. (2020). "Text-to-SQL generation for question answering on electronic medical records," in *Proceedings the International Conference on World Wide Web (WWW)*.

Wang, P., Shi, T., and Reddy, C. K. (2020). "Text-to-SQL generation for question answering on electronic medical records," in *Proceedings of the World Wide Web Conference, Association for Computing Machinery*. Taipei, Taiwan, 350–361.

Webster, J. J., and Kit, C. (1992). Tokenization as the initial phase in NLP," in *Proceedings of the 14th conference on Computational linguistics, Vol. 4*. Nantes, France, 1106–1110.

Xie, T., Wu, C. H., Shi, P., Zhong, R., Scholak, T., Yasunaga, M., et al. (2022). UnifiedSKG: unifying and multi-tasking structured knowledge grounding with text-to-text language models. *arXiv preprint* arXiv:2201.05966.

Xu, X., Liu, C., and Song, D. (2017). SQLNet: Generating structured queries from natural language without reinforcement learning. *arXiv preprint* arXiv:1711.04436.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., Le, Q. V., et al. (2019). "XLNet: generalized autoregressive pretraining for language understanding," in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Vancouver: Curran Associates Inc, 5753–5763.

Yin, W., Kann, K., Yu, M., and Schütze, H. (2017). Comparative study of CNN and RNN for natural language processing. *arXiv preprint* arXiv:1702.01923.

Yu, T., Li, Z., Zhang, Z., Zhang, R., and Radev, D. (2018c). "TypeSQL: knowledge-based type-aware neural text-to-SQL generation," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 2*, New Orleans, LA: Association for Computational Linguistics, 588–594.

Yu, T., Zhang, R., Yang, K., Yasunaga, M., Wang, D., Li, Z., et al. (2018a). "Spider: a large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-SQL task," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium, 3911–3921.

Yu, T., Zhang, R., Yang, K., Yasunaga, M., Wang, D., Li, Z., et al. (2018b). "SyntaxSQLNet: syntax tree networks for complex and cross-domain text-to-SQL task," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels: Association for Computational Linguistics, 1653–1663.

Zelle, J. M. (1996). "Learning to parse database queries using inductive logic programming," in *Proceedings of the thirteenth national conference on Artificial intelligence, Vol. 2*. Portland, Oregon, 1050–1055.

Zettlemoyer, L. S., and Collins, M. (2005). "Learning to map sentences to logical form: structured classification with probabilistic categorial grammars," in *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence*. Edinburgh, Scotland, 658–666.

Zhong, V., Xiong, C., and Socher, R. (2017). Seq2SQL: generating structured queries from natural language using reinforcement learning. *arXiv preprint* arXiv:1709.00103.

frontiers | Frontiers in Virtual Reality

TYPE Original Research
PUBLISHED 03 July 2024
DOI 10.3389/frvir.2024.1363193

OPEN ACCESS

EDITED BY
Daniele Giunchi,
University College London, United Kingdom

REVIEWED BY
Ilias Maglogiannis,
University of Piraeus, Greece
Soraia Oueida,
American University of the Middle East, Kuwait

*CORRESPONDENCE
Jacob Stuart,
jpstuar@emory.edu

RECEIVED 30 December 2023
ACCEPTED 20 May 2024
PUBLISHED 03 July 2024

CITATION
Stuart J, Stephen A, Aul K, Bumbach MD, Huffman S, Russo B and Lok B (2024), Developing augmented reality filters to display visual cues on diverse skin tones.
Front. Virtual Real. 5:1363193.
doi: 10.3389/frvir.2024.1363193

Developing augmented reality filters to display visual cues on diverse skin tones

Jacob Stuart[1]*, Anita Stephen[2], Karen Aul[3], Michael D. Bumbach[2], Shari Huffman[2], Brooke Russo[2] and Benjamin Lok[4]

[1]School of Medicine, Emory University, Atlanta, GA, United States, [2]College of Nursing, University of Florida, Gainesville, FL, United States, [3]College of Nursing, University of South Florida, Tampa, FL, United States, [4]Department of Computer and Information Science and Engineering, University of Florida, Gainesville, FL, United States

Introduction: Variations in skin tone can significantly alter the appearance of symptoms such as rashes or bruises. Unfortunately, previous works utilizing Augmented Reality (AR) in simulating visual symptoms have often failed to consider this critical aspect, potentially leading to inadequate training and education. This study seeks to address this gap by integrating generative artificial intelligence (AI) into the AR filter design process.

Methods: We conducted a 2 × 5 within-subjects study with second-year nursing students (N = 117) from the University of Florida. The study manipulated two factors: symptom generation style and skin tone. Symptom generation style was manipulated using a filter based on a real symptom image or a filter based on a computer-generated symptom image. Skin tone variations were created by applying AR filters to computer-generated images of faces with five skin tones ranging from light to dark. To control for factors like lighting or 3D tracking, 101 pre-generated images were created for each condition, representing a range of filter transparency levels (0–100). Participants used visual analog scales on a computer screen to adjust the symptom transparency in the images until they observed image changes and distinct symptom patterns. Participants also rated the realism of each condition and provided feedback on how the symptom style and skin tone impacted their perceptions.

Results: Students rated the symptoms displayed by the computer-generated AR filters as marginally more realistic than those displayed by the real image AR filters. However, students identified symptoms earlier with the real-image filters. Additionally, SET-M and Theory of Planned Behavior questions indicate that the activity increased students' feelings of confidence and self-efficacy. Finally, we found that similar to the real world, where symptoms on dark skin tones are identified at later stages of development, students identified symptoms at later stages as skin tone darkened regardless of cue type.

Conclusion: This work implemented a novel approach to develop AR filters that display time-based visual cues on diverse skin tones. Additionally, this work provides evidence-based recommendations on how and when generative AI-based AR filters can be effectively used in healthcare education.

KEYWORDS

augmented reality, visual cue training, healthcare, simulation, symptoms, fidelity, realism

1 Introduction

In healthcare simulations, the accurate representation of diverse skin tones is not merely an ethical imperative, but a medical necessity. Patients with skin of color are more likely to experience misdiagnosis or be diagnosed later in their disease's development (Narla et al., 2022). A major contributor to these disparities is the lack of adequate training for skin of color in medical education. For example, multiple works have researched the inclusion of images containing dark skin tones in medical textbooks and resources and found that dark skin tones are only represented in 4%–18% of images (Ebede and Papier, 2006; Kaundinya and Kundu, 2021; Harp et al., 2022). Further, a previous study found that only 19.5% of program directors and 25.4% chief residents reported having lectures on skin of color from an acknowledged expert (Nijhawan et al., 2008). Given the scarcity of resources for skin of color, it is unsurprising that previous research found that healthcare providers reported significantly less confidence assessing lupus-related rashes in people with skin of color than in patients with fair skin (Kannuthurai et al., 2021). However, this issue is not isolated to just Lupus as another work reports that 47% of dermatologists believed their medical training was inadequate in teaching them how to identify skin conditions for people with darker skin tones (Buster et al., 2012). Those who felt their training was inadequate stated the need for more exposure to training materials and patients with skin of color (Buster et al., 2012).

The logical approach to addressing this disparity would involve educating learners on recognizing symptoms across a broad spectrum of skin tones. Yet, the aforementioned shortage of medical imagery showcasing dark skin tones complicates this solution. However, Augmented reality (AR) filters, which digitally overlay graphics or effects onto real-world images or videos (Fribourg et al., 2021), are a promising solution to display symptoms during healthcare training. Previous works have used AR to depict a variety of medical conditions. Some examples include Noll et al. using AR based tracking to overlay melanoma onto users (Noll et al., 2017), Liang et al. overlaying a virtual head depicting stroke symptoms onto a manikin (Liang et al., 2021), Stuart et al. using AR filters to overlay allergic reaction symptoms onto a conversational agent (Stuart et al., 2022), and Stuart et al. using AR filters to develop a system that allowed students to manipulate Lupus symptoms in real-time (Stuart et al., 2023). Unfortunately, current explorations into using AR for simulating visual symptoms are still in early stages. As such, they were mainly focused on making the symptoms visible for just one person/manikin and did not investigate how AR application/development would need to differ to be applied to multiple skin tones (Noll et al., 2017; Liang et al., 2021; Stuart et al., 2022; 2023). AR filters particularly need to be tailored to skin tone for diseases like lupus and melanoma which can manifest differently depending on skin color (Gloster and Neal, 2006; Nelson, 2020; Lee et al., 2023). For example, what appears as a red rash on lighter skin may appear dark brown on darker skin (Ludmann, 2022).

To work towards a solution for having AR filters depict symptoms on a range of diverse skin tones, this work builds upon that of Stuart et al. by introducing the use of generative artificial intelligence (AI) within the AR filter design process (Stuart et al., 2023). Generative AI has been defined as the use of models, such as generative adversarial networks or encoder-decoder networks, to generate various resources (García-Peñalvo and Vázquez-Ingelmo, 2023). Specifically, this work examines using a commercial diffusion model training system (Scenario, 2023) to create an image generator that can take in a face image and output a similar face with a Malar rash, a distinct butterfly shaped (i.e., mainly covers the cheeks and nose) face rash that can develop over time by those with Systemic Lupus Erythematosus (Ludmann, 2022). This new image can then be used to produce an AR filter. This AR filter creation method is evaluated for five different skin tones using a similar web-based evaluation system to Stuart et al. (Stuart et al., 2023) gathering information on when users could identify symptoms when increasing the alpha (transparency) value of the AR filter and how realistic users thought the symptoms appeared.

This work focuses on the process of using generative AI in the design process to allow designers to create AR filters that better represent time-based visual symptoms across a broad spectrum of skin tones. By doing so, this research can help identify potential biases, address disparities in perception, and inform ethical development/deployment of AR experiences. Ultimately, these findings can contribute to industry best practices, promote diversity, equity, and inclusion in technology development, and ensure that AR technology is accessible and enjoyable for all users, regardless of their skin tone or background.

2 Materials

Five different skin tones were examined in this study. We will refer to these as light, medium-light, medium-dark, dark-light, and dark. For each skin tone, participants saw visual cues based on real images and visual cues based on generated images. Thus, participants were asked to complete visual analog scale questions for a total of 10 conditions (Skin tone X Cue type): Light-Real, Light-Generated, Medium-Light-Real, Medium-Light-Generated, Medium-Dark-Real, Medium-Dark-Generated, Dark-Light-Real, Dark-Light-Generated, Dark-Real, and Dark-Generated (Figure 1).

To create the real-image skin conditions, we selected computer-generated faces from generated. photos, a company focused on producing photorealistic images of people, with skin tones ranging from light to dark (Generated, 2023). Once images were selected, a similar process to (Stuart et al., 2023) was used to overlay the symptoms on to the faces (Section 2.2). The five faces depicted were chosen as they were good skin tone matches to existing medical imagery that could be used for the real conditions.

To create the generated conditions, image generators were trained using images depicting malar rash symptoms for each skin tone (Section 2.1). The objective of the image generator is to take an input image of an individual and output a new image. This output image depicts a new person with a similar skin tone and shows the signs of malar rash symptoms. Output images are used to develop the computer-generated AR filter conditions (Figure 2).

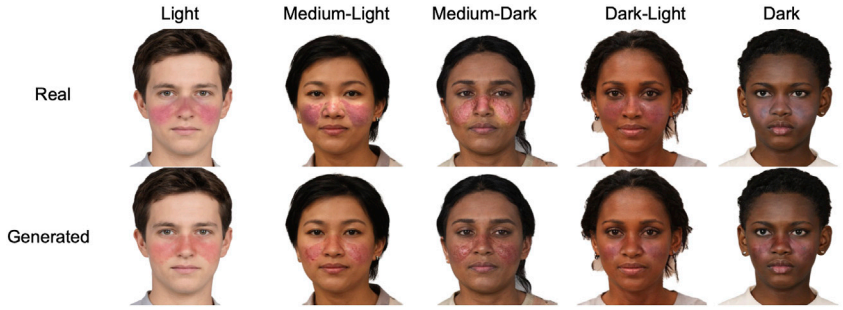

FIGURE 1
Chart showing skin tone X cue type variables. The real row shows faces with AR filters that are made using real images of people with malar rash symptoms. The generated row shows faces with AR filters that are made using images created using the generative AI platform Scenario.

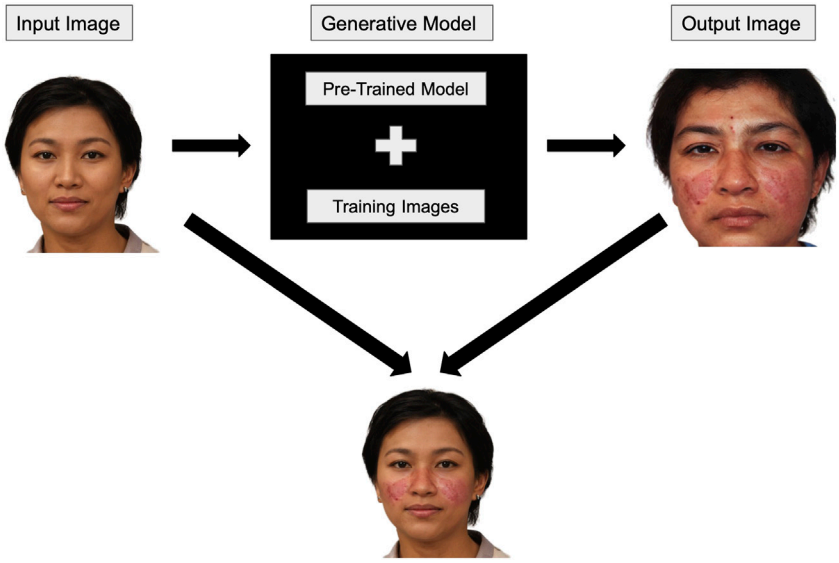

FIGURE 2
An input image is given to the model to generate a picture of a person with a similar skin tone with malar rash symptoms. The output image is then used to produce the AR filter which can be applied to the original input image.

2.1 Image generator

We utilized Scenario, a commercial platform specializing in generative image models, to create our image generator (Scenario, 2023). This platform provides pre-trained models capable of generating human figures and avatars. Users can fine-tune these pre-trained models by uploading their unique set of images. This flexibility is highly advantageous as the pre-trained model already understands human facial features, allowing the additional data to focus on the appearance of individuals with malar rash symptoms. This approach reduces the necessity for an extensive training set that portrays people exhibiting malar rash symptoms.

An iterative design process was used to create the training sets for each model. The first model included all (n = 66) images that could be found from reputable online sources of systemic lupus erythematosus patients depicting malar rash. This initial model often lightened the skin tone around the nose and cheeks, contained visual artifacts that harmed the visual quality of the malar rash visual cue, and generally lightened darker skin tones, which interfered with the visual accuracy of the malar rash (Figure 3).

After creating this initial model, several steps were taken to enhance the image quality for subsequent models, in line with established best practices (Shorten and Khoshgoftaar, 2019; Larrazabal et al., 2020; Wang et al., 2020; Maluleke et al., 2022):

1. Images suffering from poor or low-lighting conditions were eliminated as they often resulted in output images with lighter skin tones in the rash pattern area, largely due to significant specular reflections in these images.
2. Low-resolution images were removed to minimize output image artifacts.

FIGURE 3
This figure shows examples of images generated using the first model. Note that many of the images are significantly lighter than the input image. Additionally, these may end up exhibiting features that are more representative of those with lighter skin tones (e.g., different facial features).

3. The original dataset was divided into subsets representing different skin tones to counter the lightening (whitewashing) of darker skin tones caused by an overrepresentation of lighter skin tone examples.

This dataset division resulted in smaller training sets for each skin tone. While feasible in this case, such an approach might not be viable in scenarios with limited diverse images or when the visual cue is not as clearly defined.

Future improvements can be made to the dataset division and output image selection processes, which were manually executed in this project. A system that measures skin tone similarity, perhaps referencing Fitzpatrick Phototypes, might allow a more rigorous division of training sets and selection of final output images. However, designing an accurate skin tone comparison system would necessitate further research, as factors like shadows, lighting, and reflections would need to be considered. For the current iteration of this process, once the generated images passed the author's approval (i.e., did not have obvious visual artifacts or racial bias), they were reviewed by nursing collaborators for face validity before being used to create the AR filters (Figure 1).

2.2 AR filter creation and application

To create the ten AR filters, the real symptom images, the generated symptom images, and the face images from Generated. photos were uploaded to Lens Studio, a program for AR filter creation for the Snapchat platform (Snap Inc, 2021). Once uploaded, a face mask was created for each condition. Facial details that are not relevant to the Malar rash symptoms (e.g., areas of the forehead) are removed from the face mask using an opacity texture. This would result in an AR Filter that would overlay areas of the face that present Malar rash symptoms. The face images from Generated. photos could then be set as the camera source and the real and generated symptom images could be applied to a face using the face mask. The face mask would automatically track to the face in the camera source.

For each of the ten conditions, we used an AR filter to create 101 images, with the alpha (i.e., transparency) level of the Malar rash face mask ranging from 0 to 100. The images were then uploaded to Qualtrics to create visual analog questions.

2.3 Visual analog question design

Similar to Stuart et al., this work uses visual analog scales to depict patient deterioration over time by manipulating the alpha level of the AR visual cue (Stuart et al., 2023). These scales enabled students to manipulate the Malar rash symptoms in real-time. This method let us use the created AR filters while managing variables introduced by AR, such as tracking, lighting, or hardware issues.

The visual analog scales use modified versions of Qualtrics' visual analog scale question. Each question utilized 101 images (Section 2.2) with alpha levels ranging from 0% (full transparency) to 100% (full visibility). This allowed for the use of a 0–100 scale. Other benefits of this method include allowing students to provide precise points at which they noticed symptom developments, allowing students to easily control the state of the symptom that was being displayed, and go back to a previous state if they accidentally passed where they believe they noticed changes, and it helped to reduce the total time needed to complete the survey (which is vital with the limited class time allowed to complete the study).

Other options for displaying symptoms were considered. These included 1) applying the AR filters in real-time using Snap's Camera Kit an SDK that allows developers to implement Snap's AR technology into websites, and 2) creating short videos that automatically had the symptoms develop over time with students clicking when they noticed the desire stages. The slider method was chosen over the real-time application because it was unclear if all student's laptops would be capable of running Camera Kit. Additionally, the slider method was chosen instead of showing students a video of the symptom developing over time and getting an actual time amount for several reasons. Most importantly, rash symptoms vary in the severity they can reach, and the time it takes to develop (Brown, 2003). Therefore, it is more important to identify stages of symptom development, such as initial changes and pattern identification, by the variable being manipulated (alpha level) rather than the time the variables are changed over. Time as a variable can be manipulated in future works to investigate different symptom development speeds.

In addition to reviewing the output images discussed in section 2.1, the five nursing faculty also separately reviewed the visual analog questions for all conditions. The nursing collaborators were asked to evaluate the face validity of the rash and its development using the

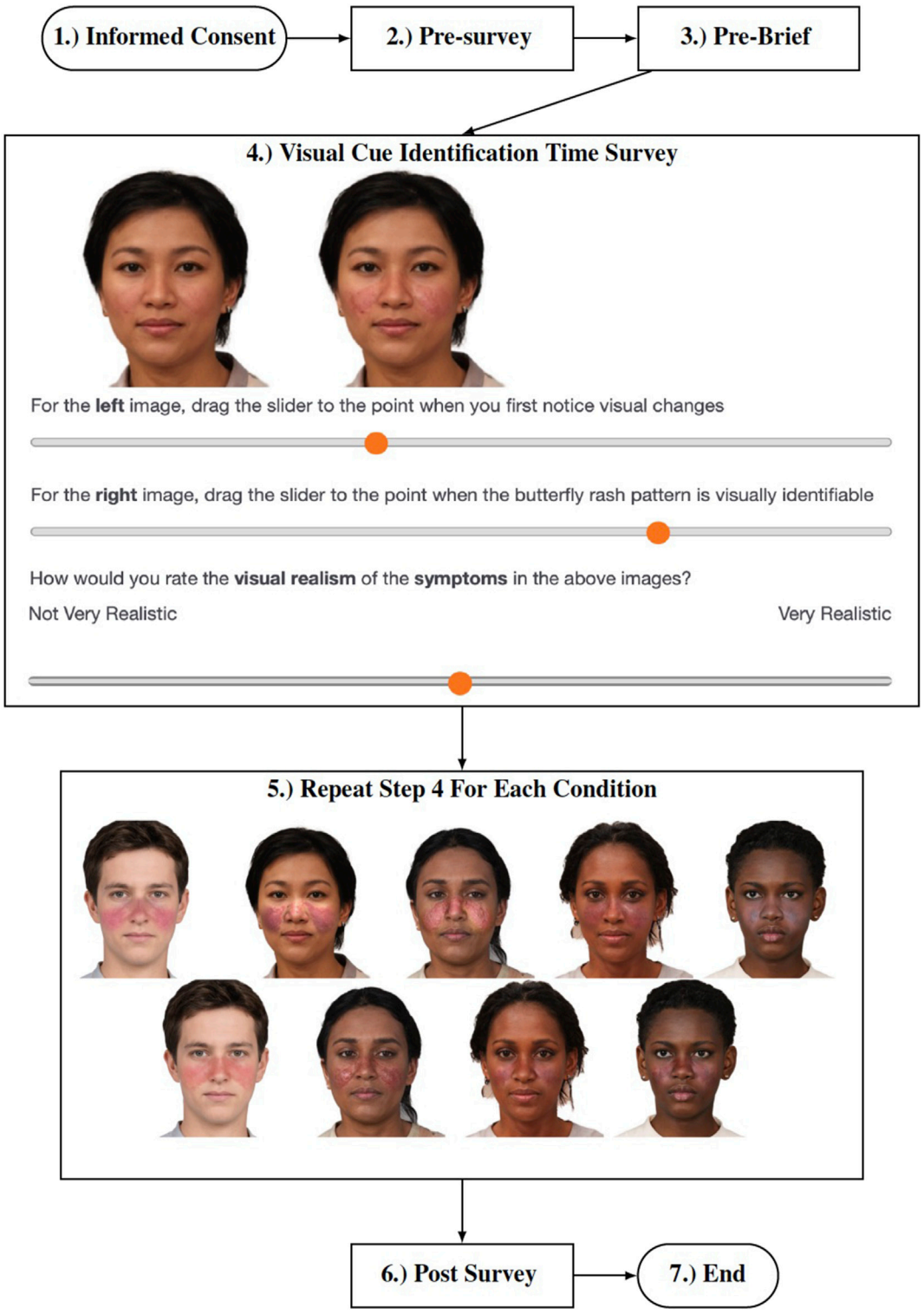

FIGURE 4
Procedure followed by participants in the study.

visual analog questions (i.e., does this look like an accurate representation of Malar rash symptoms on each of the skin tones?) and the visibility of the symptoms. Visibility was assessed because faculty wanted to see if the images would provide educational value to students. If the symptoms displayed in the images were too difficult for students to acknowledge and assess with different skin tones the educational value provided by the images would be low.

3 Methods

3.1 Participants

Second-year nursing students (N = 117) were recruited from a nursing course (Principals of Personalized Nursing Care 2) taught at the University of Florida in the Spring of 2023. Six students reported not having normal or corrected-to-normal vision and a colorblindness test determined that 13 students experienced some level of color blindness. These participants are excluded from analysis in this paper due to potential differences in visual perception of the conditions. This led to a remaining population of 98 nursing students.

Of the 98 remaining students, all were 18–24 years old. Students' self-reported genders were: 9 Males and 88 Females, and one non-binary. Students' self-reported races were: 8 Asian, 1 Asian/Other, 3 Black or African American, 1 Black or African American/Asian, 2 Other, 79 White or Caucasian, 2 White or Caucasian/Asian, 1 White or Caucasian/Other, and one preferred not to say. As for their familiarity with the malar rash symptoms presented: five were slightly familiar, one was somewhat familiar, and 92 were not familiar at all.

3.2 Study procedure

During class, participants were provided a link to the Qualtrics study in their course management software to follow the study flow shown in Figure 4. Participants began by reading and signing the informed consent. Then participants completed a pre-survey that included demographics, screen brightness, and color blindness questions. Following the pre-survey, a pre-brief section informed participants about malar rash symptoms and explained the questions they would answer regarding the malar rash visual cues. After the pre-brief, participants completed two visual analog scale questions and a semantic differential scale for each of the 10 conditions. Finally, participants completed a post-survey and ended the study.

3.3 Metrics

3.3.1 Pre-survey metrics

The pre-survey questionnaire asked seven demographics questions regarding participants' age, race, gender, vision status, malar rash familiarity, screen brightness, and color blindness.

In addition to the demographics questions, students were also asked to answer eight theory of planned behavior (TPB) intention questions before and after the intervention. These questions adapted from Ajzen et al. were used to gather changes in user perceptions regarding their intention to take patient skin characteristics into consideration the next time they perform a skin assessment (Ajzen, 2006). This survey measures three variables that influence a users intentions to perform a behavior as well as their overall perception of their own intentions. The three variables measured are behavioral beliefs, normative beliefs, and control beliefs. Each of these variables is measured using two seven-point semantic scale questions.

Behavioral beliefs are "beliefs about the likely outcomes of the behavior and the evaluations of these outcomes" (Ajzen, 2006). The two questions in the results that correspond to this belief are labeled AttitudeGood and AttitudeBeneficial. These questions are "Taking patient skin characteristics into consideration the next time I perform a skin assessment would be X for the patient's health outcome" where X was a rating from bad to good, and "Taking patient skin characteristics into consideration the next time I perform a skin assessment would be X″ where X was a rating from not beneficial to beneficial.

Normative beliefs are "beliefs about the normative expectations of others and motivation to comply with these expectations" (Ajzen, 2006). The two questions in the results that correspond to this belief are labeled NormLikeMe and NormApprove. These questions are "Most people like me take patient skin characteristics into consideration every time they perform a skin assessment" rated from unlikely to likely, and "Most people who are important to me approve of taking patient skin characteristics into consideration every time I perform a skin assessment" rated from disagree to agree.

Control beliefs are "beliefs about the presence of factors that may facilitate or impede performance of the behavior and the perceived power of these factors" (Ajzen, 2006). The two questions in the results that correspond to this belief are labeled ControlUpToMe and ControlEfficacy. These questions are "Taking patient skin characteristics into consideration the next time I perform a skin assessment is up to me" rated from disagree to agree, and "I am confident that I can take patient skin characteristics into consideration every time I perform a skin assessment" rated from true to false.

This questionnaire ends with a direct measure of user intentions with the question "I intend to take patient skin characteristics into consideration the next time I participate in a skin assessment to have a good patient health outcome" with a rating from false to true.

In addition to following the theory of planned behavior framework, these questions were also reviewed by nursing collaborators to verify that they would be coherent to those in the nursing domain.

3.3.2 Intervention metrics

Following the TPB intention questions, participants were asked to complete visual analog scale questions and the semantic differential scale for each condition, a total of 30 questions (Figure 4). The visual analog scale questions are used to measure at what point users identified different stages of patient deterioration (initial changes in appearance and the butterfly pattern appearance). The semantic differential scale question was used to measure users perceptions of symptom realism rated from 0 to 100 with 0 labeled as "Not Very Realistic" and 100 labeled as "Very Realistic". We asked users to assess the realism of the stages they identified (just noticeable and a clear butterfly pattern), to understand whether they perceive these symptoms as realistic before full development. This differs from Stuart et al. which asked about realism when the symptoms were fully developed (fully opaque symptoms) Stuart et al. (2023). Our approach in this study aimed to understand whether using transparency as a manipulative variable maintains the realism of symptoms during critical stages of their development. If realism was rated low overall or if realism was much lower for a specific skin tone, then it would have been reasonable to assume that

alpha level was not a feasible variable to manipulate when looking at symptom development over time for diverse skin tones.

3.3.3 Post-survey metrics

The study concluded with a post-survey gathering user intentions again (TPB), a subset of four questions related to confidence and self-efficacy from the simulation-effectiveness tool (SET-M) which is used to evaluate perceptions of the effectiveness of learning in the simulation environment (Leighton et al., 2015), and two questions gathering open-ended qualitative responses from participants regarding their perceptions of the visual cues and how they believe their perceptions were affected by the differences in filter and object fidelity.

4 Results

4.1 Data analysis

Descriptive and inferential statistics are reported for participants first noticing changes, noticing the malar rash pattern, and the perceived realism of the malar rash. Similar to (Stuart et al., 2022), we utilized the filter alpha level to determine when students first identified changes and noticed the malar rash pattern. In these sections, a lower score signifies students identifying changes and patterns sooner. Additionally, the units of measurement for realism is similar to the previous work and refers to the students' self-reported perceptions on a 0–100 visual analog scale. A lower score for this is interpreted as students perceiving the visual cue as less realistic.

For each visual analog scale question and the semantic differential scale, a two-way repeated measures ANOVA was performed. All questions were assessed for normality using Q-Q plots and Kolmogorov-Smirnov tests and all were found to follow normal distributions. A Mauchly test of sphericity was performed to check for sphericity assumptions. For tests that violated sphericity, the Greenhouse-Geisser and the Huynh-Feldt epsilon values are greater than 0.75. Therefore the repeated measures ANOVA results for these measures are reported based on the Huynh-Feldt corrections. When repeated measures ANOVA indicated significant differences, *post hoc* tests were performed using the Holm correction. The post hocs allowed for the analysis of the perceptual differences between skin tones and cue types.

For the Simulation Effectiveness Tool Modified (SET-M), frequencies of student responses are reported since there are no comparisons to pre-intervention data or other conditions to be made.

For the theory of planned behavior questions, a wilcoxon signed-rank test was conducted to compare the ordinal data from the pre-post responses. Additionally, the frequencies of each response are reported.

For the qualitative questions, we analyzed the responses to identify if students believed the symptom style affected their perception of realism of the symptoms and their ability to identify the symptoms. There were 89 and 86 responses for these questions respectively. A first pass of the responses was completed for each question to determine categories. During this first pass, it was determined that the majority of user responses were comparing the conditions. From this finding, it was decided to categorize the responses for the realism question as either stating the symptom realism was similar for both cue styles, more realistic for the AR-filter based on the computer-generated image, or more realistic for the AR filter based on the real symptom image. The ability to identify symptoms categories were similarly created. Responses were only counted towards categories if answers were clear and unambiguous, so no discussion between coders was necessary.

4.2 Quantitative

4.2.1 First noticing changes

A Mauchly's test of sphericity was significant for skin tones for this measure, therefore the Huynh-Feldt correction was used. A two-way repeated measures ANOVA showed that first changes ratings differed significantly between skin tone levels, $F(3.23, 313.68) = 105.02$, $p < 0.001$, $\omega^2 = 0.236$. Additionally, first changes ratings differed significantly between cue types, $F(1.00, 97.00) = 26.86$, $p < 0.001$, $\omega^2 = 0.016$.

Post hoc testing using the Holm correction revealed that the light skin tone condition was rated significantly lower in regards to first noticing changes when compared to all other skin tones, the medium-light skin tone condition was rated significantly lower in regards to first

TABLE 1 Post Hoc results for skin tone in regards to first noticing changes.

		Mean difference	SE	t	Cohen's d	p_{holm}
Light	Medium Light	−4.510	1.082	−4.168	−0.327	< 0.001
	Medium Dark	−5.117	1.082	−4.729	−0.371	< 0.001
	Dark Light	−17.485	1.082	−16.156	−1.269	< 0.001
	Dark	−16.622	1.082	−15.359	−1.206	< 0.001
Medium Light	Medium Dark	−0.607	1.082	−0.561	−0.044	0.852
	Dark Light	−12.974	1.082	−11.989	−0.942	< 0.001
	Dark	−12.112	1.082	−11.192	−0.879	< 0.001
Medium Dark	Dark Light	−12.367	1.082	−11.428	−0.897	< 0.001
	Dark	−11.505	1.082	−10.631	−0.835	< 0.001
Dark Light	Dark	0.862	1.082	0.797	0.063	0.852

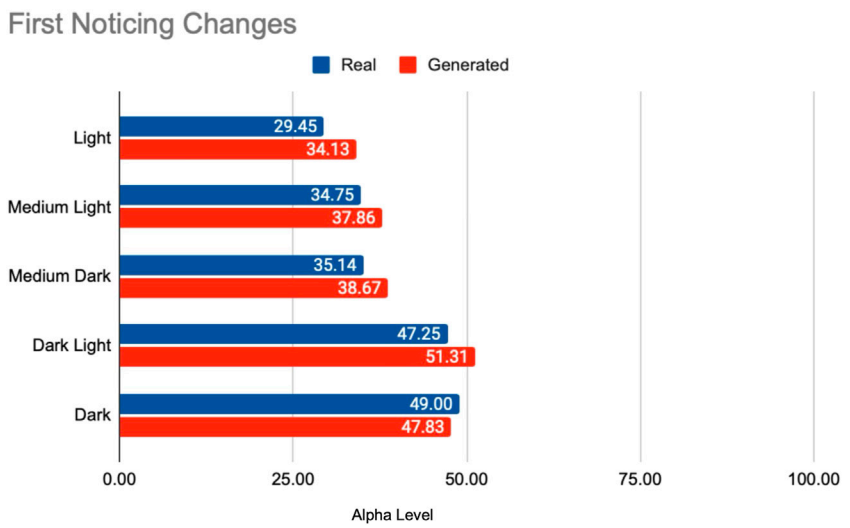

FIGURE 5
Visual depiction of means showing how as skin tone darkened, it took users longer to notice initial changes occurring with the visual cue.

TABLE 2 Post Hoc results for cue type in regards to first noticing changes.

		Mean difference	SE	t	Cohen's d	p_{holm}
Real	Generated	−2.843	0.549	−5.182	−0.206	< 0.001

TABLE 3 Descriptive Statistics for first noticing changes.

Skin-tone	Cue type	Mean	SD	N
Light	Real	29.449	9.212	98
	Generated	34.133	11.536	98
Medium Light	Real	34.745	10.112	98
	Generated	37.857	11.182	98
Medium Dark	Real	35.143	11.295	98
	Generated	38.673	12.523	98
Dark Light	Real	47.245	16.935	98
	Generated	51.306	17.637	98
Dark	Real	49.000	17.719	98
	Generated	47.827	16.046	98

noticing changes when compared to the dark-light and dark skin tones, and the medium-dark skin tone condition was rated significantly lower in regards to first noticing changes when compared to the dark-light and dark skin tone conditions (Table 1). Overall these results along with descriptive statistics results suggest that ratings on the first changes slider increased as skin tone darkened (Figure 5).

As for cue type, *post hoc* testing revealed that ratings on the first changes slider were lower for the cues based on real images (mean difference = −2.84, $p < 0.001$) (Table 2). This finding suggests that cues based on real images were noticed at earlier stages of development compared to the computer generated images. Descriptive statistics are shown in Table 3.

4.2.2 Noticing pattern

A Mauchly's test of sphericity was significant for skin tones for this measure, therefore the Huynh-Feldt correction was used. A two-way repeated measures ANOVA showed that noticing pattern ratings differed significantly between skin tone levels, $F(3.84, 372.54) = 176.95$, $p < 0.001$, $\omega^2 = 0.253$. Additionally, noticing pattern ratings differed significantly between cue types, $F(1.00, 97.00) = 96.73$, $p < 0.001$, $\omega^2 = 0.025$.

Post hoc testing using the Holm correction revealed that the light skin tone condition was rated significantly lower in regards to noticing pattern changes when compared to all other skin tones, the medium-light skin tone condition was rated significantly lower when compared to all other skin tones except for the light skin tone, and the medium-dark skin tone condition was rated significantly lower when compared to the dark-light and dark skin tone conditions (Table 4. Overall these results along with descriptive statistics results suggest that ratings on the noticing pattern slider increased as skin tone darkened (Figure 6).

As for cue type, *post hoc* testing revealed that ratings on the noticing pattern slider were lower for the cues based on real images (mean difference = −4.44, $p < 0.001$) (Table 5). This finding suggests that butterfly pattern for the cues based on real images were noticed at earlier stages of development compared to the computer generated images (Table 6).

4.2.3 Realism

A Mauchly's test of sphericity was significant for skin tones for this measure, therefore the Huynh-Feldt correction was used. A two-way repeated measures ANOVA showed that realism ratings differed significantly between skin tone levels, $F(3.73, 361.66) =$

TABLE 4 Post Hoc results for skin tone in regards to noticing pattern.

		Mean difference	SE	t	Cohen's d	p_{holm}
Light	Medium Light	−2.852	1.049	−2.719	−0.177	0.021
	Medium Dark	−5.327	1.049	−5.078	−0.330	< 0.001
	Dark Light	−20.490	1.049	−19.533	−1.268	< 0.001
	Dark	−20.327	1.049	−19.378	−1.258	< 0.001
Medium Light	Medium Dark	−2.474	1.049	−2.359	−0.153	0.038
	Dark Light	−17.638	1.049	−16.814	−1.092	< 0.001
	Dark	−17.474	1.049	−16.659	−1.082	< 0.001
Medium Dark	Dark Light	−15.163	1.049	−14.455	−0.939	< 0.001
	Dark	−15.000	1.049	−14.300	−0.928	< 0.001
Dark Light	Dark	0.163	1.049	0.156	0.010	0.876

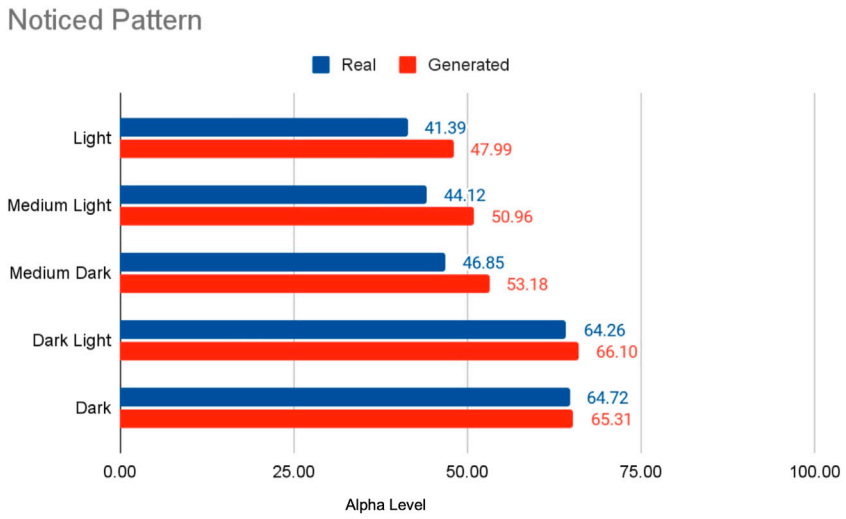

FIGURE 6
Visual depiction of means showing how as skin tone darkened, it took users longer to notice the butterfly pattern that occurs with the visual cue.

4.71, p < = 0.001, $\omega^2 = 0.003$. Additionally, realism ratings differed significantly between cue types, F (1.0, 97.0) = 4.82, p = 0.031, $\omega^2 = 0.002$.

Post hoc testing using the Holm correction revealed that ratings on the realism slider were significantly lower for the dark skin tone condition when compared to the light skin tone condition (mean difference = 2.95, p = 0.002) and the medium-dark skin tone condition (mean difference = 2.70, p = 0.005) (Table 7). As for the cue type, the results suggest that learners viewed the cues based on the real images as less realistic than the computer generated cues (mean difference = −1.47, p = 0.031) (Table 8). Descriptive statistics are shown in Table 9, Figure 7.

4.2.4 Simulation effectiveness tool modified questions (SET-M)

Analyzing the results from the SET-M questions, we focused on the frequencies of student responses rather than a pre-post comparison or contrasting against other conditions, eliminating the need for statistical inference. The data gathered suggests a positive learner reception towards the slider interface intervention. A majority of the students reported feeling more prepared, more understanding of the pathophysiology, and more confident in their assessment and teaching skills upon completing the study (Table 10). This perception of benefit gives a strong indication that the intervention was effective in enhancing student readiness and understanding.

4.2.5 Theory of planned behavior questions

A Wilcoxon's signed-rank test showed that completing the visual cue activity significantly increased ratings for the AttitudeGood (W = 0.00, p < 0.001), NormativeApprove (W = 61.50, p < 0.001), ControlUpToMe (W = 24.00, p < 0.001), and ControlEfficacy (W = 188.00, p < 0.001) questions. These results indicate improved perceptions of feelings towards the perceptions of the behavior by themselves and others (AttitudeGood, NormativeApprove) and increased perceptions of self-efficacy (ControlUpToMe, ControlEfficacy) (Tables 11, 12).

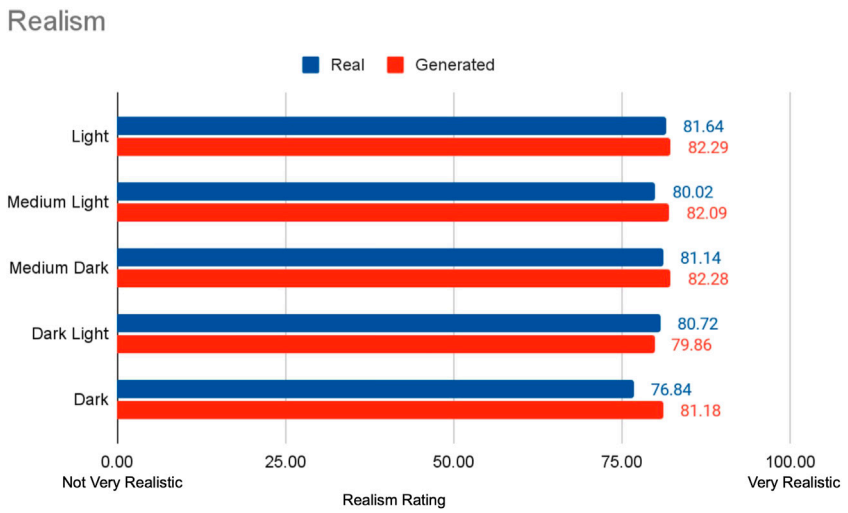

FIGURE 7
Visual depiction of user rating of realism.

TABLE 5 Post Hoc results for cue type in regards to noticing pattern.

		Mean difference	SE	t	Cohen's d	p_{holm}
Real	Generated	−4.441	0.452	−9.835	−0.275	< 0.001

TABLE 6 Descriptive Statistics for noticing pattern.

Skin-tone	Cue type	Mean	SD	N
Light	Real	41.388	16.651	98
	Generated	47.990	15.288	98
Medium Light	Real	44.122	12.764	98
	Generated	50.959	14.542	98
Medium Dark	Real	46.847	14.467	98
	Generated	53.184	16.195	98
Dark Light	Real	64.255	17.890	98
	Generated	66.102	16.970	98
Dark	Real	64.724	18.508	98
	Generated	65.306	17.404	98

4.3 Qualitative

4.3.1 Please describe how the symptom style (computer-generated vs. real symptom image) affected your perception of symptom realism?

There were 89 total responses for this question though not all responses were relevant so are not counted in the following data description. 31 stated that the symptom realism was similar or not affected by the symptom style, five stated that the symptoms appeared more realistic when the AR filter based on the computer-generated image was used, and two stated the symptoms appeared more realistic when the AR filter based on the real image was used. These responses align with the realism results which found that students perceived the computer generated images as marginally more realistic compared to the AR filter based on the real image.

4.3.2 Please describe how the symptom style (computer-generated vs. real symptom image) affected your ability to identify the symptoms?

There were 86 total responses for this question though not all responses were relevant so are not counted in the following data description. 24 stated that their ability to identify the symptoms was similar or not affected by the symptom style, eight stated that it was easier to identify the symptoms when the AR filter based on the computer-generated image was used, and 14 stated it was easier to identify the symptoms when the AR filter based on the real image was used. These responses align with the change and pattern identification results which found that students identified changes and patterns marginally sooner when the AR filter based on the real image was used.

5 Discussion

5.1 Summary of key results

In this work, we found that students rated the symptoms displayed by the computer-generated AR filters as marginally more realistic than the symptoms displayed by the real image AR filters. However, students

TABLE 7 Post Hoc results for cue skin tone in regards to realism.

		Mean difference	SE	t	Cohen's d	p_{holm}
Light	Medium Light	0.909	0.778	1.168	0.050	0.974
	Medium Dark	0.254	0.778	0.327	0.014	0.984
	Dark Light	1.670	0.778	2.147	0.092	0.227
	Dark	2.951	0.778	3.793	0.163	0.002*
Medium Light	Medium Dark	−0.655	0.778	−0.842	−0.036	0.984
	Dark Light	0.762	0.778	0.979	0.042	0.984
	Dark	2.042	0.778	2.625	0.113	0.072
Medium Dark	Dark Light	1.416	0.778	1.821	0.078	0.416
	Dark	2.696	0.778	3.466	0.149	0.005*
Dark Light	Dark	1.280	0.778	1.646	0.071	0.503

TABLE 8 Post Hoc results for cue type in regards to realism.

		Mean difference	SE	t	Cohen's d	p_{holm}
Real	Generated	−1.469	0.669	−2.194	−0.081	0.031

TABLE 9 Descriptive statistics for realism ratings.

Skin-tone	Cue type	Mean	SD	N
Light	Real	81.635	18.436	98
	Generated	82.292	16.533	98
Medium Light	Real	80.019	19.206	98
	Generated	82.090	17.720	98
Medium Dark	Real	81.142	17.618	98
	Generated	82.277	16.406	98
Dark Light	Real	80.721	17.301	98
	Generated	79.864	19.261	98
Dark	Real	76.844	19.601	98
	Generated	81.182	18.274	98

identified symptoms earlier with the real-image filters. Additionally, SET-M and theory of planned behavior questions indicate that the activity increased students feelings of confidence and self-efficacy. Finally, we found that similar to the real world, where symptoms on dark skin tones are identified at later stages of development, students identified symptoms at later stages of development as skin tone darkened regardless of cue type (Schwartz et al., 2003; Hu et al., 2006; Khan and Mian, 2020; Nelson, 2020).

5.2 Developing inclusive AR filters that display visual cues on diverse skin tones

Overall, the results indicate that this method of using AR filters to depict time-based visual cues is effective at improving student self-efficacy and confidence regarding their abilities to identify malar rash symptoms on different skin tones. This is supported by the SET-M results, which indicate that the majority of students found the activity to improve their perceptions of self-efficacy and self-confidence, and the results of the theory of planned behavior questions which indicate that the activity led to improvements in students' perceptions of using skin characteristics while performing a skin assessment and their perceived self-efficacy in their ability to use skin characteristics the next time they perform a skin assessment.

Additionally, the findings indicate that AR filters based on computer-generated images perform similarly to AR filters based on real images. This conclusion is supported by the realism and both identification ratings. The realism ratings reveal that students generally perceived the computer-generated symptoms as slightly more realistic than the real-image symptoms with a mean difference of 1.47 on a 101 point scale. The initial changes and pattern identification ratings show that students noticed the initial changes and butterfly pattern in real-image symptoms marginally sooner than in the computer-generated symptoms with mean differences of 2.84 and 4.44 on 101 point scales.

We believe the differences that are observed between the symptom styles are connected with the salience of symptoms. It appears that the computer-generated images blended better with the skin and had less specular reflections because they were taken using cameras with photographic flashes. This led to a reduced salience of the symptoms, which seemed to render the symptoms more realistic in appearance.

We also found that the real images were identified at earlier stages of development. We believe this was due to the nature of real image AR filters. Unlike the computer-generated images, they do not perfectly blend with the skin and often demonstrate specular reflections from the lighting conditions the source images were

TABLE 10 Frequencies: SET-M questions.

	Do not agree	Somewhat agree	Strongly agree
I am better prepared to respond to changes in my patient's condition	1 (1.02%)	50 (51.02%)	47 (47.96%)
I developed a better understanding of the pathophysiology	18 (18.37%)	42 (42.86%)	38 (38.78%)
I am more confident of my assessment skills	7 (7.14%)	51 (52.04%)	40 (40.82%)
I am more confident in my ability to teach patients about their illness and interventions	17 (17.35%)	50 (51.02%)	31 (31.63%)

TABLE 11 Theory of Planned Behavior: Wilcoxon signed-rank test.

Measure 1		Measure 2	W	p
Pre - AttitudeGood	-	Post - AttitudeGood	0.000	< 0.001
Pre - AttitudeBeneficial	-	Post - AttitudeBeneficial	0.000	0.004
Pre - NormLikeMe	-	Post - NormLikeMe	276.500	0.005
Pre - NormApprove	-	Post - NormApprove	61.500	< 0.001
Pre - ControlUpToMe	-	Post - ControlUpToMe	24.000	< 0.001
Pre - ControlEfficacy	-	Post - ControlEfficacy	188.000	< 0.001
Pre - Intention	-	Post - Intention	25.500	0.005

taken in. These characteristics resulted in filters that did not blend as well and thus, exhibited higher salience.

Both observations underscore the role of symptom salience in the perceived authenticity of images. The blending property of computer-generated images lent a realism to the symptoms, whereas the pronounced salience in real images made the symptoms more readily noticeable at the early stages of the disease.

Overall though, it appears that both methods used to generate the AR filters are effective when used to develop training opportunities that depict visual cues on a range of diverse skin tones, indicating the computer-generated images are a viable alternative to provide source images for AR filters. This contributes to addressing persistent disparities in healthcare education, such as the insufficient diversity and quality of medical images.

However, while these methods have shown promise, ensuring their continued accuracy and effectiveness in representing diverse skin tones is critical. Machine-generated images can lead to bias if an expert is not included in future design processes. For instance, the machine learning model may generate symptoms on minority skin tones in a way that mirrors their depiction on light skin tones, which might not be an accurate representation of the symptoms (Figure 3). To prevent such potential inaccuracies, we recommend maintaining an 'expert-in-the-loop' model for the cue design process. This would require the active participation of a specialist in the development of stimuli, ensuring that the medical images produced accurately and appropriately represent symptoms on different skin tones (Girardi et al., 2015; Guo et al., 2016; Li et al., 2020).

5.3 Design guidelines

Based on our findings, we provide the following design guidelines.

- Generative AI for visual cue training: Generative AI might be a preferable alternative for developing visual cue training using AR filters compared to medical illustrations or real images. While the computer-generated images took learners slightly longer to identify (Sections 4.2.1 and 4.2.2), the learners also perceived the computer-generated images as more realistic (Section 4.2.3). The slight delay in identification time may be an acceptable trade-off for ensuring that training can accommodate a variety of skin tones and adapt to different lighting conditions. Additionally, the use of computer-generated images helps to alleviate potential privacy concerns with overlaying portions of real patient faces onto training stimuli. Overall, we recommend future work to explore generative AI as an alternative for developing visual cue training when real medical imagery is scarce. However, the validity of the generated images should be verified by experts to ensure their suitability for training.
- Iterative Evaluation of AI-generated Images: To avoid potential inconsistencies or inaccuracies in AI-generated images, it is recommended to implement an iterative evaluation process. This process should involve experts in the field who can verify the validity of the images produced. If needed, the machine learning models used can be refined based on their feedback.
- Diversity in Training: AR visual cue training should accurately represent a diverse range of skin tones. The results of this work highlighted that learners took longer to identify the visual cues as skin tone darkened (Sections 4.2.1Sections .1 and 4.2.2). This phenomena exists in the real-world as nurses and physicians diagnose individuals with darker skin tones at a later stages than light skin tone counterparts (Schwartz et al., 2003; Hu et al., 2006; Khan and Mian, 2020; Nelson, 2020). Our solution provides an avenue to provide opportunities for practice that may not otherwise be possible due to a lack of existing resources. We recommend continuing to provide learners with stimuli depicting a variety of skin tones to help increase the number of diverse training opportunities and potentially reduce this disparity gap.
- Need for Better Metrics: While our results suggest that this intervention increased learners' feelings of self-efficacy and confidence (Sections 4.2.4Sections .4 and 4.2.5), it is unclear how this may actually improve their future performance. We recommend future work should aim to improve the metrics used to measure if learners are identifying visual cues. Ideally, future training can identify a learners current level, provide

TABLE 12 Theory of planned behavior frequencies table.

	1	2	3	4	5	6	7
Pre - AttitudeGood	0	0	1 (1.02%)	4 (4.08%)	4 (4.08%)	9 (9.18%)	80 (81.63%)
Post - AttitudeGood	0	0	0	1 (1.02%)	2 (2.04%)	4 (4.08%)	91 (92.86%)
Pre - AttitudeBeneficial	0	0	0	3 (3.06%)	4 (4.08%)	5 (5.10%)	86 (87.76%)
Post - AttitudeBeneficial	0	0	0	1 (1.02%)	2 (2.04%)	3 (3.06%)	92 (93.88%)
Pre - NormLikeMe	0	2 (2.04%)	7 (7.14%)	17 (17.35%)	26 (26.53%)	19 (19.39%)	27 (27.55%)
Post - NormLikeMe	0	3 (3.06%)	4 (4.08%)	10 (10.20%)	25 (25.10%)	16 (16.33%)	40 (40.82%)
Pre - NormApprove	1 (1.02%)	0	1 (1.02%)	11 (11.24%	14 (14.29%)	21 (21.43%)	50 (51.02%)
Post - NormApprove	1 (1.02%)	0	1 (1.02%)	2 (2.04%)	8 (8.16%)	18 (18.37%)	68 (69.39%)
Pre - ControlUpToMe	7 (7.14%)	5 (5.10%)	6 (6.12%)	5 (5.10%)	6 (6.12%)	14 (14.29%)	55 (56.12%)
Post - ControlUpToMe	7 (7.14%)	1 (1.02%)	2 (2.04%)	2 (2.04%)	4 (4.08%)	7 (7.14%)	75 (76.53%)
Pre - ControlEfficacy	0	2 (2.04%)	5 (5.10%)	10 (10.20%)	21 (21.43%)	28 (28.57%)	32 (32.65%)
Post - ControlEfficacy	1 (1.02%)	0	5 (5.10%)	0	16 (16.33%)	21 (21.43%)	55 (56.12%)
Pre - Intention	0	0	0	2 (2.04%)	4 (4.08%)	12 (12.25%)	80 (81.63%)
Post - Intention	0	0	0	0	2 (2.04%)	7 (7.14%)	89 (90.82%)

feedback on potential disparities they may contribute to, and the measure their improvement at identifying symptoms in multipart training simulations.

6 Limitations and future directions

This study faces several limitations, such as not using an AR device, varying screen brightness and color calibrations among users, using transparency as an indicator of symptom progression, and limited symptom types. These issues are discussed briefly:

1) This study does not use an AR device to display filters. This prevents the inclusion of variables like tracking, viewing angle, and device calibration. Despite some persistent screen differences (like brightness and color), our approach, alongside measures to regulate screen brightness, ensures a more controlled design. Future research can apply these findings in simulations to explore how AR display variables and tracking influence user perceptions of symptoms.
2) The alpha level slider, scaling linearly from 0 to 100, is used to approximate symptom development, but symptom progression is not necessarily linear. Symptoms may initially change rapidly before slowly reaching peak severity and can vary among patients, even for the same symptom (Brown, 2003). Future studies should explore more methods to more accurately represent symptom developments, enhancing healthcare education.
3) This study focuses solely on the malar rash due to its distinctiveness and prevalence in various conditions. It is an additive symptom, adding redness to the face, which may differ from subtractive symptoms like paleness, where color is removed. Further research is required to understand the nuances of different symptom types.

7 Conclusion

This study demonstrates the potential of utilizing generative AI in the AR filter design process. Using the evaluation tool developed by Stuart et al. (Stuart et al., 2023), our results indicate that AR filters designed using generative AI can be effective teaching tools for healthcare students to enhance their self-efficacy and confidence in identifying malar rash symptoms across a range of diverse skin tones. The positive outcomes observed in both SET-M scores and theory of planned behavior questions highlight the effectiveness of this approach in improving students' perceptions of their clinical assessment skills. Most importantly, our findings suggest that computer-generated images can be a viable alternative to real images in the development of AR filters, as they were found to be comparable in terms of realism and pattern identification, providing a potential avenue to reduce healthcare disparities. Together, the findings of this research indicate that using generative AI in the AR filter design process is a promising direction to help improve the inclusivity of healthcare training.

Data availability statement

The raw data supporting the conclusion of this article will be made available by the authors, without undue reservation.

Ethics statement

The studies involving humans were approved by University of Florida Internal Review Board. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

JS: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing–original draft, Writing–review and editing. AS: Conceptualization, Project administration, Resources, Supervision, Validation, Writing–original draft, Writing–review and editing. KA: Conceptualization, Project administration, Resources, Supervision, Validation, Writing–original draft, Writing–review and editing. MB: Conceptualization, Project administration, Resources, Supervision, Validation, Writing–original draft, Writing–review and editing. SH: Writing–original draft, Writing–review and editing. BR: Writing–original draft, Writing–review and editing. BL: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing–original draft, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was funded by the National Science Foundation award numbers 1800961 and 1800947.

Acknowledgments

We would like to thank the members of the virtual experiences research group for their assistance throughout the paper writing process.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Ajzen, I. (2006). Constructing a theory of planned behavior questionnaire.

Brown, A. F. (2003). The emergency department epidemiology of acute allergic events: can we ever compare apples with apples? *Emerg. Med. Australasia* 15, 315–317. doi:10.1046/j.1442-2026.2003.00468.x

Buster, K. J., Stevens, E. I., and Elmets, C. A. (2012). Dermatologic health disparities. *Dermatol. Clin.* 30, 53–59. doi:10.1016/j.det.2011.08.002

Ebede, T., and Papier, A. (2006). Disparities in dermatology educational resources. *J. Am. Acad. Dermatology* 55, 687–690. doi:10.1016/j.jaad.2005.10.068

Fribourg, R., Peillard, E., and McDonnell, R. (2021). "Mirror, mirror on my phone: investigating dimensions of self-face perception induced by augmented reality filters," in 2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Bari, Italy, 04-08 October 2021 (IEEE), 470–478.

García-Peñalvo, F., and Vázquez-Ingelmo, A. (2023). What do we mean by GenAI? A systematic mapping of the evolution, trends, and techniques involved in generative AI. *Int. J. Interact. Multimedia Artif. Intell.* 8, 7. doi:10.9781/ijimai.2023.07.006

Generated (2023). *Photos faces*.

Girardi, D., Kueng, J., and Holzinger, A. (2015). A domain-expert centered process model for knowledge discovery in medical research: putting the expert-in-the-loop. *Brain Inf. Health*, 389–398. doi:10.1007/978-3-319-23344-4_38

Gloster, H. M., and Neal, K. (2006). Skin cancer in skin of color. *J. Am. Acad. Dermatology* 55, 741–760. doi:10.1016/j.jaad.2005.08.063

Guo, X., Yu, Q., Li, R., Alm, C. O., Calvelli, C., Shi, P., et al. (2016). An expert-in-the-loop paradigm for learning medical image grouping. *Adv. Knowl. Discov. Data Min.*, 477–488. doi:10.1007/978-3-319-31753-3_38

Harp, T., Militello, M., McCarver, V., Johnson, C., Gray, T., Harrison, T., et al. (2022). Further analysis of skin of color representation in dermatology textbooks used by residents. *J. Am. Acad. Dermatology* 87, e39–e41. doi:10.1016/j.jaad.2022.02.069

Hu, S., Soza-Vento, R. M., Parker, D. F., and Kirsner, R. S. (2006). Comparison of stage at diagnosis of melanoma among hispanic, Black, and white patients in miami-dade county, Florida. *Archives Dermatology* 142, 704–708. doi:10.1001/archderm.142.6.704

Kannuthurai, V., Murray, J., Chen, L., Baker, E. A., and Zickuhr, L. (2021). Health care practitioners' confidence assessing lupus-related rashes in patients of color. *Lupus* 30, 1998–2002. doi:10.1177/09612033211045284

Kaundinya, T., and Kundu, R. V. (2021). Diversity of skin images in medical texts: recommendations for student advocacy in medical education. *J. Med. Educ. Curric. Dev.* 8, 238212052110258. doi:10.1177/23821205211025855

Khan, S., and Mian, A. (2020). Racism and medical education. *Lancet Infect. Dis.* 20, 1009. doi:10.1016/S1473-3099(20)30639-3

Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H., and Ferrante, E. (2020). Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proc. Natl. Acad. Sci.* 117, 12592–12594. doi:10.1073/pnas.1919012117

Lee, V., Sokumbi, O., and Onajin, O. (2023). Collagen vascular diseases. *Dermatol. Clin.* 41, 435–454. doi:10.1016/j.det.2023.02.009

Leighton, K., Ravert, P., Mudra, V., and Macintosh, C. (2015). Updating the simulation effectiveness tool: item modifications and reevaluation of psychometric properties. *Nurs. Educ. Perspect.* 36, 317–323. doi:10.5480/15-1671

Li, G., Mao, R., Hildre, H. P., and Zhang, H. (2020). Visual attention assessment for expert-in-the-loop training in a maritime operation simulator. *IEEE Trans. Industrial Inf.* 16, 522–531. doi:10.1109/TII.2019.2945361

Liang, C.-J., Start, C., Boley, H., Kamat, V. R., Menassa, C. C., and Aebersold, M. (2021). Enhancing stroke assessment simulation experience in clinical training using augmented reality. *Virtual Real.* 25, 575–584. doi:10.1007/s10055-020-00475-1

Ludmann, P. (2022). *Lupus and your skin: signs and symptoms*.

Maluleke, V. H., Thakkar, N., Brooks, T., Weber, E., Darrell, T., Efros, A. A., et al. (2022). Studying bias in GANs through the Lens of race. *arXiv:2209.02836*, 344–360. doi:10.1007/978-3-031-19778-9_20

Narla, S., Heath, C. R., Alexis, A., and Silverberg, J. I. (2022). Racial disparities in dermatology. *Archives Dermatological Res.* 315, 1215–1223. doi:10.1007/s00403-022-02507-z

Nelson, B. (2020). How dermatology is failing melanoma patients with skin of color. *Cancer Cytopathol.* 128, 7–8. doi:10.1002/cncy.22229

Nijhawan, R. I., Jacob, S. E., and Woolery-Lloyd, H. (2008). Skin of color education in dermatology residency programs: does residency training reflect the changing demographics of the United States? *J. Am. Acad. Dermatology* 59, 615–618. doi:10.1016/j.jaad.2008.06.024

Noll, C., von Jan, U., Raap, U., and Albrecht, U.-V. (2017). Mobile augmented reality as a feature for self-oriented, blended learning in medicine: randomized controlled trial. *JMIR mHealth uHealth* 5, e139. doi:10.2196/mhealth.7943

Scenario (2023). *Scenario*.

Schwartz, K. L., Crossley-May, H., Vigneau, F. D., Brown, K., and Banerjee, M. (2003). Race, socioeconomic status and stage at diagnosis for five common malignancies. *Cancer Causes Control* 14, 761–766. doi:10.1023/A:1026321923883

Shorten, C., and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *J. Big Data* 6, 60. doi:10.1186/s40537-019-0197-0

Snap Inc (2021). *Lens Studio*.

Stuart, J., Aul, K., Stephen, A., Bumbach, M. D., and Lok, B. (2022). The effect of virtual human rendering style on user perceptions of visual cues. *Front. Virtual Real.* 3. doi:10.3389/frvir.2022.864676

Stuart, J., Stephen, A., Aul, K., Bumbach, M. D., Huffman, S., Russo, B., et al. (2023). Using augmented reality filters to display time-based visual cues. *Front. Virtual Real.* 4. doi:10.3389/frvir.2023.1127000

Wang, Y., Cao, Y., Zha, Z.-J., Zhang, J., and Xiong, Z. (2020). "Deep degradation prior for low-quality image classification," in Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13-19 June 2020.

Frontiers in Virtual Reality

Explores the possibilities of virtual and extended reality

An exciting new journal in its field which advances our understanding of extended reality to develop new technologies and find applications for society.

Discover the latest Research Topics

See more →

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact