frontiers | Research Topics

Outbreak oracles: how AI's journey through COVID-19 shapes future epidemic strategy

Edited by
Dmytro Chumachenko, Jasleen Kaur and Jake Y. Chen

Published in
Frontiers in Artificial Intelligence
Frontiers in Big Data

ISSN 1664-8714
ISBN 978-2-8325-6513-1
DOI 10.3389/978-2-8325-6513-1

Generative AI statement
Any alternative text (Alt text) provided alongside figures in the articles in this ebook has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

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

Outbreak oracles: how AI's journey through COVID-19 shapes future epidemic strategy

Topic editors

Dmytro Chumachenko — National Aerospace University – Kharkiv Aviation Institute, Ukraine

Jasleen Kaur — National Research Council Canada (NRC), Canada

Jake Y. Chen — University of Alabama at Birmingham, United States

Citation

Chumachenko, D., Kaur, J., Chen, J. Y., eds. (2025). *Outbreak oracles: how AI's journey through COVID-19 shapes future epidemic strategy*.

Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-6513-1

Table of contents

frontiers | Frontiers in Artificial Intelligence

TYPE Editorial
PUBLISHED 13 June 2025
DOI 10.3389/frai.2025.1636444

OPEN ACCESS

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

*CORRESPONDENCE
Dmytro Chumachenko
✉ dichumachenko@gmail.com

RECEIVED 27 May 2025
ACCEPTED 30 May 2025
PUBLISHED 13 June 2025

CITATION
Chumachenko D, Kaur J and Chen JY (2025) Editorial: Outbreak oracles: how AI's journey through COVID-19 shapes future epidemic strategy. *Front. Artif. Intell.* 8:1636444. doi: 10.3389/frai.2025.1636444

Editorial: Outbreak oracles: how AI's journey through COVID-19 shapes future epidemic strategy

Dmytro Chumachenko[1]*, Jasleen Kaur[2] and Jake Y. Chen[3]

[1]Mathematical Modelling and Artificial Intelligence Department, National Aerospace University "Kharkiv Aviation Institute, " Kharkiv, Ukraine, [2]Faculty of Health, University of Waterloo, Waterloo, ON, Canada, [3]Department of Biomedical Informatics and Data Science, The Informatics Institute, University of Alabama at Birmingham, Birmingham, AL, United States

KEYWORDS

artificial intelligence, COVID-19, epidemic modeling, digital surveillance, pandemic preparedness, explainable AI, public health decision-making, data ethics

Editorial on the Research Topic
Outbreak oracles: how AI's journey through COVID-19 shapes future epidemic strategy

In the wake of the coronavirus disease 2019 (COVID-19) pandemic, the convergence of artificial intelligence (AI), Big Data, and *in silico* simulation has emerged as a cornerstone in pandemic surveillance and public health informatics. As early as January 2020, AI-driven tools, such as BlueDot, flagged unusual pneumonia clusters in Wuhan days before the World Health Organization (WHO) public alert (Brownstein et al., 2009). Meanwhile, platforms like CORD-19 were developed to aggregate SARS-CoV-2 literature for rapid machine-driven synthesis (Wang et al., 2020). By mid-2021, AI-based forecasting models were routinely incorporated into national response dashboards, demonstrating that algorithmic surveillance could anticipate hospitalization peaks with a lead time of up to 2 weeks (Institute for Health Metrics and Evaluation, 2019). This transformative period witnessed AI's potential to reshape public health strategies, emphasizing its significance in future epidemic preparedness. However, challenges such as data privacy, algorithmic bias, and unequal access to technological infrastructure persist as obstacles to global adoption. Addressing these limitations is essential to ensure the inclusive and ethical deployment of AI in public health.

The primary objective of this Research Topic is to collate groundbreaking research and critical reviews that highlight AI's contributions during the COVID-19 era and its implications for future epidemic strategies. We aim to foster a comprehensive understanding of the pivotal AI-driven methodologies in the pandemic response and how these innovations can be harnessed for future health crises.

The Research Topic includes original research, technology, code, as well as perspective papers. We received 16 submissions, 9 of which, after a careful review process, were accepted for publication in this Research Topic.

Drawing on Johns Hopkins surveillance data from 34 countries, a study by Nesteruk(a) evaluates how demographic structure and surveillance intensity shaped the late-phase contours. Regression analyses of 2022–2023 COVID-19 incidence, mortality, and case–fatality rates reveal median age as the primary determinant: older populations recorded higher numbers of detected cases and deaths per million. Testing density displayed a strong association with incidence, yet when insufficient, it exaggerated fatality ratios.

Song et al. introduce K-Track-COVID, an R Shiny dashboard that unifies government APIs, WorldPop demographics, and KDCA line lists to render fine-grained maps, regional time series, and SEIRD/SVEIRD scenario forecasts for all 17 South Korean provinces. Interactive controls enable users to filter dates, strata, and health-system indicators, while an embedded stochastic simulator generates animations and tabular projections 100 days ahead. Compared with the WHO, Johns Hopkins University (JHU), and the Centers for Disease Control and Prevention (CDC) dashboards, K-Track-COVID uniquely couples hotspot analytics with modifiable epidemiological models, offering a transferable template for data-driven preparedness beyond COVID-19.

Barmak et al. advance explainable AI by introducing a transition-matrix framework that maps high-dimensional embeddings onto clinician-defined features, thereby rendering model outputs transparent without retraining. The authors formalize the mapping as a pseudo-inverse solution and implement a visual analytics pipeline that scales across modalities. Validation on the Massachusetts Institute of Technology-Beth Israel Hospital Electrocardiogram (MIT-BIH ECG) arrhythmia and ACDC cardiac-MRI datasets yields Cohen's κ of 0.89 and 0.80, respectively, demonstrating strong alignment with expert annotations while preserving diagnostic accuracy.

A study by Oliveira et al. presents a cohort study spanning eight Brazilian hospitals ($n = 421$) to develop low-cost triage models for COVID-19. Using 28 collected admission variables, the authors construct seven data subsets to explore missing data strategies and train 27 classical algorithms, as well as dense neural networks, under Monte Carlo cross-validation. A random forest achieved 80% accuracy (AUC: 0.91), while an SVC yielded 87% PPV with minimal false positives. Dyspnea, general condition, SpO_2, age, and urea emerged as key predictors, underscoring the potential of pragmatic AI deployment in resource-constrained settings.

Nesteruk's(b) study generalizes the classical SIR framework by partitioning infectious and removed compartments into registered and unregistered subgroups and introducing a visibility coefficient β, which yields a five-equation system whose analytical solutions facilitate accelerated parameter estimation. Applying the model to two pertussis waves in England (2023–2024) under full case ascertainment, he reproduces the cumulative and daily incidence curves with a 17% 4-month predictive error.

Melnykova et al. developed an AI pipeline for analyzing and correcting post-traumatic dysarthric speech in military patients. Using the TORGO corpus, they benchmarked a mel-spectrogram CNN against an MFCC-based LSTM, achieving accuracies of 94% and 91%, respectively. The CNN achieves higher precision and recall for healthy and low-severity classes, whereas the LSTM performs better in detecting severe cases. Feature-saliency analysis highlights spectral cues guiding rehabilitation. Ensemble modeling and data augmentation are proposed to generalize across accents and noise, underscoring AI's promise for scalable speech therapy.

A study by Chaikovsky et al. investigates whether subtle quantitative alterations in serial 12-lead electrocardiograms (ECGs) predict outcomes in severe COVID-19. Among 26 intensive care unit (ICU) patients (six of whom died), the authors computed 240 waveforms and HRV metrics, summarized by a composite U-score. They paired them with NEWS and SpO_2 for cluster and CART modeling. T-wave singular value decomposition (SVD), R-wave amplitude (lead II), and Q-wave amplitude (lead I) proved to be the most discriminative, yielding a three-split decision tree that classified survival with 96% cross-validated accuracy, highlighting ECG micrometrics as a practical bedside prognostic tool.

Lyimo et al. surveyed 76 Environmental Health Officers across Morogoro, Ilala, and Dodoma to assess their readiness for ML-based forecasting of waterborne diseases. With a 66% response rate, respondents displayed moderate ICT competence yet limited AI literacy: only 54% had previously encountered AI/ML and 64% rated their familiarity as low. While the majority recognized ML's potential to improve outbreak prediction, they flagged infrastructure gaps, poor data quality, and skill shortages as barriers.

The perspective article by Huaiyan et al. offers a comprehensive view on how digital technologies will redefine infectious disease practice in the aftermath of COVID-19. Surveying bioinformatics, AI, big data analytics, nanotechnology vaccines, blockchain, and telemedicine, the authors map converging trends in prevention, early diagnosis, therapy, and supply–chain governance, while highlighting the ethical imperatives of data sovereignty and algorithmic bias.

In the future, we urge the community to align AI outbreak tools with the Findability, Accessibility, Interoperability, and Reusability (FAIR) Data Principles and the WHO's Preparedness Strategy (2023), ensuring that models are not only performant but interoperable and equitable across low-resource settings (Wilkinson et al., 2016). By building on the lessons of COVID-19, where transparency, open data, and cross-disciplinary collaboration proved critical, the field should coalesce around a shared framework that accelerates innovation and trust in future AI-powered epidemic responses.

Author contributions

DC: Writing – review & editing, Writing – original draft. JK: Writing – original draft, Writing – review & editing. JC: Writing – original draft, Writing – review & editing.

Acknowledgments

We are grateful to all the authors and reviewers contributing to this Research Topic.

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

Brownstein, J. S., Freifeld, C. C., and Madoff, L. C. (2009). Digital disease detection — harnessing the web for public health surveillance. *New Engl. J. Med.* 360, 2153–2157. doi: 10.1056/NEJMp0900702

Institute for Health Metrics and Evaluation (2019). *COVID-19 Projections*. Institute for Health Metrics and Evaluation. Available online at: https://www.healthdata.org/covid (accessed May 23, 2025).

Preparedness Strategy (2023). *A Fragile State of Preparedness: 2023 Report on the State of the World's Preparedness. Gpmb.org*. Available online at: https://www.gpmb.org/annual-reports/overview/item/a-fragile-state-of-preparedness-2023-report-on-the-state-of-the-worlds-preparedness (accessed May 23, 2025).

Wang, L. L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., et al. (2020). CORD-19: the COVID-19 open research dataset. *arXiv:2004.10706*.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Sci. Data* 3:160018. doi: 10.1038/sdata.2016.18

frontiers | Frontiers in Big Data

TYPE Original Research
PUBLISHED 10 January 2024
DOI 10.3389/fdata.2023.1355080

OPEN ACCESS

EDITED BY
Dmytro Chumachenko,
National Aerospace University – Kharkiv Aviation Institute, Ukraine

REVIEWED BY
Sergiy Yakovlev,
Lodz University of Technology, Poland
Viktoriia Alieksieieva,
Technical University of Applied Sciences Wildau, Germany

*CORRESPONDENCE
Igor Nesteruk
✉ inesteruk@yahoo.com

RECEIVED 13 December 2023
ACCEPTED 27 December 2023
PUBLISHED 10 January 2024

CITATION
Nesteruk I (2024) Trends of the COVID-19 dynamics in 2022 and 2023 vs. the population age, testing and vaccination levels.
Front. Big Data 6:1355080.
doi: 10.3389/fdata.2023.1355080

Trends of the COVID-19 dynamics in 2022 and 2023 vs. the population age, testing and vaccination levels

Igor Nesteruk*

Institute of Hydromechanics, National Academy of Sciences of Ukraine, Kyiv, Ukraine

Introduction: The population, governments, and researchers show much less interest in the COVID-19 pandemic. However, many questions still need to be answered: why the much less vaccinated African continent has accumulated 15 times less deaths per capita than Europe? or why in 2023 the global value of the case fatality risk is almost twice higher than in 2022 and the UK figure is four times higher than the global one?

Methods: The averaged daily numbers of cases *DCC* and death *DDC* per million, case fatality risks *DDC/DCC* were calculated for 34 countries and regions with the use of John Hopkins University (JHU) datasets. Possible linear and non-linear correlations with the averaged daily numbers of tests per thousand *DTC*, median age of population A, and percentages of vaccinations *VC* and boosters *BC* were investigated.

Results: Strong correlations between age and *DCC* and *DDC* values were revealed. One-year increment in the median age yielded 39.8 increase in *DCC* values and 0.0799 *DDC* increase in 2022 (in 2023 these figures are 5.8 and 0.0263, respectively). With decreasing of testing level *DTC*, the case fatality risk can increase drastically. *DCC* and *DDC* values increase with increasing the percentages of fully vaccinated people and boosters, which definitely increase for greater A. After removing the influence of age, no correlations between vaccinations and *DCC* and *DDC* values were revealed.

Discussion: The presented analysis demonstrates that age is a pivot factor of visible (registered) part of the COVID-19 pandemic dynamics. Much younger Africa has registered less numbers of cases and death per capita due to many unregistered asymptomatic patients. Of great concern is the fact that COVID-19 mortality in 2023 in the UK is still at least 4 times higher than the global value caused by seasonal flu.

KEYWORDS

COVID-19 pandemic dynamics, daily numbers of cases and deaths per capita, case fatality risk, mathematical modeling of infection diseases, statistical methods

Introduction

In the fourth year of the COVID-19 pandemic, the population and governments show much less interest in it. In particular, only 39% of countries reported at least one case to WHO in the period from 31 July to 27 August 2023.[1] Thus accumulated numbers of cases *CC* and deaths *DC* per million show stabilization trends (COVID-19 Data, 2023) and can be used to estimate the impact of different factors on the pandemic dynamics and to answer some important questions. In particular, why the much less vaccinated African continent has accumulated 36 times less cases and 15 times less deaths per capita than Europe (see text footnote 1, COVID-19 Data, 2023, and lines 26 and 27 in Table 1)? Why in 2023 the

1 https://covid19.who.int/data. Accessed November 23, 2023.

TABLE 1 Median age, accumulated numbers the COVID-19 cases and deaths per capita in 2021–2023.

No, i	Country or region	Median age in years (see text footnotes 13, 14), A_i	Accumulated numbers of confirmed COVID-19 cases per million (COVID-19 Data, 2023)			Accumulated numbers of COVID-19 related deaths per million (COVID-19 Data, 2023)		
			December 31, 2021, $CC_i^{(1)}$	December 31, 2022, $CC_i^{(2)}$	September 10, 2023, $CC_i^{(3)}$	December 31, 2021, $DC_i^{(1)}$	December 31, 2022, $DC_i^{(2)}$	September 10, 2023, $DC_i^{(3)}$
1	USA	38.5	158,249.8	293,865.4	305,763.9	2,421.163	3,199.789	3,331.912
2	Taiwan	42.3	712.707	370,284.7	428,515.6*	35.575	638.377	795.2*
3	Hong Kong	45.6	1,689.041	350,605	389,150.1**	28.442	1,576.608	1,895.5**
4	India	28.7	24,583.31	31,526.41	31,751.74	339.465	374.479	375.414
5	France	41.7	134,773	587,831.2	603,427.6	1,921.267	2,501.554	2,599.316
6	Germany	47.8	841,31.66	446,707.5	461,051.1	1,411.686	1,987.985	2,098.829
7	Brazil	33.2	103,401.9	168,602.6	175,183.5	2,874.028	3,221.972	3,272.712
8	South Korea	43.2	12,259.77	560,818.7	667,207.1	108.558	622.822	693.495
9	Japan	48.6	13,987.61	234,809.8	272,715.7	148.388	463.995	602.606
10	Italy	46.5	101,315.8	426,312.9	440,207.7	2,324.744	3,130.08	3,242.127
11	UK	40.6	199,110	358,390.2	366,026.1	2,619.105	3,201.28	3,396.691
12	Turkey	32.2	110,635.4	199,255.1	199,255.1	962.067	1,188.394	1,188.394
13	Mexico	29.3	31,644.64	57,030.53	60,217.33	2,382.841	2,599.908	2,623.915
14	Peru	29.1	67,181.25	130,816.2	132,711.8	5,949.676	6,407.655	6,504.19
15	Iran	31.7	69,934.03	85,386.89	85,993	1,485.84	1,633.891	1,652.796
16	Indonesia	31.1	15,472.59	24,391.22	24,730.83	523.025	582.981	587.721
17	Canada	41.8	54,674.47	116,834.2	122,158	779.054	1,268.284	1,382.081
18	South Africa	28.8	57,543.97	67,595.88	67,995.81	1,520.372	1,712.495	1,712.946
19	Egypt	24.1	3,466.327	4,644.856	4,649.271	195.756	223.461	223.714
20	Israel	30.4	146,252.2	491,245.2	511,817.9	874.061	1,235.475	1,340.353
21	Nigeria	18.6	1,105.114	1,219.221	1,311.295	13.865	14.437	14.437
22	Australia	37.5	14,004.71	412,017.8	442,814.2	93.325	680.587	900.471
23	New Zealand	37.2	2,650.961	396,786	458,570.6	9.836	449.541	635.259
24	Vietnam	31.9	17,632.27	117,379.7	118,378.2	329.922	439.835	440.039
25	European Union	44.4	119,091.4	397,673.6	408,494.6	2,033.132	2,668.518	2,763.03

(Continued)

10.3389/fdata.2023.1355080

TABLE 1 (Continued)

No, i	Country or region	Median age in years (see text footnotes 13, 14), A_i	Accumulated numbers of confirmed COVID-19 cases per million (COVID-19 Data, 2023)			Accumulated numbers of COVID-19 related deaths per million (COVID-19 Data, 2023)		
			December 31, 2021, $CC_i^{(1)}$	December 31, 2022, $CC_i^{(2)}$	September 10, 2023, $CC_i^{(3)}$	December 31, 2021, $DC_i^{(1)}$	December 31, 2022, $DC_i^{(2)}$	September 10, 2023, $DC_i^{(3)}$
26	Europe	42	116,033.4	325,338.4	334,708.4	2,102.787	2,700.244	2,788.427
27	Africa	18	6,904.119	9,111.662	9,201.963	160.421	181.186	181.545
28	Asia	31	17,928.47	58,153.69	63,664.54	266.155	325.065	345.988
29	North America	35	106,779.6	199,059.8	207,318.4	2,040.378	2,580.998	2,670.316
30	South America	31	91,280.98	152,849.9	157,558.3	2,730.093	3,071.714	3,104.744
31	High income countries	No data	107,262	322,668.1	340,229.5	1,632.993	2,214.069	2,320.501
32	Upper middle income countries	No data	33,103.88	89,499.82	96,734.09	871.032	1,016.661	1,055.513
33	Lower middle income countries	No data	19,114.67	28,098.46	28,385.07	345.868	387.735	390.133
34	The world	30.5	35,792.14	91,468.7	96,645.07	686.398	842.509	872.575

*Values calculated with the use of Worldometers (see text footnotes 10, 11).

**Values calculated with the use of Worldometers (see text footnotes 12, 13).

10.3389/fdata.2023.1355080

global value of the case fatality risk is almost twice higher than in 2022 and the UK figure is four times higher than the global one (Nesteruk, 2023a)?

We will apply the linear and non-linear correlation analysis using accumulated relative characteristics: the numbers of cases and deaths per million (*CC* and *DC*), numbers of fully vaccinated people and boosters per hundred (*VC* and BC), tests per thousand (*TC*) available in files of John Hopkins University (JHU) (COVID-19 Data, 2023).

The impact of various factors on the COVID-19 pandemic dynamics was estimated in many papers. Some examples can be found in Byass (2020), Davies et al. (2020), Distante et al. (2020), Fanelli and Piazza (2020), Hamzah et al. (2020), Ng and Gui (2020), Chintala et al. (2021), Mohammadi et al. (2021), Nesteruk and Rodionov (2021, 2022a,b), Pardhan and Drydakis (2021), Rossman et al. (2021), Statsenko et al. (2021), Nesteruk (2021a, 2022), Nesteruk et al. (2022), and Nesteruk and Keeling (2023). In particular, *CC* values accumulated as of December 23, 2021 in Ukrainian regions and European countries showed no correlations with the size of population, its density, and the urbanization level, while *DC* and $CFR = DC/CC$ values reduce with the increase of the urbanization level in European countries (Nesteruk et al., 2022). The increase of income (Gross Domestic Product per capita) leads to increase in *CC*, *VC*, *BC*, and *TC* values, but *DC* and *CFR* demonstrate opposite trend in European countries (Nesteruk and Rodionov, 2022b).

Many asymptomatic COVID-19 cases (Shang et al., 2022; Schreiber et al., 2023)[2,3,4] can cause a big difference between visible and real pandemic dynamics (Nesteruk, 2021a,b,c). That is why the higher testing level can increase the numbers of registered cases. Sometimes the testing level is too low to reveal all the cases predicted by theory. Probably such situation occurred in Japan in summer 2022 (Nesteruk, 2023b). It was shown in Nesteruk (2022), that the test per case ratio *TC/CC* (or test positivity rate *CC/TC*) is very important characteristic to control the pandemic. Very strong correlation between *CC* and *TC* was revealed for values accumulated before August 1, 2022 in European and African countries (Nesteruk and Rodionov, 2022b). Since the *TC* values have stopped to be updated by JHU in different days of 2022 (Nesteruk and Rodionov, 2022b; COVID-19 Data, 2023), in this study, we will investigate a correlation between the averaged daily numbers of cases *DCC* and tests per capita *DTC*.

The severity of SARS-CoV-2 infection increases for older patients (Statsenko et al., 2021); almost half of the infected children can be asymptomatic (Fowlkes et al., 2022). With the use of the statistical analysis of 2020 datasets it was shown that younger populations have less clinical cases per capita and it was predicted that "without effective control measures, regions with relatively older populations could see disproportionally more cases of COVID-19, particularly in the later stages of an unmitigated epidemic" (Davies et al., 2020). This forecast was confirmed in Nesteruk and Keeling (2023) with the use of *CC* and *DC* datasets for 79 countries and regions including 10 so-called Zero-COVID countries.[5] It was shown that 1-year increment in the median age yields 12,000–18,000 increase in *CC* values and 52–83 increase in *DC* values. In this study, we will investigate correlations between median age *A* and the averaged daily numbers of cases and deaths per capita *DCC* and *DDC*, respectively.

The high numbers of circulating SARS-CoV-2 variants[6,7,8] and re-infected persons[9] (Flacco et al., 2022; Guedes et al., 2023) raise questions about the effectiveness of vaccinations. In particular, many scientists are inclined to think that the pandemic will not be stopped only through vaccination (Lazarus et al., 2022). The non-linear correlations show that *CC* and *DC* values increase with the growth of the vaccination level *VC*, while *CFR* decreases (Nesteruk and Rodionov, 2022b). In this study, we will investigate correlations between *VC* and *BC* values registered in 2022 and 2023 and *DCC*, *DDC*, and *CFR* figures. We will try also to answer the question why the number of cases and death per capita are higher in more vaccinated countries.

Materials and methods

We will use the accumulated numbers of laboratory-confirmed COVID-19 cases CC_i and deaths DC_i per million, accumulated numbers of tests per thousand TC_i, accumulated numbers of fully vaccinated people VC_i and boosters BC_i per hundred for 33 countries (shown in Tables 1, 2) and the world ($i = 1, 2, .., 34$). The have chosen the countries with highest numbers of accumulated cases and death [according to the recent WHO reports (see text footnote 1)], some other countries and regions listed in COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU) (COVID-19 Data, 2023) (version of file updated on September 28, 2023). Since Chinese statistics shows some contradictions (see, e.g., Nesteruk, 2023c or compare JHU files updated on September 28 and March 9, 2023), we have used only figures for Taiwan and Hong Kong. In particular, in Table 1 we show corresponding CC_i and DC_i values from the March-9-version of JHU file (not available on September 28). The CC_i and DC_i values for Taiwan

2 https://edition.cnn.com/2020/11/02/europe/slovakia-mass-coronavirus-test-intl/index.html. Accessed November 23, 2023.

3 https://www.voanews.com/covid-19-pandemic/slovakias-second-round-coronavirus-tests-draws-large-crowds. Accessed November 23, 2023.

4 https://podillyanews.com/2020/12/17/u-shkolah-hmelnytskogo-provely-eksperyment-z-testuvannyam-na-covid-19/. Accessed November 23, 2023.

5 https://en.wikipedia.org/wiki/Zero-COVID. Accessed November 23, 2023.

6 https://www.who.int/activities/tracking-SARS-CoV-2-variants. Accessed November 30, 2023.

7 https://en.wikipedia.org/wiki/Variants_of_SARS-CoV-2. Accessed September 13, 2023.

8 https://www.cdc.gov/coronavirus/2019-ncov/variants/variant-classifications.html. Accessed September 13, 2023.

9 https://coronavirus.health.ny.gov/covid-19-reinfection-data. Accessed September 13, 2023.

and Hong Kong for 2023 we have calculated with the use of Worldometers.[10,11,12,13] We ignore data from Ukraine and Russia to exclude the influence of military operations on the COVID-19 statistics. To take into account the average age of population, we have used the information about the median ages A_i from Earthly Data[14] and Visual Capitalist[15] (see Table 1).

To calculate the averaged daily numbers of cases *DCC* and deaths *DDC* per million in 2022 and 2023 we will use simple formulas:

$$DCC_i^{(j)} = \frac{CC_i^{(j+1)} - CC_i^{(j)}}{T_j}; \quad i = 1, 2, .., 34; \quad j = 1, 2 \qquad (1)$$

$$DDC_i^{(j)} = \frac{DC_i^{(j+1)} - DC_i^{(j)}}{T_j}; \quad i = 1, 2, .., 34; \quad j = 1, 2 \qquad (2)$$

where T_1 =365 and T_2 =252. The *CFR* values corresponding to 2022 and 2023 can be calculated as follows:

$$CFR_i^{(j)} = \frac{DC_i^{(j+1)} - DC_i^{(j)}}{CC_i^{(j+1)} - CC_i^{(j)}}; \quad i = 1, 2, .., 34; \quad j = 1, 2 \qquad (3)$$

The total average CFR_i^* levels (during the entire period of the COVID-19 pandemic) can be calculated for every country and region:

$$CFR_i^* = \frac{DC_i^{(3)}}{CC_i^{(3)}}; \quad i = 1, 2, .., 34 \qquad (4)$$

To estimate the average daily numbers of tests per thousand, we will use the formula:

$$DTC_i^{(j)} = \frac{TC_i^{(2)} - TC_i^{(1)}}{T_i^{(TC)}}; \quad i = 1, 2, .., 34 \qquad (5)$$

Durations of corresponding periods of time $T_i^{(TC)}$ are listed in Table 2. Unfortunately, the testing data is not available for many countries and regions.

We will use the linear regression to calculate the regression coefficients r and the optimal values of parameters a and b for corresponding best fitting straight lines (Draper and Smith, 1998):

$$y = a + bx \qquad (6)$$

where explanatory variables x will be A, DTC, VC, and BC and dependent variables y will be DCC, DDC, CFR, DTC, VC, and BC.

We will use also the F-test for the null hypothesis that says that the proposed linear relationship (6) fits the data sets. The experimental values of the Fisher function can be calculated using the formula:

$$F = \frac{r^2(n-m)}{(1-r^2)(m-1)} \qquad (7)$$

where n is the number of observations (number of countries and regions taken for statistical analysis); m=2 is the number of parameters in the regression equation (Draper and Smith, 1998). The corresponding experimental values F have to be compared with the critical values $F_C(k_1, k_2)$ of the Fisher function at a desired significance or confidence level ($k_1 = m - 1$, $k_2 = n - m$, see, e.g., Appendix[16]). If $F/F_C(k_1, k_2) < 1$, the correlation is not supported by the results of observations. The highest values of $F/F_C(k_1, k_2)$ correspond to the most reliable correlation.

We will use also non-linear regression:

$$y = cx^{\gamma} \qquad (8)$$

which can be reduced to the linear one by introducing new variables (Nesteruk and Rodionov, 2022b):

$$z \equiv \log y = \log c + \gamma w, \quad w \equiv \log x \qquad (9)$$

Unfortunately, the testing data are almost not available in 2023, since the population and governments show much less interest in the COVID-19 pandemic. This fact and scattered dates of fixing the *VC* and *BC* values (see Table 2) complicate a full-fledged statistical analysis.

Results

The results of calculations with the use of Equations (1–5) are listed in Table 3 and shown in Figures 1–3 vs. median age, testing level *DTC*, percentage of fully vaccinated persons $VC_i^{(1)}$ (for 2022) and $VC_i^{(2)}$(for 2023), and numbers of boosters per hundred $BC_i^{(1)}$ (for 2022) and$BC_i^{(2)}$ (for 2023). The averaged daily numbers of cases decreased drastically in 2023 in comparison with corresponding values in 2022 (compare $DCC^{(2)}$ and $DCC^{(1)}$ values in Table 3 or "triangles" and "circles" in Figure 1, the only exception is Nigeria). The global figure of average daily cases has diminished 7.4 times in 2023 (see the last row of Table 3). Turkey has stopped to show new cases in 2023. USA, China, Japan do not report any COVID-19 cases and related deaths since May 15, 2023 (see text footnote 1).

In 2023 the averaged daily numbers of deaths significantly decreased in all countries and regions (compare $DDC^{(2)}$ and $DDC^{(1)}$ values in Table 3 or "triangles" and "circles" in Figure 2), yielding 3.6 times decrease in global *DDC* figures (see the last row of

10 https://www.worldometers.info/coronavirus/country/taiwan/. Accessed September 30, 2023.

11 https://www.worldometers.info/world-population/taiwan-population/. Accessed September 30, 2023.

12 https://www.worldometers.info/world-population/china-hong-kong-sar-population/. Accessed September 30, 2023.

13 https://www.worldometers.info/coronavirus/country/china-hong-kong-sar/. Accessed September 30, 2023.

14 https://database.earth/population. Accessed September 30, 2023.

15 https://www.visualcapitalist.com/mapped-the-median-age-of-every-continent/. Accessed September 30, 2023.

16 https://onlinepubs.trb.org/onlinepubs/nchrp/cd-22/manual/v2appendixc.pdf. Accessed November 23, 2023.

TABLE 2 Accumulated numbers the tests per capita and percentage of fully vaccinated people and boosters in 2021-2023 (COVID-19 Data, 2023).

No, i	Country or region	Accumulated numbers of tests per thousand $TC_i^{(j)}$, corresponding dates and number of days $T_i^{(TC)}$					Numbers of fully vaccinated people per hundred as of:		Numbers of boosters per hundred as of:	
		$TC_i^{(1)}$	Date in 2021	$TC_i^{(2)}$	Date in 2022	$T_i^{(TC)}$, days	July 1. 2022 $VC_i^{(1)}$	July 1. 2023 $VC_i^{(2)}$	July 1. 2022 $BC_i^{(1)}$	July 1. 2023 $BC_i^{(2)}$
1	USA	2,155.384	Dec 31	2,708.533	Jun 18	169	67.33	69.47[23]	37.91	40.08[8]
2	Taiwan	208.248	Dec 31	545.749	Jun 22	173	81.52	87	72.89	106.4
3	Hong Kong	4,615.889	Dec 31	6,594.993	May 24	144	86.21	90.81	60.36	94.91
4	India	481.597	Dec 31	609.938	Jun 21	172	64.58	67.17	3.15	16.04
5	France	2,908.831	Dec 31	4,126.754	Jun 18	169	78.14	78.44	58.86	70.34
6	Germany	1,108.669	Dec 26	1,574.021	Jun 12	168	76.04	76.24[20]	68.6	77.72[20]
7	Brazil	308.411	Dec 31	330.912	March 11	70	78.52	81.82[17]	49.7	58.7[17]
8	South Korea	874.47	Dec 31	1,934.578	Jun 15	166	85.48	85.64[25]	73.01	79.76[14]
9	Japan	224.641	Dec 31	429.37	Jun 22	173	82.6	83.4[22]	64.04	141.72[22]
10	Italy	2,365.578	Dec 31	3,795.998	Jun 22	173	81.14	81.22	69.66	80.88
11	UK	5,845.841	Dec 31	7,480.121	May 19	139	74.36	75.19[10]	59.26	59.81[9]
12	Turkey	1,403.53	Dec 31	1,924.668	May 31	151	62.21[6]	62.31[13]	43.22[6]	48.54[13]
13	Mexico	95.052	Dec 31	122.879	Jun 18	169	62.7[3]	64.19[11]	41.65[3]	44.73[12]
14	Peru	646.299	Dec 31	859.283	April 05	95	81.44	84.21	61.41	90.41
15	Iran	477.494	Dec 31	594.485	Jun 01	152	65.34[2]	66.15[29]	31.11[2]	32.27[29]
16	Indonesia	155.199	Dec 31	217.363	March 21	80	61.18[30]	63.48[27]	17.86[30]	24.99[27]
17	Canada	1,380.218	Dec 31	1,629.606	Jun 06	157	81.79	82.6[15]	57.41	79.14[15]
18	South Africa	357.471	Dec 31	431.667	Jun 22	173	31.83	35.13[26]	5.79	7.36[26]
19	Egypt	No data		109.415	May 01		33.6[5]	38.15[24]	5.03[5]	13.71[24]
20	Israel	3,637.259	Dec 31	5,573.137	Jun 22	173	65.09	65.19[21]	56.45	61.03[21]
21	Nigeria	17.916	Dec 26	24.74	Jun 22	178	9.6[4]	31.94[16]	0.54[4]	5.63[16]
22	Australia	2,120.241	Dec 31	2,830.525	Jun 22	173	82.7[3]	82.7[18]	62.19	75.6[15]
23	New Zealand	1,087.221	Dec 31	1,416.09	Jun 23	174	79.34	80.66[19]	52.95	68.81[19]
24	Vietnam	765.759	Dec 30	880.538	Jun 20	172	81.67[7]	87.55[28]	57.03[1]	59.05[28]

(Continued)

10.3389/fdata.2023.1355080

TABLE 2 (Continued)

No, i	Country or region	Accumulated numbers of tests per thousand $TC_i^{(j)}$, corresponding dates and number of days $T_i^{(TC)}$					Numbers of fully vaccinated people per hundred as of:		Numbers of boosters per hundred as of:	
		$TC_i^{(1)}$	Date in 2021	$TC_i^{(2)}$	Date in 2022	$T_i^{(TC)}$, days	July 1. 2022 $VC_i^{(1)}$	July 1. 2023 $VC_i^{(2)}$	July 1. 2022 $BC_i^{(1)}$	July 1. 2023 $BC_i^{(2)}$
25	European Union	No data		No data			72.56	72.86	52.15	62.09
26	Europe	No data		No data			65.26	66.21	40.44	48.22
27	Africa	No data		No data			17.89	31.36	2.1	6.4
28	Asia	No data		No data			70.16	73.22	28.35	38.23
29	North America	No data		No data			63.72	65.7	37.23	41.9
30	South America	No data		No data			74.75	77.12	47.5	58.31
31	High income countries	No data		No data			73.18	74.3	51.83	66.37
32	Upper middle income countries	No data		No data			77.15	78.74	45.55	49.97
33	Lower middle income countries	No data		No data			53.28	59.36	9.17	19.35
34	The world	No data		No data			60.07	64.66	26.58	34.93

Figures corresponding to different days in 2022:
May: [1]—18.
June: [2]—1; [3]—17; [4]—19; [5]—29; [6]—30; [30]—21.
July: [7]—7.
September: [8]—1; [9]—4; [10]—11.
October: [11]—7.
November: [12]—18; [13]—22.
December: [14]—12.
Figures corresponding to different days in 2023:
February: [15]—2.
March: [16]—19; [17]—22; [18]—24.
April: [19]—4; [20]—7.
May: [21]—4; [22]—7; [23]—9; [24]—21; [25]—26.
June: [26]—4; [27]—6; [28]—30.
July: [29]—4.

10.3389/fdata.2023.1355080

TABLE 3 The results of calculations of average daily characteristics with the use of Equations (1–4).

No, i	Country or region	Average daily numbers of COVID-19 cases per million, $DCC^{(j)}$		Average daily numbers of deaths per million, $DDC^{(j)}$		Case fatality risks			Average daily numbers of tests per thousand, DTC_i
		2022	2023	2022	2023	2022, $CFR^{(1)}$	2023, $CFR^{(2)}$	Total, CFR^*, Equation (4)	
1	USA	371.55	47.22	2.1332	0.5243	0.00574	0.0111	0.0109	3.27
2	Taiwan	1,012.53	231.08	1.6515	0.6223	0.00163	0.00269	0.00186	1.95
3	Hong Kong	955.93	152.96	4.2416	1.2654	0.00444	0.00827	0.00487	13.74
4	India	19.02	0.89	0.0959	0.00371	0.00504	0.00415	0.0118	0.75
5	France	1,241.26	61.89	1.5898	0.3879	0.00128	0.00627	0.00431	7.21
6	Germany	993.35	56.92	1.5789	0.4399	0.00159	0.00772	0.00455	2.77
7	Brazil	178.63	26.11	0.9533	0.2013	0.00534	0.00771	0.0187	0.32
8	South Korea	1,502.90	422.18	1.4089	0.2804	0.000937	0.000664	0.00104	6.39
9	Japan	604.99	150.42	0.8647	0.5500	0.00142	0.00365	0.00221	1.18
10	Italy	890.40	55.14	2.2064	0.4446	0.00248	0.00806	0.00736	8.27
11	UK	436.38	30.30	1.5950	0.7754	0.00366	0.02559	0.00928	11.76
12	Turkey	242.79	0	0.6201	0	0.00255	–	0.00596	3.45
13	Mexico	69.55	12.65	0.5947	0.09527	0.00855	0.00753	0.0436	0.16
14	Peru	174.34	7.52	1.2547	0.3831	0.00720	0.05093	0.0490	2.24
15	Iran	42.34	2.41	0.4056	0.07502	0.00958	0.03119	0.0192	0.77
16	Indonesia	24.43	1.35	0.1643	0.01881	0.00672	0.01396	0.0238	0.78
17	Canada	170.30	21.13	1.3404	0.4516	0.00787	0.02138	0.0113	1.59
18	South Africa	27.53	1.59	0.5264	0.00179	0.01911	0.00113	0.0252	0.43
19	Egypt	3.22	0.0175	0.0759	0.001004	0.02351	0.05730	0.0481	–
20	Israel	945.19	81.64	0.9902	0.4161	0.00105	0.00510	0.00262	11.19
21	Nigeria	0.31	0.365	0.001567	0	0.00501	0	0.0110	0.0383
22	Australia	1,090.44	122.21	1.6089	0.8726	0.00148	0.00713	0.00203	4.11
23	New Zealand	1,079.82	245.18	1.2047	0.7370	0.00112	0.00301	0.00139	1.89
24	Vietnam	273.28	3.96	0.3011	0.0008095	0.00110	0.000204	0.00371	0.67
25	European Union	763.24	42.94	1.7408	0.3750	0.00228	0.00873	0.00676	–

(Continued)

10.3389/fdata.2023.1355080

TABLE 3 (Continued)

No, i	Country or region	Average daily numbers of COVID-19 cases per million, $DCC^{(j)}$		Average daily numbers of deaths per million, $DDC^{(j)}$		Case fatality risks			Average daily numbers of tests per thousand, DTC_i
		2022	2023	2022	2023	2022, $CFR^{(1)}$	2023, $CFR^{(2)}$	Total, CFR^*_i, Equation (4)	
26	Europe	573.43	37.18	1.6369	0.3499	0.00285	0.00941	0.00833	–
27	Africa	6.05	0.358	0.05689	0.001425	0.00941	0.00398	0.0197	–
28	Asia	110.21	21.86	0.1613	0.08303	0.00146	0.00380	0.00543	–
29	North America	252.82	32.77	1.4811	0.3544	0.00586	0.01082	0.0129	–
30	South America	168.68	18.68	0.9360	0.1311	0.00555	0.00702	0.0197	–
31	High income countries	590.15	69.68	1.5920	0.4223	0.00270	0.00606	0.00682	–
32	Upper middle income countries	154.51	28.71	0.3990	0.1541	0.00258	0.00537	0.0109	–
33	Lower middle income countries	24.61	1.14	0.1147	0.009516	0.00466	0.00837	0.0137	–
34	The world	152.54	20.54	0.4277	0.1193	0.00280	0.00581	0.00903	–

Table 3). Seasonal global influenza mortality is between 294 and 518 thousand in the period from 2002 to 2011 (Paget et al., 2019). After dividing the presented figures over the world population 8,060.5 million (see text footnote 14) and 365 days, the corresponding averaged daily number of deaths per million $DDC_{(infl)}$ will range between 0.1 and 0.18. The global value of $DDC^{(2)}_{34} = 0.1193$ is comparable with the influenza mortality, but in 2023 in many countries (including the UK) the corresponding $DDC^{(2)}$ values are much higher than $DDC_{(infl)}$ (see Table 3).

The global case fatality risk in 2023 is approximately twice higher than in 2022 despite of the increase in percentages of fully vaccinated people and boosters (see the last rows of Tables 2, 3). In 2023 the *CFR* values were lower only in India, South Korea, Mexico, South Africa, Nigeria, Vietnam and Africa (compare corresponding columns in Table 3). There are countries with drastic growth of the *CFR* values in 2023 in comparison with 2022 (for example, almost seven times for the UK and Peru). In 2023 only Egypt, Peru and Iran have higher case fatality risks than in the UK (see Table 3; the markers corresponding to the UK are placed inside red circles in Figures 1–3).

There are countries with traditional high levels of *CFR*. To smooth temporarily fluctuations, the total average CFR^*_i values (during the entire period of the COVID-19 pandemic) were calculated with the use of Equation (4), listed in Table 3 and shown in Figure 3 by blue "stars." The value CFR^*_{11} corresponding to the UK is only slightly higher than the global one. Many countries (e.g., the US, India, Mexico, Peru, Iran, Indonesia, Canada, South Africa, Egypt, Nigeria) have higher CFR^*_i values.

To remove the influence of the seasonal factors in the UK, let us calculate the values of $DCC^{(1*)}_{11}$, $DDC^{(1*)}_{11}$ and $CFR^{(1*)}_{11}$ for the period January 1, 2022–May 19, 2022 and the same values $DCC^{(2*)}_{11}$, $DDC^{(2*)}_{11}$ and $CFR^{(2*)}_{11}$ and for the period January 1, 2023–May 19, 2023 with the use of Equations (1–3) and JHU datasets:

$$DCC^{(1*)}_{11} = (329{,}252.5 - 199{,}110)/139 = 936.28;$$
$$DDC^{(1*)}_{11} = (2{,}944.618 - 2{,}619.105)/139 = 2.34;$$
$$CFR^{(1*)}_{11} = (2{,}944.618 - 2{,}619.105)/(329{,}252.5 - 199{,}110) = 0.0025;$$
$$DCC^{(2*)}_{11} = (364{,}587.9 - 358{,}390.2)/139 = 44.59;$$
$$DDC^{(2*)}_{11} = (3{,}370.043 - 3{,}201.28)/139 = 1.214;$$
$$CFR^{(2*)}_{11} = (3{,}370.043 - 3{,}201.28)/(364{,}587.9 - 358{,}390.2) = 0.0272.$$

In 2023 we can see huge decrease in the number of cases and moderate diminishing in the number of death. As the result, the case fatality risk in the beginning of 2023 exceeded the 2022 figure around 11 times. The reason could be explained by the fact that testing and reporting most cases we stopped in UK since April 2022. The only people who get recorded as cases are those that are tested in hospital, and even there not everyone with respiratory infections gets tested. The death data may include all those where COVID-19 is listed on the death certificate, which might even include those that have not tested positive but where the doctors suspect COVID-19.

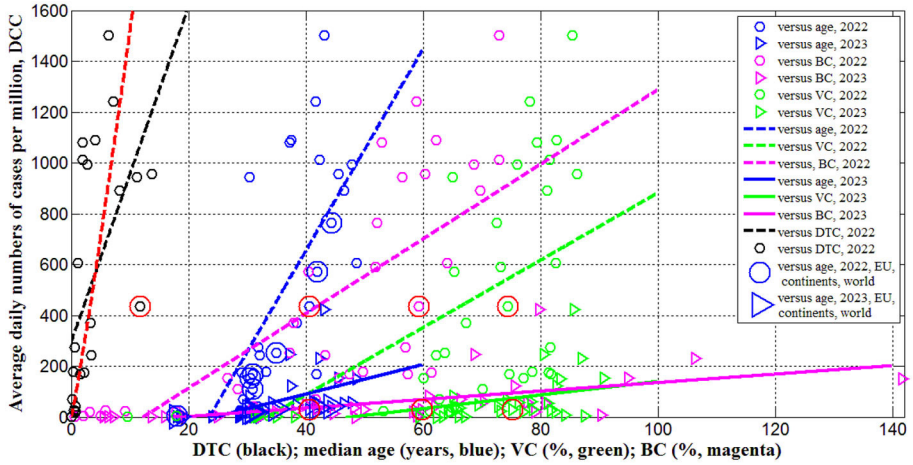

FIGURE 1
Averaged daily numbers of COVID-19 cases per million in 2022 ("circles") and 2023 ("triangles") vs. median age (blue) and levels of vaccinations (green), boosters (magenta) and testing (black). Best fitting lines are solid for 2023 and dashed for 2022. The UK data are located in red circles. The values corresponding to EU, continents and the world are duplicated by lager markers. Red curve represents the results of non-linear correlation (Equation 10).

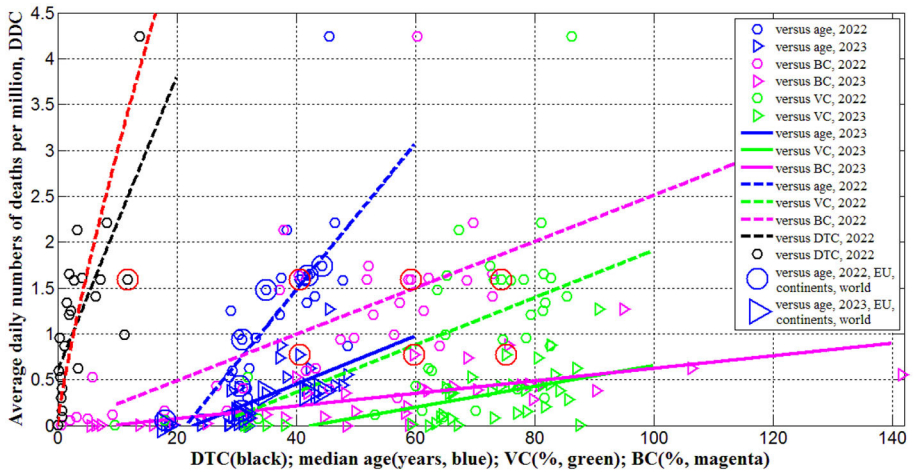

FIGURE 2
Averaged daily numbers of COVID-19 related deaths per million in 2022 ("circles") and 2023 ("triangles") vs. median age (blue) and levels of vaccinations (green), boosters (magenta) and testing (black). Best fitting lines are solid for 2023 and dashed for 2022. The UK data are located in red circles. The values corresponding to EU, continents and the world are duplicated by lager markers. Red curve represents the results of non-linear correlation (Equation 11).

It must be noted, the growth of *CFR* in the UK occurred in the period of increasing the vaccination and booster levels (COVID-19 Data, 2023):

	VC	BC
December 31, 2021	70.26%	50.5%
May 19, 2022	73.92%	58.67%
September 11, 2022	79.71%	75.19%

To investigate possible correlations between *DCC, DDC* and *CFR* values and explanatory variables *A, DTC*, $VC^{(1)}$, $VC^{(2)}$, $BC^{(1)}$, and $BC^{(2)}$, the linear regression (Equation 6) and Fisher test were used. The results of calculations of optimal values of parameters *a* and *b*, correlation coefficients and experimental values of the Fisher function *F* (Equation 7) are listed in Table 4. The *F* values were compared with the critical ones $F_C(1, n-2)$ at the confidence level 0.01. The numbers of observations *n* are different for different correlations due to the absence of data for some countries and regions.

Since many COVID-19 patients are asymptomatic (see text footnotes 2–4) (Nesteruk, 2021b,c, 2023b; Fowlkes et al., 2022), the high testing level (*DTC* or *TC*) could help to reveal more cases and COVID-19 related deaths. This trend was supported statistically (see rows 4 and 11 in Table 4 and black lines in Figures 1, 2). Nevertheless, the linear regression yields unacceptable non-zero values of parameter *a*, which mean that some cases and deaths could be revealed at zero testing level. To remove this discrepancy,

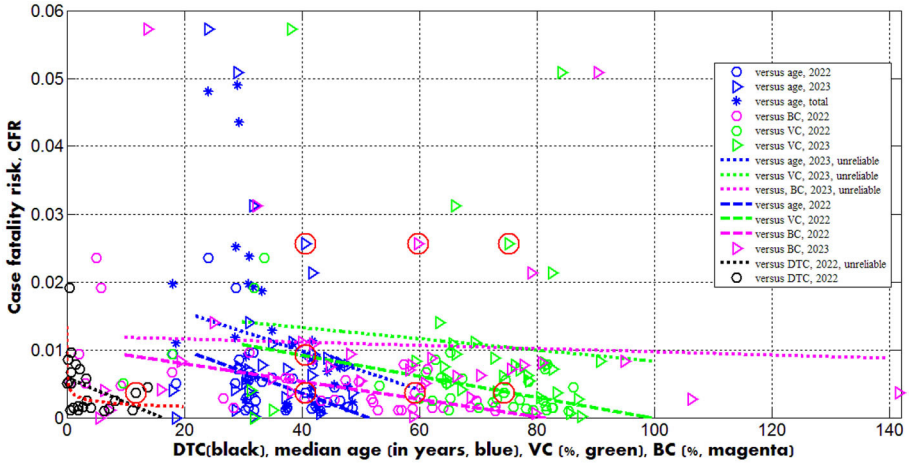

FIGURE 3
Case fatality risks in 2022 ("circles") and 2023 ("triangles") vs. median age (blue) and levels of vaccinations (green), boosters (magenta) and testing (black). Best fitting lines are solid for 2023 and dashed for 2022. The dotted lines correspond to the correlations that are not supported at the significance level 0.01. The UK data are located in red circles. Red curve represents the results of non-linear correlation (Equation 12).

the non-linear approach (Equations 8, 9) was applied for the same countries listed in Table 3 ($n = 23$).

Equations (10–12) represent the best fitting curves (see red lines in Figures 1–3), correlation coefficients and experimental values of the Fisher function:

$$DCC^{(1)} = 118.7499 \cdot \mathrm{DTC}^{1.106479}; r = 0.81835; F = 42.58 \quad (10)$$

$$DDC^{(1)} = 0.4428336 \cdot \mathrm{DTC}^{0.828329}; r = 0.76624; F = 29.86 \quad (11)$$

$$CFR^{(1)} = \frac{0.00372913}{\mathrm{DTC}^{0.27815}}; r = -0.46942; \ F = 5.94 \quad (12)$$

The values of r^2 and F for variables z and w are higher than for y and x [compare corresponding values in Equations (10–12) and rows 4, 11, and 18 in Table 4]. The relationships (10) and (11) are supported at significance level 0.001 $F_C(1, 21) = 14.6[F_C(1, 21) = 14.6]$. The similar very strong correlation between the numbers of cases and tests per capita accumulated in European and African countries as of August 1, 2022 was found in Nesteruk and Rodionov (2022b):

$$CC = 80.099 \cdot \mathrm{TC}^{1.02755}; n = 89; r = 0.9496; F = 767.6 \quad (13)$$

Nevertheless, 16 European countries with the highest testing level ($TC > 3{,}000$) have demonstrated no correlation between CC and TC even at the significance level 0.05 (Nesteruk and Rodionov, 2022b).

Only 5 countries and territories listed in Table 2 (Hong Kong, France, Italy, the UK, and Israel) had TC values higher than 3,000 in 2022. The $DCC^{(1)}$ values are rather high and vary from 436 to 1,241 in these countries (see Table 3). Nevertheless, many infectious persons were not detected. This is evidenced not only by the higher numbers of cases per capita in South Korea ($DCC_8^{(1)} = 1{,}502.9$, Table 3) at lower testing level (compare corresponding DTC values in Table 3), but also by the results of total testing in some countries and institutions, which revealed many previously unregistered COVID-19 patients (see text footnotes 2–4). Taking the maximum $DCC_8^{(j)}$ values (corresponding to South Korea) as estimations real number of cases per capita in 2022 and 2023, we can calculate the visibility coefficients

$$\beta_i^{(j)} = \frac{\mathrm{DCC}_8^{(\mathrm{j})}}{\mathrm{DCC}_\mathrm{i}^{(\mathrm{j})}}; \ j = 1, 2; \ i = 1, 2, ..., 34 \quad (14)$$

as the ratios of real and registered numbers of cases [similar relationship can be obtained using the accumulated numbers of cases per capita (Nesteruk and Rodionov, 2022b)]. For example, figures corresponding to the UK are $\beta_{11}^{(1)} = 3.4$; $\beta_{11}^{(2)} = 13.9$; Europe—$\beta_{26}^{(1)} = 2.6$; $\beta_{26}^{(2)} = 11.4$, and Africa—$\beta_{27}^{(1)} = 248.4$; $\beta_{27}^{(2)} = 1{,}179.3$.

An experimental estimation of the visibility coefficient can be obtained from the results of total testing in Slovakia [89.5% of population was tested on October 31–November 7, 2020 and a number of previously undetected cases, equal to about 1.63% of the population was revealed (see text footnotes 2, 3)]. Since the number of detected cases in Slovakia was approximately 1% of population (COVID-19 Data, 2023), we can estimate the visibility coefficient $\beta \approx 2.63$ for that period. As of September 10, 2023 the ratio of CC values for South Korea and Slovakia [667,207.1/330,868.413 (COVID-19 Data, 2023)] yields the visibility coefficient 2.02. The results of a random testing in two kindergartens and two schools in Chmelnytskii (Ukraine) revealed the value of visibility coefficient 3.9 in December 2020 (see text footnote 4).

The generalized SIR models and algorithms of their parameter identification (Nesteruk, 2021b,d, 2023b) allowed theoretical estimating of the visibility coefficients. In particular, values from 3.7 to 20.4 were obtained for Ukraine (Nesteruk, 2021a,b) and 5.4 for Qatar (Nesteruk, 2021c) in different periods of the COVID-19 pandemic. The lack of appropriate testing did not allowed detecting the first SARS-CoV-2 cases, which probably appeared long before December 2019 (Weinberger et al., 2020). In particular, theoretical

TABLE 4 Optimal values of parameters in Equation (1), correlation coefficients and the results of Fisher test applications.

Number	Variable x	Number of observations n	Correlation coefficient R	Optimal values of parameter a in Equation (6)	Optimal values of parameter b in Equation (6)	Experimental value of the Fisher function F, Equation (7), $m = 2$	Critical value of Fisher function $F_c(1,n\text{-}2)$ for the confidence level 0.01 (Appendix[16])	F/F_c
Correlations for the averaged daily numbers of COVID-19 cases in 2022, $DCC^{(1)}$								
1	Versus age	31	0.7108	−937.43	39.767	29.616	7.77	3.81
2	Versus $VC^{(1)}$	34	0.56202	−444.10	13.252	14.775	7.74	1.91
3	Versus $BC^{(1)}$	34	0.7459	−181.28	14.687	40.136	7.74	5.19
4	Versus DTC	23	0.5613	292.96	66.052	9.66	7.85	1.2
Correlations for the averaged daily numbers of COVID-19 cases in 2023, $DCC^{(2)}$								
5	Versus age	31	0.5009	−143.583	5.8231	9.711	7.77	1.25
6	Versus $VC^{(2)}$	34	0.4696	−128.811	2.6796	9.051	7.74	1.17
7	Versus $BC^{(2)}$	34	0.5674	−32.3957	1.6692	15.195	7.74	1.96
Correlations for the averaged daily numbers of deaths in 2022, $DDC^{(1)}$								
8	Versus age	31	0.7324	−1.7244	0.07993	33.553	7.77	4.32
9	Versus $VC^{(1)}$	34	0.5553	−0.6756	0.02582	14.264	7.74	1.84
10	Versus $BC^{(1)}$	34	0.6513	−0.02176	0.02529	23.579	7.74	3.05
11	Versus DTC	23	0.7138	0.59477	0.16076	21.8	7.85	2.8
Correlations for the averaged daily numbers of deaths in 2023, $DDC^{(2)}$								
12	Versus age	31	0.6793	−0.6040	0.02626	24.843	7.77	3.20
13	Versus $VC^{(2)}$	34	0.5927	−0.4892	0.0114	17.332	7.74	2.24
14	Versus $BC^{(2)}$	34	0.6903	−0.06489	0.006845	29.131	7.74	3.76
Correlations for the case fatality risks in 2022, $CFR^{(1)}$								
15	Versus age	31	−0.5006	0.01631	−0.0003177	9.698	7.77	1.25
16	Versus $VC^{(1)}$	34	−0.5901	0.01538	−0.0001553	17.092	7.74	2.21
17	Versus $BC^{(1)}$	34	−0.5937	0.01053	−0.0001305	17.423	7.74	2.25
18	Versus DTC	23	−0.3566	0.005909	−0.00036525	3.059	7.85	0.39
Correlations for the case fatality risks in 2023, $CFR^{(2)}$								
19	Versus age	30	−0.1700	0.02122	−0.00028515	0.834	7.78	0.11
20	Versus $VC^{(2)}$	33	−0.1008	0.01656	−8.2800e−05	0.318	7.75	0.041
21	Versus $BC^{(2)}$	33	−0.0560	0.01203	−2.36499e−05	0.0975	7.75	0.013

10.3389/fdata.2023.1355080

estimates give the date of the appearance of the first case at the beginning of August 2019 (Nesteruk, 2021d).

Dependence (12) can be accepted at significance level 0.05 [$F_C(1, 21) = 4.43$; a similar equation can be obtained by dividing (11) over (10)] and shows that the case fatality risk increases with diminishing of the testing level even in the period of the high interest in the SARS-CoV-2 infection (as it was in 2022). In 2023, when the people paid attention to severe cases only and make tests correspondingly, *CFR* values can increase drastically. Therefore, one should probably not be afraid of a significant increase of the case fatality risk in the UK in 2023. Of much greater concern is the fact that COVID-19 mortality in this country ($DDC_{11}^{(2)} = 0.7754$, see Table 3) is still at least 4 times higher than the global value caused by seasonal flu (Paget et al., 2019).

Equations (10, 13) may give the illusion that the low number of cases per capita in Africa is due only to the low testing level typical for low-income countries (see Nesteruk and Rodionov, 2022b). The visibility coefficients and another characteristic—the ratio of the number of tests to the number of cases *DTS*—will allow us to understand the situation and draw the right conclusions. High *DTS* values mean that many persons surrounding the detected infectious patient (e.g., family members, colleagues, neighbors) were tested and isolated (this causes a decrease in the number of new infections, i.e., *DCC*). For example, very high tests per case ratios ($DTS > 100$) in Hong Kong in 2020 and 2021 allowed controlling the COVID-19 epidemic completely (Nesteruk, 2022) [the smoothed daily numbers of new cases per million did not exceed 20 (COVID-19 Data, 2023)]. After January 18, 2022, the daily numbers of new cases started to increase, but the daily numbers of tests remained almost constant yielding drastically diminishing of the daily tests per case ratio (Nesteruk, 2022) and very high *DCC* values in February-March 2022 (COVID-19 Data, 2023).

It follows from Equation (10) that the averaged daily test per case ratio:

$$DTS \equiv \frac{1000 \cdot DTC}{DCC^{(1)}} = \frac{13.34}{\left[DCC^{(1)}\right]^{0.0962}}$$

increases for countries with low *DCC* figures (in particular, for African ones, see Table 3). The similar relationship follows from Equation (13) for the accumulated characteristic $TS = 1{,}000^{*}TC/CC$. For example, *DTS* values (calculated using the information available in Table 3) are equal to 26.9 (the UK); 39.4 (India); 123.5 (Nigeria); 4.3 (South Korea); 1.95 (Japan) and demonstrate that the probability to miss an infectious person due to the lack of tests is much higher in Japan or South Korea than in Nigeria or India. During the severe pandemic wave in Japan in summer 2022, the daily numbers tests probably were not enough to confirm COVID-19 in patients with symptoms (Nesteruk, 2023b).

Therefore, the reason for the low number of registered cases per capita in Africa or in India should not be found in insufficient testing, but in large values of the visibility coefficients (Equation 14), which attribute to large numbers of asymptomatic infections. Since the severity of SARS-CoV-2 infection increases for older patients (Davies et al., 2020; Statsenko et al., 2021) and almost half of the infected children can be asymptomatic (Fowlkes et al., 2022), the regions with older population are expected to have much higher accumulated numbers of cases per capita (Davies et al., 2020). It was shown that 1-year increment in the median age yields 12,000–18,000 increase in *CC* values (Nesteruk and Keeling, 2023). Rows 1 and 5 in Table 4 and blue lines in Figure 1 illustrate the same trend for *DCC* values. One-year increment in the median age increased *DCC* values by 39.8 in 2022 and by 5.8 in 2023.

The stronger correlations and same trends were obtained for the averaged daily numbers of deaths per capita *DDC* vs. median age of population *A* (see rows 8 and 12 in Table 4 and blue lines in Figure 2). One-year increment in the median age increases the *DDC* values by 0.0799 in 2022 and by 0.0263 in 2023. The characteristics calculated for large regions (EU, continents and the world) are very close to the best fitting blue lines (see large markers in Figures 1, 2). We can conclude that the young age of Africa ($A_{27} = 18$, see Table 1) is the main reason of very low numbers of cases and death per capita registered on this continent.

Opposite and much weaker age trends we can see for the case fatality risks (lines 15 and 19 in Table 4). The decrease of *CFR* values with increase of the age (supported only by the 2022 dataset) looks unexpected [especially taking into account the fact that in 2020 younger populations had less clinical cases per capita (Davies et al., 2020)]. Probably, the reason is better medical treatment in the reach countries with the high median age.

The numbers of cases and deaths per capita increase with increasing the percentages of fully vaccinated people and boosters (see rows 2, 3, 6, 7, 9, 10, 13, 14 in Table 4 and green and magenta best fitting lines in Figures 1, 2). Re-infections in vaccinated persons are common (Flacco et al., 2022; Guedes et al., 2023), but a very clear uprising trend with increasing *VC* and *BC* values is unexpected despite the similar result for smoothed daily numbers of cases reported in Nesteruk and Rodionov (2022a) (JHU datasets with 7-days-smoothing corresponding to September 1, 2021 and February 1, 2023 were used for statistical analysis). Obtained trends could be a result of age influence, since the most vaccinated countries have higher A_i values (see Tables 1, 2). We will discuss this correlation in the next Section. Another reason could be the introduction of special passports that removed restrictions for vaccinated persons. Many vaccinated people in countries with high *VC* and *BC* values started to visit crowded places, travel despite they can spread the infection. In many countries (in particular, in Ukraine) the vaccination procedure was associated with overcrowding in hospitals, which could contribute to the spread of the infection too.

As expected, the case fatality risks decrease with increasing the percentages of fully vaccinated people and boosters (see rows 16, 17, 20, 21 in Table 4 and green and magenta best fitting lines in Figure 3). Similar result was obtained in Nesteruk and Rodionov (2022a) with the use of JHU datasets for European and some other countries. In 2023, the decreasing trend was not supported by Fisher test. Probably, this is due to the more chaotic data. In particular, the different days correspond to $VC_i^{(2)}$ and $BC_i^{(2)}$ values listed in Table 2, no *CFR* value can be calculated for Turkey.

Discussion

The explanatory variables *A*, *DTC*, $VC^{(1)}$, $VC^{(2)}$, $BC^{(1)}$, and $BC^{(2)}$, used in our analysis can be also dependent on each other. We have used the linear regression (Equation 6) and Fisher test to find correlations between *DTC*, $VC^{(1)}$, $VC^{(2)}$, $BC^{(1)}$, and $BC^{(2)}$ values and explanatory variable *A*. The results of calculations are listed in

TABLE 5 Correlations vs. the median age of populations and purified levels of vaccinations.

Number	Variable y	Number of observations n	Correlation coefficient R	Optimal values of parameter a in Equation (1)	Optimal values of parameter b in Equation (1)	Experimental value of the Fisher function F, Equation (3), $m = 2$	Critical value of Fisher function $F_c(1,n\text{-}2)$ for the confidence level 0.01 (Appendix[16])	F/F_c
Correlations vs. median age of population, A								
1	DTC	23	0.4389	−4.7762	0.23290	5.011	7.85	0.64
2	$VC^{(1)}$	31	0.7673	2.3288	1.83706	41.515	7.77	5.34
3	$VC^{(2)}$	31	0.7348	17.5274	1.49084	34.036	7.77	4.38
4	$BC^{(1)}$	31	0.7933	−34.8469	2.21980	49.246	7.77	6.34
5	$BC^{(2)}$	31	0.7810	−51.3696	3.03961	45.339	7.77	5.84
Correlations vs. “purified” numbers of fully vaccinated persons per hundred, VP								
6	$DCC^{(1)}$	31	0.03726	463.8161	1.35771	0.0403	7.77	0.0052
7	$DDC^{(1)}$	31	0.00318	1.0919	0.00022628	0.000294	7.77	3.8e-5
8	$CFR^{(1)}$	31	−0.3236	0.005114	−0.00013372	3.391	7.77	0.43
Correlations vs. “purified” numbers of boosters per hundred, BP								
9	$DCC^{(1)}$	31	0.2969	463.792	9.75216	2.804	7.77	0.36
10	$DDC^{(1)}$	31	0.0989	1.0919	0.0063360	0.2864	7.77	0.037
11	$CFR^{(1)}$	31	−0.3625	0.005117	−0.00013502	4.386	7.77	0.56

Optimal values of parameters in Equation (1), correlation coefficients and the results of Fisher test applications.

10.3389/fdata.2023.1355080

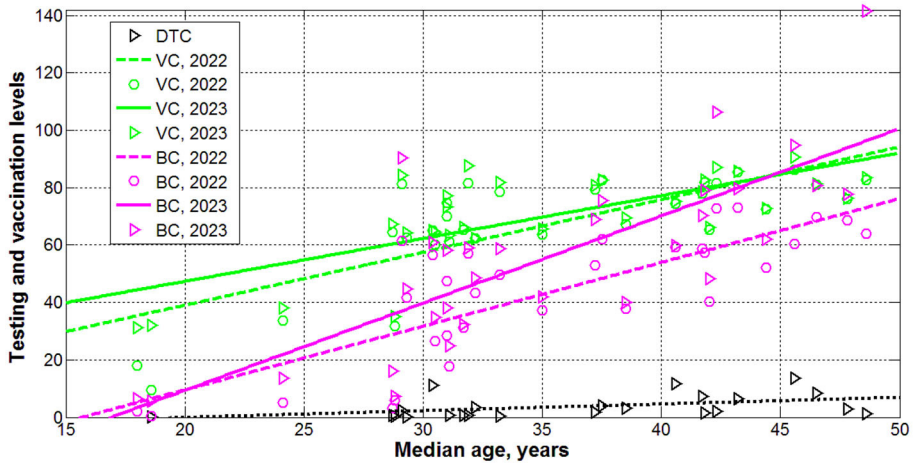

FIGURE 4
Levels of testing (black), vaccinations (green) and boosters (magenta) in 2022 ("circles") and 2023 ("triangles") vs. median age. Best fitting lines are solid for 2023 and dashed for 2022. The dotted line corresponds to the *DTC* correlation (supported at the significance level 0.05).

Table 5 and displayed in Figure 4. We can see strong correlations between $VC^{(1)}$, $VC^{(2)}$, $BC^{(1)}$, and $BC^{(2)}$ vs. median age of population A (see rows 2–5 in Table 5; green and magenta best fitting lines in Figure 4). The correlation between A and DTC is supported at the confidence level 0.05 (see the first row in Table 5 and the black best fitting line in Figure 4). The growth of the median age leads to the increase of testing level and the percentage of vaccinations and boosters. These correlations can be a result of higher incomes in aged countries and more vaccinations and boosters in older people.

Now we can explain why the numbers of cases and deaths per capita can increase with increasing the percentages of fully vaccinated people and boosters (see rows 2, 3, 6, 7, 9, 10, 13, 14 in Table 4 and green and magenta best fitting lines in Figures 1, 2)? Values $VC^{(1)}$, $VC^{(2)}$, $BC^{(1)}$, and $BC^{(2)}$ are not independent and definitely increase with the age. On the other hand, DCC and DDC values also increase with growth of A_i (see rows 1, 5, 8, 12 in Table 4 and blue best fitting lines in Figures 1, 2). To remove the influence of age in correlations between vaccinations and DCC and DDC values, let us consider the "purified" variations of $VC^{(1)}$and $BC^{(1)}$ (we limited ourselves only to 2022 with more reliable statistical data):

$$VP_i = VC_i^{(1)} - (2.3288 + 1.8376A_i)\,; \qquad i = 1, 2, .., 34$$
$$BP_i = BC_i^{(1)} - (-34.8469 + 2.2198A_i)\,; \qquad i = 1, 2, .., 34$$

To obtain the "purified" variations the percentages of vaccinations VP_i and boosters BP_i, we have excluded from variations $VC^{(1)}$and $BC^{(1)}$ the values predicted by the by the best fitted lines listed in Table 5 (rows 2 and 4).

We have used the linear regression (Equation 6) and Fisher test to find correlations between DCC, DDC and CFR values and explanatory variables VP and BP. The results of calculations are listed in Table 5 (lines 6–11). No correlations were revealed at the confidence level 0.01. Thus, the vaccinations and booster themselves do not increase the numbers of cases and death per capita. No correlations between VC and the numbers of cases and deaths per capita accumulated in 15 European countries with the highest testing level as of August 1, 2022 were revealed at the confidence level 0.05 (Nesteruk and Rodionov, 2022b). The lack of decreasing trends and fact that severe pandemic waves occurred in countries with high vaccination levels [e.g., Israel, Hong Kong and Japan (Nesteruk, 2021a, 2022)] call into question the effectiveness of vaccinations due to coronavirus mutations (see text footnotes 5–8) and large numbers of re-infections (see text footnote 9, Flacco et al., 2022; Guedes et al., 2023).

As expected, the case fatality risks reduce for higher values of VP and BP (see lines 8, 11 in Table 5), but at lower confidence level than vs. $VC^{(1)}$and $BC^{(1)}$ (see lines 16 and 17 in Table 4).

Conclusions

The averaged daily numbers of cases DCC and death DDC per million, case fatality risks DDC/DCC were calculated for 34 countries and regions with the use of John Hopkins University (JHU) datasets for numbers per capita accumulated in 2022 and 2023. Linear and non-linear approaches were used to find correlations with the averaged daily numbers of tests per thousand DTC, median age of population A, and percentages of vaccinations VC and boosters BC.

One-year increment in the median age yielded 39.8 increase in DCC values and 0.0799 DDC increase in 2022 (in 2023 these figures are 5.8 and 0.0263, respectively). With decreasing of testing level DTC the case fatality risk can increase drastically. DCC and DDC values increase with increasing the percentages of fully vaccinated people and boosters. Since VC and BC values definitely increase with at higher A, the corrected variations of VC and BC were introduced, which showed no correlations with DCC and DDC values.

The presented analysis demonstrates that age is a pivot factor in visible (registered) part of the COVID-19 pandemic dynamics. Much younger Africa has registered less numbers of cases and death per capita due to many unregistered asymptomatic patients. Of

great concern is the fact that COVID-19 mortality in 2023 in the UK is still at least 4 times higher than the global value caused by seasonal flu.

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.

Author contributions

IN: Writing—original draft, Writing—review & editing, Data curation, Investigation, Methodology, Software.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. The study was supported by the Solidarity Satellite Programme of Isaac Newton Institute for Mathematical Sciences, Cambridge, UK.

Acknowledgments

The author was grateful to Professor Robin Thompson, Professor Matt Keeling, and Oleksii Rodionov for their support and providing very useful information.

Conflict of interest

The 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.

Publisher's note

References

Byass, P. (2020). Eco-epidemiological assessment of the COVID-19 epidemic in China, January-February 2020. *medRxiv* [Preprint]. doi: 10.1101/2020.03.29.20046565

Chintala, S., Dutta, R., and Tadmor, D. (2021). COVID-19 spatiotemporal research with workflow-based data analysis. *Infect. Genet. Evol.* 88, 104701. doi: 10.1016/j.meegid.2020.104701

COVID-19 Data (2023). *COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU).* Available online at: https://github.com/owid/covid-19-data/tree/master/public/data (accessed November 23, 2023).

Davies, N. G., Klepac, P., Liu, Y., Prem, K., Jit, M., CMMID COVID-19 working group, et al. (2020). Age-dependent effects in the transmission and control of COVID-19 epidemics. *Nat. Med.* 26, 1205–1211. doi: 10.1038/s41591-020-0962-9

Distante, C., Piscitelli, P., and Miani, A. (2020). COVID-19 outbreak progression in italian regions: approaching the peak by the end of march in northern italy and first week of April in Southern Italy. *Int. J. Environ. Res. Public Health* 17, 3025. doi: 10.3390/ijerph17093025

Draper, N. R., and Smith, H. (1998). *Applied Regression Analysis*, 3rd ed. Hoboken, NJ: John Wiley. doi: 10.1002/9781118625590

Fanelli, D., and Piazza, F. (2020). Analysis and forecast of COVID-19 spreading in China, Italy and France. *Chaos Solitons Fractals* 134, 109761. doi: 10.1016/j.chaos.2020.109761

Flacco, M. E., Soldato, G., Acuti Martellucci, C., Di Martino, G., Carota, R., Caponetti, A., et al. (2022). Risk of SARS-CoV-2 reinfection 18 months after primary infection: population-level observational study. *Front. Public Health* 10, 884121. doi: 10.3389/fpubh.2022.884121

Fowlkes, A. L., Yoon, S. K., Lutrick, K., Gwynn, L., Burns, J., Grant, L., et al. (2022). Effectiveness of 2-dose BNT162b2 (Pfizer BioNTech) mRNA vaccine in preventing SARS-CoV-2 infection among children aged 5-11 years and adolescents aged 12–15 years -PROTECT Cohort, July 2021–February 2022. *MMWR Morb. Mortal. Wkly. Rep.* 71, 422–428. doi: 10.15585/mmwr.mm7111e1

Guedes, A. R., Oliveira, M. S., Tavares, B. M., Luna-Muschi, A., Lazari, C. D. S., Montal, A. C., et al. (2023). Reinfection rate in a cohort of healthcare workers over 2 years of the COVID-19 pandemic. *Sci Rep.* 13, 712. doi: 10.1038/s41598-022-25908-6

Hamzah, F., Binti, A., Lau, C., Nazri, H., Ligot, D. V., Lee, G., et al. (2020). CoronaTracker: worldwide COVID-19 outbreak data analysis and prediction. *Bull. World Health Organ* 1, 32. doi: 10.2471/BLT.20.255695

Lazarus, J. V., Romero, D., Kopka, C. J., Karim, S. A., Abu-Raddad, L. J., Almeida, G.,et al. (2022). A multinational Delphi consensus to end the COVID-19 public health threat. *Nature* 611, 332–345. doi: 10.1038/s41586-022-05398-2

Mohammadi, A., Meniailov, I., Bazilevych, K., Yakovlev, S., and Chumachenko, D. (2021). Comparative study of linear regression and SIR models of COVID-19 propagation in Ukraine before vaccination. *Radioelectron. Comput. Syst.* 5–18. doi: 10.32620/reks.2021.3.01

Nesteruk, I. (2021a). Influence of possible natural and artificial collective immunity on new COVID-19 pandemic waves in Ukraine and Israel. *Explor. Res. Hypothesis Med.* 7, 8–18. doi: 10.14218/ERHM.2021.00044

Nesteruk, I. (2021b). Visible and real sizes of new COVID-19 pandemic waves in Ukraine. *Innov. Biosyst. Bioeng.* 5, 85–96. doi: 10.20535/ibb.2021.5.2.230487

Nesteruk, I. (2021c). The real COVID-19 pandemic dynamics in Qatar in 2021: simulations, predictions and verifications of the SIR model. *Semin. Ciênc. Exatas Tecnol.* 42(1Supl), 55–62. doi: 10.5433/1679-0375.2021v42n1Suplp55

Nesteruk, I. (2021d). *COVID19 Pandemic Dynamics*. Berlin: Springer Nature. doi: 10.1007/978-981-33-6416-5

Nesteruk, I. (2022).Vaccination and testing as a means of ending the COVID-19 pandemic: comparative and statistical analysis. *medRxiv* [Preprint]. doi: 10.2139/ssrn.4161720

Nesteruk, I. (2023a). Can we start to ignore the SARS-CoV-2 disease? *medRxiv* [Preprint]. doi: 10.1101/2023.09.18.23295709

Nesteruk, I. (2023b). Improvement of the software for modeling the dynamics of epidemics and developing a user-friendly interface. *Infect. Dis. Model.* 8, 806–821. doi: 10.1016/j.idm.2023.06.003

Nesteruk, I. (2023c). What is wrong with Chinese COVID-19 statistics? *Epidemiol. Biostat. Public Health* 18, 9–11. doi: 10.54103/2282-0930/20637

Nesteruk, I., and Keeling, M. (2023). Population age as a key factor in the COVID-19 pandemic dynamics. *Res. Sq.* doi: 10.1101/2023.11.30.23299229

Nesteruk, I., and Rodionov, O. (2021). Impact of vaccination and testing levels on the dynamics of the COVID-19 pandemic and its cessation. *J. Biomed. Res. Environ. Sci.* 2, 1141–1147. doi: 10.37871/jbres1361

Nesteruk, I., and Rodionov, O. (2022a). Omicron waves of the COVID-19 pandemic and efficacy of vaccinations and testing. *J. Biomed. Res. Environ. Sci.* 3, 588–594. doi: 10.37871/jbres1484

Nesteruk, I., and Rodionov, O. (2022b). The COVID-19 pandemic in rich and poor countries. *Res. Sq*. doi: 10.21203/rs.3.rs-2348206/v1

Nesteruk, I., Rodionov, O., and Walczak, S. (2022). "Comparative and statistical analysis of the COVID-19 pandemic dynamics," in *Proceedings of the 2022 IEEE 41th International Conference on Electronics and Nanotechnology (ELNANO). October 10-14, 2022* (Kyiv: IEEE), 379–384. doi: 10.1109/ELNANO54667.2022.9927116

Ng, K. Y., and Gui, M. M. (2020). COVID-19: development of a robust mathematical model and simulation package with consideration for ageing population and time delay for control action and resusceptibility. *Phys. D: Nonlinear Phenom.* 411, 132599. doi: 10.1016/j.physd.2020.132599

Paget, J., Spreeuwenberg, P., Charu, V., Taylor, R. J., Iuliano, A. D., Bresee, J., et al. (2019). Global mortality associated with seasonal influenza epidemics: new burden estimates and predictors from the GLaMOR Project. *J. Glob. Health* 9, 020421. doi: 10.7189/jogh.09.020421

Pardhan, S., and Drydakis, N. (2021). Associating the change in new COVID-19 cases to GDP per capita in 38 European Countries in the first wave of the pandemic. *Front. Public Health* 8, 582140. doi: 10.3389/fpubh.2020.582140

Rossman, H., Shilo, S., Meir, T., Gorfine, M., Shalit, U., and Segal, E. (2021). COVID-19 dynamics after a national immunization program in Israel. *Nat. Med.* 27, 1055–1061. doi: 10.1038/s41591-021-01337-2

Schreiber, P. W., Scheier, T., Wolfensberger, A., Saleschus, D., Vazquez, M., Kouyos, R., et al. (2023). Parallel dynamics in the yield of universal SARS-CoV-2 admission screening and population incidence. *Sci. Rep.* 13, 7296. doi: 10.1038/s41598-023-33824-6

Shang, W., Kang, L., Cao, G., Wang, Y., Gao, P., Liu, J., et al. (2022). Percentage of asymptomatic infections among SARS-CoV-2 omicron variant-positive individuals: a systematic review and meta-analysis. *Vaccines* 10, 1049. doi: 10.3390/vaccines10071049

Statsenko, Y., Al Zahmi, F., Habuza, T., Almansoori, T. M., Smetanina, D., Simiyu, G. L., et al. (2021). Impact of age and sex on COVID-19 severity assessed from radiologic and clinical findings. *Front. Cell. Infect. Microbiol.* 11, 777070. doi: 10.3389/fcimb.2021.777070

Weinberger, D. M., Cohen, T., Crawford, F. W., Mostashari, F., Olson, D., Pitzer, V. E., et al. (2020). Estimating the early death toll of COVID-19 in the United States. *medRXiv* [preprint]. doi: 10.1101/2020.04.15.20066431

frontiers | Frontiers in Public Health

TYPE Technology and Code
PUBLISHED 26 April 2024
DOI 10.3389/fpubh.2024.1347862

OPEN ACCESS

EDITED BY
Dmytro Chumachenko,
National Aerospace University – Kharkiv Aviation Institute, Ukraine

REVIEWED BY
Sean Mark Patrick,
University of Pretoria, South Africa
Eustachio Cuscianna,
University of Bari Aldo Moro, Italy
Amit Kumar Banerjee,
Indian Institute of Chemical Technology (CSIR), India
Samson Malwa Haumba,
Georgetown University Medical Center, United States

*CORRESPONDENCE
Taesung Park
✉ tspark@stats.snu.ac.kr

RECEIVED 01 December 2023
ACCEPTED 12 April 2024
PUBLISHED 26 April 2024

CITATION
Song H, Han K, Park J, Liu Z, Goo T, Krishnamurthy A and Park T (2024) K-Track-Covid: interactive web-based dashboard for analyzing geographical and temporal spread of COVID-19 in South Korea.
Front. Public Health 12:1347862.
doi: 10.3389/fpubh.2024.1347862

K-Track-Covid: interactive web-based dashboard for analyzing geographical and temporal spread of COVID-19 in South Korea

Hanbyul Song[1], Kyulhee Han[1], Jiwon Park[1], Zhe Liu[1], Taewan Goo[1], Ashok Krishnamurthy[2] and Taesung Park[3]*

[1]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, Republic of Korea, [2]Department of Mathematics and Computing, Mount Royal University, Calgary, AB, Canada, [3]Department of Statistics, Seoul National University, Seoul, Republic of Korea

The COVID-19 pandemic has necessitated the development of robust tools for tracking and modeling the spread of the virus. We present 'K-Track-Covid,' an interactive web-based dashboard developed using the R Shiny framework, to offer users an intuitive dashboard for analyzing the geographical and temporal spread of COVID-19 in South Korea. Our dashboard employs dynamic user interface elements, employs validated epidemiological models, and integrates regional data to offer tailored visual displays. The dashboard allows users to customize their data views by selecting specific time frames, geographic regions, and demographic groups. This customization enables the generation of charts and statistical summaries pertinent to both daily fluctuations and cumulative counts of COVID-19 cases, as well as mortality statistics. Additionally, the dashboard offers a simulation model based on mathematical models, enabling users to make predictions under various parameter settings. The dashboard is designed to assist researchers, policymakers, and the public in understanding the spread and impact of COVID-19, thereby facilitating informed decision-making. All data and resources related to this study are publicly available to ensure transparency and facilitate further research.

KEYWORDS

COVID-19, South Korea, R shiny, interactive dashboard, mathematical model, infectious disease

Introduction

Although the COVID-19 pandemic is no longer acute, its effects on health, economy, and society remain significant, highlighting the need to understand various aspects of its recovery and analyze its impacts (1). Governments and institutions worldwide are in need of sophisticated tools to track, understand, and manage the spread of the virus (1). Numerous epidemiological dashboards like those from the World Health Organization (WHO) (2), Johns Hopkins University's (JHU) Center for Systems Science and Engineering (CSSE) (3), and the Centers for Disease Control and Prevention (CDC) (4, 5) provide valuable insights into global COVID-19 trends.

The WHO dashboard, for instance, offers daily global counts of cases, deaths, and vaccine distribution focusing more on the global scale. Similarly, the JHU CSSE Dashboard stands out

for its geospatial representation and data granularity, focused on the United States (3). In addition to cases, deaths, and recoveries data, it has incorporated other metrics like case-fatality ratios and testing data. Although it offers three levels of administrative data granularity—countries, and states/provinces—this fine-grained detail is primarily available only for the United States. Furthermore, while the dashboard provides valuable data, its map representation is limited to the most recent date, without displaying changes over time (3). The CDC's COVID-NET dashboard offers different topics to see specific data trends. These topics include rates by age group, sex, race, ethnicity, and clinical characteristics. While the data can be displayed in a graph or as a table, a map representation is not provided (4). Lastly, the Africa CDC's COVID-19 Africa Hotspot Dashboard helps to identify countries that have growing or widespread outbreaks of COVID-19. It is tailored to be highly sensitive in detecting potential hotspots, acknowledging the trade-off in specificity (5).

Existing dashboards often lack forecasting capabilities, essential for policymakers and health professionals to prepare and strategize based on potential future trends. Most current tools fail to offer these predictive insights, highlighting a limitation that needs to be addressed. To address the limitations of current analytical tools, we present 'K-Track-Covid,' a web-based interactive dashboard developed with the R Shiny package (6). This dashboard offers insights into the epidemiological situation and features a robust forecasting function. Its predictive power comes from incorporating advanced mathematical models to forecast the pandemic's trajectory, which we will outline briefly in the Methods section.

The 'K-Track-Covid' dashboard focuses on South Korea—a nation that has been at the forefront of proactive and robust public health interventions, such as comprehensive contact tracing and testing (7). This targeted approach allows the dashboard to delve into complexities often glossed over in more broad-scope dashboards (2, 3). Unlike global dashboards that provide periodic global snapshots, 'K-Track-Covid' allows for detailed examination of South Korea-specific data across various times, facilitating time- and location-based analyses. This feature supports stakeholders, from researchers to policymakers, in making informed decisions. 'K-Track-Covid' represents a significant advancement in digital epidemiological tools, offering a detailed lens on South Korea's COVID-19 situation.

Furthermore, the 'K-Track-Covid' dashboard features a highly modular and scalable architecture, allowing easy customization to various countries or regions. The key to this adaptability is access to robust and reliable local data. With the integration of such data, the dashboard transforms into a versatile instrument for conducting detailed, localized COVID-19 analysis, regardless of location.

The dashboard offers four main functions: (1) it automatically updates COVID-19 data in South Korea, including case counts (confirmed and deceased), according to the available data released by the government. (2) It allows for daily and cumulative tracking, helping to project short and long-term trends. This feature is designed to resume full functionality should the government reinitiate the release of COVID-19 data. (3) The dashboard presents data at various geographical resolutions, enhanced with socio-demographic metrics such as medical resources and population profiles, for detailed spatial analyses. (4) Most notably, it integrates a mathematical model to forecast and visually illustrate the disease's spread over time on a dynamic MP4 animated map. This unique feature allows users to anticipate hotspots and track regional progression, providing an engaging experience. This combination of features makes it a comprehensive tool for academics, policymakers, and the public interested in South Korea's pandemic trajectory.

In a time of abundant yet complex data, 'K-Track-Covid' aims to be a reliable, user-friendly resource for all audiences. The remainder of this paper is structured to offer a detailed account of 'K-Track-Covid.' Following this Introduction, the Methods section will elucidate the technical architecture, data sources, and the overall structure of the dashboard. The Results section will focus on the dashboard's functionalities, user engagement, and impact, including a comparison with other available dashboards to contextualize its utility and performance. In the Discussion, we will explore broader implications for pandemic management in South Korea, as well as the adaptability of such a tool for other regions or future public health crises.

Materials and methods

Data sourcing

Regional data

Data Sourcing is an essential part of our dashboard, ensuring accurate COVID-19 information for South Korea. We use an Application Programming Interface (API) from the South Korean government's data portal, provided by the Korea Disease Control and Prevention Agency (KDCA) (8). To manage this data, we utilize R libraries, employing 'httr' (9) for Hyper Text Transfer Protocol (HTTP) requests and 'xml2' (10) for parsing Extensible Markup Language (XML) data. This combination ensures consistent and standardized data that is ready for analysis.

Our dataset covers the period from the pandemic's start in early 2020 to September 2023, offering a broad temporal range for analysis. Initially, data updates occurred daily until May 2023, transitioning to weekly updates from that point forward. On September 1, 2023, the Korean government reclassified COVID-19 as a Class 4 infectious disease, leading to the discontinuation of regular updates to the provincial incidence data (8). We primarily use the daily data to ensure detailed analyses. This dataset spans all 17 of South Korea's major administrative regions, like Seoul, Busan, and Daegu, covering metrics such as confirmed cases, deaths, and outbreak status per area. Such granularity provides a precise regional perspective on the pandemic, which is crucial for academic study and policymaking.

In addition to the COVID-19 data, we also incorporate regional socio-demographic data obtained from the Korean Statistical Information Service (KOSIS) (11). These data provide demographic insights such as age distribution, gender, and smoking rate, which are crucial for a nuanced understanding of the pandemic's impact. Including KOSIS data enables a more comprehensive analysis, as socio-demographic factors are significant determinants of infection rates and health outcomes. This multi-dimensional approach not only informs targeted public health strategies but also enriches academic discourse on the effects of COVID-19 across diverse population strata within South Korea.

Spatial demographic data

Our dashboard enhances epidemiological analysis with spatial demographic data from WorldPop (12). This source offers detailed population density maps at 1 km × 1 km and 100 m x 100 m resolutions globally, useful for addressing census gaps and supporting UN agencies' planning efforts. In 'K-Track-Covid,' this data helps analyze

COVID-19 spread in relation to population densities and demographics, especially in densely populated areas like Seoul, Busan, and Daegu. Integrating these maps provides richer, more contextually informed insights, aiding in developing effective public health strategies.

Architecture

The 'K-Track-Covid' dashboard is developed using the RStudio Shiny framework for its reliability, user-friendliness, and advanced data handling capabilities. We employ key R packages for efficient operations: dplyr (13) and tidyverse (14) for data management, sf (15) and raster (16) for geospatial data, plotly (17) for interactive, and ggplot2 (18) for static visualizations. Geospatial interactivity is enhanced with leaflet (19) and leaflet. Extras (20). The architecture combines Shiny with shinydashboard (21) for web interactivity, dividing into a User Interface (UI) and a server-side function. While primarily R-based, we incorporate JavaScript, Cascading Style Sheets (CSS), and Hyper Text Markup Language (HTML) to augment functionality and design.

Dashboard structure

The 'K-Track-Covid' dashboard is structured based on three main tabs: Map, Regional Trend, and Simulation. Table 1 contains the description of each main tab.

The dashboard's user interface is crafted for easy navigation, featuring dropdown menus for region selection, timeframe buttons, and interactive graphs and maps. Tooltips offer extra details when hovering over data points.

In the development phase of the 'K-Track-Covid' dashboard, we integrated a user feedback mechanism to refine and enhance the platform's usability and functionality continually. This feedback mechanism, in the form of a user rating form, is modeled on the Mobile App Rating Scale (21), a validated tool for assessing the quality of health mobile apps. Our adaptation of this scale allows users to rate their experience across multiple dimensions: ease of use, reliability, quality, scope of information, and aesthetics. This approach enables us to gather actionable insights into the dashboard's performance from the user's perspective, guiding future improvements.

TABLE 1 K-Track-Covid dashboard structure.

Main tab	Description
Map	The Map tab incorporates leaflet to display geographical information. A leaflet map is presented on this tab, which covers the entire window size.
Regional Trend	Focuses on visual analytics at a more granular level. It utilizes the plotly package to render the plots.
Simulation	Integrates mathematical models to simulate COVID-19 scenarios based on user-defined parameters.

Map tab

The Map tab (Figure 1) in 'K-Track-Covid' employs Leaflet technology to create an interactive geospatial display. Situated on the left-hand side, a fixed yet draggable absolute panel hosts various control elements, such as a date range slider, a dropdown for overlay options, and selection boxes for additional output. The date range slider allows users to navigate different time periods, providing a temporal dimension often missing from other dashboards. When a user positions their cursor over a particular region on the map, the system displays detailed data about the region. This information can include metrics like daily cases and deaths, each available in raw counts, proportions, and per 100 k population formats.

One of the standout features of the Map tab is its capability to show the spread of the disease with the regional socio-demographic data, including the total number of beds, doctors, seniors, and foreigners per 1,000 people, alongside smoking rates and the total population segmented by age group and gender. Integrating these metrics allows for identifying hotspots and high-risk zones by setting thresholds on metrics like senior population percentages, smoking rates, and healthcare availability. Users can select these metrics to visualize risk levels across regions, enhancing resource allocation and intervention strategies. This feature complements existing COVID-19 data, offering a fuller risk assessment tool for public health decision-making.

The leaflet R package supports the dashboard's geospatial visualizations, utilizing OpenStreetMap (22) for base maps. The interface has been enhanced with custom CSS styling for intuitive navigation and aesthetic appeal, improving the control panel's functionality and user interaction with the map.

Regional trend tab

The Regional Trend tab (Figure 2) offers users a diverse range of time-series visualization options rendered using the plotly package (17). It provides a dynamic look at the temporal trends of COVID-19 metrics for individual regions within South Korea. Unlike the Map tab, which shows a single moment, the Regional Trend tab provides a comprehensive historical perspective, highlighting changes and trends in the pandemic's progression across different regions. This approach helps identify patterns, seasonal variations, and outliers in data such as new cases, new deaths, and cumulative figures. The underlying data for these visualizations, drawn from KDCA, has been pre-processed to reflect accurate, region-specific statistics.

The sidebar panel includes a region and an outcome selector dropdown, acting as auxiliary tools to refine the time-series plot according to user preferences. This feature allows users to concentrate on particular regions and outcomes, adding depth and offering a detailed view essential for scientific research and policy development. Consequently, the Regional Trend tab provides a deeper, time-sensitive insight into the pandemic's effects across the varied landscapes of South Korea.

Simulation tab

The Simulation tab (Figure 3) stands out as a pivotal aspect of our dashboard, delivering simulations for projecting COVID-19 scenarios through shinyJS (23) for enhanced UI functionality. It includes a customization sidebar, allowing users to choose between the Susceptible-Exposed-Infectious-Recovered-Deceased (SEIRD) (24, 25) and Susceptible-Vaccinated-Exposed-Infectious-Recovered-Deceased (SVEIRD) (26, 27) models for forecasting. This capability lets users explore future pandemic trends and outcomes based on

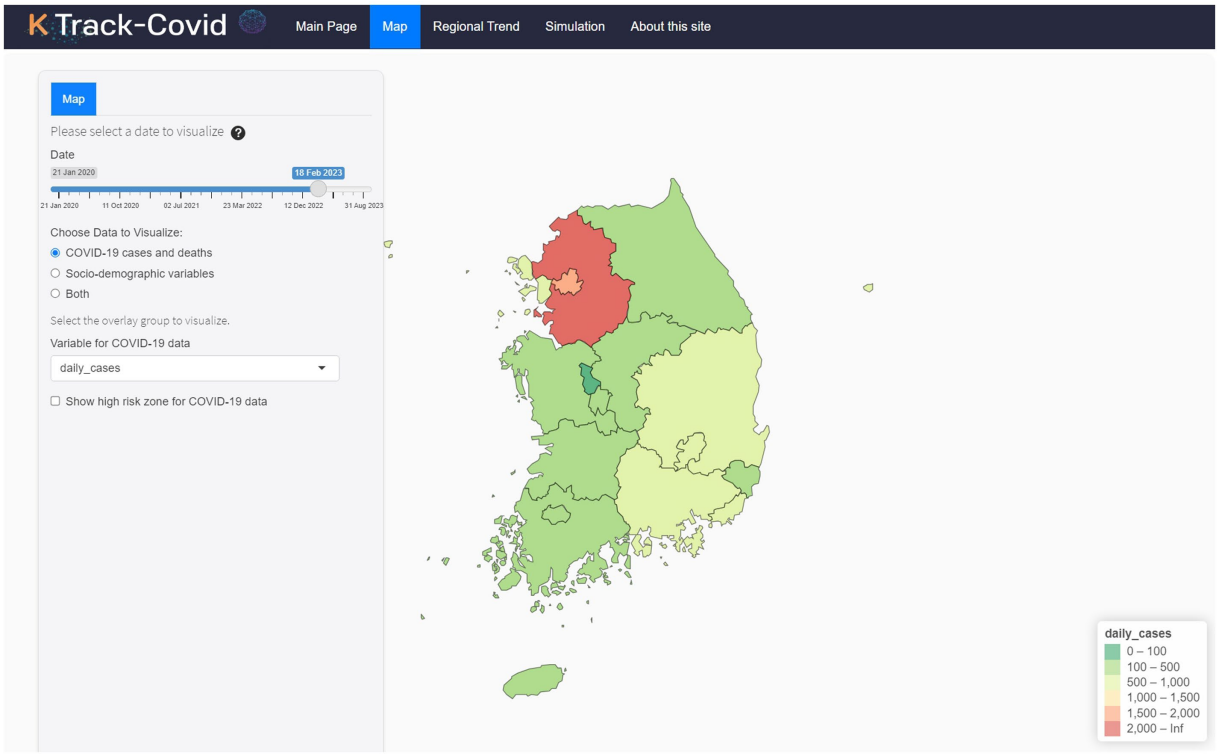

FIGURE 1
Overview of the K-Track-Covid dashboard's Map tab.

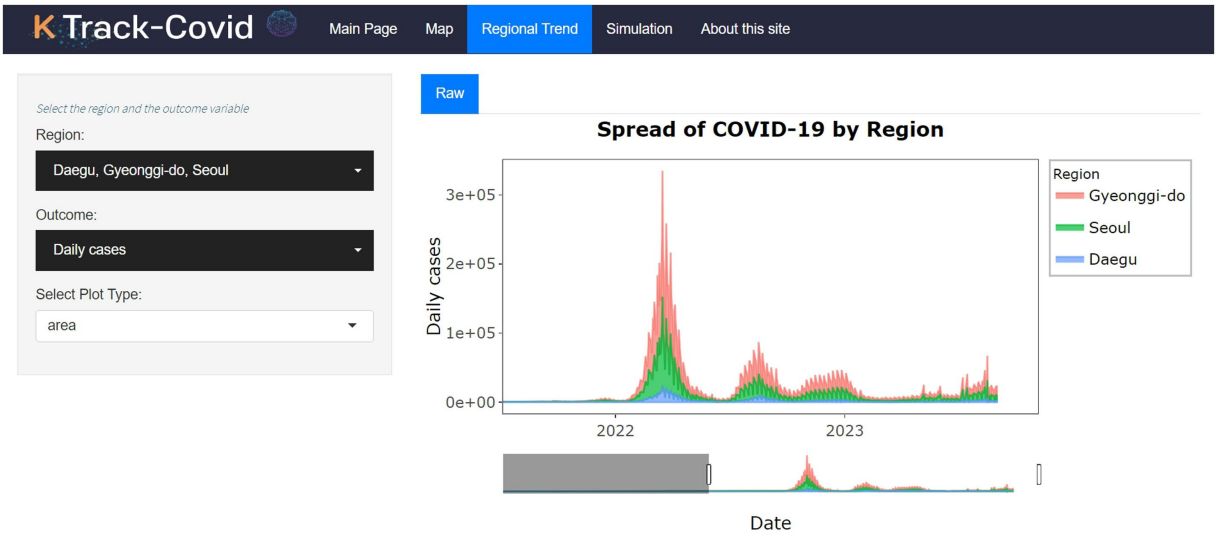

FIGURE 2
User interface and features of the Regional Trend tab.

user-defined parameters, providing a dynamic, personalized simulation experience. This unique functionality, not found in similar tools, significantly enhances our dashboard's analytical depth, offering a nuanced approach to disease spread prediction.

The SEIRD model enriches the traditional Susceptible-Infectious-Recovered (SIR) framework by integrating 'exposed' and 'deceased' compartments, allowing for a more detailed understanding of disease progression (28). Expanding upon the SEIRD framework, the SVEIRD model introduces a 'vaccinated' compartment to capture the dynamics of vaccination within the population (26). This addition allows for the simulation of vaccination outcomes alongside natural disease progression, providing a comprehensive view of potential interventions. A crucial aspect of both models is incorporating a weight function

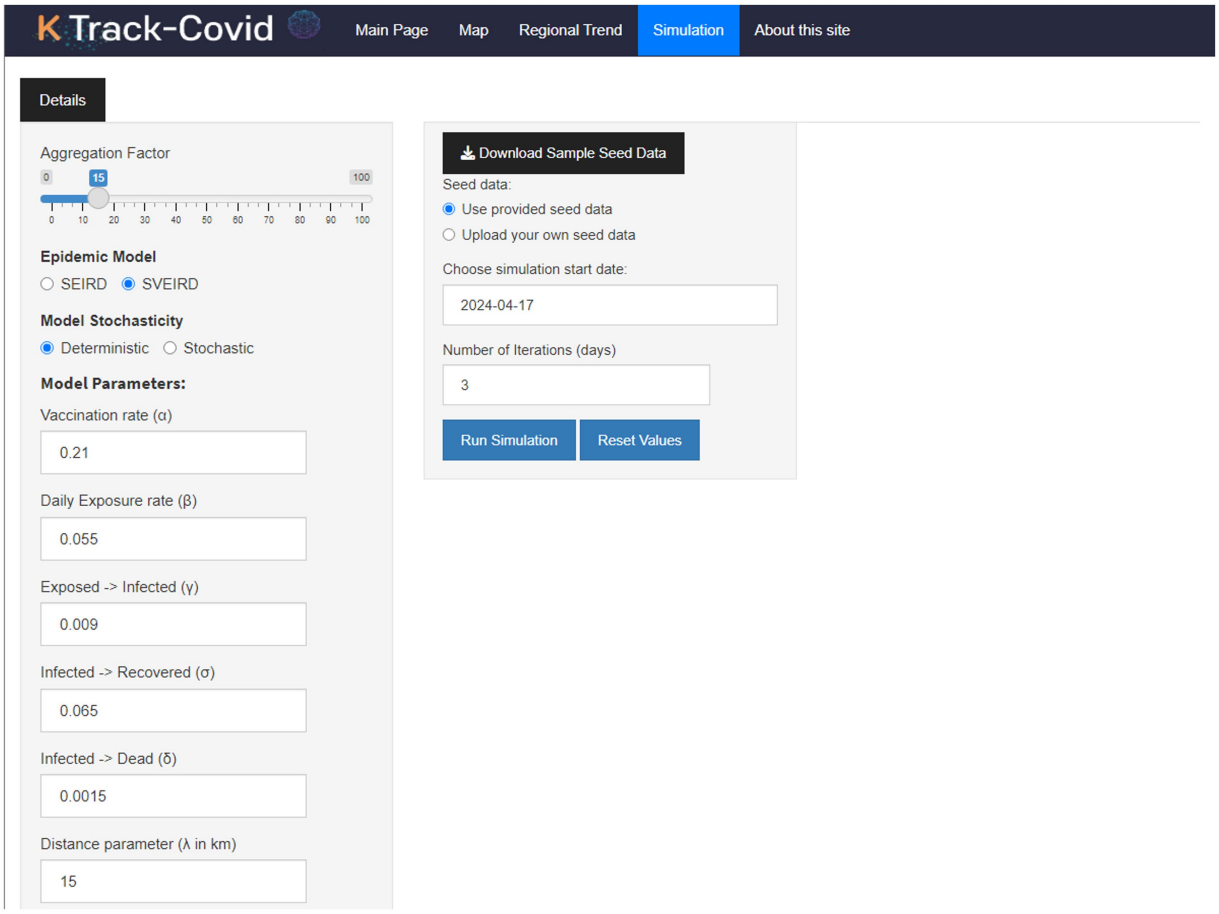

FIGURE 3
Initial overview of the Simulation tab, displaying the control panel and default outputs.

that calculates the impact of infectious individuals on susceptible individuals based on spatial distribution (29).

The dashboard also includes a feature for stochastic simulation, which introduces probabilistic variations to the model outcomes. This is facilitated by adjustable parameters such as transmission rate (β), recovery rate (σ), and the rate at which exposed individuals become infectious (γ). Users have the flexibility to customize their simulations further by uploading their seed data or utilizing pre-loaded datasets.

In the output section of the Simulation tab, users are provided with textual and tabular summaries that detail the selected epidemiological model's characteristics. This section aims to offer a clear and structured understanding of the model's implications without delving into the mathematical equations, focusing instead on the outcomes and interpretations crucial for public health planning and response strategies.

Results

The multi-dimensional analysis capabilities incorporated into our COVID-19 dashboard serve to elucidate various facets of the pandemic within South Korea. Our dashboard integrates epidemiological, temporal, and geographical data, offering insights that enhance our understanding of the virus's spread and impact. This innovation is at the core of our project, and it significantly advances current methodologies for COVID-19 analytics. As highlighted in the Materials and Methods section, please be advised that since the Korean government's cessation of regular updates of regional incidence data as of September 1, 2023, the Map tab and the Regional Trend tab report the results up to this date.

The 'K-Track-Covid' dashboard can be accessed online (https://k-track-covid.shinyapps.io/k-track-covid/, accessed on 12 April 2024). As demonstrated in Figure 1, the dashboard is designed to offer an intuitive and interactive user experience, divided into distinct modules accessible via individual tabs.

Map tab

As shown in Figure 4, the Map tab offers geospatial analytics across multiple levels, allowing users to explore daily case numbers or proportions and available hospital beds at the district level. It highlights at-risk populations, enhancing localized response strategies. The map facilitates temporal analysis through selectable date ranges and enables comparisons of current statistics against historical trends with data overlay. This integration provides a comprehensive view of tracking pandemic dynamics and healthcare capacity.

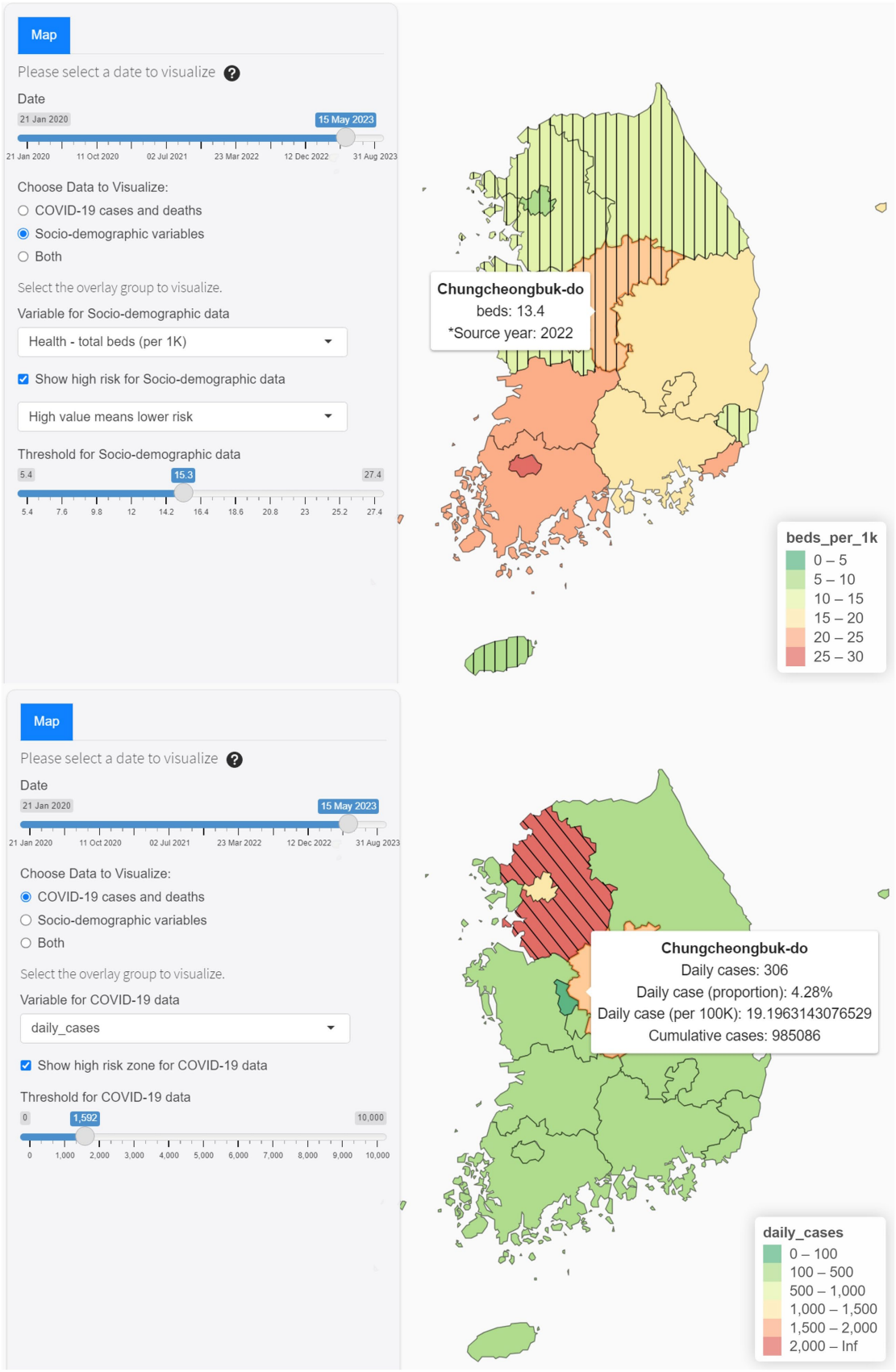

FIGURE 4
Interactive features of Map tab.

Regional trend tab

The Regional Trend tab, illustrated in Figure 2, focuses on more localized data, presenting time-series data visualization at the regional level. Users can customize the visualizations through drop-down menus, choosing from different COVID-19 metrics like daily new cases, cumulative cases, and deaths. The analytics in this tab facilitate comparisons within regions, providing a clearer perspective on the pandemic's localized impact as it unfolds over time.

Simulation tab

The Simulation tab, illustrated in Figure 3, describes the dashboard's advanced features. It uses mathematical models to predict COVID-19. The tab allows for significant customization, including the ability for users to adjust parameters to meet specific analytical needs. A user manual is available to assist new users.

Input summary panel

Figure 5 provides a snapshot of the Input Summary panel. This panel delivers a comprehensive summary of all user-defined parameters. It serves as a legend and a quick reference, allowing users to verify their configurations before interpreting the simulations.

Schematic diagram panel

The Schematic Diagram panel in Supplementary Figure S1 shows a graphical representation of the model's structure. This diagram outlines the compartmental relationships and flows among Susceptible, Exposed, Infected, Recovered, and Deceased compartments, enhancing understanding of the model's complexity.

Initial seed data panel

Figure 6 illustrates the initial seed data panel, which visually represents the model input. The seed data encompasses detailed location information, including the region and its corresponding latitude and longitude, followed by initial compartmental values such as vaccinated, exposed, infected, recovered, and deceased individuals. Users have the flexibility to either use the provided seed data or upload their own adjusted seed data. By uploading their own seed data, users can customize the initial conditions for each compartment, tailoring the model to specific regional dynamics.

MP4 animation panel

In Figure 7, the dashboard features an MP4 animation panel that visually demonstrates the epidemiological states' progression from day one to day 100, covering the entire forecast period. This animation offers a dynamic representation of the model's predictions over time, making the data more accessible and understandable. Additionally, the animation can be used to forecast and visually illustrate the disease's spread across the timeframe. For convenience, users can download the animation by clicking on the vertical ellipsis in the right corner of the video.

Output summary panel

The Output Summary panel in Supplementary Figure S2 is designed to present a comprehensive tabulation of the model's predictions, generated based on user-defined inputs and parameters. This is particularly useful for expert users or researchers requiring precise data points for further analysis or reporting. In addition, having a tabular format enables easy data export into other statistical software for subsequent analyses.

Plot panel

The Plot panel (Supplementary Figure S3) within the Simulation tab serves as a visual analytics environment, offering a dynamic and interpretable representation of the model's output metrics over time. Users can download the figure by clicking the 'Save Image' button.

Comparison to existing dashboards

Following the detailed exploration of the 'K-Track-Covid' dashboard's features and capabilities, a comparative analysis is conducted to place its functionalities in context with four other leading COVID-19 dashboards. This comparison is summarized in a comprehensive table (Table 2) that delineates the available functions across all platforms, offering a clear side-by-side view of each dashboard's unique features and shared capabilities. 'K-Track-Covid' provides a comprehensive toolset for regional and socio-demographic analysis with the unique addition of simulation capabilities. However, it underperforms in the global analysis compared to the WHO, JHU, and CDC dashboards. Its commitment to data accessibility and open-source development is on par with the leading dashboards, making it a strong contender in existing COVID-19 dashboards.

Discussion

Our study demonstrates the effectiveness of the 'K-Track-Covid' in analyzing COVID-19's trajectory and impact in South Korea. Its interactive design and analytical capabilities make it useful for researchers, policymakers, and the public. Integrating regional, epidemiological, and spatial demographic data, the dashboard offers a complete view of the pandemic's spread and effects in South Korea. This localized approach enables an in-depth understanding of COVID-19 dynamics within the country's specific public health and societal context.

Global platforms like WHO and JHU provide valuable pandemic data on a global scale, while our 'K-Track-COVID' dashboard offers in-depth, temporal analysis specific to South Korea. Additionally, our dashboard enhances the socio-demographic analysis, similar to the COVID-NET approach, by displaying it on an interactive map, thus improving the identification of potential high-risk zones. Unlike other platforms, it integrates simulation models for forecasting, which is essential for proactive public health interventions. It also features hotspot identification similar to the COVID-19 Africa dashboard, with adjustable thresholds for cases or deaths to pinpoint emerging risk areas effectively.

An essential feature of 'K-Track-Covid' is its integration of socio-demographic data, enabling the identification of potential hotspots and high-risk zones. By incorporating metrics such as the regional distribution of healthcare resources and age demographics, users can pinpoint areas that may face medical constraints or have higher vulnerability. Utilizing these socio-demographic indicators enhances analytical depth and supports targeted interventions for regions needing immediate resources. Another feature of the dashboard is its applicability for future pandemic prevention, preparedness, and

Input Summary | Mathematical Model | Schematic Diagram | Initial Seed Data

MP4 Animation | Output Summary | Plot

Variable	Value
Country	South Korea
WorldPop Raster Dimension	661 rows x 704 columns = 465344 grid cells
Aggregation Factor	5
Aggregated Raster Dimension	133 rows x 141 columns = 18753 grid cells
Compartmental Model	SVEIRD
Model Parameters	Alpha: 0.21 Beta: 0.055 Gamma: 0.009 Sigma: 0.065 Delta: 0.0015
Average Distance Travelled/Day (in km)	15
Radius (1 = Moore neighbourhood)	3
Uploaded Seed Data	seeddata_KOR_initialSeedData_default.csv
Number of iterations (days)	100

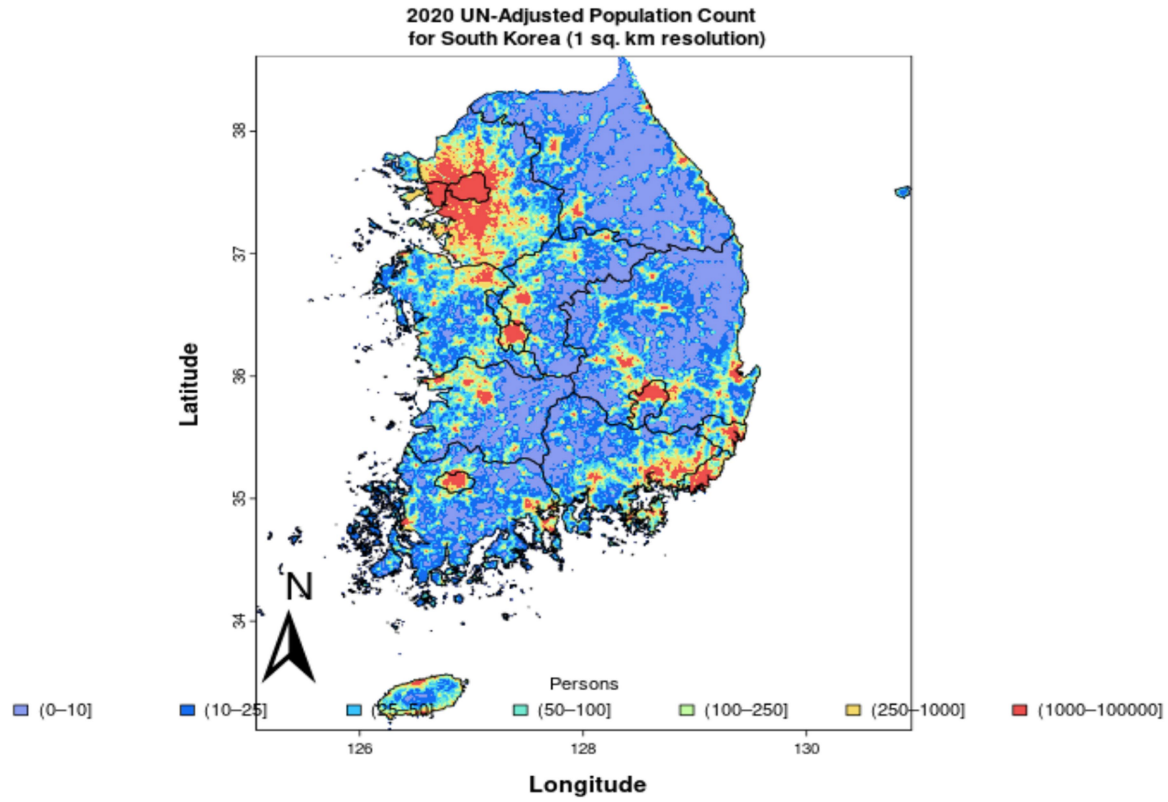

FIGURE 5
Input summary panel showing user-defined parameters.

response. The Simulation tab can be adjusted to model various infectious diseases by modifying its parameters. For example, the dashboard could be recalibrated to track and analyze the spread of influenza. This flexibility allows health authorities to adapt our models for different infectious diseases rapidly. Furthermore, the tab uses spatial demographic data from WorldPop, which offers population density information at a 1 km by 1 km resolution, enabling an analysis of disease spread relative to population count.

Despite its contributions, we acknowledge the 'K-Track-Covid' dashboard's limitations. Our current version of 'K-Track-Covid' is

Input Summary | Mathematical Model | Schematic Diagram | Initial Seed Data | MP4 Animation | Output Summary | Plot

Show 10 entries Search:

	Location	lat	lon	InitialVaccinated	InitialExposed	InitialInfections	InitialRecovered	InitialDead
1	Busan	35.21141793	129.0912098	283948	236333	2127	2089651	2941
2	Chungcheongbuk-do	36.70060634	127.8182155	146078	235667	2121	1074363	1096
3	Chungcheongnam-do	36.52544702	126.8431605	195142	254444	2290	1389103	1645
4	Daegu	35.78116616	128.579926	156270	270889	2438	1514327	2069
5	Daejeon	36.31868326	127.3785945	105885	158000	1422	1012260	1011
6	Gangwon-do	37.71681608	128.2809986	146730	190667	1716	1004411	1416
7	Gwangju	35.13442287	126.8551262	111908	177222	1595	1017606	875
8	Gyeonggi-do	37.49093933	127.1860658	891168	1484333	13359	9257912	8643
9	Gyeongsangbuk-do	36.3096476	128.7406821	244145	296000	2664	1579006	2182
10	Gyeongsangnam-do	35.34559463	128.2745798	257965	293889	2645	2073868	2102

Showing 1 to 10 of 17 entries Previous 1 2 Next

FIGURE 6
Initial seed data represented by the first 10 rows.

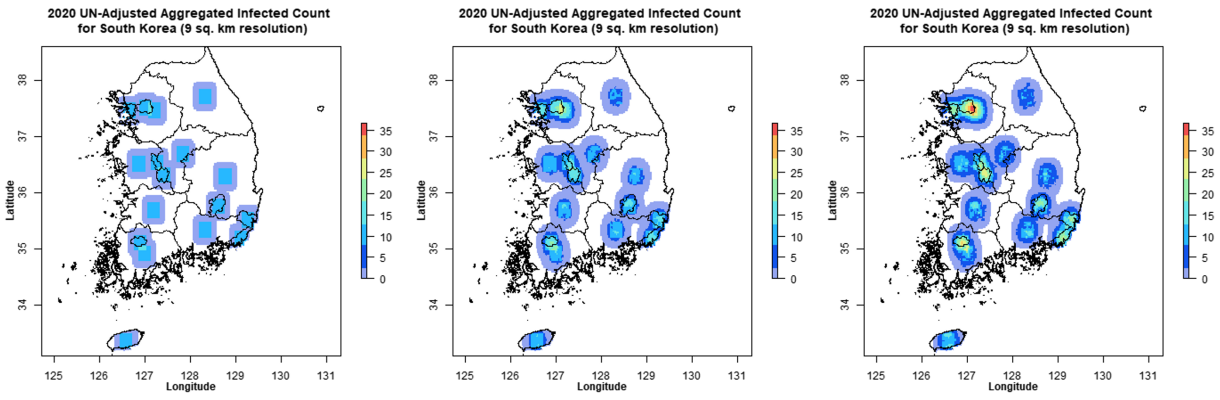

FIGURE 7
MP4 animation showing the spread of the disease over time.

TABLE 2 Comparison with publicly available COVID-19 dashboards.

Dashboard	Map representation	Regional analysis	Global analysis	Socio-demographic analysis	Simulation analysis	Data accessibility	Open-source replicability
WHO dashboard	O	X	O	O	X	O	O
JHU dashboard	O	O	O	X	X	O	O
COVID-NET dashboard	X	O	X	O	X	O	X
COVID-19 Africa dashboard	O	X	O	X	X	X	X
K-Track-Covid	O	O	X	O	O	O	O

limited to COVID-19 in South Korea. However, it possesses the potential for broader applications. The dashboard can be adapted to model various infectious diseases by incorporating additional geographical and regional data. This adaptability is illustrated through the Map and Regional Trend tabs, which can be recalibrated for other diseases with the necessary data. Furthermore, with access to international geographical information, 'K-Track-Covid' could offer insights on a global scale, enhancing infectious disease management worldwide.

In the future, expanding the 'K-Track-Covid' dashboard to include other infectious diseases and broadening its scope to various countries is a promising avenue for research. This extension will enable a comprehensive analysis of disease spread and containment strategies across diverse geographic and demographic contexts. Moreover, by integrating real patient data with the enactment of regional policies, mainly through metrics like the Korea Stringency Index (KSI) (30), the dashboard can provide detailed insights into the effectiveness of various public health strategies. Such analysis is pivotal for refining pandemic preparedness and response, enabling the development of targeted, evidence-based policies. Future research will focus on leveraging the KSI to evaluate the relationship between policy stringency and infection trends, aiming to optimize public health interventions for better disease management and prevention.

In conclusion, 'K-Track-Covid' is an innovative tool in epidemiological analysis, offering comprehensive insights into the spread and impact of COVID-19 in South Korea. Its predictive capabilities and user-friendly interface make it an asset for a wide range of stakeholders. By facilitating a deeper understanding of the pandemic's dynamics, 'K-Track-Covid' contributes significantly to informed decision-making in public health. As the global community continues to navigate the challenges of the COVID-19 pandemic and future public health crises, tools like 'K-Track-Covid' will be essential in guiding effective and timely responses.

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.

Author contributions

HS: Writing – original draft, Writing – review & editing, Data curation, Formal analysis, Software, Visualization. KH: Formal analysis, Visualization, Writing – review & editing. JP: Writing – review & editing. ZL: Writing – review & editing. TG: Writing – review & editing, Writing – original draft. AK: Writing – review & editing, Methodology. TP: Writing – review & editing, Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing – original draft.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This research was supported by research grants from the Ministry of Science and ICT, South Korea (No. 2021M3E5E3081425).

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

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2024.1347862/full#supplementary-material

References

1. Coronavirus disease (COVID-19) pandemic. Available at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019.

2. World Health Organization. WHO COVID-19 dashboard. Available at: https://covid19.who.int/ (2020)

3. Dong E, Du H, Gardner L. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect Dis.* (2020) 20:533–4. doi: 10.1016/S1473-3099(20)30120-1

4. CDC Center for Disease Control and Prevention. COVID-NET interactive dashboard. Available at: https://www.cdc.gov/coronavirus/2019-ncov/covidnetdashboard/de/powerbi/dashboard.html (2023)

5. CDC A. COVID-19 Africa Hotstpot dashboard. Available at: https://africacdccovid.org/ (2024)

6. Chang W, Chung J, Allaire J, Sievert C, Schloerke B, Xie Y, et al. Web application framework for R. (2023) Available at: https://cran.r-project.org/web/packages/shiny/shiny.pdf

7. Kim J-H, An JA-R, Min P-k, Bitton A, Gawande AA. How South Korea responded to the Covid-19 outbreak in Daegu. *NEJM Catalyst*. (2020) 1:4 doi: 10.1056/CAT.20.0159

8. Agency KDCaP. Korea centers for disease control and prevention. Available at: https://www.kdca.go.kr/ (2023)

9. Wickham H. Tools for working with URLs and HTTP. Available at: https://CRAN.R-project.org/package=httr (2023)

10. Wickham HHJ, Ooms J. Parse XML. Available at: https://CRAN.R-project.org/package=xml2 (2023)

11. KOrean statistical information service, KOSIS. Available at: https://kosis.kr/eng/.

12. World Pop. South Korea 2020 UN adjusted. In: School of Geography and Environmental Science UoSDoGaG, University of Louisville; Departement de Geographie, Universite de Namur) and Center for International Earth Science Information Network (CIESIN), Columbia University, editor (2018).

13. Wickham HFR, Henry L. Dplyr: a grammar of data manipulation. Available at: https://CRAN.R-project.org/package=dplyr (2023)

14. Wickham HAM, Bryan J, Chang W, McGowan LD, François R, et al . Welcome to the tidyverse. (2023) Available at: https://cran.r-project.org/web/packages/tidyverse

15. Pebesma E. Simple features for R: Standardized support for spatial vector data. Available at: https://journal.r-project.org/archive/2018/RJ-2018-009/RJ-2018-009.pdf (2018)

16. Hijmans RJ. Geographic data analysis and modeling. Available at: https://CRAN.R-project.org/package=raster (2021)

17. Sievert C. Interactive web-based data visualization with R, plotly, and shiny. Available at: https://plotly-r.com/ (2019)

18. Wickham H, Chang W, Henry L, Pedersen TL, Takahashi K, Wilke C, et al. Create elegant data visualisations using the grammar of graphics. (2023) Available at: https://cran.r-project.org/web/packages/ggplot2/index.html

19. Cheng J, Schloerke B, Karambelkar B, Xie Y, Wickham H, Russell K et al. Create interactive web maps with the JavaScript 'Leaflet' Library. (2023) Available at: https://CRAN.R-project.org/package=leaflet

20. Karambelkar B, Schloerke B, Zheng B, Cura R, Voge M, Dumke M, et al. Extra functionality for 'leaflet' package. (2022) Available at: https://CRAN.R-project.org/package=leaflet.extras

21. Stoyanov SR, Hides L, Kavanagh DJ, Zelenko O, Tjondronegoro D, Mani M. Mobile app rating scale: a new tool for assessing the quality of health mobile apps. *JMIR Mhealth Uhealth*. (2015) 3:e27. doi: 10.2196/mhealth.3422

22. Contributors O. *Open street map foundation*. Available at: https://www.openstreetmap.org (2018)

23. Attali D. Easily improve the user experience of your shiny apps in seconds. Available at: https://CRAN.R-project.org/package=shinyjs (2022)

24. Korolev I. Identification and estimation of the SEIRD epidemic model for COVID-19. *J Econom*. (2021) 220:63–85. doi: 10.1016/j.jeconom.2020.07.038

25. Viguerie A, Lorenzo G, Auricchio F, Baroli D, Hughes TJ, Patton A, et al. Simulating the spread of COVID-19 via a spatially-resolved susceptible–exposed–infected–recovered–deceased (SEIRD) model with heterogeneous diffusion. *Appl Math Lett*. (2021) 111:106617. doi: 10.1016/j.aml.2020.106617

26. Kannan D, Gurusriram R, Banerjee R, Bhattacharjee S, Varadwaj PK. Will there be a third COVID-19 wave? A SVEIRD model-based study of India's situation. *Indian J Phys Proc Indian Assoc Cultiv Sci (2004)*. (2021) 95:2513–21. doi: 10.1007/s12648-021-02196-w

27. Ye Y, Zhang Q, Wei X, Cao Z, Yuan H-Y, Zeng DD. Equitable access to COVID-19 vaccines makes a life-saving difference to all countries. *Nat Hum Behav*. (2022) 6:207–16. doi: 10.1038/s41562-022-01289-8

28. Kermack WO, McKendrick AG. A contribution to the mathematical theory of epidemics. *Proc. R. Soc. London, Ser. A*. (1927) 115:700–21. doi: 10.1098/rspa.1927.0118

29. Cobb L, Krishnamurthy A, Mandel J, Beezley JD. Bayesian tracking of emerging epidemics using ensemble optimal statistical interpolation. *Spat. Spatiotemporal Epidemiol.* (2014) 10:39–48. doi: 10.1016/j.sste.2014.06.004

30. Apio C, Han K, Lee D, Lee B, Park T. Development of new stringency indices for nonpharmacological social distancing policies implemented in Korea during the COVID-19 pandemic: random Forest approach. *JMIR Public Health Surveill*. (2024) 10:e47099. doi: 10.2196/47099

frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 25 November 2024
DOI 10.3389/frai.2024.1482141

OPEN ACCESS

EDITED BY
Jasleen Kaur,
University of Waterloo, Canada

REVIEWED BY
Antonio Sarasa-Cabezuelo,
Complutense University of Madrid, Spain
Tetyana Romanova,
University of Leeds, United Kingdom
Andrzej Kotyra,
Lublin University of Technology, Poland

*CORRESPONDENCE
Sergiy Yakovlev
✉ s.yakovlev@khai.edu

RECEIVED 17 August 2024
ACCEPTED 06 November 2024
PUBLISHED 25 November 2024

CITATION
Barmak O, Krak I, Yakovlev S, Manziuk E, Radiuk P and Kuznetsov V (2024) Toward explainable deep learning in healthcare through transition matrix and user-friendly features.
Front. Artif. Intell. 7:1482141.
doi: 10.3389/frai.2024.1482141

Toward explainable deep learning in healthcare through transition matrix and user-friendly features

Oleksander Barmak[1], Iurii Krak[2,3], Sergiy Yakovlev[4,5]*, Eduard Manziuk[1], Pavlo Radiuk[1] and Vladislav Kuznetsov[2,3]

[1]Department of Computer Science, Khmelnytskyi National University, Khmelnytskyi, Ukraine, [2]Department of Theoretical Cybernetics, Taras Shevchenko National University of Kyiv, Kyiv, Ukraine, [3]Laboratory of Communicative Information Technologies, V.M. Glushkov Institute of Cybernetics, Kyiv, Ukraine, [4]Department of Mathematical Modeling and Artificial Intelligence, National Aerospace University "Kharkiv Aviation Institute", Kharkiv, Ukraine, [5]Institute of Computer Science and Artificial Intelligence, V.N. Karazin Kharkiv National University, Kharkiv, Ukraine

Modern artificial intelligence (AI) solutions often face challenges due to the "black box" nature of deep learning (DL) models, which limits their transparency and trustworthiness in critical medical applications. In this study, we propose and evaluate a scalable approach based on a transition matrix to enhance the interpretability of DL models in medical signal and image processing by translating complex model decisions into user-friendly and justifiable features for healthcare professionals. The criteria for choosing interpretable features were clearly defined, incorporating clinical guidelines and expert rules to align model outputs with established medical standards. The proposed approach was tested on two medical datasets: electrocardiography (ECG) for arrhythmia detection and magnetic resonance imaging (MRI) for heart disease classification. The performance of the DL models was compared with expert annotations using Cohen's Kappa coefficient to assess agreement, achieving coefficients of 0.89 for the ECG dataset and 0.80 for the MRI dataset. These results demonstrate strong agreement, underscoring the reliability of the approach in providing accurate, understandable, and justifiable explanations of DL model decisions. The scalability of the approach suggests its potential applicability across various medical domains, enhancing the generalizability and utility of DL models in healthcare while addressing practical challenges and ethical considerations.

KEYWORDS

healthcare, artificial intelligence, deep learning, medical signal processing, medical image analysis, model interpretability

1 Introduction

The rapid development of AI has made it important to explain the decisions made by AI systems, a concept known as explainable artificial intelligence (XAI) (Confalonieri et al., 2020). Within AI, machine learning (ML) encompasses various algorithms and models, including traditional ML methods and DL techniques. DL, a subset of ML, utilizes neural networks with multiple layers to model complex patterns in data. However, DL models often suffer from the "black box" problem, where their internal decision-making processes are not transparent, limiting their trustworthiness in critical applications like healthcare (Hassija et al., 2024).

It is also worth noting that XAI implements the "right to explanation" (Vredenburgh, 2022), that is, the right to have a clear explanation of the result of the algorithm's work. This right applies to each of us when the algorithm's decision directly affects a person. Such rights are already being developed, although the general «right to explanation» is still under

discussion. In the information society, the "right to explanation" is becoming an extremely important concept, as digital technologies, AI, and ML will continue to be actively applied to solving various problems of human activity (Venkatesan et al., 2023; Longo et al., 2024).

Our study enhances the role of AI in resource distribution and strategic decision-making by making DL model decisions more interpretable for healthcare providers. This interpretability is crucial for effective decision-making and resource management in health crises like the COVID-19 pandemic (Zaoui et al., 2023). Moreover, the paper addresses the ethical, legal, and societal dimensions of AI by emphasizing transparency and trustworthiness in AI applications. The proposed methods ensure that AI decisions are accurate, understandable, and justifiable by establishing clear criteria and metrics. We define "understandable" as the degree to which healthcare professionals can comprehend the model's decision-making process through interpretable features that are directly related to clinical knowledge. "Justifiable" refers to the model's ability to provide explanations that are supported by clinical guidelines and empirical evidence. These criteria are quantitatively assessed using statistical metrics such as Cohen's Kappa coefficient to measure agreement between model explanations and expert annotations, and by evaluating the consistency of the model's decisions with established medical standards.

In this study, we aim to address the issue of explaining decisions made by AI. Previously, in Radiuk et al. (2024), we proposed an approach to explain the results of a DL model by mapping its decisions to those of a traditional ML model using a transition matrix. This approach requires both a DL model and a corresponding ML model trained on the same data. However, in practice, we often have a DL model without an equivalent ML model. In this case, we cannot apply the specified approach to explaining the decisions made by the DL model.

Therefore, the objective of this study is to apply our approach to the case of medical data processing, where there is no ML model corresponding to the DL model. Instead, we consider a set of features that are understandable to healthcare experts. Using these features, we propose interpreting the decisions obtained by the DL model. To fulfill the study's goal, it is essential to develop a new scalable visual analytics approach. The scalability of our approach lies in its ability to be applied across different types of medical data and tasks without the need for retraining the DL model or developing new ML models for each case. By utilizing a transition matrix to bridge the DL model's decision-making with expert-defined features, the method can be adapted to various medical signal and image processing applications. This adaptability allows for efficient extension to new datasets and clinical problems, thereby enhancing its practical utility in diverse healthcare settings. Finally, the main contribution of this work is the scalable approach to explain the results obtained by DL models, based on features understandable to healthcare experts for medical signal and image processing tasks.

The structure of the article is as follows. The following section presents an analysis of the current state of the problem under study. Section 3 describes the proposed scalable approach to presenting decisions made by DL models using features understandable to the physician, given the solution of medical signal and image processing problems. Section 4 presents the results of computational experiments and their interpretation. Finally, Section 5 concludes the results obtained and suggests further directions of this research.

2 State of the arts

In general, it is believed that XAI adheres to three principles: transparency, interpretation, and explanation (Phillips et al., 2021). We can talk about the inherent transparency of XAI if the developer can describe and explain how the model forms and updates parameters from statistical training data and how it makes predictions on new data (Pääkkönen and Ylikoski, 2021). By interpretation of XAI, we mean understanding how the AI model forms its output data and explaining its decisions to people (Räuker et al., 2023). Explanation in XAI is an important concept but without a clear definition. It is believed that AI explanation in a broad sense is a set of features that influenced the decision (i.e., classification or prediction) for a specific case (Notovich et al., 2023). If AI-based approaches meet these requirements, then they are said to provide the basis for justifying decisions, tracking and verifying them, and improving and researching new facts (Kim et al., 2023).

Explainable artificial intelligence issues are especially critical in areas such as medicine, defense, finance, and law, where it is important to understand AI decisions and trust them (Manziuk et al., 2021; Mora-Cantallops et al., 2024). Today there are many approaches that provide decent results in various tasks of such sensitive areas of human activity (Wang and Chung, 2021). In general, DL methods provide better results compared to traditional ML methods for solving problems with heterogeneous data (Krak et al., 2023). In particular, convolutional neural network (CNN) models (Radiuk et al., 2021) are state-of-the-art for computer vision tasks (Smith et al., 2021), and transformer models are state-of-the-art for natural language processing tasks (Khurana et al., 2023). However, as already mentioned, decisions made by DL methods are not always transparent and understandable.

The field of XAI is experiencing significant advancements, particularly in the development of methods to enhance the transparency of AI models in the healthcare domain. Researchers are actively exploring various approaches, including the construction of feature models and the use of manually crafted features to provide clearer explanations of AI decisions. As an example, Bassiouny et al. (2021) present an innovative approach to diagnosing neonatal lung diseases by training an object detection model, faster-RCNN, to identify seven key lung ultrasound features rather than making direct diagnostic predictions. This methodology enhances the interpretability of the results and keeps clinicians in control by providing annotated images to support their diagnostic decisions. The study demonstrates that the model surpasses single-stage detectors like RetinaNet, achieving high mean average precision, thus balancing performance with trustworthiness in medical practice.

In their review, Salahuddin et al. (2022) explore various interpretability methods for deep neural networks in medical image analysis, emphasizing that these methods aim to enhance transparency and trust in AI systems. They highlight that while these interpretability techniques provide valuable insights, they are often approximations and may not fully capture the true decision-making processes of the models, necessitating cautious application in clinical settings. In addition, Chan et al. (2022) developed and compared three ML

models to predict long-term mortality in critically ill ventilated patients, finding that boosting algorithms and logistic regression achieved similar performance.

Similarly, Lu et al. (2023) propose a comprehensive workflow that includes a step where medical professionals label differential diagnosis features according to medical guidelines, effectively blacklisting irrelevant features extracted from electronic health records. This approach aims to "reduce workloads of clinicians in human-in-loop data mining" by focusing on feature oversight rather than full prediction, thus enhancing the trustworthiness and efficiency of the AI model.

In Moreno-Sánchez (2023), a heart failure survival prediction model is enhanced by integrating explainable AI techniques, aiming to balance predictive performance and interpretability. This approach provides transparency by explaining feature contributions to predictions, making the model's decision-making process clearer for clinicians. Consequently, it fosters greater trust and practical adoption in clinical settings.

Pintelas et al. (2023) introduce a novel framework for 3D image recognition that utilizes interpretable features such as lines, vertices, and contours to enhance explainability. This approach is particularly promising for medical imaging, achieving performance comparable to state-of-the-art black-box models while maintaining transparency. However, the development of interpretable methodologies for 3D image segmentation remains an emerging area of research, with most existing techniques originally designed for 2D image classification tasks.

Based on the analysis of existing literature, we identified a lack of clear methodologies for constructing feature models that enhance the interpretability of DL models in medical applications. The primary goal of this study is to enhance the decision-making processes of DL models in processing medical signals and images by introducing a novel scalable approach that translates complex model outputs into interpretable features understandable to healthcare professionals.

The main scientific contributions of this work are:

- We introduced a new scalable visual analytics approach that utilizes a transition matrix to bridge the DL model's decision-making with interpretable features defined by experts.
- Our approach systematically incorporates clinical guidelines and expert rules into the feature selection and model development process.
- We applied and validated our approach on two distinct medical datasets–ECG signals for arrhythmia detection and MRI scans for heart disease classification–achieving strong agreement with expert annotations (Cohen's Kappa coefficients of 0.89 and 0.80, respectively).

3 Materials and methods

3.1 Basic approach

In Radiuk et al. (2024), we addressed the problem of explaining decisions made by DL models by establishing a relationship between the features learned by a DL model and those used in a traditional ML model. This idea is illustrated in Figure 1.

The process described above involves the formation of ML models, which have all the necessary features of understandable AI: transparency, interpretability, and explainability. Otherwise, these features (areas of attention) are formed according to certain algorithms (DL models) and, as a result, are not entirely clear, or not at all clear to the end user.

It is worth noting that there are also intermediate cases when the signs are "in the middle" between the indicated cases, such as:

- Decomposition of understandable areas of attention into "incomprehensible" signs, both with the possibility of reverse transformation and without such possibility.
- In addition to features understandable to the public of experts, there may be separate features (or combinations of previously obtained ones) that are understandable to more experienced experts or are based on intuition; here, for each specific case, the community decides whether these cases are transparent or not.

As a result of the above, let us have one training sample and two models with features. One model is built by the DL model, and the other model has features formed by an expert.

Next, we formalize the problem under consideration. We represent the features of the DL model in the form of matrix A of dimension $m \times k$ and the features of the other model in the form of matrix B of dimension $m \times l$ as follows:

$$m\left\{\underbrace{\begin{pmatrix} a_1^1 & a_2^1 & \cdots & a_k^1 \\ a_1^2 & a_2^2 & \cdots & a_k^2 \\ \cdots & \cdots & \cdots & \cdots \\ a_1^m & a_2^m & \cdots & a_k^m \end{pmatrix}}_{k}\right. = A \tag{1}$$

$$m\left\{\underbrace{\begin{pmatrix} b_1^1 & b_2^1 & \cdots & b_l^1 \\ b_1^2 & b_2^2 & \cdots & b_l^2 \\ \cdots & \cdots & \cdots & \cdots \\ b_1^m & b_2^m & \cdots & b_l^m \end{pmatrix}}_{l}\right. = B \tag{2}$$

where m is the number of vectors obtained from the training sample during DL model training, k represents the number of features, and l stands for the number of features.

We emphasize once again that the features formalized in formulas (1, 2) are obtained from the same training sample. We also note that in general k can be equal to, less than, or greater than l.

In practical problems that are modeled in this way, that is, in the presence of two mappings for the same objects for different sets of features, it is often necessary to express feature vectors of different dimensions through each other. In other words, consider the problem

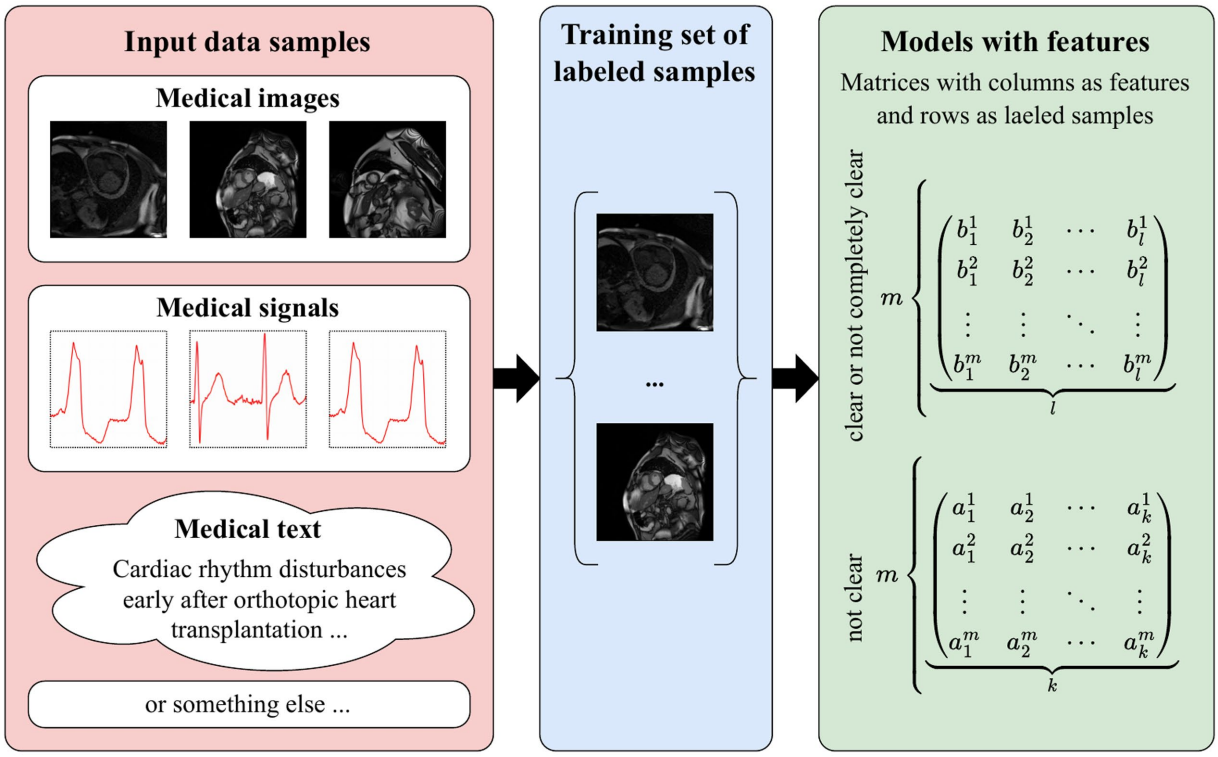

FIGURE 1
The basic idea of our approach: The process of converting various medical data types–such as images, signals, and text–into a labeled training dataset, which is subsequently transformed into feature matrices where rows represent labeled samples and columns correspond to extracted features, organized by clarity for model development.

where for different matrices A and B it is necessary to find such a matrix T that the following equality holds:

$$B = TA \tag{3}$$

where T is the transition matrix between matrices A and B.

Note that in linear algebra, formula (3) is a usual change of basis of a vector space, and if the condition $m = k = l$ is met, finding matrix T is trivial, that is:

$$T = BA^{-1} \tag{4}$$

For the case under consideration, $m \neq k \neq l$, the inverse matrix does not exist, and therefore it is proposed to apply a generalization of the inverse matrix—the pseudo-inverse matrix (Cvetković Ilić and Wei, 2017). We propose to find such a matrix T of dimension $k \times l$, which provides the transition between matrices A and B:

$$AT \approx B \tag{5}$$

Note that the approximation in formula (5) is established with respect to the Euclidean norm in the feature space of the matrices. It is proposed to find matrix T as follows:

$$T \approx A^{+}B \tag{6}$$

In practice, it is proposed to define A^{+} using SVD decomposition (Kalman, 2002), even though other methods are described in Krak et al. (2020):

$$A^{+} = V\Sigma^{+}U^{T} \tag{7}$$

where $A = V\Sigma U^{T}$ is the singular value decomposition of matrix A, the matrix Σ^{+} is formed by transposing the matrix Σ and replacing all its non-zero values of diagonal elements with inverse ones:

$$\Sigma^{+} = \left[D^{+} : 0 \right], \quad m \geq k;$$

$$\Sigma^{+} = \begin{bmatrix} D^{+} \\ \cdots \\ 0 \end{bmatrix}, \quad m < k;$$

$$D^{+} = diag\left(\sigma_1^{+}, \sigma_2^{+}, \sigma_3^{+}, \ldots, \sigma_l^{+}\right), \quad \sigma_i^{+} = \begin{cases} \dfrac{1}{\sigma_i^{+}}, \sigma_i > 0; \\ 0, \sigma_i = 0. \end{cases}$$

Therefore, for an arbitrary row vector of features a_j^{*}, $i = \overline{1,k}$, obtained by the model defined by matrix A, the corresponding row vector of features b_i^{*}, $j = \overline{1,l}$, by the model defined by matrix B, is determined using the obtained transition matrix T as follows:

$$b_i^{*} = a_j^{*}T, \quad i = \overline{1,k}, \quad j = \overline{1,l} \tag{8}$$

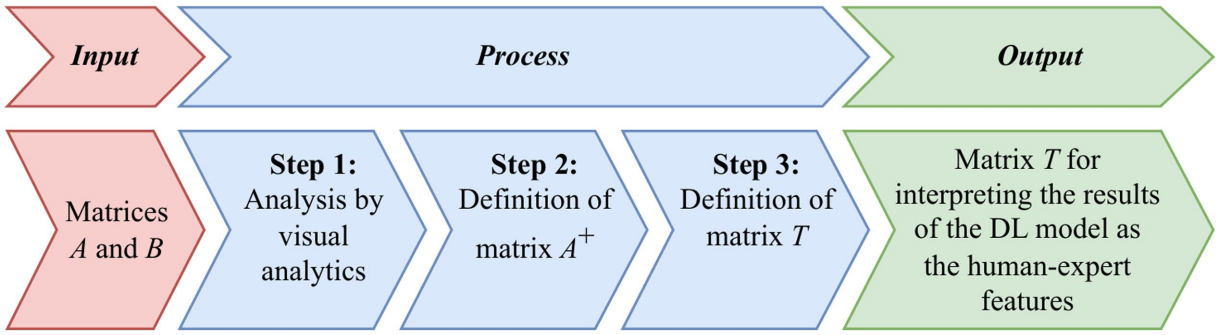

FIGURE 2
Diagram of the basic approach to derive the transition matrix T from input matrices A and B, involving three main steps: visual analytics of the matrices, defining matrix A^+, and establishing matrix T, which serves to interpret DL model outputs in terms of human-expert features.

The approach described above by formulas (1–8) somehow correlates with approximation, that is, the description by one function, even given in tabular form, of a given form of another function, perhaps also in tabular form.

There are several approaches to data approximation. One of them consists in approximating a complex function with a simpler function, which is used for all tabular values, but it is not necessary that it passes through all points. This approach is also called curve fitting, which is sought to be carried out so that its deviation from the tabular data is minimal. The authors propose to use the transition matrix T according to formula (4) between two feature models, presented in the form of matrices, for the same set of input data as such a function.

Figure 2 briefly shows the main steps of the basic approach, first proposed in Radiuk et al. (2024) to obtaining the transition matrix T.

First, we extract two matrices, A from the DL model and B from the ML model, both representing the same data samples.

Step 1: Use a visual analytics tool, either Principal Component Analysis (PCA) by Pearson (1901), or t-distributed Stochastic Neighbor Embedding (t-SNE) by Hinton and Roweis (2002), to map these high-dimensional feature vectors onto a two-dimensional plane.

A dimensionality reduction technique is aimed at preserving the local structure of the data and reveal clusters, allowing us to visually compare the feature vectors from the DL model and the interpretable feature set. By ensuring their relative positions match across models, we facilitate the accurate computation of the transition matrix T.

Step 2: Compute the pseudo-inverse matrix A^+.

Step 3: Calculate the transition matrix T.

Finally, use matrix T to translate DL results into features understandable by the ML model.

3.2 The proposed scalable approach

To overcome the absence of a corresponding ML model, we propose a scalable approach that constructs matrix B using expert-defined interpretable features. This approach allows us to apply the transition matrix method to enhance the interpretability of DL models.

The proposed scalable approach is aimed at simplifying complex, hard-to-understand features from a DL model into a more user-friendly form, making the results easier to interpret. The extracted feature vector, which is the penultimate layer in a DL model, is transformed using a transition matrix T by formula (6) to produce results that are understandable to the end user.

Suppose there is an expert in the subject area of the problem under consideration (i.e., the end-user) who compiles an exhaustive list of features by which they determine the belonging of an object to a particular class. Further, for each feature from the list of features, the expert indicates the numerical intervals into which the value of the feature should fall for the classes under consideration. Finally, for each instance (object) from the training dataset, the value of each feature is calculated.

The values of features can be determined in several ways, namely:

- Empirically, using the expert's knowledge of the subject area of the problem under consideration.
- Using formulas or statistical indicators that are understandable to the end user.
- By visual representation (in various ways) of a fragment of a signal or image, in comparison with similar fragments from labeled training data.
- Utilizing visual analytics.
- Using ML models specially built for this case.
- Using DL models specially built for this case.

The selection of interpretable features is guided by the following criteria:

- Clinical relevance: Features must be directly related to clinical indicators that healthcare professionals use for diagnosis and treatment decisions.
- Measurability: Features should be quantifiable using available tools or methods, ensuring consistent measurement across different samples.
- Distinctiveness: Selected features should provide unique information about the data, minimizing redundancy and multicollinearity.
- Expert Consensus: Features should be agreed upon by a panel of experts to reflect standard clinical understanding and practices.

Figure 3 shows the main steps of the method for constructing matrix B, according to the proposed scalable approach.

Below we provide the following steps to build matrix B.

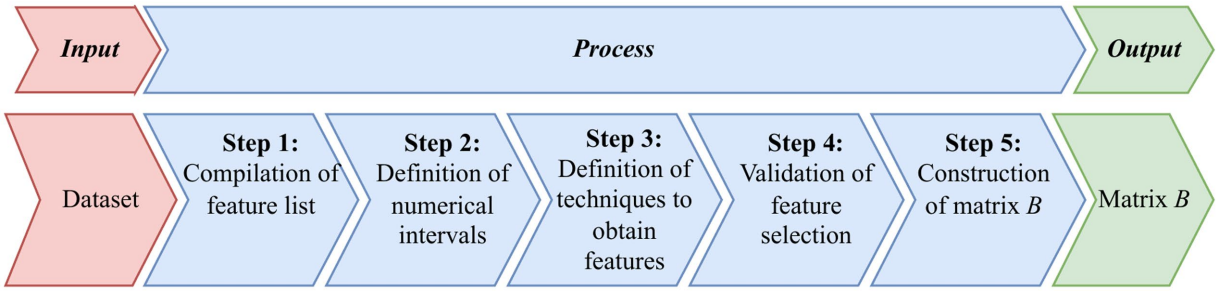

FIGURE 3
Diagram that outlines the proposed sequential approach to constructing matrix *B* from an initial dataset, through steps including compiling a feature list, defining numerical intervals, selecting techniques to extract features, validating feature selection, and ultimately constructing matrix *B* based on the defined transition matrix from the deep learning model to the feature model.

Input information: An expert panel in the subject area of the problem under consideration and the dataset on which the DL model was trained.

Step 1: Compilation of feature list. An expert panel comprising cardiologists and medical imaging specialists compiles a comprehensive list of interpretable features based on clinical guidelines, literature, and diagnostic practices. The goal is to cover all relevant aspects that could influence the classification while ensuring the features meet the criteria above.

Step 2: Definition of numerical intervals. For each feature, the experts define numerical intervals that correspond to different classes or pathological conditions. These intervals are based on clinical thresholds and empirical data, providing a clear delineation between classes.

Step 3: Definition of techniques to obtain features. We establish standardized methods for quantifying each feature:

- Direct measurement: Using signal processing techniques or image analysis tools to extract quantitative values from the data.
- Computational algorithms: Applying validated algorithms (e.g., peak detection algorithms for ECG) to automate feature extraction.
- Statistical analysis: Employing statistical methods to calculate features such as mean values, variances, and ratios that are clinically significant.

Step 4: Validation of feature selection. The selected features undergo validation to ensure reliability and consistency:

- Pilot testing: Features are tested on a subset of the data to assess their discriminative power and measurement consistency.
- Inter-rater reliability: Multiple experts independently measure the features on the same samples to calculate agreement levels, using statistical metrics.
- Refinement: Based on validation results, features may be refined, or additional features may be included to enhance interpretability and accuracy.

Step 5: Construction of matrix *B*. Using the validated methods, we extract feature values for each sample in the training set, forming matrix *B*. This matrix represents the data in terms of interpretable features aligned with expert understanding.

Output information: Matrix *B*.

It should be also noted that the scalability within the proposed scalable approach refers to the following characteristics:

- Adaptability to various medical domains: The method can be efficiently extended to different types of medical data and DL models without substantial changes to the core methodology.
- Ease of integration with expert-defined features: By utilizing a transition matrix and mapping DL outputs to interpretable features, the approach can be applied across different clinical problems with minimal adjustments.

3.3 Evaluation criterion

In this work, Cohen's Kappa coefficient (κ) is used to evaluate the quality of the proposed approach. The κ coefficient is a reliable statistical indicator for evaluating inter-expert reliability for qualitative (categorical) elements. It quantifies the level of agreement between two experts beyond chance.

The formula for Cohen's Kappa coefficient κ is as follows:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{9}$$

where P_o is the level of observed (empirical) agreement between two experts, and P_e is the level of expected (calculated) agreement between the same experts.

For the problem of binary classification of medical signals and/or images with a confusion matrix consisting of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), the elements of formula (9) have the following form:

$$P_o = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}} \tag{10}$$

$$P_e = \left(\frac{(\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TP} + \mathrm{FN}) + (\mathrm{TN} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP})}{(\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN})^2} \right) \tag{11}$$

In formula (10), P_o is the proportion of cases in which the DL model and the human expert come to a consistent decision. Instead, in formula (11), P_e is calculated based on the marginal sum values of the decisions of the two experts.

The value of the κ coefficient according to formula (9) is in the range [−1;1], where 1 denotes perfect agreement, 0 stands for no agreement, and negative values reflect less agreement than expected by chance.

4 Results and discussion

Experimental evaluation of the proposed scalable approach was performed by solving two problems with DL models:

- ECG analysis: Utilized a deep CNN based on the architecture proposed by Kovalchuk et al. (2024).
- MRI analysis: Employed another deep CNN with a U-Net architecture for image segmentation and classification as described in Slobodzian et al. (2023).

Next, we present the results and discussion of the application of the proposed approach to explaining the decisions made by DL models.

4.1 Detection of pathologies of heart activity based on ECG

The proposed approach was validated by the constructed DL_{ECG} model for the problem of detecting pathologies of heart activity (arrhythmias) based on ECG in Kovalchuk et al. (2024). Below we describe the training dataset, the DL_{ECG} model, and the set of features that explained the decisions and results of the proposed approach (the value of κ).

4.1.1 Training dataset and DL model

The problem of detecting pathologies of heart activity (arrhythmias) based on ECG was solved using the reference dataset MIT-BIH Arrhythmia Database (MIT-BIH) (Moody and Mark, 2005). The training of the DL_{ECG} model was performed on 80% of the data from MIT-BIH. Given the annotations of the MIT-BIH set, the following classes/pathologies were selected for the classification problem:

- Normal beat.
- Premature ventricular contraction.
- Paced beat.
- Right bundle branch block beat.
- Left bundle branch block beat.
- Atrial premature beat.
- Fusion of ventricular and normal beat.
- Fusion of paced and normal beat.
- Others.

Input information for training and testing the DL_{ECG} model is presented as a triad of cardiac cycles—in the center is the main cardiac cycle, to which the previous and next cardiac cycles were added (Figure 4).

In this work, the DL_{ECG} model was created based on the modified architecture from Kovalchuk et al. (2024). The classification accuracy for the training set was 99.95%, for the test set—99.13%.

The penultimate layer of the DL_{ECG} model contained 8,192 neurons, and the number of samples in the training sample was 52,180. Accordingly, the size of the A_{ECG} matrix was $m_{ECG} = 52{,}180$—the number of objects from the training subsample of the MIT-BH dataset, $k_{ECG} = 8{,}192$—the number of features formed by the DL_{ECG} model.

4.1.2 Features on ECG for explanation

For the experiment, we focused on detecting PVCs or Ventricular Extrasystole as defined by established clinical guidelines, such as the American Heart Association's recommendations on ECG interpretation. Cardiologists integrated these guidelines to identify key ECG features for PVC detection, applying expert rules to enhance model explanations and performance:

1. Absence of the P wave:

- Integration of clinical guidelines: According to clinical standards, the absence of a P wave preceding a QRS complex suggests ectopic ventricular activity, characteristic of PVCs.
- Expert rules applied: If the P wave is absent or not temporally associated with the QRS complex, it indicates a PVC.
- Method of measurement: Used the NeuroKit2 toolkit (Makowski et al., 2021) to detect P wave presence, ensuring compliance with guidelines for accurate P wave identification.

2. Expanded and deformed QRS complex:

- Integration of clinical guidelines: Clinical guidelines state that PVCs present with widened (≥120 ms) and abnormally shaped QRS complexes due to aberrant conduction pathways.
- Expert rules applied: A QRS duration exceeding 120 ms with atypical morphology is indicative of a PVC.
- Method of measurement: Employed a shallow neural network trained on data annotated per these guidelines to detect QRS abnormalities.

3. Full compensatory pause:

- Integration of clinical guidelines: A full compensatory pause following a PVC is a diagnostic criterion, where the sum of the pre-and post-PVC RR intervals equals twice the normal RR interval.
- Expert rules applied: Applied the rule that $RR_{\text{prev}} + RR_{\text{next}} \approx 2 \times \mathrm{RR}_n$, within a clinically acceptable tolerance.
- Method of measurement: Calculated RR intervals using NeuroKit2, adhering to guidelines for RR interval measurement. In the following subsection, we provide a detailed description of the measurement process.

By integrating clinical guidelines and expert rules into feature selection and measurement, we enhanced the model's explanations

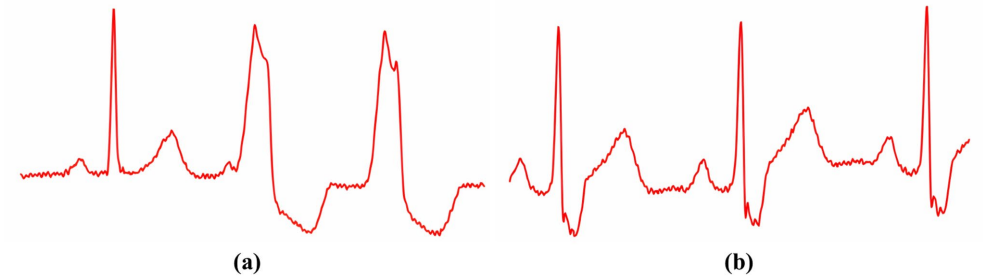

FIGURE 4
This figure presents an ECG signal fragment displaying a triad of cardiac cycles, where panel (A) illustrates a heterogeneous appearance of the R peak, while panel (B) shows a homogeneous expression of the R peak across cycles.

and performance, aligning the DL model's outputs with clinical practice.

4.1.3 Statistical analysis for ECG classification

Given the significant amount of training and the significant time of experts regarding filling in the values of features, non-empirical methods of determining the value of features were used in this work.

For the "Absent P peak" feature, PCA was used. The application of PCA and the reduction of data dimension to 3 made it possible to make sure that the signal fragment with the presence and absence of P peaks is separate. Given this, the presence/absence of the P peak was determined using the Neurokit2.

Visualization of dimensionality reduction by PCA for the "Absent P peak" feature is shown in Figure 5.

For the "Expanded and deformed QRS complex" feature, due to the complexity of its detection by other methods, it is proposed to use a specially trained neural network.

The "Full compensatory pause" feature. A compensatory pause is the time elapsed after an extrasystole until the occurrence of a normal contraction. Therefore, in the case when the extrasystole is located between other extrasystoles, this calculation is not performed and is calculated only for the last case of extrasystole in the sequence.

The presence or absence of this feature was checked as follows:

- Using the Neurokit2 package, the average RR interval between normal cardiac cycles (RR_n) was determined.
- The interval between the R peak with extrasystole and the R peak of the previous cycle (RR_{prev}) was determined.
- The interval between the R peak with extrasystole and the R peak of the next normal cycle ($RR_{next}t$) was determined.

A full compensatory pause was determined under the following conditions:

$$RR_{prev} + RR_{next} = 2 \times RR_n,$$
$$\left(2 \times RR_n - \left(RR_{prev} + RR_{next}\right)\right) < \text{tolerance}$$

According to the above rules, the values of features were determined for each sample from the training set, and, in this way, matrix B was obtained. Further, according to formula (6), the transition matrix T was determined.

Coefficient κ_1 was calculated to evaluate the agreement between the class annotations in the test set and the class predictions made by the DL_{ECG} model. Coefficient κ_2 was calculated to determine the agreement between the class annotations obtained by the DL_{ECG} model and those obtained by the approximated feature values.

The resulting κ_1 was 0.98. To assess the precision of this estimate, a 95% confidence interval (CI) was computed, resulting in a CI of 0.96–1.00. Additionally, the associated p value was calculated to be <0.001, indicating that the observed agreement is highly unlikely to be due to chance. The high κ_1 value, combined with the narrow confidence interval, signifies an almost perfect agreement between the expert annotations and the DL_{ECG} model's predictions. This strong agreement is further supported by the p value, which confirms the statistical significance of the result.

The resulting κ_2 was 0.89, with a 95% CI of 0.85–0.93, and a p value of <0.001. This κ_2 value indicates a strong agreement between the model's predictions and the approximated feature-based annotations, although it is slightly lower than κ_1. The slightly broader confidence interval reflects a bit more variability in the agreement, which could be attributed to the approximation process. Nonetheless, the p value still indicates that this agreement is highly significant, and the κ_2 value demonstrates that the expert's features can reliably replicate the decisions made by the DL_{ECG} model.

The comparison between κ_1 and κ_2, along with their respective confidence intervals and p-values, provide a comprehensive understanding of the model's performance. The high κ_1 value, coupled with a narrow confidence interval and a significant p-value, confirms the DL_{ECG} model's capability to accurately classify ECG in alignment with expert annotations. The slightly lower, yet still strong, κ_2 value suggests that while the approximated feature values can effectively mirror the model's decisions, there is a slight decrease in agreement, which may warrant further investigation into the approximation methods, or the features used.

4.2 Detection of pathologies of heart activity based on MRI

The proposed scalable approach was also validated by the DL_{MRI} model for the problem of detecting pathologies of heart activity based on MRIs (Slobodzian et al., 2023).

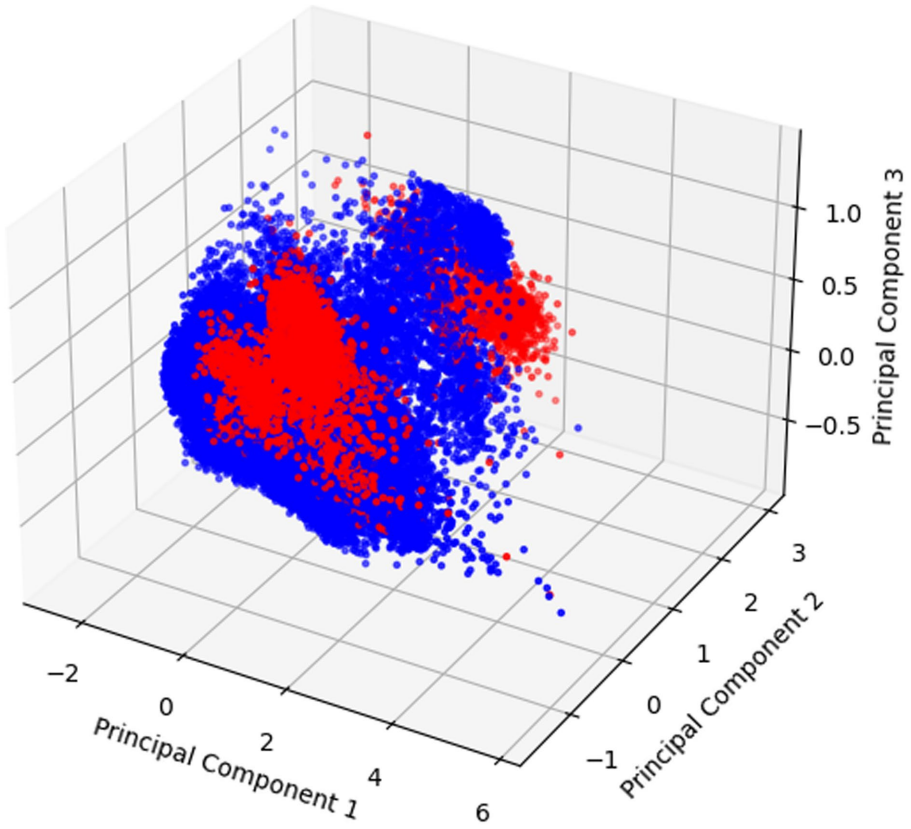

FIGURE 5
The results of applying PCA to classify data based on the "Absent P peak" feature, showing data points distributed across three principal components for visualization of feature separation.

Next, we briefly describe the training dataset of MRIs, the DL_{MRI} model, the set of features that explained the decisions, and the results of the proposed approach (the value of κ).

4.2.1 Training dataset and DL model

For the problem of detecting pathologies of heart activity based on MRIs, a modified dataset of the Automatic Cardiac Diagnosis Challenge (ACDC) (Bernard et al., 2018) was used. Samples of the ACDC set of 100 and 50 patients were used for training and testing the network, respectively. Given the annotations to the ACDC set, the following classes/pathologies were selected for classification:

- Normal condition.
- Dilated cardiomyopathy (DCM).
- Hypertrophic cardiomyopathy (HCM).
- Myocarditis (MINF).
- Arrhythmogenic right ventricular cardiomyopathy (ARV).

An example of presenting input data to the DL model according to the ACDC dataset is illustrated in Figure 6.

The DL_{MRI} model was created based on the modified architecture from Slobodzian et al. (2023). The classification accuracy for the test set of MRIs was over 96.5%. The size of the A_{MRI} matrix was m_{MRI} = 100 – the number of objects from the training subsample of the ACDC dataset, k_{MRI} = 1,024 – the number of features formed by the DL_{MRI} model.

4.2.2 Features in MRI for explanation

For the classification task, 20 features were considered. At the same time, according to the features of identifying pathologies identified by the doctor, the following set of geometric features was formed for further classification:

- The ratio of the volume of the left ventricle to the volume of the right ventricle at the end of systole.
- The volume of the left ventricle at the end of the systole.
- The ratio of the volume of the left ventricle to the volume of the right ventricle at end-diastole.
- The volume of the left ventricle at end-diastole.
- The volume of the right ventricle at end-systole.
- The volume of the right ventricle at end-diastole.
- Ejection fraction of the left ventricle.
- Ejection fraction of the right ventricle.
- The ratio of myocardial volume to left ventricular volume at the end of systole.
- Myocardial mass at the end of diastole.
- Myocardial volume at the end of systole.
- The ratio of myocardial mass to left ventricular volume at the end of diastole.

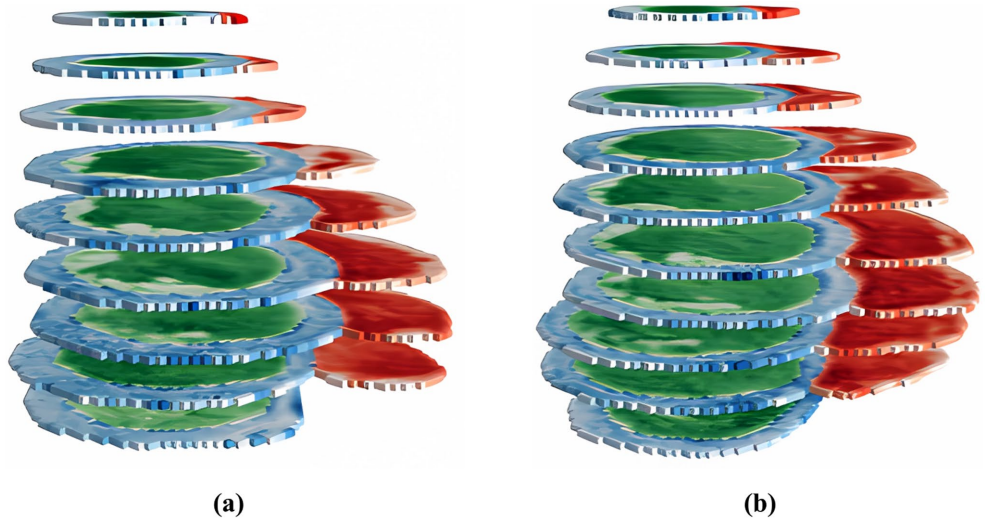

FIGURE 6
A 3D visualization of segmented MRI data prepared for classification, with images from the ES phase in panel (A) and the ED phase in panel (B), while preserving MRI signal intensity values across layers for enhanced phase differentiation.

- Maximum average myocardial wall thickness at end-diastole.
- Maximum average myocardial wall thickness at end systole.
- Mean standard deviation of myocardial wall thickness at end-systole.
- Mean standard deviation of myocardial wall thickness at end-diastole.
- Standard deviation of the standard deviation of myocardial wall thickness at end-diastole.
- Standard deviation of the standard deviation of myocardial wall thickness at end-systole.
- Standard deviation of mean myocardial wall thickness at end-diastole.
- Standard deviation of mean myocardial wall thickness at end-systole.

For the interpretation task, we utilized 20 features selected based on clinical guidelines for diagnosing DCM, such as those provided by the American College of Cardiology and the European Society of Cardiology. Cardiologists integrated these guidelines and applied expert rules to identify key geometric features, enhancing model explanations and performance:

1. Left ventricular end-diastolic volume (LVEDV):

- Integration of Clinical Guidelines: Elevated LVEDV is a primary diagnostic criterion for DCM according to clinical guidelines.
- Expert Rules Applied: An LVEDV exceeding clinically established thresholds, adjusted for body surface area, indicates DCM.
- Method of Measurement: Calculated LVEDV from segmented MRI images using volumetric analysis, following standardized protocols.

2. Left ventricular ejection fraction (LVEF):

- Integration of Clinical Guidelines: Reduced LVEF (<45%) is a key indicator of systolic dysfunction in DCM patients.
- Expert Rules Applied: An LVEF below clinical thresholds signifies impaired cardiac function consistent with DCM.
- Method of Measurement: Derived from LVEDV and end-systolic volume (LVESV) using the formula:

$$\mathrm{LVEF} = \frac{\mathrm{LVEDV} - \mathrm{LVESV}}{\mathrm{LVEDV}} \times 100\%.$$

3. Myocardial mass at end-diastole:

- Integration of Clinical Guidelines: Changes in myocardial mass are significant in DCM diagnosis, as per clinical standards.
- Expert Rules Applied: Increased myocardial mass beyond normal ranges for a given body size indicates pathological remodeling associated with DCM.
- Method of Measurement: Estimated myocardial mass using standardized techniques, ensuring compliance with clinical measurement guidelines.

By incorporating clinical guidelines and expert rules into the feature selection and quantification process, we ensured that the features are clinically meaningful, thereby enhancing the interpretability and performance of the DL_{MRI} model.

4.2.3 Statistical analysis for MRI classification

For further experiments, the DCM class was chosen. Since the volume of the training set was small, it did not take much time for experts to fill in the values of features. Cardiologists as experts determined the values of features for each sample from the training set and, in this way, matrix B was formed. Further, according to formula (6), the transition matrix T was determined.

For each object from the test set, the values of features were approximated according to formula (8). Subsequently, two key values of κ were also calculated to assess the agreement between the predictions DL_{MRI} of and the expert annotations. These κ values

provide a quantitative measure of the reliability and consistency of the DL_{MRI} model's performance, and they were computed using formula (9).

Coefficient κ_1 was calculated to evaluate the agreement between the class annotations in the test set and the classifications made by the DL_{MRI} model.

Coefficient κ_2 was calculated to determine the agreement between the class annotations obtained by the DL_{MRI} model and those obtained through the approximated feature values.

The resulting κ_1 value was 0.87, indicating a very good level of agreement. To further substantiate this finding, a 95% confidence interval was computed, yielding a range of 0.83–0.91. Additionally, the associated p-value was determined to be <0.001, confirming the statistical significance of this agreement. The relatively narrow confidence interval suggests that the κ_1 value is a stable estimate, and the low p-value strongly supports that the observed agreement is unlikely due to random chance. This high κ_1 value aligns well with the model's overall performance, highlighting the DL_{MRI} model's robust capability in accurately classifying MRI data per expert labels.

The resulting κ_2 value was 0.80, with a 95% confidence interval of 0.76–0.84, and a p value of <0.001. This κ_2 value indicates a significant match, though slightly lower than κ_1, reflecting a strong agreement but with slightly more variability. The broader confidence interval compared to κ_1 suggests some degree of uncertainty, possibly arising from the approximation process or the inherent variability in the feature values. Nonetheless, the significant p value still confirms that this agreement is meaningful and not due to random variation.

The comparison between κ_1 and κ_2, along with their respective confidence intervals and p-values, provides a detailed insight into the DL_{MRI} model's performance and the reliability of the feature approximation approach. The κ_2 value of 0.87, with a narrow confidence interval and a highly significant p-value, underscores the strong alignment between the model's predictions and expert annotations. Meanwhile, the slightly lower κ_2 value of 0.80 suggests that while the approximation method is effective, there is a minor decrease in agreement that could be attributed to the complexity of the MRI data or the approximation method itself.

Overall, the inclusion of these detailed statistical indicators, i.e., κ values, confidence intervals, and p values, adds robustness to the analysis, strengthening the reliability and validity of both DL_{MRI} and DL_{ECG} model's performance and the transparency of the proposed scalable approach.

4.3 Discussion and limitations of the proposed scalable approach

While the proposed scalable visual analytics approach has demonstrated high reliability and enhanced interpretability by showing strong agreement between the DL_{ECG} and DL_{MRI} models and expert annotations, several practical challenges and limitations must be acknowledged when considering real-world healthcare applications.

Firstly, the heterogeneity inherent in real-world healthcare settings presents a significant limitation. Diverse patient populations, varying data quality, and differing clinical protocols across institutions can affect the generalizability and robustness of the model. Implementing the approach across multiple institutions may be challenging due to inconsistencies in data formats, acquisition techniques, and labeling standards. This diversity can lead to models that are tailored to specific datasets and may not perform effectively on unseen data from different sources or populations. Variations in patient demographics and disease prevalence can also impact on the model's ability to generalize, potentially limiting its clinical utility. Addressing this drawback requires extensive validation and potential customization for each setting, which can be resource-intensive.

Secondly, integrating the proposed method into existing clinical workflows poses practical challenges. The approach requires careful planning to ensure it does not disrupt standard practices or burden healthcare professionals. Clinicians may need additional training to interpret the model outputs effectively, and the time required to compute interpretable features and generate explanations must be minimized to be practical in fast-paced clinical environments. This limitation could hinder the adoption of the approach, as any increase in workload or decrease in efficiency is a significant drawback in clinical settings where time and resources are limited. Future research should focus on streamlining the computation processes and developing user-friendly interfaces to facilitate seamless integration.

Data privacy and security concerns present another critical limitation. Utilizing patient data for model training and feature extraction raises significant privacy issues, and ensuring compliance with regulations such as HIPAA (Modifications to the HIPAA Privacy, Security, Enforcement, and Breach Notification Rules, 2013) or GDPR (European Parliament and Council of the European Union, 2016) is essential. There is a risk that the transparency provided by the approach may inadvertently reveal sensitive information or biases present in the data. This limitation is a significant drawback because it can lead to ethical and legal consequences if patient confidentiality is compromised. Implementing secure data handling protocols, anonymization techniques, and establishing ethical guidelines for the use of AI-generated explanations are necessary steps to mitigate these concerns.

The reliance on expert-defined features introduces potential biases and inconsistencies, which is another notable limitation. Creating a comprehensive and universally accepted list of interpretable features is challenging, as experts may have differing opinions on which features are most relevant. This subjectivity can lead to models that only partially capture the diversity of clinical presentations and may overlook subtle but significant patterns present in the data. Additionally, features may exhibit high multicollinearity, where multiple features are correlated with each other, reducing the clarity and effectiveness of the model's explanations. This limitation can result in oversimplified diagnoses and potentially unreliable model explanations, which is a drawback for clinical decision-making. To resolve this, future research should incorporate data-driven feature selection methods alongside expert input, such as PCA, t-SNE or mutual information techniques, to identify relevant features that enhance the model's robustness.

Consistency in determining numerical feature values is also a limitation. Experts may struggle to quantify features invariably, especially those requiring subjective assessment or intricate measurement protocols. This inconsistency can introduce variability into the training data, leading to potentially unreliable model explanations and affecting the model's performance on new data. This is a drawback because it reduces trust in the system's outputs and can hinder clinical adoption. Developing clear guidelines and automated measurement tools can reduce variability in feature quantification.

Leveraging objective, reproducible measurement techniques minimizes reliance on subjective assessments, enhancing consistency across diverse users and settings.

The inherent complexity of DL models poses challenges for interpretability, despite using the transition matrix approach. High-dimensional feature spaces and nonlinear relationships make it challenging to interpret how input data influences the model's predictions fully. This complexity can hinder the ability to provide clear and actionable explanations to healthcare professionals, which is a significant drawback since interpretability is crucial for trust and acceptance in clinical practice. Utilizing model-agnostic interpretability tools, such as SHapley Additive exPlanations (SHAP) values or Local Interpretable Model-agnostic Explanations (LIME), can provide insights into model predictions even in complex models. Simplifying model architectures where possible and focusing on key features can make interpretations more accessible to clinicians.

Finally, the scalability of our approach suggests its potential applicability to a wide range of medical domains beyond ECG and MRI analysis. By mapping complex DL model outputs to interpretable features defined by experts, the method can be adapted to other types of medical data, such as histopathology images, genomic data, or medical text analysis. For instance, in histopathology, features like cell morphology, tissue patterns, and staining intensities could be used to explain DL models classifying cancer subtypes. In genomics, gene expression levels, mutation frequencies, or pathway activations might serve a similar purpose. Moreover, the approach could help interpret DL models processing clinical notes in medical text analysis by linking model outputs to medically relevant terms and concepts.

Addressing these limitations requires a multidisciplinary effort involving clinicians, data scientists, ethicists, and policymakers. We can enhance the model's robustness and interpretability by incorporating data-driven feature selection methods alongside expert input, implementing regularization techniques, and utilizing model-agnostic interpretability tools. Developing clear guidelines and automated tools for consistent feature quantification will improve reliability. Establishing ethical frameworks and ensuring compliance with data privacy regulations will mitigate legal and ethical concerns. Enhancing the scalability and generalization of the approach through flexible frameworks and adaptation techniques is essential for its practical implementation. Future research should focus on these areas to overcome the identified drawbacks and facilitate broader implementation of the approach in real-world healthcare applications.

5 Conclusion

In this study, we introduced a scalable approach designed to make DL model decisions more explainable by mapping them to interpretable features defined by healthcare experts. The criteria for selecting these features were clearly established, integrating clinical guidelines and expert rules to ensure that the features are clinically relevant, measurable, distinctive, and agreed upon by professionals. The approach was rigorously tested on two distinct medical datasets: ECG signals for detecting arrhythmias and MRI scans for classifying heart diseases. The DL models achieved Cohen's Kappa coefficients of 0.89 for the ECG and 0.80 for the MRI datasets, demonstrating strong agreement with expert annotations. These results underscore the reliability of the proposed method in providing accurate, understandable, and justifiable explanations of DL model decisions.

Addressing potential limitations, our approach acknowledges the challenges of feature selection biases, generalization to unseen data, and interpretability in complex models. By incorporating data-driven feature selection methods alongside expert input–such as PCA, t-SNE or mutual information techniques–we can reduce biases and enhance the model's robustness. Implementing regularization techniques, cross-validation, and testing on external datasets can improve generalizability. Utilizing model-agnostic interpretability tools like SHAP values or LIME can provide insights even in complex models, making interpretations more accessible to clinicians.

Overall, our scalable approach enhances the interpretability of DL models in medical applications by providing accurate, understandable, and justifiable explanations according to established medical standards. This positions the method as a valuable tool for integrating AI into diverse areas of healthcare, potentially improving diagnostics, treatment planning, and patient outcomes across various specialties while addressing practical challenges and ethical considerations.

Future work should focus on integrating clinical guidelines and expert rules more systematically into the feature selection and model development process. This integration will enhance model explanations and performance by ensuring that the features and model outputs align with established medical standards. Moreover, improving the feature selection process through standardized and automated methods, along with enhancing the scalability of the approach for adaptation to various medical datasets and clinical environments, will further strengthen the utility and applicability of the method in real-world healthcare settings.

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.

Author contributions

OB: Formal analysis, Writing – review & editing, Methodology, Project administration, Conceptualization. IK: Writing – review & editing, Methodology, Conceptualization, Supervision. SY: Writing – review & editing, Methodology, Conceptualization, Project administration. EM: Writing – original draft, Software, Investigation, Validation. PR: Formal analysis, Investigation, Software, Writing – original draft, Validation. VK: Data curation, Resources, Writing – original draft.

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

References

Bassiouny, R., Mohamed, A., Umapathy, K., and Khan, N. (2021). An interpretable object detection-based model for the diagnosis of neonatal lung diseases using ultrasound images. *Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.* 2021, 3029–3034. doi: 10.1109/EMBC46164.2021.9630169

Bernard, O., Lalande, A., Zotti, C., Cervenansky, F., Yang, X., Heng, P.-A., et al. (2018). Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE Trans. Med. Imaging* 37, 2514–2525. doi: 10.1109/TMI.2018.2837502

Chan, M.-C., Pai, K.-C., Su, S.-A., Wang, M.-S., Wu, C.-L., and Chao, W.-C. (2022). Explainable machine learning to predict long-term mortality in critically ill ventilated patients: a retrospective study in Central Taiwan. *BMC Med. Inform. Decis. Mak.* 22:75. doi: 10.1186/s12911-022-01817-6

Confalonieri, R., Coba, L., Wagner, B., and Besold, T. R. (2020). A historical perspective of explainable artificial intelligence. *WIREs Data Mining Knowledge Discov.* 11:e1391. doi: 10.1002/widm.1391

Cvetković Ilić, D. S., and Wei, Y. (2017). "Completions of operator matrices and generalized inverses" in Algebraic Properties of Generalized Inverses. eds. D. S. Cvetković-Ilić and Y. Wei (Singapore: Springer), 51–88.

European Parliament and Council of the European Union (2016). Regulation (EU) 2016/679 of the European Parliament and of the council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (general data protection regulation). Available online at: https://eur-lex.europa.eu/eli/reg/2016/679/oj (Accessed July 29, 2024).

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., et al. (2024). Interpreting black-box models: a review on explainable artificial intelligence. *Cogn. Comput.* 16, 45–74. doi: 10.1007/s12559-023-10179-8

Hinton, G., and Roweis, S. (2002). "Stochastic neighbor embedding" in *Proceedings of the 15th International Conference on Neural Information Processing Systems*, Cambridge, MA, USA, MIT Press, 857–864.

Kalman, D. (2002). A singularly valuable decomposition: the SVD of a matrix. *Coll. Math. J.* 27, 2–23. doi: 10.2307/2687269

Khurana, D., Koli, A., Khatter, K., and Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. *Multimed. Tools Appl.* 82, 3713–3744. doi: 10.1007/s11042-022-13428-4

Kim, D., Song, Y., Kim, S., Lee, S., Wu, Y., Shin, J., et al. (2023). How should the results of artificial intelligence be explained to users?—research on consumer preferences in user-centered explainable artificial intelligence. *Technol. Forecast. Soc. Chang.* 188:122343. doi: 10.1016/j.techfore.2023.122343

Kovalchuk, O., Barmak, O., Radiuk, P., and Krak, I., (2024). "ECG arrhythmia classification and interpretation using convolutional networks for intelligent IoT healthcare system" in *1st International Workshop on Intelligent & CyberPhysical systems (ICyberPhyS-2024)*. CEUR-WS.org, Aachen, pp. 47–62. Available online at: https://ceur-ws.org/Vol-3736/paper4.pdf (Accessed July 30, 2024).

Krak, I., Barmak, O., Manziuk, E., and Kulias, A. (2020). Data classification based on the features reduction and piecewise linear separation. *Adv. Intellig. Syst. Comput.* 1072, 282–289. doi: 10.1007/978-3-030-33585-4_28

Krak, I., Kuznetsov, V., Kondratiuk, S., Azarova, L., Barmak, O., and Radiuk, P. (2023). "Analysis of deep learning methods in adaptation to the small data problem solving" in Lecture Notes in Data Engineering, Computational Intelligence, and Decision Making. eds. S. Babichev and V. Lytvynenko (Cham: Springer), 333–352.

Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Ser, J. D., et al. (2024). Explainable artificial intelligence (XAI) 2.0: a manifesto of open challenges and interdisciplinary research directions. *Inform. Fusion* 106:102301. doi: 10.1016/j.inffus.2024.102301

Lu, K., Tong, Y., Yu, S., Lin, Y., Yang, Y., Xu, H., et al. (2023). Building a trustworthy AI differential diagnosis application for Crohn's disease and intestinal tuberculosis. *BMC Med. Inform. Decis. Mak.* 23:160. doi: 10.1186/s12911-023-02257-6

Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., et al. (2021). NeuroKit2: a Python toolbox for neurophysiological signal processing. *Behav. Res. Methods* 53, 1689–1696. doi: 10.3758/s13428-020-01516-y

Manziuk, E., Barmak, O., Krak, I., Mazurets, O., and Skrypnyk, T., (2021). "Formal model of trustworthy artificial intelligence based on standardization" in *Proceedings of the 2nd International Workshop on Intelligent Information Technologies & Systems of Information Security (IntelITSIS-2022)*, Khmelnytskyi, Ukraine, 24–26 March 2021. CEUR-WS.org, Aachen, pp. 190–197. Available online at: https://ceur-ws.org/Vol-2853/short18.pdf (Accessed July 30, 2024).

Modifications to the HIPAA Privacy, Security, Enforcement, and Breach Notification Rules (2013). Under the Health Information Technology for Economic and Clinical Health Act and the Genetic Information Nondiscrimination Act; Other Modifications to the HIPAA Rules; Final Rule, 78 Fed. Reg. 5566 (Jan. 25, 2013) (to be codified at 45 C.F.R. pts. 160 & 164).

Moody, G.B., and Mark, R.G. (2005). MIT-BIH arrhythmia database, Software, v. 1.0.0, Data Collection. Available online at: physionet.org (Accessed July 30, 2024).

Mora-Cantallops, M., García-Barriocanal, E., and Sicilia, M.-Á. (2024). Trustworthy AI guidelines in biomedical decision-making applications: a scoping review. *Big Data Cogn. Comput.* 8:73. doi: 10.3390/bdcc8070073

Moreno-Sánchez, P. A. (2023). Improvement of a prediction model for heart failure survival through explainable artificial intelligence. *Front. Cardiovasc. Med.* 10:1219586. doi: 10.3389/fcvm.2023.1219586

Notovich, A., Chalutz-Ben Gal, H., and Ben-Gal, I. (2023). "Explainable artificial intelligence (XAI): motivation, terminology, and taxonomy" in Machine Learning for Data Science Handbook: Data Mining and Knowledge Discovery Handbook. eds. L. Rokach, O. Maimon and E. Shmueli (Cham: Springer), 971–985.

Pääkkönen, J., and Ylikoski, P. (2021). Humanistic interpretation and machine learning. *Synthese* 199, 1461–1497. doi: 10.1007/s11229-020-02806-w

Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. *London Edinburgh Dublin Philos. Magaz. J. Sci.* 2, 559–572. doi: 10.1080/14786440109462720

Phillips, P. J., Hahn, C. A., Fontana, P. C., Yates, A. N., Greene, K., Broniatowski, D. A., et al. (2021). Four Principles of Explainable Artificial Intelligence. Commerce Department, Gaithersburg, MD, USA: NIST, 8312.

Pintelas, E., Livieris, I. E., and Pintelas, P. (2023). Explainable feature extraction and prediction framework for 3D image recognition applied to pneumonia detection. *Electronics* 12:2663. doi: 10.3390/electronics12122663

Radiuk, P., Barmak, O., and Krak, I. (2021). An approach to early diagnosis of pneumonia on individual radiographs based on the CNN information technology. *Open Bioinform. J.* 14, 93–107. doi: 10.2174/1875036202114010093

Radiuk, P., Barmak, O., Manziuk, E., and Krak, I. (2024). Explainable deep learning: a visual analytics approach with transition matrices. *Mathematics* 12:1024. doi: 10.3390/math12071024

Räuker, T., Ho, A., Casper, S., and Hadfield-Menell, D. (2023). "Toward transparent AI: a survey on interpreting the inner structures of deep neural networks" in *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE, New York, NY, USA, pp. 464–483.

Salahuddin, Z., Woodruff, H. C., Chatterjee, A., and Lambin, P. (2022). Transparency of deep neural networks for medical image analysis: a review of interpretability methods. *Comput. Biol. Med.* 140:105111. doi: 10.1016/j.compbiomed.2021.105111

Slobodzian, V., Radiuk, P., Zingailo, A., Barmak, O., and Krak, I., (2023). "Myocardium segmentation using two-step deep learning with smoothed masks by Gaussian blur" in *6th International Conference on Informatics & Data-Driven Medicine (IDDM-2023)*. CEUR-WS.org, Aachen, 77–91. Available online at: https://ceur-ws.org/Vol-3609/paper7.pdf (Accessed July 30, 2024).

Smith, M. L., Smith, L. N., and Hansen, M. F. (2021). The quiet revolution in machine vision—a state-of-the-art survey paper, including historical review, perspectives, and future directions. *Comput. Ind.* 130:103472. doi: 10.1016/j.compind.2021.103472

Venkatesan, V. K., Ramakrishna, M. T., Izonin, I., Tkachenko, R., and Havryliuk, M. (2023). Efficient data preprocessing with ensemble machine learning technique for the early detection of chronic kidney disease. *Appl. Sci.* 13:2885. doi: 10.3390/app13052885

Vredenburgh, K. (2022). The right to explanation. *J Polit Philos* 30, 209–229. doi: 10.1111/jopp.12262

Wang, Y., and Chung, S. H. (2021). Artificial intelligence in safety-critical systems: a systematic review. *Ind. Manag. Data Syst.* 122, 442–470. doi: 10.1108/IMDS-07-2021-0419

Zaoui, S., Foguem, C., Tchuente, D., Fosso-Wamba, S., and Kamsu-Foguem, B. (2023). The viability of supply chains with interpretable learning systems: the case of COVID-19 vaccine deliveries. *Glob. J. Flex. Syst. Manag.* 24, 633–657. doi: 10.1007/s40171-023-00357-w

frontiers | Frontiers in Artificial Intelligence

TYPE Technology and Code
PUBLISHED 18 December 2024
DOI 10.3389/frai.2024.1495074

OPEN ACCESS

EDITED BY
Dmytro Chumachenko,
National Aerospace University – Kharkiv Aviation Institute, Ukraine

REVIEWED BY
Jayakumar Kaliappan,
Vellore Institute of Technology (VIT), India
Siddharth Gosavi,
Sai Gosavi Speciality Clinic, India

*CORRESPONDENCE
Yuri Oliveira
yuriaoliv@gmail.com

RECEIVED 12 September 2024
ACCEPTED 27 November 2024
PUBLISHED 18 December 2024

CITATION
Oliveira Y, Rios I, Araújo P, Macambira A, Guimarães M, Sales L, Rosa Júnior M, Nicola A, Nakayama M, Paschoalick H, Nascimento F, Castillo-Salgado C, Ferreira VM and Carvalho H (2024) Artificial intelligence in triage of COVID-19 patients. *Front. Artif. Intell.* 7:1495074. doi: 10.3389/frai.2024.1495074

Artificial intelligence in triage of COVID-19 patients

Yuri Oliveira[1]*, Iêda Rios[2], Paula Araújo[3], Alinne Macambira[4], Marcos Guimarães[5], Lúcia Sales[6], Marcos Rosa Júnior[7], André Nicola[1], Mauro Nakayama[8], Hermeto Paschoalick[8], Francisco Nascimento[9], Carlos Castillo-Salgado[10], Vania Moraes Ferreira[1] and Hervaldo Carvalho[1]

[1]School of Medicine, University of Brasilia, Brasilia, Brazil, [2]School of Health Sciences, University of Brasilia, Brasilia, Brazil, [3]University Hospital of Brasilia, University of Brasilia, Brasilia, Brazil, [4]Hospital of Tropical Diseases, Federal University of Tocantins, Araguaína, Brazil, [5]University Hospital, Federal University of Vale do São Francisco, Petrolina, Brazil, [6]Institute of Health Sciences, Federal University of Pará, Belém, Brazil, [7]University Hospital Cassiano Antônio de Moraes, Federal University of Espírito Santo, Vitória, Brazil, [8]University Hospital, Federal University of Grande Dourados, Dourados, Brazil, [9]Department of Electrical Engineering, University of Brasilia, Brasilia, Brazil, [10]School of Public Health, Johns Hopkins University, Baltimore, MD, United States

In 2019, COVID-19 began one of the greatest public health challenges in history, reaching pandemic status the following year. Systems capable of predicting individuals at higher risk of progressing to severe forms of the disease could optimize the allocation and direction of resources. In this work, we evaluated the performance of different Machine Learning algorithms when predicting clinical outcomes of patients hospitalized with COVID-19, using clinical data from hospital admission alone. This data was collected during a prospective, multicenter cohort that followed patients with respiratory syndrome during the pandemic. We aimed to predict which patients would present mild cases of COVID-19 and which would develop severe cases. Severe cases were defined as those requiring access to the Intensive Care Unit, endotracheal intubation, or even progressing to death. The system achieved an accuracy of 80%, with Area Under Receiver Operating Characteristic Curve (AUC) of 91%, Positive Predictive Value of 87% and Negative Predictive Value of 82%. Considering that only data from hospital admission was used, and that this data came from low-cost clinical examination and laboratory testing, the low false positive rate and acceptable accuracy observed shows that it is feasible to implement prediction systems based on artificial intelligence as an effective triage method.

KEYWORDS

artificial intelligence, machine learning, clinical data, COVID-19, outcome prediction, prediction algorithms, triage

1 Introduction

Initiated in 2019, Coronavirus Disease (or COVID-19), a respiratory infection caused by Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), rapidly spread worldwide, reaching pandemic status as early as March 2020 (Sott et al., 2022; Mohamed et al., 2020). According to the World Health Organization (WHO), by October 4, 2023, there were already 771,151,224 confirmed cases and 6,960,783 deaths globally (World Health Organization, 2023). COVID-19 is primarily characterized by pneumonia symptoms, including fever, fatigue, and dry cough. However, other symptoms such as gastrointestinal alterations, anosmia, or even ophthalmological changes can also be observed (Luo et al., 2022). Severe cases require intensive care, with the need for endotracheal intubation and mechanical ventilation, leading

to death in extreme cases (Luo et al., 2022; Ou et al., 2020; Martono and Mulyanti, 2023; Fang et al., 2020).

The rapid increase in infection numbers highlighted the unpreparedness of healthcare systems, affecting even developed countries. There were shortages of qualified professionals (Buonsenso et al., 2021; Yoshioka-Maeda et al., 2020), medications (Kanji et al., 2020), equipment (Sandhu et al., 2022), and reagents for laboratory tests (Brinati et al., 2020; Arpaci et al., 2021). To address resource deficiencies, new strategies emerged, including changes in management approaches (Sott et al., 2022; Mohamed et al., 2020; Buonsenso et al., 2021; Yoshioka-Maeda et al., 2020), vaccine distribution (Mohamed et al., 2022), and the development of alternative techniques for treatments and diagnostics (Kanji et al., 2020; Brinati et al., 2020; Arpaci et al., 2021; Kim, 2022). Social impacts were also notable, including worsened economic situations, increased cases of domestic violence, and disruptions in the treatment of other diseases (Sott et al., 2022).

The global crisis and its devastating consequences, along with resource scarcity, have motivated the development of several artificial intelligence (AI) based tools. Algorithms have been implemented to assist in COVID-19 diagnosis, interpretation of imaging exams, prediction of case variations, drug and vaccine development, and prognosis prediction for infected patients (Brinati et al., 2020; Arpaci et al., 2021; Rahman et al., 2021; Yu et al., 2021; Isgut et al., 2023; Shakibfar et al., 2023; Wang et al., 2021; Napolitano et al., 2022; Ramírez-Del Real et al., 2022; Kamel et al., 2023).

Developing robust predictors of morbidity and mortality during a pandemic could aid in strategic healthcare planning, providing early indications, for example, of the need for hospitalization for a group of infected patients (Shakibfar et al., 2023). The quality of data used to construct a predictor is crucial, as incomplete, or poorly processed clinical data can introduce biases that hinder implementation (Isgut et al., 2023).

Being the first major pandemic of the digital era, an enormous amount of information has been collected (Isgut et al., 2023; Napolitano et al., 2022). With the goal of identifying risk factors most associated with COVID-19 severity, dozens of cohort studies, case–control studies, and case series have been conducted (Ou et al., 2020; Martono and Mulyanti, 2023; Fang et al., 2020). Based on this information, AI-based predictors of mortality and morbidity can be developed (Shakibfar et al., 2023).

Artificial Intelligence (AI) is a term that has been widely used in literature, but often with distinct definitions. This becomes more evident when this term is used in conjunction with Machine Learning (ML). Although there is no consensus, it is safe to define AI as a system that seeks to mimic human intelligence. ML, on the other hand, is considered a subset of AI and refers to systems capable of learning from datasets without explicit programming on how to make decisions (Kühl et al., 2022). Another frequently used concept is that of Artificial Neural Networks (ANNs), systems inspired by the functioning of the nervous system and capable of learning from presented data. ANNs can have various architectures and can be defined as a subset of ML (Wang et al., 2021; Haglin et al., 2019).

There are various ML methods, each more suitable for different tasks. Learning can be supervised, semi-supervised, or unsupervised, and tasks include classification, clustering, regression, localization, among others. In a classification task, such as predicting COVID-19 morbidity based on clinical data, several supervised algorithms can be used. These algorithms are called supervised because clinical data is presented with labels, i.e., in conjunction with the outcome for each patient (Uddin et al., 2019).

Within the family of supervised classification methods, different approaches can be taken, such as logistic regression, decision trees, support vector machines, Random Forest, artificial neural networks, among others. Each method has a different approach to data analysis, but they all work by finding parameters that minimize classification error. These parameters represent the learning acquired from data analysis (Uddin et al., 2019).

Artificial Neural Networks (ANNs) have gained popularity due to increased computer processing power and the availability of digital information. These algorithms are inspired by the functioning of the nervous system, where a network of artificial neurons can receive input data, such as laboratory test results, and deliver an output indicating the class to which the individual belongs—such as the presence or absence of a disease. This method involves successive nonlinear transformations to determine whether each unit (or artificial neuron) in the network will be activated, simulating neuronal depolarization (Grossi and Buscema, 2007; Rodvold et al., 2001; Buscema, 2002).

These artificial networks can learn input data patterns so well that they often memorize the labels of the presented data, resulting in overfitting and low generalization capacity for unseen data. Overfitting is a major challenge when developing a machine learning (ML)-based classifier because an overfitted system performs well on training data but poorly when used to classify new patients. An overfitted system can be likened to a doctor who correctly diagnoses only cases for which they have seen the answer before but struggles to recognize diseases in new patients. Various techniques are used to prevent overfitting, including controlling the neural network's size, limiting learning iterations (early stopping), dropout regularization, and others (Grossi and Buscema, 2007; Rodvold et al., 2001; Buscema, 2002; Pansambal and Nandgaokar, 2023; Salehin and Kang, 2023).

Combining the extensive data collection carried out during the pandemic with artificial intelligence methods, several published studies aim at predicting patient outcomes for COVID-19. Fernandes et al. (2021) utilized data from 1,040 patients, incorporating laboratory, clinical, and demographic information. As in this study, only data collected at the time of hospital admission were considered; however, the data were collected from a single hospital. A total of 57 variables were used after excluding those with over 90% missing data and those with a correlation above 0.9. The implemented models demonstrated high predictive capability, with Area Under the Receiver Operating Characteristic Curve (AUC) values exceeding 0.91 in identifying various adverse outcomes.

Kwok et al. (2023) also achieved favorable results in developing a predictor for adverse outcomes in COVID-19 hospitalizations. Using retrospective data from 16 different hospitals, they trained ML models with 92 variables, achieving an AUC of 0.852 and an accuracy of 89%. Notably, they utilized the 6 variables that contributed most to the AI model to construct a new risk score.

In another retrospective study, this time including results from radiological examinations, Tenda et al. (2024) developed another predictor with an AUC of 0.815 in the dataset used to train the classifier and 0.770 in a dataset collected solely for model validation. In another study with retrospective multicenter data, Kamel et al. (2023) were able to predict progression to adverse outcomes. By

comparing different subsets, they determined that hematological variables had the highest predictive capacity for outcomes, but the combination of all collected data resulted in better accuracy. Various tested algorithms showed 90% accuracy, with excellent sensitivity, specificity, and AUC.

Jimenez-Solem et al. (2021) also investigated the predictive capability of severe outcomes using ML. This study stands out for its analysis of different outcomes at various points during patient evaluation. Additionally, they performed external validation, demonstrating that these ML models can be utilized in patients with COVID-19.

Heldt et al. (2021) also achieved satisfactory results using only data collected during the initial moments of medical care. Additionally, Hao et al. (2022) analyzed similar outcomes and raised an important question regarding racial bias in the classifiers, with high false positive rates for hospitalization risk in Black patients.

We identified some research gaps in our literature review. The first observation relates to data collection methods. Most published studies relied on retrospective cohorts, which can introduce biases and confounding factors (Kamel et al., 2023). Additionally, most studies gathered data from a single healthcare center. Single-center studies may limit the generalizability of classification models (Kwok et al., 2023; Tenda et al., 2024; Hao et al., 2022). Possibly due to publication bias, the studies we reviewed did not thoroughly discuss data availability, as we noticed that a large portion of the analyzed variables were available for over 90% of study participants. These near-ideal datasets fail to reflect the reality of many healthcare systems during the health crisis, where multiple tests were not performed due to limited availability. This resource scarcity results in missing-not-at-random (MNAR) data, introducing biases and confounding factors that must be addressed (Isgut et al., 2023).

The main objective of this study is to assess the performance of ML algorithms as outcome predictors during COVID-19 hospitalization using only hospital admission data. Throughout this process, we aim to:

Evaluate the availability of data in real-world scenarios, particularly in resource-constrained environments where access to advanced diagnostic tests is limited, leading to a significant amount of missing data. To achieve this, we conducted a prospective multicenter cohort study, tracking patients admitted with suspected COVID-19 and collecting clinical, laboratory, and demographic information. This approach aimed to minimize the risk of bias and potential confounders typically associated with retrospective or single-center studies.

Discuss different methods for handling far from ideal datasets, with a great quantity of MNAR data and the potential biases they may introduce, covering the entire process from data preprocessing to AI model training.

Analyze the performance of classification algorithms on a dataset that has been preprocessed to reflect the clinical significance of variables, particularly by categorizing physical and laboratory examination results based on established reference values and grouping variables with similar medical significance to reduce sparsity.

Assess the capability of ML algorithms to triage hospitalized COVID-19 patients by predicting which individuals are likely to experience significant deterioration, such as requiring intubation, admission to the Intensive Care Unit (ICU), or facing a fatal outcome.

Assess the impact of different approaches to handling missing data, such as selecting variables with the highest completion rates while excluding subjects lacking data for all selected variables or using mean imputation.

Validate a methodology for the development of tools to assist in combating new diseases, starting from data collection, preprocessing, and experimenting with different algorithms.

The contributions of this research are multifaceted and include de following key points:

1. The study demonstrates the feasibility of using ML algorithms to predict a severe outcome of COVID-19 based on hospital admission data. Utilizing data from a prospective multicenter cohort, the study achieved 80% accuracy and 91% AUC. By using only hospital admission data, we developed systems capable of making early predictions, enabling important conclusions to be drawn at the beginning of a patient's hospitalization. This early identification of at-risk patients can significantly enhance triage processes and improve clinical decision-making, ultimately contributing to better patient outcomes.
2. The article emphasizes the importance of data quality and preprocessing in developing AI predictors, addressing challenges posed by missing data and sparse variables, which are common in clinical studies, especially during crises in healthcare systems. The study employed mean imputation for missing data but acknowledged that other methods may yield better results, while utilizing medical knowledge-based strategies to group and reduce the dimensionality of sparse variables.
3. The research compares the performance of various ML algorithms, including support vector machines, random forests, and dense neural networks, providing a comprehensive analysis of accuracy, sensitivity, specificity, and AUC for each algorithm. Although these techniques are well-established and widely used, the study focuses on their performance on this specific dataset, which has its limitations and underwent a different preprocessing approach.
4. The study contributes to the literature on AI predictors for COVID-19 outcomes, particularly by utilizing data from prospective multicenter studies, which improves generalizability and applicability compared to single-center retrospective studies that may have limited external validity. Moreover, the prospective data ensures better quality control and reduces potential bias.
5. This study develops predictive models using a dataset that mirrors the practical constraints faced by healthcare systems, particularly during times of crisis. Unlike studies that rely on idealized datasets and often overlook challenges in access to diagnostic tests and other resources, our approach incorporates these limitations directly. This enables the creation of models that are not only predictive but also adaptable to real-world healthcare environments where resources are constrained.
6. The article underscores the necessity of integrating medical knowledge into the development and evaluation of AI systems, arguing that AI should complement rather than replace human expertise, while highlighting the importance of considering the feasibility, applicability, and clinical validity of AI predictors.
7. Finally, the study provides a replicable methodology for developing morbidity and mortality predictors that can

be applied to other diseases. This framework, from data collection and preprocessing to algorithm selection and evaluation, can be adapted to different clinical scenarios and contribute to advancing AI-based decision support tools in healthcare.

2 Methods

Our study involved several key steps in the methodology. Initially, we collected data from patients hospitalized with respiratory syndrome. Following data collection, we performed extensive preprocessing, which included cleaning the data and handling missing values. Subsequently, we composed a primary dataset along with various subsets tailored for specific analyses. These subsets were utilized to train different machine learning classifiers, enabling us to assess their performance and effectiveness in predicting outcomes related to the condition under investigation. Figure 1 illustrates this process.

2.1 Data collection

We conducted a prospective, multicenter, concurrent cohort study that included all patients over 18 years of age admitted for respiratory syndrome, with confirmed or suspected COVID-19, at one of the participating hospitals: University Hospital of Brasília (HUB), Brasília-DF, Brazil; Regional Hospital of Asa Norte (HRAN), Brasília-DF, Brazil; Hospital das Clínicas of the Federal University of Minas Gerais (UFMG), Belo Horizonte-MG, Brazil; Cassiano Antônio Moraes University Hospital (HUCAM), Vitória-ES, Brazil; João de Barros Barreto University Hospital of the Federal University of Pará (UFPA), Belém-PA, Brazil; University Hospital of the Federal University of São Francisco Valley (Univasf), Petrolina-PE, Brazil; University Hospital of the Federal University of Grande Dourados (UFGD), Dourados-MS, Brazil; and Tropical Diseases Hospital (HDT-UFT), Araquaína-TO, Brazil. The data collection period was from June 2020 to January 2021 and only pregnant patients were excluded from the study. Cases of respiratory syndrome were defined by the association of general symptoms such as dyspnea, decreased oxygen saturation, cyanosis, and/or radiographic lung changes.

Patients or their legal representatives authorized data collection through an informed consent form, where they were informed about all risks associated with the study, as well as the measures taken to ensure data privacy and protection. They were also notified that they could withdraw from the study at any time. To protect data privacy, identifying features were removed before preprocessing and training, with only an automatically assigned random identification number retained. Access to potentially identifying data was limited to a few

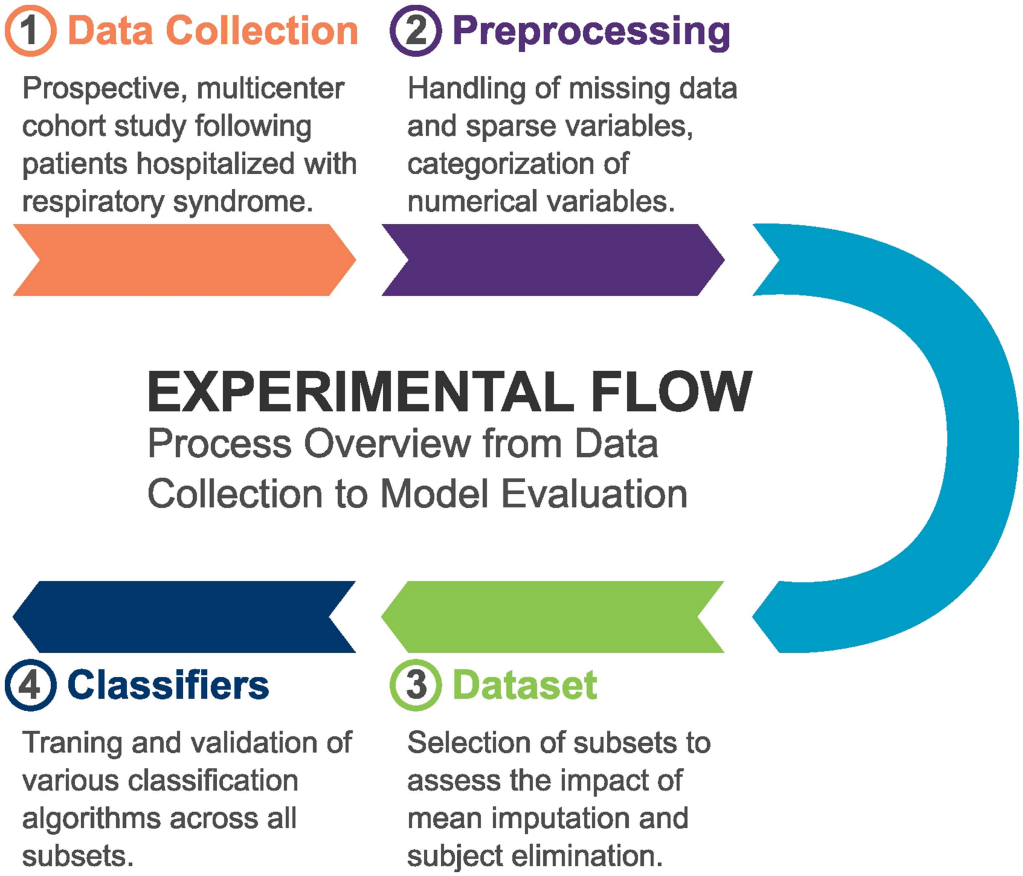

FIGURE 1
Overview of the methodology, illustrating the steps involved in data collection, preprocessing, dataset composition, and the training of machine learning classifiers using various subsets.

team members who ensured its security. The study adhered to the ethical principles outlined in the Declaration of Helsinki and the Belmont Report, and it was approved by an independent ethics committee (Research Ethics Committee from the University of Brasília School of Medicine), under the Certificate of Ethical Appreciation Presentation number 31941420.4.1001.5558 and approval number 4.054.462. No additional interventions were performed beyond those necessary for the proper monitoring of hospitalized patients, such as blood tests and clinical examinations.

Clinical data were collected within the first 72 h of hospitalization, recording the initial clinical examination (including vital signs and identification of comorbidities), demographic information, imaging, and laboratory tests. These tests included Polymerase Chain Reaction (PCR) on sputum, nasal swabs, and serology for SARS-CoV-2, aiming to identify the presence of the coronavirus. After data collection, patients were divided into two groups: those with confirmed SARS-CoV-2 infection (either by PCR or serology) and patients with negative test results. All patients, whether COVID-19 positive or not, were further categorized into those who experienced severe complications and those who did not. Severe complications included patients requiring mechanical ventilation, endotracheal intubation, or ICU admission, as well as those who experienced cardiac arrest, severe sepsis, or who died.

2.2 Data preprocessing

For the development of severity prediction algorithms, the data underwent preprocessing. Due to resource scarcity during the pandemic, several tests could not be conducted on all patients, resulting in a high number of missing data points among individuals. Missing data or incompleteness is a challenge in various studies conducted during the pandemic (Isgut et al., 2023). Initially, variables collected in less than 50% of participants and constant variables (those with the same value for all participants) were removed. For example, the presence of jugular turgor, that wasn't identified in any patient, was excluded.

When dealing with incomplete data, two approaches are possible (Isgut et al., 2023). The first involves excluding individuals who lack data for the algorithm's variables of interest. The second approach is data imputation, where missing values are estimated to retain all individuals in the study. Various imputation methods exist; in this study, data were imputed using the means. For instance, the average hemoglobin value across all patients was used to fill in missing hemoglobin measurements for those who did not undergo the test. Both elimination and mean imputation approaches were employed in this work.

In addition to missing data, clinical conditions that were not observed in a significant number of subjects, sparse data, also pose challenges in developing predictors using data from clinical studies (Tipirneni and Reddy, 2022). Because these clinical variables are positive in a very small proportion of individuals, they do not provide much information to classifiers and increase complexity by adding to the data's dimensionality. Several variables collected in our cohort, particularly prior diseases, symptoms, or physical examination findings, were positive in less than 1% of patients. Although they are sparse, these variables might be important to the prediction and eliminating them from the study might decrease system's accuracy. To reduce data dimensionality without disregarding the importance of these alterations, we cataloged variables suitable for grouping or clustering based on medical knowledge. Thus, variables with similar clinical significance or those composing the same syndrome were combined using a logical "OR" operation, such that the grouping variable would be negative only if all the originating variables were also negative. With the operation, the union of previous Chronic Obstructive Pulmonary Disease, asthma or other chronic lung conditions resulted in a single variable of presence of pre-existing lung diseases. The same approach was applied to symptoms of Upper Respiratory Tract Infection, grouping presence of runny nose, postnasal drip, nasal congestion, sneezing, sore throat, difficulty swallowing and presence of facial sinus compression pain. Nonspecific viral infection symptoms were the union of muscle pain, dizziness, joint pain, skin lesions, diarrhea, abdominal pain, mental confusion, chills, enlarged lymph nodes, nausea, vomiting, loss of appetite and pale mucous membranes. Signs of pulmonary involvement were the union of presence of productive cough, chest pain, cyanosis, chest compression pain, respiratory effort, increased respiratory rate, irregular respiratory pattern, decreased chest expansion, snoring, wheezing, crackles, pinkish sputum, and pulmonary edema.

After this grouping, variables with low representation (less than 5% positive cases) were excluded from the analysis. Variables whose clinical meaning is related to another alteration, such as loss of taste and anosmia, were considered repeated and the ones with lower frequency of positive cases were removed. The distribution of variables for each class (or outcome) was not considered during variable selection to avoid bias in classifier analysis.

All numerical data was categorized. Age was categorized based on percentiles, using 0, 10, 25, 50, 75, 90 and 100% as separators. Thus, patients whose age falls at or below the 10th percentile will be assigned the value 0. Those whose age is above the 10th percentile, but equal or less than the 25th percentile will receive the value 1. Similarly, value 2 corresponds to ages between the 25th and 50th percentiles, and so on. This approach transforms age from years to a class ranging from 0 to 5.

The hemoglobin levels were categorized for female patients using the following intervals: greater than 0 and lower or equal to 6.5 g/dL; greater than 6.5 and lower or equal to 8 g/dL; greater than 8 and lower or equal to 10 g/dL; greater than 10 and lower or equal to 12 g/dL; greater than 12 and lower or equal to 16.5 g/dL; and above 16.5 g/dL. The same was done for male patients, but the value 12 was substituted to 13 g/dL and 16.5 to 18 g/dL. Consequently, a woman with a hemoglobin level of 8 g/dL was assigned category 1, while a man with a hemoglobin level of 13.7 g/dL was assigned category 4. This range of values was inspired by the National Health Institute (NSH) and National Cancer Institute (NCI) classifications for anemia and its severity.

Leukocytes, lymphocytes, platelets, and urea were categorized as low, normal, and high, also inspired by the reference values defined by the National Health Institute. Leukocytes were considered low when equal or lower than 4,000 cells/ mm^3, high when higher than 11,000 cells/ mm^3 and normal in between. For lymphocytes the lower boundary was 1 cell/mm^3 and the higher was 4 cells/mm^3. For platelets the values were 150,000 cells/mm^3 and 400,000 cells/mm^3, respectively, and for urea it was 15.02 mg/dL and 46.84 mg/dL.

Although measured upon patient admission, systolic and diastolic blood pressures were categorized using criteria inspired by the American Heart Association (AHA) classification for hypertension

and hypotension. The range between 140 and 180 mmHg for systolic blood pressure was excluded (as there were only 2 cases above 180 mmHg). This classification approach was chosen solely to use the values as reference points, not for diagnosing or classifying arterial hypertension. Thus, the systolic pressure ranges were the following: greater than 0 and lower or equal to 90 mmHg; greater than 90 and lower or equal to 120 mmHg; greater than 120 and lower or equal to 130 mmHg; greater than 130 and lower or equal to 140 mmHg; and above 140 mmHg. For diastolic pressure, the ranges were: greater than 0 and lower or equal to 60 mmHg; greater than 60 and lower or equal to 80 mmHg; greater than 800 and lower or equal to 90 mmHg; greater than 90 and lower or equal to 120 mmHg; and above 120 mmHg.

The heart rate was categorized using 60 bpm as the cutoff value for bradycardia and 100 bpm as the cutoff value for tachycardia. The respiratory rate was defined as normal up to 16 breaths per minute (bpm), borderline between 16 and 20 bpm, elevated between 20 and 24 bpm, and very high above 24 bpm. Finally, oxygen saturation (SpO2) values equal to or below 88% were considered extremely low, between 88 and 92% inclusive as very low, between 92 and 96% inclusive as low, and above 96% as normal.

This categorization was performed to align the numerical values across different variables. For instance, hemoglobin levels varied around the number 10, while platelet counts ranged around 200,000. The difference in dynamic range and magnitude among variables could potentially lead to poorer performance in certain algorithms. Some algorithms might assign greater importance to specific variables not due to their predictive significance but based on their absolute values. Another approach that could have been used is normalization, which scales all numerical variables between 0 and 1. However, we opted for categorization because it also aligned numerical variables (such as laboratory test results) with categorical ones (such as overall health status). For example, the general condition at admission was classified as follows: 0: good, 1: mildly compromised, 2: moderately compromised, 3: severely compromised, and 4: critical. This categorization also considers how physicians analyze clinical examination results.

2.3 Dataset development

To assess the impact of eliminating individuals with missing data and imputation on algorithm training, we developed a dataset by selecting the 28 most complete variables—those collected from most patients. Using this dataset, we constructed 7 different subsets. One subset included all 421 patients and all 28 variables, with missing values imputed using the mean. In the remaining subsets, no imputation was performed, resulting in data loss due to incomplete records. These 6 subsets were created by varying the number of variables included, ranging from the 23 most complete variables to a set containing all 28 variables. The 7 subsets will be named A to G. The variables and number of subjects of each subset is described below:

1. Subset A, 355 subjects. Contains 23 variables:
   a) Age.
   b) Hemoglobin.
   c) Leukocytes count.
   d) Lymphocytes count.
   e) Platelets count.
   f) Heart rate.
   g) SpO2.
   h) Absence of previous disease.
   i) Systemic Arterial Hypertension.
   j) Diabetes Mellitus.
   k) Obesity.
   l) Renal insufficiency.
   m) Fever.
   n) Dry cough.
   o) Headache.
   p) Weakness.
   q) Anosmia.
   r) Sex.
   s) Upper Respiratory Tract Infection signs.
   t) Dyspnea.
   u) Signs of pulmonary involvement.
   v) Presence of nonspecific viral infection symptoms.
   w) Pre-existing lung diseases.
2. Subset B, 335 subjects. Contains 24 variables:
   a) All 23 variables from subset A.
   b) General condition at admission.
3. Subset C, 309 subjects. Contains 25 variables:
   a) All 24 variables from subset B.
   b) Respiratory rate at admission.
4. Subset D, 285 subjects. Contains 26 variables:
   a) All 25 variables from subset C.
   b) Systolic blood pressure at admission.
5. Subset E, 285 subjects. Contains 27 variables:
   a) All 26 variables from subset D.
   b) Diastolic blood pressure at admission.
6. Subset F, 246 subjects. Contains 28 variables:
   a) All 27 variables from subset E.
   b) Urea.
7. Subset G, 421 subjects. Contains 28 variables:
   a. Same 28 variables from subset F, with data imputation to fill missing values.

The advantage of constructing the subsets in this manner lies in its consideration of data availability when selecting variables, thereby aiming to preserve the maximum number of individuals in the groups while minimizing the elimination of patients with missing data. By reducing the amount of missing data during classifier development, we aimed to mitigate the risk of bias introduced by MNAR data while preserving as much data as possible.

By focusing on the 28 most complete variables, we ensured that the selected variables could be collected across a wide range of patients, even during a healthcare system crisis. This approach suggests that these variables represent more readily available and easily applicable examinations, ultimately enhancing the applicability of the model in real-world clinical settings. Additionally, we aimed to evaluate the balance between maintaining more variables in the subset, which could potentially increase accuracy, and the loss of individuals due to elimination, which could reduce both accuracy and generalization.

2.4 Classifiers development

The data was then split into training and test groups, comprising 80 and 20% of the total patients, respectively. These subsets were used

to train ML algorithms. Random selection between the groups can introduce bias in accuracy analysis; for instance, when individuals that are easier to classify end up in the test group, artificially inflating the system's accuracy. To mitigate this impact, each algorithm underwent training through 5,000 simulations, with each simulation representing a random resampling of the groups and retraining from scratch. This validation technique, known as Monte Carlo Cross-Validation, shows more accurate result even with subsets containing over 1,000 observations (Shan, 2022). During each simulation, accuracy, sensitivity, specificity, negative predictive value, positive predictive value, F1-Score, and the AUC were calculated. The minimum and maximum values, mean, and standard deviation of each metric were evaluated. The libraries used for algorithm training were Lazy Predict 0.2.12, Scikit-Learn 1.2.0 (Buitinck et al., 2011), and Tensorflow 2.10.1 (Abadi et al., 2016), executed in Python 3.9.

The first step involved running 100 simulations using the LazyClassifier function from the Lazy Predict 0.2.12 library, which automates the training of 27 different ML-based algorithms. These algorithms include: Stochastic Gradient Descent Classifier (SGD), Linear Support Vector Classification (Linear SVC), Logistic Regression, Linear Discriminant Analysis, Ridge Classifier with Cross Validation (CV), Ridge Classifier, AdaBoost Classifier, Nearest Centroid, Gaussian Naïve Bayes, Passive Aggressive Classifier, Nu Support Vector Classification (NuSVC), Light Gradient Boosting Machine Classifier (LGBM classifier), Quadratic Discriminant Analysis, Support Vector Classification (SVC), XGBoost Classifier, Bernoulli Naïve Bayes, KNeighbors Classifier, Random Forest Classifier, Bagging Classifier, Perceptron, Extra Trees Classifier, Extra Tree Classifier, Calibrated Classifier with CV, Label Spreading, Label Propagation, Decision Tree Classifier, Dummy Classifier.

The mean accuracy for each algorithm across each subset was recorded. This data-driven approach allowed for a comprehensive analysis to identify the most effective algorithms. Only the seven algorithms that demonstrated the highest mean accuracy were selected for further investigation. Similarly, the single subset that yielded the best accuracy across all algorithms was chosen for use in subsequent steps.

With the selected subset and 7 best-performing algorithms, 5,000 new simulations were conducted using the Scikit-Learn 1.2.0 library. Additionally, the same subset was employed to train 11 distinct dense neural networks architectures. Those networks differ in terms of the number of layers, the number of units per layer, and the dropout rate. By randomly excluding some units, dropout distributes learning across connections and helps reduce overfitting, improving model generalization and neural network performance when faced with unseen data (Salehin and Kang, 2023). The neural networks were built, trained and evaluated using the Tensorflow 2.10.1 library. To test the architecture with best performance, 100 simulations were conducted for each network. The hidden layers of the tested architectures were the following:

1. Single layer: 32 units with 20% dropout.
2. Two-layer: First layer with 32 units and 20% dropout, followed by a second layer with 16 units and 20% dropout.
3. Three-layer: Sequential layers with 32 units (20% dropout), 16 units (20% dropout), and 8 units (20% dropout).
4. Single layer: 16 units with 20% dropout.
5. Single layer: 8 units with 20% dropout.
6. Two-layer: First layer with 8 units and 20% dropout, followed by a second layer with 4 units and 20% dropout.
7. Four-layer: Layers configured with 64 units (20% dropout), 32 units (20% dropout), 16 units (10% dropout), and a final layer of 8 units without dropout.
8. Single layer: 8 units without dropout.
9. Two-layer: First layer with 16 units without dropout, followed by a second layer with 8 units without dropout.
10. Three-layer: Layers arranged with 16 units (no dropout), 8 units (no dropout), and a final layer of 4 units without dropout.
11. Two-layer: First layer with 8 units without dropout, followed by a second layer with 4 units without dropout.

All neural networks were trained using the same subset and therefore shared the same input layer configuration. Their hidden layers uniformly utilized ReLU activation, while their output layers consisted of a single unit with Sigmoid activation. The training process employed Adam optimization and Binary Cross-Entropy as the loss function over 20 epochs. The architecture that yielded the highest validation accuracy was selected for an additional 5,000 simulations.

3 Results

3.1 Cohort outcomes analysis

Out of the 537 study participants, 421 were confirmed to have a diagnosis of COVID-19. Among them, 133 experienced an outcome classified as severe (31.5%). It's important to note that a single patient may have had multiple severe outcomes during the follow-up period, but they were treated the same way as the ones that had a single severe outcome.

3.2 Demographic analysis

The demographic profile of patient groups that progressed with severe and non-severe forms was analyzed, as described in Tables 1, 2. It can be observed that there is a predominance of female patients, of mixed ethnicity, and aged over 50 years old.

3.3 Classifier results

During 100 simulations, the 7 subsets constructed in this study were used to train 27 ML algorithms using the Lazy Predict library. In each simulation, the training and test groups were randomly separated in an 80 to 20% ratio, respectively, and the accuracy of each simulation was computed. The average accuracy was then calculated for each classifier and each subset, resulting in the values shown in Table 3. It was possible to determine that the highest average accuracy was achieved with the Support Vector Classifier (SVC) for the 28-variable subset (subset F, without imputation), and the seven best classifiers for this subset were SVC, NuSVC, Random Forest Classifier, Ridge Classifier CV, Ridge Classifier, Extra Trees Classifier, and Linear Discriminant Analysis.

TABLE 1 Gender, race, and marital status distributions.

	Non severe (absolute risk)	Severe (absolute risk)	Total
Gender			
Male	128 (71.91%)	50 (28.09%)	178 (42.28%)
Female	160 (65.84%)	83 (34.16%)	243 (57.72%)
Race			
White	53 (68.83%)	24 (31.17%)	77 (18.29%)
Black	17 (73.91%)	6 (26.09%)	23 (5.46%)
Mixed	189 (66.55%)	95 (33.45%)	284 (67.46%)
Asian	5 (62.50%)	3 (37.50%)	8 (1.90%)
Indigenous	2 (66.67%)	1 (33.33%)	3 (0.71%)
Chose not to declare	19 (86.36%)	3 (13.64%)	22 (5.32%)
Blank/unfilled	3 (75.00%)	1 (25.00%)	4 (9.50%)
Marital status			
Single	72 (69.23%)	32 (30.77%)	104 (24.70%)
Married	150 (73.17%)	55 (26.83%)	205 (48.69%)
Divorced	26 (72.22%)	10 (27.78%)	36 (8.55%)
Widows	12 (57.14%)	9 (42.86%)	21 (4.99%)
Other	24 (52.17%)	22 (47.83%)	46 (10.93%)
Blank/unfilled	4 (44.44%)	5 (55.56%)	9 (2.14%)

TABLE 2 Age distribution.

	Non severe	Severe	Total
Mean	55.60	61.82	57.57
Median	55	62	57
Standard deviation	15.35	16.54	15.98
Blank/unfilled	1	0	1

Using the Scikit-Learn library, these seven classifiers were employed for 5,000 new simulations, once again with random redistributions between the training and test groups (Monte Carlo cross-validation). This library combines algorithm implementations based on previous works. In these simulations, only subset F was used (which includes 28 variables and has no imputation). Evaluation metrics were computed to calculate the mean and margin of error for a 95% Confidence Interval (CI 95).

In addition to the previously mentioned algorithms, 11 different configurations of dense neural networks were tested on this subset through 100 simulations with random resampling of the training and test groups. These simulations facilitated the identification of the most effective topology or neural network configuration. The classifier's error can be described as a function of the chosen hyperparameters, such as the number of layers and units per layer. The objective is to minimize this error; however, due to computational constraints, it's not feasible to exhaustively test all possible combinations. Consequently, 11 configurations with slight variations were selected, and their average accuracies were computed. The optimal configuration emerged as one with a single hidden layer containing 32 units and a 20% dropout rate. This configuration was subjected to the same training regimen as the seven highest-performing classical ML classifiers, undergoing a Monte Carlo Cross-Validation with 5,000 simulations.

Table 4 describes the results of each algorithm after 5,000 trainings. Very similar values were observed among different algorithms, with a very small margin of error around the average accuracy (CI 95). The high specificity, combined with a significant positive predictive value of the SVC, indicates a high success rate when the algorithm classifies patients as severe. On the other hand, the Ridge Classifier, Ridge CV, and Linear Discriminant Analysis algorithms maintain accuracy while achieving better results in identifying patients who will not progress to the severe form. The neural network, however, exhibited low sensitivity and lower accuracy compared to the other methods. An important metric for classifier evaluation, the AUC, indicates that the tested classifiers had less satisfactory performance in distinguishing between classes. An exception was the Random Forest, with an average AUC of 0.91 and an average accuracy of 80%.

In a context of multiple hospitalized patients and resource scarcity, such as during the COVID-19 pandemic, the achieved predictive values would be extremely useful. The 87% positive predictive value attained by the SVC, coupled with a small number of false positives, allows for better resource allocation and interventions, such as surveillance or transfer to another healthcare unit. It is essential to note that only admission data were used for this analysis, and clinical follow-up during hospitalization, along with additional tests, could lead to patient reclassification and enhance system performance. Also, since the data originated from a prospective cohort, in newer studies the classifiers could be implemented during the cohort, guiding data collection toward most important variables.

Classifiers for COVID-19 severity prediction have been published with accuracies ranging from 74.4 to 95.20% and AUC values between 0.66 and 0.997 (Shakibfar et al., 2023; Wang et al., 2021). However, a significant portion of this research relied on data from retrospective and/or single-center studies, limiting the generalizability of the findings (Wang et al., 2021). In our work, we conducted a prospective multicenter study aimed at ensuring that this methodology can be applied in real-time during new episodes of public health crises.

The availability of data during public health crises poses significant challenges. Resource scarcity in various healthcare centers limits not only the number of clinical variables that can be collected but also the ability to leverage predictive models that require these variables. Previous studies often utilized data collected from healthcare services with greater resource availability, resulting in a higher number of tests performed and fewer missing data. This situation, however, does not reflect the reality of many healthcare centers in underdeveloped or developing countries. By utilizing only the most readily available variables in healthcare settings of different complexity levels, we enable the application of the predictor even in adverse situations.

Ramírez-Del Real et al. (2022) also achieved satisfactory results by using demographic, clinical, and laboratory data to predict COVID-19 mortality. Data were collected from healthy individuals who were subsequently monitored to evaluate outcomes if they contracted COVID-19. They achieved an accuracy of 90.41%, with positive and negative predictive values of 94.28 and 87.36%, respectively. This study suggests the possibility of identifying

TABLE 3 Average accuracy after 100 simulations with each subset.

	Average accuracy per subset						
Classifier	A	B	C	D	E	F	G
XGBoost classifier	74.62%	77.25%	77.27%	77.32%	77.63%	77.20%	72.87%
Decision tree classifier	70.13%	70.13%	69.94%	70.19%	70.82%	71.44%	66.07%
Logistic regression	74.83%	77.51%	78.06%	77.63%	77.65%	77.72%	75.02%
AdaBoost classifier	73.77%	77.22%	75.87%	76.32%	76.30%	76.32%	72.72%
Bagging classifier	75.14%	77.33%	78.21%	77.54%	77.84%	77.08%	74.55%
Linear discriminant analysis	74.70%	77.90%	78.76%	78.16%	78.40%	78.38%	75.46%
Linear SVC	74.80%	77.63%	78.32%	77.68%	77.77%	77.34%	75.09%
SVC	76.30%	79.12%	79.90%	79.44%	80.12%	80.78%	77.27%
Passive aggressive classifier	67.48%	67.96%	70.24%	70.93%	69.42%	69.94%	67.05%
NuSVC	75.07%	78.84%	79.52%	79.09%	79.75%	80.14%	76.35%
Nearest centroid	68.04%	70.99%	71.95%	71.04%	70.58%	71.54%	69.78%
LGBM classifier	74.49%	77.33%	77.23%	78.26%	78.40%	77.46%	73.36%
Ridge classifier	74.68%	78.45%	78.84%	78.32%	78.51%	78.74%	75.52%
Ridge classifier CV	74.75%	78.42%	78.84%	78.35%	78.77%	79.04%	75.75%
Random forest classifier	76.20%	78.97%	80.18%	79.30%	80.33%	80.10%	76.74%
Gaussian NB	70.03%	71.70%	72.94%	71.00%	71.56%	71.78%	69.52%
Calibrated classifier CV	74.41%	77.88%	78.08%	76.67%	77.68%	76.90%	75.65%
Bernoulli NB	73.07%	74.36%	76.16%	75.68%	74.84%	76.10%	72.64%
Extra trees classifier	74.83%	77.72%	78.61%	78.49%	79.51%	79.56%	75.60%
Quadratic discriminant analysis	71.46%	72.51%	73.31%	72.04%	70.96%	68.82%	70.29%
Extra tree classifier	65.07%	69.19%	70.56%	68.60%	67.40%	67.74%	65.91%
Label spreading	67.75%	70.66%	74.06%	73.26%	73.35%	74.24%	67.69%
Label propagation	67.72%	70.70%	74.03%	73.26%	73.35%	74.24%	67.68%
KNeighbors classifier	73.00%	76.34%	77.65%	75.82%	75.70%	75.80%	73.19%
Perceptron	68.89%	70.13%	71.35%	70.82%	69.88%	70.80%	67.05%
SGD classifier	68.28%	71.54%	72.08%	72.18%	71.98%	71.38%	68.38%
Dummy classifier	68.68%	70.76%	71.16%	69.82%	71.30%	69.56%	68.32%

Subsets A-F address missing data through subject elimination, while subset G uses mean imputation.

individuals more vulnerable to unfavorable outcomes even before contracting the disease.

3.4 Contributing features

After training, we selected the two models that demonstrated the best performance, Random Forest and SVC, to extract the features that contributed most significantly to the classification process. For the Random Forest model, we averaged the "feature importance" values for each variable across the 5,000 simulations. In the case of the SVC model, which was implemented using the Radial Basis Function kernel, we employed the Permutation Feature Importance technique, as implemented in Scikit-Learn. The Permutation Importance values for each variable were computed for each simulation, and we used the average values to identify the most significant features. To conclude our analysis, we performed logistic regression on the same subset used to train the models and calculated the *p*-values associated with each variable. These values are presented in Table 5.

The most statistically significant variables for predicting severe outcomes, with a p-value less than 0.05, included dyspnea, the patient's general condition upon hospital admission, peripheral oxygen saturation, urea levels, platelet count, and signs of upper respiratory tract infection. In the Random Forest model, the most important features were dyspnea, general condition of the patient upon admission, age, oxygen saturation upon admission, respiratory rate upon admission, and urea. For the SVC model, the significant features included dyspnea, general condition of the patient upon admission, oxygen saturation upon admission, urea, age, and platelet count.

These results share some similarities with findings from previous studies, where age and platelet count were identified as significant predictors of prognosis in severely hospitalized patients (Martono and Mulyanti, 2023; Fang et al., 2020; Kamel et al., 2023). Additionally, other important features, such as different hematological variables, D-dimer and C-reactive protein, could further enhance the predictive power of the system. However, these tests were unavailable for a significant portion of the studied population.

TABLE 4 Average performance metrics from 5,000 simulations using subset F (with 28 variables, no imputation).

Classifier	Accuracy	F1-score	Positive predicted value	Negative predicted value	Sensitivity	Specificity	AUC
SVC	0.79 ± 0.0014	0.54 ± 0.0031	0.87 ± 0.0036	0.78 ± 0.0016	0.40 ± 0.0031	0.97 ± 0.0008	0.79 ± 0.0020
NuSVC	0.79 ± 0.0015	0.56 ± 0.0032	0.83 ± 0.0037	0.79 ± 0.0018	0.44 ± 0.0036	0.96 ± 0.0010	0.78 ± 0.0022
Random forest classifier	0.80 ± 0.0014	0.58 ± 0.0031	0.81 ± 0.0039	0.79 ± 0.0017	0.46 ± 0.0034	0.95 ± 0.0011	0.91 ± 0.0014
Extra trees classifier	0.79 ± 0.0014	0.57 ± 0.0031	0.77 ± 0.0039	0.79 ± 0.0017	0.47 ± 0.0035	0.94 ± 0.0012	0.78 ± 0.0019
Ridge classifier CV	0.80 ± 0.0015	0.63 ± 0.0028	0.75 ± 0.0037	0.82 ± 0.0017	0.56 ± 0.0034	0.91 ± 0.0014	0.77 ± 0.0020
Ridge classifier	0.78 ± 0.0015	0.62 ± 0.0028	0.69 ± 0.0038	0.82 ± 0.0017	0.57 ± 0.0033	0.88 ± 0.0016	0.77 ± 0.0020
Linear discriminant analysis	0.78 ± 0.0015	0.61 ± 0.0027	0.67 ± 0.0036	0.82 ± 0.0017	0.58 ± 0.0034	0.87 ± 0.0017	0.77 ± 0.0020
Artificial neural network	0.76 ± 0.0016	0.45 ± 0.0034	0.81 ± 0.0050	0.76 ± 0.0018	0.33 ± 0.0033	0.96 ± 0.0012	0.77 ± 0.0018

4 Discussion

An extensive data preprocessing process was carried out to identify as many filling errors or duplicates as possible. Additionally, the variable selection process involved not only searching for better classifier performance but also minimizing the loss of individuals after data cleaning. Constructing seven different subsets, each encompassing varying numbers of variables and consequently individuals, allowed us to observe the effect on classifier outcomes.

When assessing the cost-effectiveness of using the developed classifiers, it is crucial to consider that laboratory tests and clinical evaluations are indispensable for proper inpatient monitoring. Therefore, the use of the classifier would not incur additional expenses beyond an informatics system capable of generating results. The utilization of these classifiers does not require additional tests beyond those essential for usual medical follow-up.

During preprocessing, a test was conducted to assess the impact of imputation for filling missing data. The method used was to fill in missing data with the means calculated for all individuals. In other words, those who did not have their hemoglobin measured received the average hemoglobin value from all individuals who had the test collected. Imputation enables certain machine learning algorithms that do not accept incomplete input information to be applied to all study participants. However, it was observed that mean imputation reduced classifier accuracy. It is understood that clinical data variables are not independent of each other, and therefore, alternative data imputation methods could yield better results. Through multivariate analysis, the tests collected for a specific individual could be used to estimate the missing values. For instance, hemoglobin, respiratory rate, and clinical signs of dyspnea could be utilized to estimate peripheral oxygen saturation (SpO2) for patients who could not undergo oximetry. Further dedicated studies are recommended to identify optimal clinical data imputation methods for classifier development.

It has been observed that reducing input variables, despite decreasing the loss of individuals (due to incompleteness or missing data) and the complexity of classifier systems, compromises the accuracy of classification systems. Throughout this work, experimentation was necessary to find the best solution for this problem. Tests with different subsets helped strike the right balance between individual losses and the amount of clinical data used. A high loss of individuals would hinder classifier learning, as well as the reliability and generalizability of results. The reliance on a large amount of clinical data, coupled with the inability of most algorithms to handle incompleteness, may indicate inferior performance compared to a trained professional. This reinforces the idea that AI should be viewed as a tool rather than a replacement for human expertise.

The high number of sparse variables collected during the cohort study draws some attention. Several variables were positive in less than 1% of study participants, such as the presence of arrhythmia or prior valvular disease. This characteristic poses a particular challenge when developing a classifier. The more variables used as input, the more complex the classifier becomes. Even if we imagine a simple questionnaire for diagnosing a specific disease, the more questions required for the diagnosis, the more challenging its implementation becomes, during real healthcare situations. Also, complex classifiers are more prone to overfitting, which reduces their ability to generalize to new data and patients (Pansambal and Nandgaokar, 2023; Salehin and Kang, 2023). Simultaneously, some clinical data, despite being rare, can serve as excellent markers for unfavorable outcomes in certain pathologies. Excluding these rare data points from the analysis could compromise the quality of the final analysis. Considering this, clinical knowledge of signs and symptoms, as potential severity predictors (Ou et al., 2020; Martono and Mulyanti, 2023; Fang et al., 2020), was considered during variable selection, along with combining different variables with the same clinical significance (such as a variable representing the presence of any prior lung disease).

Regarding AI-based classifiers, it is essential to consider that a significant class imbalance can hinder the development of classification systems (Isgut et al., 2023). In this study, approximately 69% of participants did not experience severe complications. Consequently, a classifier that labeled all individuals as "non-severe" would achieve a considerable accuracy rate. However, based on the results obtained, it is possible to affirm that satisfactory accuracy models were developed, particularly when classifying an individual as belonging to the "severe" group. This is evident from the high positive predictive value and specificity.

Specificity and sensitivity are crucial for determining the quality of a medical screening test. When both are high, it indicates a low rate of classification errors. Despite the low sensitivity of the developed

TABLE 5 Feature importance and *p*-values for classification models.

Feature	Importance RF	Importance SVC	*p*-values (logistic regression)
General condition of the patient upon admission	0.1158	0.0298	< 0.05
Age (years)	0.0696	0.0040	0.1607
Number of leukocytes (number/mm^3)	0.0324	0.0009	0.7348
Number of lymphocytes (number/mm^3)	0.0185	0.0015	0.3840
Number of platelets (number/mm^3)	0.0388	0.0037	< 0.05
Urea (mg/dL)	0.0478	0.0108	< 0.05
Systolic blood pressure upon admission (mmHg)	0.0435	−0.0022	0.0565
Diastolic blood pressure upon admission (mmHg)	0.0396	0.0013	0.4980
Heart rate upon admission (bpm/min)	0.0302	0.0031	0.4154
Respiratory Rate UPON admission (breaths/min)	0.0559	0.0027	0.7092
Oxygen saturation upon admission (%)	0.0621	0.0144	< 0.05
No pre-existing conditions	0.0203	0.0002	0.4987
Hypertension	0.0247	−0.0006	0.2366
Diabetes mellitus	0.0306	0.0032	0.1365
Obesity	0.0086	−0.0008	0.6146
Kidney failure	0.0045	−0.0001	0.6052
Fever	0.0221	0.0017	0.1207
Dry cough	0.0218	−0.0009	0.0562
Headache	0.0201	0.0033	0.0757
Weakness	0.0162	0.0004	0.3476
Anosmia	0.0103	0.0002	0.5427
Gender	0.0215	−0.0009	0.4630
Signs of upper respiratory tract infection (URTI)	0.0239	0.0034	< 0.05
Dyspnea	0.1334	0.0656	< 0.05
Signs of pulmonary involvement	0.0214	−0.0006	0.7351
Constitutional symptoms	0.0226	0.0002	0.7766
Previous pulmonary diseases	0.0070	−0.0007	0.1377
Hemoglobin (g/dL)	0.0368	−0.0005	0.7695

The table displays features, their importance in Random Forest, importance in SVC, and *p*-values from logistic regression.

classifiers, the high specificity suggests a small number of false positives. Few false positives are generally preferable, even at the expense of reduced sensitivity, especially when a positive test result could lead to unnecessary and dangerous interventions (Herman, 2006; Trevethan, 2017).

Using data collected from a multicenter study, it would not be inaccurate to assume that the developed system would have greater generalization capacity. This is because the study involved healthcare units with varying capabilities, technological density, and resource access. Consequently, the system could be used in both less-structured hospitals and more comprehensive units. It is also worth emphasizing that the data were collected from a prospective cohort, ensuring greater reliability regarding data quality, despite adding technical limitations, such as collecting all variables from all participants. No other studies were found that solely utilized clinically and demographically collected information in a prospective manner. Furthermore, due to its adaptation to real-world constraints and its inclusion of individuals from diverse socioeconomic backgrounds and demographic profiles, the developed system has greater practical usability and can span different levels of healthcare units.

In recent years, several studies have been published addressing the prediction of unfavorable outcomes in COVID-19 using AI systems (Kim, 2022; Yu et al., 2021; Shakibfar et al., 2023; Wang et al., 2021; Ramírez-Del Real et al., 2022; Kamel et al., 2023). One of the challenges encountered in the literature on ML and AI for disease diagnosis or prognosis evaluation is the choice of performance metrics. Few studies have reported positive and negative predictive values, and some have not even published sensitivity and specificity values (Wang et al., 2021). While a new clinical study is necessary to precisely determine these metrics using data from new, previously unseen patients, calculating these probabilities was considered essential to assess the viability of the classifiers. It is considered that AI-based algorithms for diagnosis should be evaluated similarly to new laboratory and imaging tests.

Other studies have also demonstrated satisfactory results by developing predictors that incorporate imaging exams, such as Computed Tomography (CT), alongside clinical data to classify COVID-19 severity (Yu et al., 2021). The cohort described in this study also collected bedside chest radiographies, which could be combined with clinical data to enhance results in future work. However, by utilizing only a few variables resulting from simple tests, the prediction system could be used even in healthcare units with low technological density. It is estimated that in 2020, only approximately 15% of Brazilian municipalities had access to Computed Tomography (Pereira and Tomógrafos, 2020).

This work considers healthcare systems with resource scarcity and the reality of developing and underdeveloped countries. Nevertheless, this methodology can be adjusted to meet the needs of different countries, such as including collaboration between various governments and organizations in data collection, as well as steering data collection toward more widely available tests.

It is important to emphasize that the objective of these AI-based systems is not to replace healthcare professionals but to assist in the triaging process during high-demand situations. By providing timely and accurate support, these systems can help alleviate the workload of medical staff, allowing them to focus on critical decision-making and patient care. The collaboration between AI and healthcare providers can enhance overall efficiency and effectiveness in delivering quality care.

This classifier was designed to operate specifically in moderate and severe cases of COVID-19 that required hospitalization. Mild cases of COVID-19 typically do not require the tests we used and often do not seek healthcare services. While this classification could potentially be applied to other respiratory diseases, further validation with new data will be necessary to confirm its effectiveness in those contexts. Ongoing research and data collection will be crucial to refining and adapting the classifier for broader applications in managing other diseases.

5 Conclusion and future work

It is understood that vaccination and the emergence of new SARS-CoV-2 variants alter the disease's behavior and the number of severe complications. A reduction in severe COVID-19 cases has been observed in vaccinated individuals, while the transmission has increased with the emergence of new variants (Martono and Mulyanti, 2023). Consequently, collecting new data and updating classifiers becomes necessary.

Given its multidisciplinary nature, it is essential to emphasize the importance of going beyond preprocessing and classification algorithms by incorporating medical expertise. The fusion of medical and computational perspectives allows for a more comprehensive analysis of the feasibility, applicability, and validity of the classification system.

This work leaves a few gaps that must be addressed in future research before implementing this type of classification system in healthcare. Firstly, although mean imputation was utilized to manage missing data, more sophisticated techniques, such as multivariate analysis, may yield improved results. Secondly, the prevalence of numerous sparse variables – those that are positive in only a small proportion of patients – should be considered in future data collection efforts. By adjusting the form and method of data collection to focus on gathering more general information rather than numerous specific variables, we can reduce the complexity of the cohort study while addressing the challenges AI algorithms encounter with sparse data. Additionally, the study revealed challenges related to class imbalance, with a greater number of patients classified as "non-severe," which may hinder the classifiers' ability to accurately identify severe cases. This suggests the need for further exploration of sampling techniques or specialized algorithms. Moreover, it is essential to investigate the potential benefits of incorporating imaging data, such as chest X-rays or computed tomography scans, as well as the impact of vaccination and emerging SARS-CoV-2 variants on prediction accuracy. Finally, external validation using diverse datasets and assessing the performance of the developed system in prospective studies is crucial before it can be effectively implemented in clinical practice.

This study validated the methodology for developing predictors of morbidity and mortality, from data collection to training machine learning algorithms. This methodology can be applied to other diseases, especially those where outcomes depend on the interaction of multiple variables.

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

The studies involving humans were approved by the Ethics Committee of Faculty of Medicine of the University of Brasília (CEP-FM/UnB). Certificate of Ethical Appreciation Presentation number 31941420.4.1001.5558 and approval number 4.054.462. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin.

Author contributions

YO: Conceptualization, Data curation, Formal analysis, Methodology, Software, Validation, Writing – original draft. IR: Conceptualization, Data curation, Investigation, Writing – review & editing. PA: Data curation, Investigation, Writing – review & editing. AM: Data curation, Investigation, Writing – review & editing. MG: Data curation, Investigation, Writing – review & editing. LS: Data curation, Investigation, Writing – review & editing. MR: Data curation, Investigation, Writing – review & editing. AN: Data curation, Investigation, Writing – review & editing. MN: Data curation, Investigation, Writing – review & editing. HP: Data curation, Investigation, Writing – review & editing. FN: Methodology, Software, Validation, Writing – review & editing. CC-S: Conceptualization, Methodology, Supervision, Writing – review & editing. VF: Conceptualization, Methodology, Supervision, Writing – review &

editing. HC: Conceptualization, Funding acquisition, Investigation, Methodology, Project administration, Software, Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported in part by the University of Brasilia Foundation under Grant FINATEC/GEPRO 7153, through the Committee for Research, Innovation, and Extension in the Fight against COVID-19 (COPEI/UnB) and the Foundation for Scientific and Technological Enterprises (Finatec).

Acknowledgments

We would like to thank Eliane V. Mancuzo, from Universidade Federal de Minas Gerais (UFMG) for her contributions during the data collection phase of this study. Her efforts have significantly enriched our research.

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

Abadi, M, Agarwal, A, Barham, P, Brevdo, E, Chen, Z, Citro, C, et al. TensorFlow: a system for large-scale machine learning. *12th USENIX Symposium on Operating Systems Design and Implementation* (2016).

Arpaci, I., Huang, S., Al-Emran, M., Al-Kabi, M. N., and Peng, M. (2021). Predicting the COVID-19 infection with fourteen clinical features using machine learning classification algorithms. *Multimed. Tools Appl.* 80, 11943–11957. doi: 10.1007/s11042-020-10340-7

Brinati, D., Campagner, A., Ferrari, D., Locatelli, M., Banfi, G., and Cabitza, F. (2020). Detection of COVID-19 infection from routine blood exams with machine learning: a feasibility study. *J. Med. Syst.* 44:135. doi: 10.1007/s10916-020-01597-4

Buitinck, L., Louppe, G., Blondel, M., Fabien, P., Mueller, A., Olivier, G., et al. (2011). Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* 12, 2825–2830. http://scikit-learn.sourceforge.net

Buonsenso, D., De Rose, C., and Pierantoni, L. (2021). Doctors' shortage in adults COVID-19 units: a call for pediatricians. *Eur. J. Pediatr.* 180, 2315–2318. doi: 10.1007/s00431-021-03995-3

Buscema, M. (2002). A brief overview and introduction to artificial neural networks. *Subst. Use Misuse* 37, 1093–1148. doi: 10.1081/JA-120004171

Fang, X., Li, S., Yu, H., Wang, P., Zhang, Y., Chen, Z., et al. (2020). Epidemiological, comorbidity factors with severity and prognosis of COVID-19: a systematic review and meta-analysis. *Aging* 12, 12493–12503. doi: 10.18632/aging.103579

Fernandes, FT, de Oliveira, TA, Teixeira, CE, Batista, AFMde, Dalla Costa, G, and Chiavegatto Filho, ADP. (2021). A multipurpose machine learning approach to predict COVID-19 negative prognosis in São Paulo, Brazil. *Sci. Rep.* 11. doi: 10.1038/s41598-021-82885-y:3343

Grossi, E., and Buscema, M. (2007). Introduction to artificial neural networks. *Eur. J. Gastroenterol. Hepatol.* 19, 1046–1054. doi: 10.1097/MEG.0b013e3282f198a0

Haglin, J. M., Jimenez, G., and Eltorai, A. E. M. (2019). Artificial neural networks in medicine. *Health Technol.* 9, 1–6. doi: 10.1007/s12553-018-0244-4

Hao, B., Hu, Y., Sotudian, S., Zad, Z., Adams, W. G., Assoumou, S. A., et al. (2022). Development and validation of predictive models for COVID-19 outcomes in a safety-net hospital population. *J. Am. Med. Inform. Assoc.* 29, 1253–1262. doi: 10.1093/jamia/ocac062

Heldt, F. S., Vizcaychipi, M. P., Peacock, S., Cinelli, M., McLachlan, L., Andreotti, F., et al. (2021). Early risk assessment for COVID-19 patients from emergency department data using machine learning. *Sci. Rep.* 11:4200. doi: 10.1038/s41598-021-83784-y

Herman, C. (2006). What makes a screening exam "good"? *Ethics. JAMA J. Am. Med. Assoc.* 8, 34–37. doi: 10.1001/virtualmentor.2006.8.1.cprl1-0601

Isgut, M., Gloster, L., Choi, K., Venugopalan, J., and Wang, M. D. (2023). Systematic review of advanced AI methods for improving healthcare data quality in post COVID-19 era. *IEEE Rev. Biomed. Eng.* 16, 53–69. doi: 10.1109/RBME.2022.3216531

Jimenez-Solem, E., Petersen, T. S., Hansen, C., Hansen, C., Lioma, C., Igel, C., et al. (2021). Developing and validating COVID-19 adverse outcome risk prediction models from a bi-national European cohort of 5594 patients. *Sci. Rep.* 11:3246. doi: 10.1038/s41598-021-81844-x

Kamel, F. O., Magadmi, R., Qutub, S., Badawi, M., Badawi, M., Madani, T. A., et al. (2023). Machine learning-based prediction of COVID-19 prognosis using clinical and hematologic data. *Cureus* 15:e50212. doi: 10.7759/cureus.50212

Kanji, S., Burry, L., Williamson, D., Pittman, M., Dubinsky, S., Patel, D., et al. (2020). Therapeutic alternatives and strategies for drug conservation in the intensive care unit during times of drug shortage: a report of the Ontario COVID-19 ICU drug task force. *Can. J. Anesth.* 67, 1405–1416. doi: 10.1007/s12630-020-01713-5

Kim, D. K. (2022). Prediction models for COVID-19 mortality using artificial intelligence. *J. Pers. Med.* 12, 1–3. doi: 10.3390/jpm12091522

Kühl, N., Schemmer, M., Goutier, M., and Satzger, G. (2022). Artificial intelligence and machine learning. *Electron. Mark.* 32, 2235–2244. doi: 10.1007/s12525-022-00598-0

Kwok, S. W. H., Wang, G., Sohel, F., Kashani, K. B., Zhu, Y., Wang, Z., et al. (2023). An artificial intelligence approach for predicting death or organ failure after hospitalization for COVID-19: development of a novel risk prediction tool and comparisons with ISARIC-4C, CURB-65, qSOFA, and MEWS scoring systems. *Respir. Res.* 24:79. doi: 10.1186/s12931-023-02386-6

Luo, X., Lv, M., Zhang, X., Estill, J., Yang, B., Lei, R., et al. (2022). Clinical manifestations of COVID-19: an overview of 102 systematic reviews with evidence mapping. *J. Evid. Based Med.* 15, 201–215. doi: 10.1111/jebm.12483

Martono, F. F., and Mulyanti, S. (2023). Risk factors associated with the severity of COVID-19. *Malaysian. J. Med. Sci.* 30, 84–92. doi: 10.21315/mjms2023.30.3.7

Mohamed, K., Rodríguez-Román, E., Rahmani, F., Zhang, H., Ivanovska, M., Makka, S. A., et al. (2020). Borderless collaboration is needed for COVID-19—a disease that knows no borders. *Infect. Control Hosp. Epidemiol.* 41, 1245–1246. doi: 10.1017/ice.2020.162

Mohamed, K., Rzymski, P., Islam, M. S., Makuku, R., Mushtaq, A., Khan, A., et al. (2022). COVID-19 vaccinations: the unknowns, challenges, and hopes. *J. Med. Virol.* 94, 1336–1349. doi: 10.1002/jmv.27487

Napolitano, F., Xu, X., and Gao, X. (2022). Impact of computational approaches in the fight against COVID-19: an AI guided review of 17 000 studies. *Brief. Bioinform.* 23, 1–19. doi: 10.1093/bib/bbab456

Ou, M., Zhu, J., Ji, P., Li, H., Zhong, Z., Li, B., et al. (2020). Risk factors of severe cases with COVID-19: a Meta-analysis. *Epidemiol. Infect.* 148:e175. doi: 10.1017/S095026882000179X

Pansambal, B. H., and Nandgaokar, A. B. (2023). Integrating dropout regularization technique at different layers to improve the performance of neural networks. *Int. J. Adv. Comput. Sci. Appl.* 14:716–722. doi: 10.14569/IJACSA.2023.0140478

Pereira, E, and Tomógrafos, X. COVID-19 Análise da distribuição espacial de Tomógrafos no Brasil em tempos de pandemia. (2020). 1–89.

Rahman, M. D. A., Hossain, M. S., Alrajeh, N. A., and Gupta, B. B. (2021). A multimodal, multimedia point-of-care deep learning framework for COVID-19 diagnosis. *ACM Trans. Multimed. Comput. Commun. Appl.* 17, 1–24. doi: 10.1145/3421725

Ramírez-Del Real, T., Martínez-García, M., Márquez, M. F., López-Trejo, L., Gutiérrez-Esparza, G., and Hernández-Lemus, E. (2022). Individual factors associated with COVID-19 infection: a machine learning study. *Front. Public Health* 10:912099. doi: 10.3389/fpubh.2022.912099

Rodvold, D. M., McLeod, D. G., Brandt, J. M., Snow, P. B., and Murphy, G. P. (2001). Introduction to artificial neural networks for physicians: taking the lid off the black box. *Prostate* 46, 39–44.

Salehin, I., and Kang, D. K. (2023). A review on dropout regularization approaches for deep neural networks within the scholarly domain. *Electronics* 12, 1–15. doi: 10.3390/electronics12143106

Sandhu, P., Shah, A. B., Ahmad, F. B., Kerr, J., Demeke, H. B., Graeden, E., et al. (2022). Emergency department and intensive care unit overcrowding and ventilator shortages in US hospitals during the COVID-19 pandemic, 2020-2021. *Public Health Rep.* 137, 796–802. doi: 10.1177/00333549221091781

Shakibfar, S., Nyberg, F., Li, H., Zhao, J., Nordeng, H. M. E., Sandve, G. K. F., et al. (2023). Artificial intelligence-driven prediction of COVID-19-related hospitalization and death: a systematic review. *Front. Public Health* 11, 1–15. doi: 10.3389/fpubh.2023.1183725

Shan, G. (2022). Monte Carlo cross-validation for a study with binary outcome and limited sample size. *BMC Med. Inform. Decis. Mak.* 22:270. doi: 10.1186/s12911-022-02016-z

Sott, M. K., Bender, M. S., and da Silva, B. K. (2022). Covid-19 outbreak in Brazil: health, social, political, and economic implications. *Int. J. Health Serv.* 52, 442–454. doi: 10.1177/00207314221122658

Tenda, E. D., Henrina, J., Setiadharma, A., Aristy, D. J., Romadhon, P. Z., Thahadian, H. F., et al. (2024). Derivation and validation of novel integrated inpatient mortality prediction score for COVID-19 (IMPACT) using clinical, laboratory, and AI—processed radiological parameter upon admission: a multicentre study. *Sci. Rep.* 14:2149. doi: 10.1038/s41598-023-50564-9

Tipirneni, S., and Reddy, C. K. (2022). Self-supervised transformer for sparse and irregularly sampled multivariate clinical time-series. *ACM Trans. Knowl. Discov. Data* 16, 1–17. doi: 10.1145/3516367

Trevethan, R. (2017). Sensitivity, specificity, and predictive values: foundations, Pliabilities, and pitfalls in research and practice. *Front. Public Health* 5, 1–7. doi: 10.3389/fpubh.2017.00307

Uddin, S., Khan, A., Hossain, M. E., and Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Med. Inform. Decis. Mak.* 19:281. doi: 10.1186/s12911-019-1004-8

Wang, L., Zhang, Y., Wang, D., Tong, X., Liu, T., Zhang, S., et al. (2021). Artificial intelligence for COVID-19: a systematic review. *Front. Med.* 8, 1–15. doi: 10.3389/fmed.2021.704256

World Health Organization. WHO coronavirus (COVID-19) dashboard. (2023). Available at: https://covid19.who.int/ (Accessed October 7, 2023).

Yoshioka-Maeda, K., Iwasaki-Motegi, R., and Honda, C. (2020). Preventing the dysfunction of public health centres responding to COVID-19 by focusing on public health nurses in Japan. *J. Adv. Nurs.* 76, 2215–2216. doi: 10.1111/jan.14409

Yu, L., Shi, X., Liu, X., Jin, W., Jia, X., Xi, S., et al. (2021). Artificial intelligence Systems for Diagnosis and Clinical Classification of COVID-19. *Front. Microbiol.* 12, 1–9. doi: 10.3389/fmicb.2021.729455

frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 24 February 2025
DOI 10.3389/frai.2025.1559880

OPEN ACCESS

EDITED BY
Dmytro Chumachenko,
National Aerospace University – Kharkiv Aviation Institute, Ukraine

REVIEWED BY
Tetyana Chumachenko,
Kharkiv National Medical University, Ukraine
Sergiy Yakovlev,
Lodz University of Technology, Poland

*CORRESPONDENCE
Igor Nesteruk
✉ inesteruk@yahoo.com

RECEIVED 13 January 2025
ACCEPTED 31 January 2025
PUBLISHED 24 February 2025

CITATION
Nesteruk I (2025) General SIR model for visible and hidden epidemic dynamics. *Front. Artif. Intell.* 8:1559880. doi: 10.3389/frai.2025.1559880

General SIR model for visible and hidden epidemic dynamics

Igor Nesteruk*

Institute of Hydromechanics, National Academy of Sciences of Ukraine, Kyiv, Ukraine

To simulate hidden epidemic dynamics connected with asymptomatic and unregistered patients, a new general SIR model was proposed. For some cases, the analytical solutions of the set of 5 differential equations were found, which allow simplifying the parameter identification procedure. Two waves of the pertussis epidemic in England in 2023 and 2024 were simulated with the assumption of zero hidden cases. The accumulated and daily numbers of cases and the duration of the second wave were predicted with rather high accuracy. If the trend will not change, the monthly figure of 9 new pertussis cases (as it was in January–February 2023) can be achieved only in May 2025. The proposed approach can be recommended for both simulations and predictions of different epidemics.

KEYWORDS

mathematical modeling of infection diseases, SIR model, parameter identification, pertussis epidemic in England, hidden epidemic dynamics

1 Introduction

Asymptomatic and unregistered cases are characteristic of almost all infectious diseases, in particular, SARS-CoV-2 (Mass coronavirus testing in Slovakia-a, 2020; Mass coronavirus testing in Slovakia-b, 2020; An experiment with mass testing for COVID-19 was conducted in Khmelnytsky, n.d.; Schreiber et al., 2023; Fowlkes et al., 2022; Shang et al., 2022) and pertussis (Craig et al., 2020) are no exception. The percentage of asymptomatic patients can be age dependent and lead to huge differences in registered numbers of cases for countries with young and old population (Davies et al., 2020; Nesteruk, 2024b; Nesteruk and Keeling, 2023). Some theoretical estimations of the visibility coefficient β—the ratio of real infections to the registered ones can be found in Nesteruk (2024b), Nesteruk (2021e), Nesteruk (2021c), and Nesteruk (2021d). In this study we will use the concepts of the classical SIR (susceptible-infectious-removed) model (Kermack and McKendrick, 1927; Weiss, 2013; Daley and Gani, 2005; Keeling and Rohani, 2008; Cherniha, 2020; Mohammadi et al., 2021; Nesteruk, 2021a), its generalization for simulations of different epidemic waves (Nesteruk, 2021a; Nesteruk, 2021b; Nesteruk, 2023b) and procedures of parameter identification (Nesteruk, 2023b; Nesteruk, 2017). Numerous improvements of SIR model (see, e.g., Hethcote, 2000; Nakamura et al., 2020; Britton, 2004; Pesco et al., 2014; Nesteruk, 2023a) do not take into account the visibility coefficient.

The obtained theoretical results will be applied for simulations of the pertussis (whooping cough) epidemic in England in 2023 and 2024 (Confirmed cases of pertussis in England by month, n.d.). This disease increases the risk of infant fatality and became a serious problem in many countries including the developed ones (Pesco et al., 2014; Confirmed cases of pertussis in England by month, n.d.). Numerical differentiation of the monthly numbers of new cases revealed two waves of the epidemic in England (before and after November 2023) (Nesteruk, 2024a). Due to the absence of necessary amount of observations, SIR simulations were performed in Nesteruk (2024a) only for the first wave. In this study we will use the new approach for simulation of both waves of the pertussis epidemic and compare the predictions with the recent statistical data.

2 Differential equations and initial conditions

For every epidemic wave i, let us divide the compartment of infectious persons $I(t)$ (t is time) into visible (registered) and hidden (invisible/asymptomatic and unregistered) parts $I = I^{(v)} + I^{(h)}$ and suppose that these persons are appearing according to the visibility coefficient $\beta_i \geq 1$ and removing with rates $\rho_i^{(v)} I^{(v)}$and $\rho_i^{(h)} I^{(h)}$. Then the general SIR model (Nesteruk, 2021a; Nesteruk, 2021b; Nesteruk, 2023b) takes the following form:

$$\frac{dS}{dt} = -\alpha_i S\left(I^{(v)} + I^{(h)}\right) \tag{1}$$

$$\frac{dI^{(v)}}{dt} = \frac{\alpha_i}{\beta_i} S\left(I^{(v)} + I^{(h)}\right) - \rho_i^{(v)} I^{(v)} \tag{2}$$

$$\frac{dI^{(h)}}{dt} = \frac{(\beta_i - 1)\alpha_i}{\beta_i} S\left(I^{(v)} + I^{(h)}\right) - \rho_i^{(h)} I^{(h)} \tag{3}$$

$$\frac{dR^{(v)}}{dt} = \rho_i^{(v)} I^{(v)} \tag{4}$$

$$\frac{dR^{(h)}}{dt} = \rho_i^{(h)} I^{(h)} \tag{5}$$

The compartment of removed persons $R(t)$ is also divided into visible (registered) and hidden parts $R = R^{(v)} + R^{(h)}$. Infection and removal rates (α_i, $\rho_i^{(v)}$, $\rho_i^{(h)}$) and the visibility coefficient β_i are supposed to be constant for every epidemic wave, i.e., for the time periods: $t_i^* \leq t \leq t_{i+1}^*$, $i = 1, 2, 3, \ldots$ Summarizing Equations 1–5 yields zero value of the derivative $d\left(S + I^{(v)} + I^{(h)} + R^{(v)} + R^{(h)}\right)/dt$.

Then the sum:

$$N_i = S + I^{(v)} + I^{(h)} + R^{(v)} + R^{(h)} \tag{6}$$

must be constant for every epidemic wave. We will consider the value N_i to be an unknown parameter of the model corresponding to the *i-th* wave, which is not equal to the known volume of population and must be estimated by observations. There is no need to assume that before the outbreak all people are susceptible, since many of them are protected by their immunity, distance, lockdowns, etc. Thus, we will not reduce the problem to a 4-dimensional one. It means that the solution can be obtained by numerical integration of the set of 5 differential Equations 1–5. Nevertheless, there are some separate cases, when analytical solutions are possible (see next Section).

Taking into account Equation 6, the initial conditions for the set of Equations 1–5 at the beginning of every epidemic wave t_i^* can be written as follows:

$$I^{(v)}\left(t_i^*\right) = I_{vi},\ I^{(h)}\left(t_i^*\right) = I_{hi},\ R^{(v)}\left(t_i^*\right) = R_{vi},\ R^{(h)}\left(t_i^*\right) = R_{hi}\ S\left(t_i^*\right) = N_i - I_{vi} - I_{hi} - R_{vi} - R_{hi} \tag{7}$$

If at moment t_i^* all previously infected persons are removed, we can take into account only cases starting to appear during *i-th* wave and use the initial conditions:

$$I_{vi} = 1,\ I_{hi} = \beta_i - 1,\ R_{vi} = 0,\ R_{hi} = 0 \tag{8}$$

3 Examples of analytical solutions

Let us introduce the functions corresponding to the accumulated numbers of visible and hidden cases:

$$V^{(v)} = I^{(v)} + R^{(v)},\ V^{(h)} = I^{(h)} + R^{(h)} \tag{9}$$

Then it follows from Equations 2–5 that

$$\frac{dV^{(v)}}{dt} = \frac{\alpha_i}{\beta_i} S\left(I^{(v)} + I^{(h)}\right), \frac{dV^{(h)}}{dt} = \frac{(\beta_i - 1)\alpha_i}{\beta_i} S\left(I^{(v)} + I^{(h)}\right) \tag{10}$$

Dividing Equation 10 by Equation 1 yeilds:

$$\frac{dV^{(v)}}{dS} = -\frac{1}{\beta_i}, \frac{dV^{(h)}}{dS} = -\frac{(\beta_i - 1)}{\beta_i} \tag{11}$$

and simple linear solutions taking into account initial Equation 7:

$$V^{(v)} = -\frac{S}{\beta_i} + \frac{N_i - I_{vi} - I_{hi} - R_{vi} - R_{hi}}{\beta_i} + I_{vi} + R_{vi}, \tag{12}$$

$$V^{(h)} = -\frac{(\beta_i - 1)S}{\beta_i} + \frac{(\beta_i - 1)(N_i - I_{vi} - I_{hi} - R_{vi} - R_{hi})}{\beta_i} + I_{hi} + R_{hi}. \tag{13}$$

Equations 12, 13 allow obtaining simple linear relationship:

$$V^{(h)} = (\beta_i - 1)V^{(v)} - (\beta_i - 1)(I_{vi} + R_{vi}) + I_{hi} + R_{hi}$$

which demonstrates that the ratio of total accumulated cases $V = V^{(v)} + V^{(h)}$ to the registered ones:

$$\frac{V}{V^{(v)}} = \beta_i - \frac{(\beta_i - 1)(I_{vi} + R_{vi}) - I_{hi} - R_{hi}}{V^{(v)}} \tag{14}$$

is not constant and equals β_i only approximately at large $V^{(v)}$ numbers. Equation 14 limits the accuracy of the approach used in Nesteruk (2021e), Nesteruk (2021c), and Nesteruk (2021d).

Introducing

$$I = I^{(v)} + I^{(h)},\ I^* = \rho_i^{(v)} I^{(v)} + \rho_i^{(h)} I^{(h)} \tag{15}$$

summarizing Equations 2, 3 and dividing by Equation 1 yield the following differential equation:

$$\frac{dI}{dS} = -1 + \frac{I^*}{\alpha_i IS} \tag{16}$$

In 3 separate cases:

I $\rho_i^{(v)} = \rho_i^{(h)} = \rho_i$
II $I^{(v)} >> I^{(h)},\ I^{(v)} \approx I',\ \rho_i = \rho_i^{(v)}$
III $I^{(v)} << I^{(h)},\ I^{(h)} \approx I,\ \rho_i = \rho_i^{(h)}$

Equation 16 simplifies and has an analitycal solution taking into account the initial Equation 7:

$$\frac{dI}{dS} \cong -1 + \frac{\rho_i}{\alpha_i S}, \tag{17}$$

$$I \cong -S + N_i - R_{vi} - R_{hi} + \frac{\rho_i}{\alpha_i} \ln \frac{S}{N_i - I_{vi} - I_{hi} - R_{vi} - R_{hi}} \tag{18}$$

Equations 17, 18 exact in the case (I) and approximate in cases (II) and (III).

Putting Equation 18 into Equation 1 and integration yield:

$$F(S) \cong \alpha_i \left(t - t_i^*\right), \tag{19}$$

$$F(S) \equiv \int_{N_i - I_{vi} - I_{hi} - R_{vi} - R_{hi}}^{S} \frac{dU}{U\left(U - N_i + R_{vi} + R_{hi} - \frac{\rho_i}{\alpha_i} \ln \frac{U}{N_i - I_{vi} - I_{hi} - R_{vi} - R_{hi}}\right)} \tag{20}$$

It follows from Equations 1, 2, 15 that:

$$\frac{dI^{(v)}}{dS} = -\frac{1}{\beta_i} + \frac{\rho_i^{(v)} I^{(v)}}{\alpha_i SI}, \tag{21}$$

Taking into account that

$$\frac{dF}{dS} \cong -\frac{1}{SI}$$

(see Equations 18, 20), the solution of the non-homogenous linear Equation 21 satisfying the first initial Equation 7 can be written as follows:

$$I^{(v)}(S) \cong I_{vi} \exp\left(-\frac{\rho_i^{(v)} F(S)}{\alpha_i}\right) - \frac{1}{\beta_i} \exp\left(-\frac{\rho_i^{(v)} F(S)}{\alpha_i}\right) \times \int_{N_i - I_{vi} - I_{hi} - R_{vi} - R_{hi}}^{S} \exp\left(\frac{\rho_i^{(v)} F(U)}{\alpha_i}\right) dU \tag{22}$$

With the use of Equations 9, 15 it is possible to express other functions as follows:

$$\begin{aligned} I^{(h)}(S) &\cong I(S) - I^{(v)}(S); \\ R^{(v)}(S) &\cong V^{(v)}(S) - I^{(v)}(S); \\ R^{(h)}(S) &\cong V^{(h)}(S) - I^{(h)}(S) \end{aligned} \tag{23}$$

Then for every value of S, all unknown functions can be calculated with the use of Equations 12, 13, 18, 22, 23. Corresponding moments of time can be found with the use of Equations 19, 20. Thus, Equations 12, 13, 18–20, 22, 23 yield an approximate analytical solution of the set of differential Equations 1–5 with the initial Equation 7. In the case (I) and when $I^{(v)} = I$, this solution is exact. For $I^{(v)} = I$, there is no need in Equations 22, 23 and corresponding formulas obtained in Nesteruk (2021a), Nesteruk (2021b), Nesteruk (2023b) are also valid.

4 Examples of parameter identifications and predictions

The analytical solution simplifies the procedure of identification of unknown parameters, since there is no need in numerical integration of differential Equations 1–5. It particular, having the set of accumulated cases $V_j^{(v)}$ registered at moments t_j, we can calculate corresponding values S_j with the use of the linear Equation 12 for any values of unknown constant parameters appearing in Equations 1–7. Then Equation 20 allows calculating values $F_j = F(S_j)$. Due to the linear relationship (Equation 19 shows that there is a linear dependence between time and the function F (Equation 20), which depends on the accumulated numbers of cases), standard linear regression formulas (Draper and Smith, 1998) can be used to calculate the correlation coefficient r and values of parameters α_i and t_i^*. The optimal values of model parameters (providing the best fitting between the theoretical $V^{(v)}(t)$ curves and the results of observations $V_j^{(v)}$) correspond to the maximum value of the correlation coefficient r. Thus, the parameter identification problem can be reduced to the problem of searching the maximum of complicated but analytical function r. For $I^{(v)} = I$ (a completely visible epidemic), such approach was successfully used to simulate and predict the dynamics of mysterious children disease (Nesteruk, 2017), COVID-19 pandemic (Nesteruk, 2021a; Nesteruk, 2021b; Nesteruk, 2023b) and the pertussis epidemic in England (Nesteruk, 2024a).

Let us illustrate the parameter identification procedure for two waves of the pertussis epidemic in England in 2023 and 2024 discussed in (Nesteruk, 2024a). The accumulated confirmed numbers of cases ($V_j^{(v)}$) and corresponding moments of time t_j are

TABLE 1 Accumulated numbers of confirmed pertussis cases in England in 2023 and 2024 and estimations of the average daily numbers of visible cases.

Months, starting with January 2023	Days, starting with 1 January 2023, t_j	Accumulated numbers of visible cases, $V_j^{(v)}$, (Confirmed cases of pertussis in England by month, n.d.)	Calculated daily numbers of new visible cases at moments t_j dV/dt, equation 1 in Nesteruk (2024a)
1	31	9	–
2	59	18	0.35426
3	90	30	0.52688
4	120	50	0.86559
5	151	83	1.41559
6	181	136	2.04462
7	212	208	2.66129
8	243	301	3.20
9	273	403	3.34516
10	304	505	3.47849
11	334	615	5.75269
12	365	858	12.87096
13	396	1,413	24.79644
14	425	2,332	38.86096
15	456	3,759	58.23280
16	486	5,872	84.44247
17	517	8,924	89.67581
18	547	11,351	67.65968
19	578	13,038	44.27419
20	609	14,096	28.79785
21	639	14,800	19.94301
22	670	15,309	–

listed in Table 1 according to the official site of UK government [(Confirmed cases of pertussis in England by month, n.d.), version available on 10 August 2024, the last 4 figures were taken from 10 December 2024 version]. The values $V_j^{(v)}$ were used to calculate approximate daily numbers of new cases dV/dt at moments t_j according to the Equation 1 from (Nesteruk, 2024a) (see the last column in Table 1).

Since the general problem contains 10 unknown parameters, their identification needs high performance computing and applying AI methods even for the analytical solution Equations 12, 13, 18–20, 22, 23. When this solution is approximate, the optimal values of parameters will contain discrepancies, which can reduce the accuracy of predictions. For our example, let us take the case of exact solution $I^{(v)} = I(\beta = 1)$ and assume that at the beginning of every new epidemic wave all infectious persons from the previous waves are removed. Then we can use initial Equation 8, perform simulations and then add cases accumulated at moments when the monthly numbers of visible cases started to increase. For every wave we will have only four unknown parameters $N_i, \alpha_i, \rho_i^{(v)}$ and t_i^*. Due to linear relation 19 and using linear regression, only two of these parameters are independent.

The first and second epidemic waves were simulated with the use of $V_j^{(v)}$ and t_j corresponding to j = 3–9 and j = 11–18, respectively. The optimal values of parameters (corresponding to the maximum of correlation coefficients r = 0.999721131761662 (0.999822556136920)) are:

Ni = 50739.992 (3657890.47292358);
α_i =1.50488802805402e-05 (2.11266650706657e-06) [day]$^{-1}$;
$\rho_i^{(v)}$=0.752503596035381 (7.7088126606047) [day]$^{-1}$;
t_i^*=89.1566573802040 (333.429529143536) days.

(figures in brackets correspond to the second wave). It should be noted that these optimal values are very different for the first and second waves and differ from the figures in Nesteruk (2024a) for the simulation of the first wave using observations with j = 1–10.

Using these optimal values in the analytical solution Equations 10, 12, 18–20, 22, 23 yielded the theoretical curves shown in Figure 1 (solid and dashed for the first and the second wave, respectively). The predicted values (see blue and black "crosses" for j = 19–22, July–October 2024) are in good agreement with the theoretical blue and black curves. To estimate the accuracy of 4-month prediction, let us take the accumulated number of visible cases $V_{22}^{(v)}$ =15,309 corresponding to the end of October 2024 (see Table 1) and compare with the theoretical value $V^{(v)}\big|_{t=670}$ = 17,430 corresponding to the blue dashed line in Figure 1. After adding 505 cases accumulated at t_{10} and extracted for the simulation of the second wave (compare blue "crosses" and "circles" in Figure 1), we obtain the accuracy (17,935-15,309)/15,309 around 17%. Since the final number of visible cases decreases with

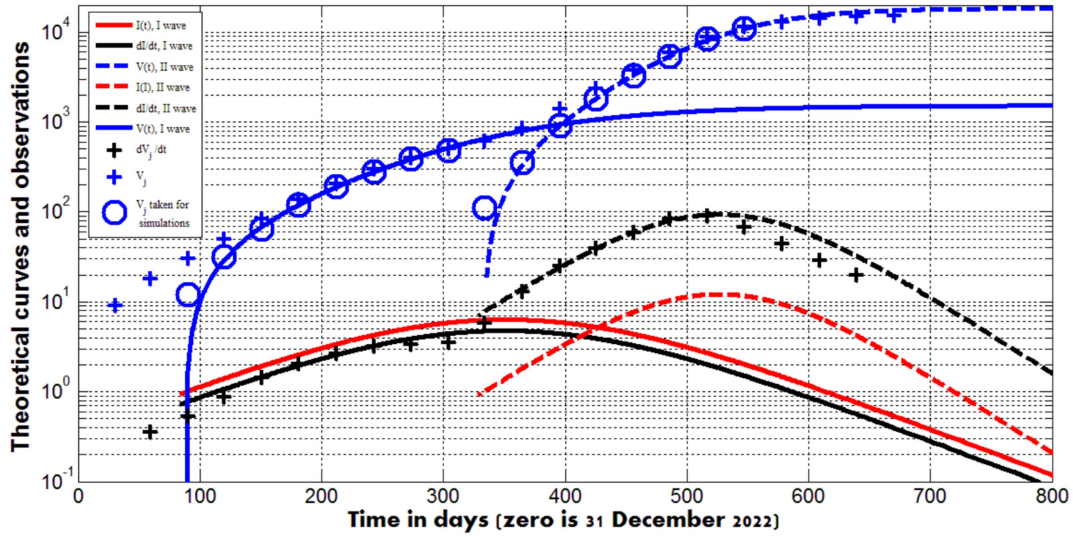

FIGURE 1
Accumulated numbers of visible pertussis cases (blue curves, the first Equation 9); the average daily numbers of new cases (black curves, the first Equation 10); numbers of infectious persons (red curves, Equation 18). "Circles" represent the confirmed numbers of cases $V_j^{(v)}$ taken for identification of parameters of the first (j = 3–10) and second (j = 11–18) waves; blue "crosses" – all confirmed numbers of cases $V_j^{(v)}$ listed in Table 1; black "crosses" – results of calculations of approximate daily numbers of new cases at moments t_j listed in the last column of Table 1.

the increase of the visibility coefficient (see the next Section), we can expect to obtain a lower theoretical value $V^{(v)}\big|_{t=670}$ and better accuracy after the real visibility coefficient will be calculated and taken into account. The accuracy of 17% is comparable with the long-time predictions for the first waves of the COVID-19 in different countries (Nesteruk, 2021a) and even for the case $\beta = 1$ is likely to allow healthcare professionals to develop the right strategy. The average daily numbers of new cases will be less than 1.0 only in March 2025. If trend will not change, the monthly figure of 9 new cases (as it was before starting the first wave, see Table 1) can be achieved only in May 2025.

5 Examples of exact solutions at different values of the visibility coefficient

The use of initial Equation 8 allows reducing the numbers of unknown parameters by 4. Then, Equations 12, 14 yield

$$V^{(v)} = \frac{N_i - S}{\beta_i}, \tag{24}$$

$$V = \beta_i V^{(v)} \tag{25}$$

According to Equation 25 the real accumulated numbers of new cases V are β_i times higher than visible figures $V^{(v)}$registered during the fixed epidemic wave, if all infectious patients were removed before this wave started.

In the case (I), i.e., equal removing rates for visible and hidden patients $\rho_i^{(v)} = \rho_i^{(h)} = \rho_i$, another simplification can be obtained with the use of Equation 18:

$$I = -S + N_i + \frac{\rho_i}{\alpha_i} \ln \frac{S}{N_i - \beta_i} \tag{26}$$

Assuming that spreading the infection stops when the real number of infectious I (visible and hidden) is less than 1.0, the corresponding final number of susceptible S_f can be obtained as a solution of the non-linear equation:

$$S_f + 1 = N_i + \frac{\rho_i}{\alpha_i} \ln \frac{S_f}{N_i - \beta_i} \tag{27}$$

following from Equation 26 and allowing us to calculate the corresponding final accumulated numbers of visible and total cases with the use of Equations 24, 25.

Figure 2 represents the results of calculations for the optimal values of parameters corresponding to two pertussis waves in England (see previous Section) and the first COVID-19 pandemic waves in Austria and the UK (in brackets), (Nesteruk, 2021a):

N_i = 75176.032 (479782.4);
α_i =1.924971386379e-05 (9.1371956639e-07) [day]$^{-1}$;
$\rho_i^{(v)}$=1.29635017900866 (0.330545378991741) [day]$^{-1}$.

Dashed lines demonstrate that the final accumulated numbers of all cases (Equations 25, 26) very slightly depend on the visibility coefficient. The final accumulated numbers of visible cases (Equations 24, 26) diminish with the increase of β_i (see solid curves). The values of other parameters are fixed and

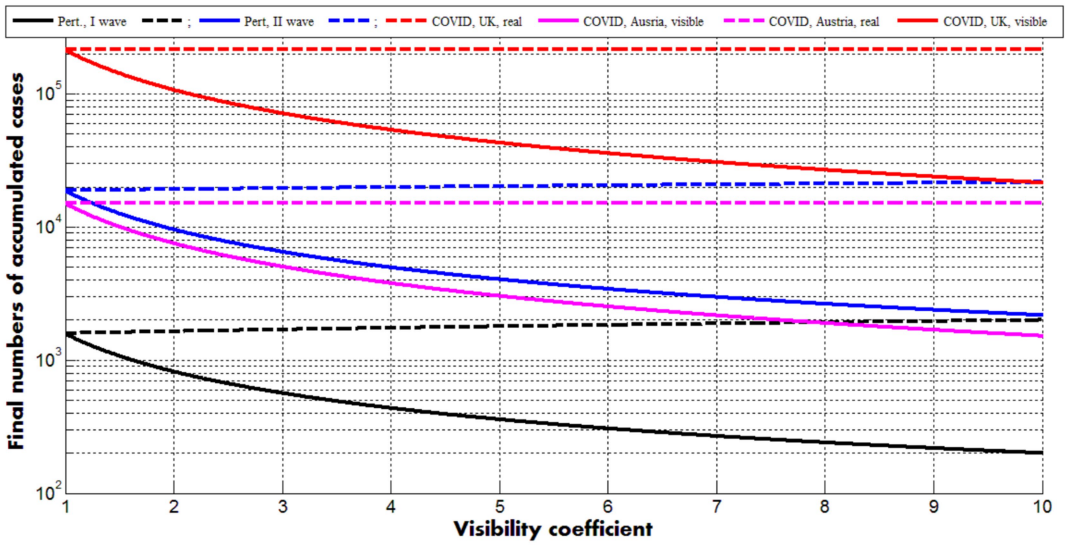

FIGURE 2
Solid curves represent final numbers of visible (registered) cases (Equations 24, 27) dashed ones–final numbers of all cases (registered and unregistered, Equations 25, 27). Back and blue lines correspond to the optimal values of parameters (listed in previous Section) for the first and second pertussis waves in England, respectively. Magenta and red curves show the results for the first COVID-19 waves in Austria and the UK (Nesteruk, 2021a), respectively.

correspond to the case β_i=1. The values $V_j^{(v)}$ for j = 20, 21, 22, which are smaller then theoretical prediction for the second pertussis wave (compare blue "crosses" and the blue dashed line in Figure 1), reflect reducing the final value of $V^{(v)}$for β_i>1. Nevertheless, good estimations of the visibility coefficient can be obtained only with the use of all parameters. Since removing rates can be different for symptomatic and asymptomatic patients, a general parameter identification procedure needs a numerical solution of the set of differential Equations 1–5 and huge numbers of calculations, which can be performed only with the use of high performance computing and AI methods. The full parameter sensitivity analysis will be considered in future research.

With the use of Equations 24, 10, 26 can be rewritten as follows:

$$I = \beta_i V^{(v)} + \frac{\rho_i}{\alpha_i} \ln \frac{N_i - \beta_i V^{(v)}}{N_i - \beta_i}, \tag{28}$$

$$\frac{dV^{(v)}}{dt} = \frac{\alpha_i}{\beta_i}\left(N_i - \beta_i V^{(v)}\right) I, \quad \frac{dV}{dt} = \alpha_i \left(N_i - \beta_i V^{(v)}\right) I \tag{29}$$

Figure 3 represents the calculations of real number of infectious I (visible and hidden, Equation 28) and visible and real numbers of new daily cases (Equation 29) versus accumulated numbers of visible cases $V^{(v)}$ for different values of the visibility coefficient. Values of other parameters correspond to the second pertussis wave in England (listed in the previous Section).

The maximum values of infectious slightly increase with the increase of the visibility coefficient (compare red lines in Figure 3). The black and the solid blue curves show the same but more pronounced trend for the real numbers of new daily cases. The positions of the maxima on these lines are very close to ones on the red curves. Corresponding moments of time can be calculated with the use of Equations 19, 20, 24. Thus, the average daily numbers of visible cases reflects trends in real numbers of infectious persons (symptomatic and asymptomatic) and can be used to control epidemics. The numbers of new visible daily cases decrease with the increase of visibility coefficient (for fixed values of other parameters and $\rho_i^{(v)} = \rho_i^{(h)} = \rho_i$, see blue curves in Figure 3). The final numbers of visible cases demonstrate the same trend (see solid curves in Figure 2).

6 Conclusion

To simulate hidden epidemic dynamics connected with asymptomatic and unregistered patients, a new general SIR model was proposed containing 5 unknown functions. For some cases, the analytical solutions of the set of 5 differential equations were found which allow simplifying the parameter identification procedure. Two waves of the pertussis epidemic in England in 2023 and 2024 were simulated for the case of zero hidden cases. Observations of accumulated visible numbers of cases during 4 months revealed rather high accuracy of predictions. If trend will not change, the monthly figure of 9 new pertussis cases (as it was in January–February 2023) can be achieved only in May 2025. The proposed approach can be recommended both for preliminary simulations of different epidemics (supposing zero hidden cases) and for further research, using presented analytical solutions or numerical integration of differential equations. The theoretical estimations of numbers of hidden cases will allow healthcare professionals to

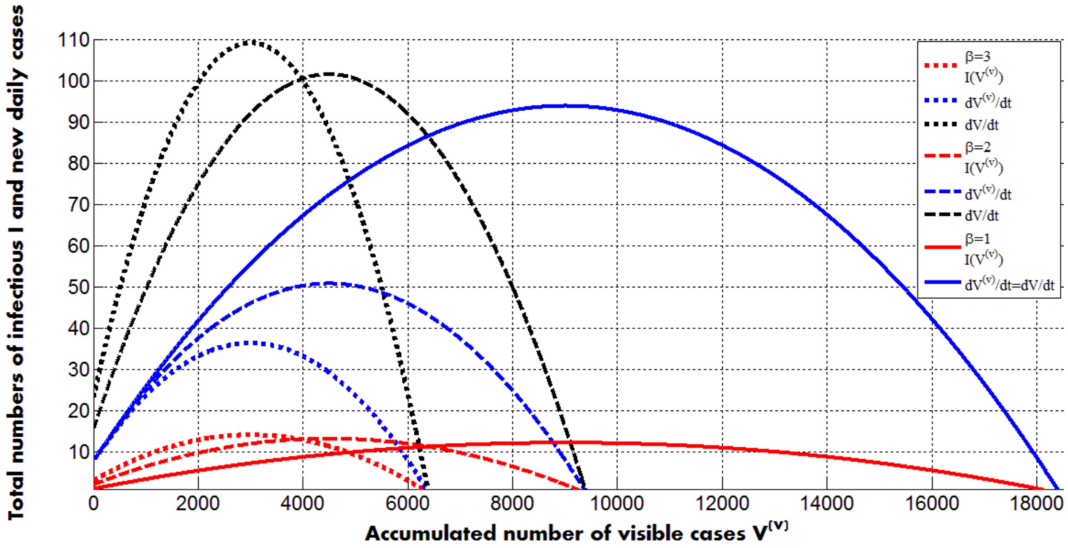

FIGURE 3
Solid, dashed and dotted curves represent calculations for values of the visibility coefficient 1; 2 and 3, respectively. Other values of parameters correspond to the second pertusis wave in England in 2023 and 2024 (listed in the previous Section). Red color corresponds to the real number of infectious persons (symptomatic and asymptomatic, Equation 28). Blue lines represent visible numbers of new daily cases; black ones - real (registered and hidden) numbers of new daily cases (Equations 28, 29).

know the real sizes of epidemics and to develop the right strategy without expensive mass testing.

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

IN: Conceptualization, Data curation, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing, Validation.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article.

Acknowledgments

The author is grateful to Ulrike Tillmann, James Robinson, Robin Thompson, Matt Keeling, Paul Brown, and Oleksii Rodionov for their support and providing very useful information. This paper was written with the support of the INI-LMS Solidarity Programme at the University of Warwick, UK.

Conflict of interest

The 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.

Generative AI statement

The authors declare that no Generative AI was used in the creation of this manuscript.

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.

References

An experiment with mass testing for COVID-19 was conducted in Khmelnytsky. Podillya news. Available at: https://podillyanews.com/2020/12/17/u-shkolah-hmelnytskogo-provely-eksperyment-z-testuvannyam-na-covid-19/ (Accessed November 23, 2024)

Britton, N. F. (2004). Essential mathematical biology. India: Springer (India) Pvt. Limited, 352.

Cherniha, V. (2020). Davydovych, a mathematical model for the COVID-19 outbreak and its applications. *Symmetry* 12:990. doi: 10.3390/sym12060990

Confirmed cases of pertussis in England by month. GOV.UK. Available at: www.gov.uk. (Accessed November 23, 2024)

Craig, R., Kunkel, E., Crowcroft, N. S., Fitzpatrick, M. C., de Melker, H., Althouse, B. M., et al. (2020). Asymptomatic infection and transmission of pertussis in households: a systematic review. *Clin. Infect. Dis.* 70, 152–161. doi: 10.1093/cid/ciz531

Daley, D. J., and Gani, J. (2005). Epidemic modeling: An introduction. eds. C. Cannings, F. C. Hoppensteadt, and L. A. Segel (New York: Cambridge University Press).

Davies, N. G., Klepac, P., Liu, Y., Prem, K., and Jit, M.CMMID COVID-19 working group; Eggo RM. (2020). Age-dependent effects in the transmission and control of COVID-19 epidemics. *Nat. Med.* 26, 1205–1211. doi: 10.1038/s41591-020-0962-9

Draper, N. R., and Smith, H. (1998). Applied regression analysis. *3rd* Edn: John Wiley.

Fowlkes, A. L., Yoon, S. K., Lutrick, K., Gwynn, L., Burns, J., Grant, L., et al. (2022). Effectiveness of 2-dose BNT162b2 (Pfizer BioNTech) mRNA vaccine in preventing SARS-CoV-2 infection among children aged 5-11 years and adolescents aged 12–15 years -PROTECT cohort, July 2021–February 2022. *MMWR Morb. Mortal Wkly. Rep.* 71, 422–428. doi: 10.15585/mmwr.mm7111e1

Hethcote, H. W. (2000). The mathematics of infectious diseases. *SIAM Rev.* 42, 599–653. doi: 10.1137/S0036144500371907

Keeling, M., and Rohani, P. (2008). Modeling infectious diseases in humans and animals. Princeton, NJ: Princeton University Press.

Kermack, W. O., and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *J. Royal Stat. Soc. Ser A.* 115, 700–721.

Mass coronavirus testing in Slovakia-a. (2020). Available at: https://edition.cnn.com/2020/11/02/europe/slovakia-mass-coronavirus-test-intl/index.html (Accessed November 23, 2024)

Mass coronavirus testing in Slovakia-b. (2020). Available at: https://www.voanews.com/covid-19-pandemic/slovakias-second-round-coronavirus-tests-draws-large-crowds (Accessed November 23, 2024)

Mohammadi, A., Meniailov, I., Bazilevych, K., Yakovlev, S., and Chumachenko, D. (2021). Comparative study of linear regression and SIR models of COVID-19 propagation in Ukraine before vaccination. *Radioelectron. Computer Syst.* 3, 5–18. doi: 10.32620/reks.2021.3.01

Nakamura, G. M., Cardoso, G. C., and Martinez, A. S. (2020). Improved susceptible–infectious–susceptible epidemic equations based on uncertainties and autocorrelation functions. *R. Soc. Open Sci.* 7:191504. doi: 10.1098/rsos.191504

Nesteruk, I. (2017). Statistics based models for the dynamics of Chernivtsi children disease. *Naukovi Visti NTUU KPI.* 5, 26–34. doi: 10.20535/1810-0546.2017.5.108577

Nesteruk, I. (2021e). Visible and real sizes of new COVID-19 pandemic waves in Ukraine. *Innov Biosyst Bioeng.* 5, 85–96. doi: 10.20535/ibb.2021.5.2.230487

Nesteruk, I. (2021c). Influence of possible natural and artificial collective immunity on new COVID-19 pandemic waves in Ukraine and Israel. *Exploratory Res. Hypothesis Med.* doi: 10.14218/ERHM.2021.00044

Nesteruk, I. (2021d). The real COVID-19 pandemic dynamics in Qatar in 2021: simulations, predictions and verifications of the SIR model. *Semina: Ciências Exatas Tecnol.* 42, 55–62. doi: 10.5433/1679-0375.2021v42n1Suplp55

Nesteruk, I. (2021a). COVID19 pandemic dynamics. Singapore: Springer Nature.

Nesteruk, I. (2021b). Detections and SIR simulations of the COVID-19 pandemic waves in Ukraine. *Comput. Math. Biophys.* 9, 46–65. doi: 10.1515/cmb-2020-0117

Nesteruk, I. (2023b). Improvement of the software for modeling the dynamics of epidemics and developing a user-friendly interface. *Infectious Disease Modelling* 8, 806–821. doi: 10.1016/j.idm.2023.06.003

Nesteruk, I. (2023a). Endemic characteristics of SARS-CoV-2 infection. *Sci. Rep.* 13:14841. doi: 10.1038/s41598-023-41841-8

Nesteruk, I. (2024b). Trends of the COVID-19 dynamics in 2022 and 2023 vs. the population age, testing and vaccination levels. *Front. Big Data* 6:1355080. doi: 10.3389/fdata.2023.1355080

Nesteruk, I. (2024a). Mathematical simulations of the pertussis epidemic in England. International workshop ProfIT AI 2024. Cambridge, MA, USA.

Nesteruk, I., and Keeling, M. (2023). Population age as a key factor in the COVID-19 pandemic dynamics. *Research Square* 30:2023. doi: 10.21203/rs.3.rs-3682693/v1

Pesco, P., Bergero, P., Fabricius, G., and Hozbor, D. (2014). Modelling the effect of changes in vaccine effectiveness and transmission contact rates on pertussis epidemiology. *Epidemics* 7, 13–21. doi: 10.1016/j.epidem.2014.04.001

Schreiber, P. W., Scheier, T., Wolfensberger, A., Saleschus, D., Vazquez, M., Kouyos, R., et al. (2023). Parallel dynamics in the yield of universal SARS-CoV-2 admission screening and population incidence. *Sci. Rep.* 13:7296. doi: 10.1038/s41598-023-33824-6

Shang, W., Kang, L., Cao, G., Wang, Y., Gao, P., Liu, J., et al. (2022). Percentage of asymptomatic infections among SARS-CoV-2 omicron variant-positive individuals: a systematic review and Meta-analysis. *Vaccines* 10:1049. doi: 10.3390/vaccines10071049

Weiss, H. (2013). The SIR model and the foundations of public health. *MatMat* 3, 1–17.

frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 12 March 2025
DOI 10.3389/frai.2025.1561079

OPEN ACCESS

EDITED BY
Jasleen Kaur,
University of Waterloo, Canada

REVIEWED BY
Ellen Cohn,
University of Pittsburgh, United States
Halyna Padalko,
University of Waterloo, Canada

*CORRESPONDENCE
Illya Chaikovsky
illya.chaikovsky@gmail.com

RECEIVED 15 January 2025
ACCEPTED 27 February 2025
PUBLISHED 12 March 2025

CITATION
Chaikovsky I, Dziuba D, Kryvova O, Marushko K, Vakulenko J, Malakhov K and Loskutov O (2025) Subtle changes on electrocardiogram in severe patients with COVID-19 may be predictors of treatment outcome.
Front. Artif. Intell. 8:1561079.
doi: 10.3389/frai.2025.1561079

Subtle changes on electrocardiogram in severe patients with COVID-19 may be predictors of treatment outcome

Illya Chaikovsky[1,2]*, Dmytro Dziuba[1], Olga Kryvova[3], Katerina Marushko[4], Julia Vakulenko[4], Kyrylo Malakhov[5] and Oleg Loskutov[1,4]

[1]Department of Anaesthesiology and Intensive Care, Shupyk National Healthcare University, Kyiv, Ukraine, [2]Department of Contactless Control Systems, V.M. Glushkov Institute of Cybernetics of the National Academy of Sciences, Kyiv, Ukraine, [3]Department of Medical Information Technologies, International Research and Training Center of the National Academy of Sciences, Kyiv, Ukraine, [4]Department of Anaesthesiology and Intensive Care for Infectious Diseases, Kyiv City Clinical Hospital No. 4, Kyiv, Ukraine, [5]Microprocessor Technology Lab, V.M. Glushkov Institute of Cybernetics of the National Academy of Sciences, Kyiv, Ukraine

Background: Two years after the COVID-19 pandemic, it became known that one of the complications of this disease is myocardial injury. Electrocardiography (ECG) and cardiac biomarkers play a vital role in the early detection of cardiovascular complications and risk stratification. The study aimed to investigate the value of a new electrocardiographic metric for detecting minor myocardial injury in patients during COVID-19 treatment.

Methods: The study was conducted in 2021. A group of 26 patients with verified COVID-19 diagnosis admitted to the intensive care unit for infectious diseases was examined. The severity of a patient's condition was calculated using the NEWS score. The digital ECGs were repeatedly recorded (at the beginning and 2–4 times during the treatment). A total of 240 primary and composite ECG parameters were analyzed for each electrocardiogram. Among these patients, 6 patients died during treatment. Cluster analysis was used to identify subgroups of patients that differed significantly in terms of disease severity (NEWS), SpO_2 and integral ECG index (an indicator of the state of the cardiovascular system).

Results: Using analysis of variance (ANOVA repeated measures), a statistical assessment of changes of indicators in subgroups at the end of treatment was given. These subgroup differences persisted at the end of the treatment. To identify potential predictors of mortality, critical clinical and ECG parameters of surviving (S) and non-surviving patients (D) were compared using parametric and non-parametric statistical tests. A decision tree model to classify survival in patients with COVID-19 was constructed based on partial ECG parameters and NEWS score.

Conclusion: A comparison of potential mortality predictors showed no significant differences in vital signs between survivors and non-survivors at the beginning of treatment. A set of ECG parameters was identified that were significantly associated with treatment outcomes and may be predictors of COVID-19 mortality: T-wave morphology (SVD), Q-wave amplitude, and R-wave amplitude (lead I).

KEYWORDS

electrocardiography, myocardial injury, severity, mortality, COVID-19, machine learning, cloud computing

1 Introduction

Experience with the pandemic has shown that the disease can pose a severe threat to the lives of patients. The main danger of the disease is acute respiratory syndrome and lung injury. However, patients may experience damage to other organs and systems: the cardiovascular system, the immune system, the liver and the kidneys. Myocardial injury occurred in at least 10% of unselected COVID-19 cases and up to 41% in critically ill patients or those with comorbidities (Cameli et al., 2020).

In the survivors, the majority showed long-term symptoms, now often referred to as long COVID-19 (Garg et al., 2021; Crook et al., 2021). One of the critical long-term clinical consequences of COVID-19 seems to be myocardial injury (Guzik et al., 2020; Italia et al., 2021; Akhmerov and Marbán, 2020).

Signs and symptoms of possible myocardial injury after COVID-19 may include severe fatigue, palpitations, chest pain, shortness of breath, postural orthostatic tachycardia syndrome (POTS) due to neurologic disturbances, post-exertional fatigue, and higher troponin levels (Lovell et al., 2022; Bandyopadhyay et al., 2020; Xie et al., 2022; Matsumori et al., 2022).

In addition, COVID-19 appears to cause severe myocarditis. It can affect the myocardium and pericardium, causing severe fatigue without other apparent symptoms (Lovell et al., 2022). Diagnosis of myocarditis is relatively inaccurate because both tests and diagnostic protocols lack accuracy. Some reports showed that symptoms persisted for an average of 47 days before being diagnosed by cardiac magnetic resonance (CMR) imaging (Ojha et al., 2021).

Therefore, it is extremely important to identify critical factors for assessing COVID-19 severity, predicting treatment outcomes, and optimizing individual treatment strategies (Izcovich et al., 2020; Fathi et al., 2021). It is known that 49 variables can provide valuable prognostic information about mortality and disease severity in patients with COVID-19 (Izcovich et al., 2020).

Numerous studies have confirmed that cardiac (Mueller et al., 2021) and other biomarkers may reflect cardiovascular injury and inflammation in COVID-19 and are strongly associated with poor prognosis and mortality (Battaglini et al., 2022; Wang et al., 2020). In addition, some electrocardiographic (Bergamaschi et al., 2021) and echocardiographic alterations (Long et al., 2021) appear to have prognostic implications for patients with COVID-19.

Several prognostic models have been developed to assess disease severity in patients with COVID-19 and predict mortality (Bertsimas et al., 2020; Pourhomayoun and Shakibi, 2021; Liang et al., 2020; Zhou et al., 2021; Yan et al., 2020; Kang et al., 2021).

Such classification models have usually been developed using various machine learning (ML) algorithms. For example, one neural network model has demonstrated 93% accuracy in predicting mortality based on patients' physiological status, symptoms, and demographic information (Pourhomayoun and Shakibi, 2021).

A multivariable logistic regression model and an online risk calculator based on 10 clinical indicators were proposed to predict critical illness development among hospitalized patients with COVID-19 (Liang et al., 2020). A support vector machine (SVM) model based on 11 routine clinical parameters was developed to assess the severity of COVID-19 patients (Zhou et al., 2021).

An interpretable mortality prediction model for COVID-19 patients was proposed by Yan et al. (2020) where the XGBoost ML algorithm was used to select predictors. The interpretable decision tree and the decision rule for 3 biomarkers that predict the survival of individual patients with more than 90% accuracy were obtained.

It should be noted that in one of the ML models for predicting the severity of COVID disease, among the 33 analyzed signs and indicators, there was the cardiac functional grading (according to New York Heart Association functional classification) (Kang et al., 2021). However, this cardiac indicator was excluded from the model because of its weak positive correlation with the severity of COVID-19.

In this context, the advanced analysis of ECG is highly demanded. This is especially true for patients with a normal or slightly changed electrocardiogram, i.e., if the analysis did not reveal any "major" category according to the, for example, Minnesota coding system. The only way to increase the diagnostic value of ECG examination is to develop proper information technology (IT)—a combination of up-to-date methods and equipment bound into a chain that provides collection, storage, pre-processing, interpretation, conclusion and dissemination of information (Chaikovsky, 2020).

At the same time, the advancement of diagnostic methods, especially instrumental ones (i.e., methods of functional diagnostics), primarily entails a constant increase of their "distributive capacity"—the ability to detect more and more and subtler changes in the function examined by one method or another. Such opportunities emerge due to progress in technical measurement tools of a specific function and even more due to the development of informational technologies. In other words, due to the creation of new metrics—numerical parameters using which one can assess the aspects of the functioning of various human organs and systems that were inaccessible before.

As a result, new ways of improving the diagnostic accuracy of a particular method within its traditional application scenarios are discovered. Additionally, familiar methods find unconventional uses in new areas.

Everything mentioned above fully applies to the new informational technologies for cardiac electrical activity assessment developed at V.M. Glushkov Institute of Cybernetics of the National Academy of Sciences of Ukraine.

The main goal set by the developers in this context was to make any electrocardiography informative. Routine ECG analysis is based on specific ECG syndromes or phenomena defined within one of the existing visual ECG analysis algorithms. However, in most cases, no ECG syndrome can be identified during the analysis of an individual electrocardiogram, at least not one that reflects cardiac pathology, i.e., belongs to the "major" category according to the Minnesota coding system, for example. During the routine analysis, one is forced to assign a single class to all these electrocardiograms—electrocardiograms with no primary ECG syndrome identified. However, the question arises: are all these electrocardiograms the

same in terms of their relative "distance" to the "ideal" electrocardiogram of a healthy human? They are not. Depending on the myocardial condition, this "distance" can be further or closer. Moreover, there is a reasonable hypothesis that this "distance" reflects the likelihood of serious cardiovascular events. This is where routine analysis of an electrocardiogram is uninformative.

That is why the Universal Scoring system method and software for ECG scaling that can provide the quantitative evaluation of the slightest changes in ECG signal were developed (Chaikovsky, 2020; Chaykovskyy et al., 2019). This approach is based on, first of all, measuring the maximum number of ECG parameters and heart rate variability and, secondly, on positioning each parameter value on a scale between the absolute norm and extreme pathology. The suggested approach follows a popular Z-scoring ideology, where quantitative, usually point-based assessment of test results is determined using a unique scale containing data about intra-group test results variation. To calculate the Z-score mean, the test value of the group and its standard deviation are needed (Colan, 2013).

This study aimed to investigate the value of a new electrocardiographic metric for detecting subtle myocardial injury in patients during COVID-19 treatment. And also, to test the hypothesis about the prognostic value of myocardial injury on the treatment outcome.

2 Materials and methods

2.1 Study design and patient characteristics

The study was conducted in 2021. A total of 26 patients with confirmed COVID-19 were monitored while on treatment in the intensive care unit (ICU) of the Kyiv Clinical Hospital #4. The hospitalization duration ranged from 5 to 27 days. All the patients were initially in a severe condition.

The vital signs were documented to evaluate the course of the illness: heart rate, blood oxygen saturation, blood pressure, body temperature, and respiration rate. Based on them, the severity of a patient's condition was calculated using a widely accepted NEWS score (Wang et al., 2020). Severe COVID-19 condition was defined as meeting NEWS aggregate score of 7 or over.

The process of patients' enrollment is presented on Figure 1.

Most patients had a history of chronic diseases, first of all CVD, but these diseases were in remission. Thus, only patients with no signs of instability in relation to heart disease and no gross changes in the electrocardiogram in accordance with Minnesota coding were included in the study. Based on this logic, as shown in Figure 1, four patients were excluded due to instability of their conditions in the context of comorbid heart diseases, and eight other patients were excluded because they had major pathological changes on the ECG according to Minnesota coding.

In 26 patients, an ECG in 12 leads by serial digital ECG device (Solvaig Ltd., Ukraine) was repeatedly recorded (at the beginning and several times during the treatment—from 2 to 4 times). Among these patients, six patients died during the treatment. The main characteristics of the patient's condition were recorded several times. The integral indicators were used to calculate the patient's severity according to the NEWS scale (National Early Warning Score) (Baker et al., 2021) and SAPS II (The Simplified Acute Physiology II Scale) (Allyn et al., 2016). A demographic and anthropometric values, clinical parameters and ICU characteristics are presented in Table 1.

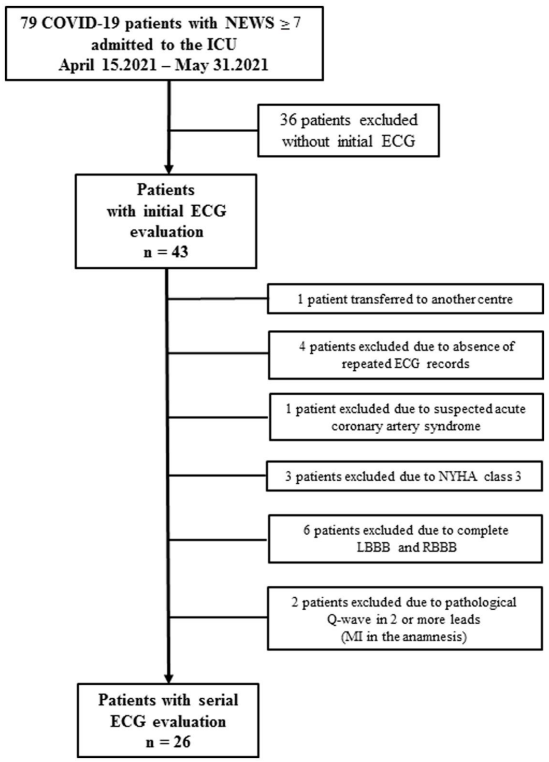

FIGURE 1
Flow-chart of the enrollment of COVID-19 patients with serial ECG evaluations.

2.2 Statistical analysis

Data are presented as means ± standard deviation (SD) or median (interquartile range, IQR) for continuous variables, based on normality and as percentages for categorical variables. A two-sample *t*-test compared the baseline characteristics of subjects within each group with unequal variances for continuous variables. Mann–Whitney *U* test was performed for variables that were not normally distributed. Two-tailed $p < 0.05$ was considered statistically significant.

The expectation–maximization (EM) clustering algorithm with 10-fold cross-validation was used to identify homogeneous groups. Homogenous groups were formed based on disease severity and integral index of patients at the beginning of treatment. As a result, two subgroups were identified that were significantly different in the severity of the disease and the integral indicator of the state of the cardiovascular system of patients at the start of treatment. Repeated measures ANOVA was used to evaluate statistical differences in the main clinical parameters in these subgroups at the beginning and the end of treatment.

We used machine learning algorithms (CART) such as Decision Trees to construct a model for classifying patient mortality. Statistical analysis was performed using Statistica 12.0 software.

3 Results

3.1 Correlation and cluster analysis, changes of integral parameters in the course of treatment

Previous studies have shown heterogeneity in clinical manifestations, severity and outcomes in COVID-19 patients. Our task was to study the heterogeneity of patients, taking into account the vital signs, the severity of the disease, and the state of the cardiovascular system (CVS). In addition, it was necessary to determine the influence of these factors on the treatment outcome.

We calculated correlation coefficients for all monitoring data to study the relationship between the CVS state's integral index and patients' vital signs during treatment. Mean values and Spearman correlation coefficients for the vital signs and the composite index of ECG U-score are shown in Table 2.

As one can see, a weak but statistically significant correlation exists between U-score composite ECG index and blood oxygen saturation, body temperature and NEWS score (in the last two cases - negative correlation).

Table 3 shows the distribution of patients according to the NEWS score, indicating the heterogeneity group of patients at the beginning of treatment.

To identify homogeneous subgroups (clusters), the sample of 26 patients was analyzed using EM cluster analysis with 10-fold cross-validation.

Hence, the problem of choosing the best parameters for clustering arises. In the general case, the task of selecting features for clustering consists of choosing the "best" set of features that helps to identify natural clusters according to the criterion. In our case, the criterion is the treatment outcome. Preliminary, feature dimensionality reduction methods were used to collapse features, in other words, to form composite features. Based on this logic, the electrocardiographic composite score called U-score, as well as composite scores NEWS and SAPS II for assessing the patient's condition in the intensive care unit, were primarily selected as features for clustering. Then, SAPS II was rejected due to its minimal variations during treatment. In addition, the SpO_2 value was chosen as a clustering parameter, since this independent feature is the most important vital sign in assessing the condition of patients with respiratory distress. Other vital signs such as HR, body temperature, as well age was also preliminarily investigated as potential parameters for clustering, but were rejected because the clusters resulting from the use of these parameters were clearly unrelated to the target criterion.

As a result, the NEWS score (as a categorial variable), SpO2 and U-score composite ECG index at the beginning of treatment have been taken for clusterization. NEWS and U-score were taken for clustering as the most integral indexes, and SpO_2 - since this is the vital sign for patients in the ISU for COVID-19.

As a result, two subgroups significantly differing from each other in SpO_2, NEWS score and U-score values were identified among these patients:

Cluster 1 included 19 patients with mean NEWS = 7.1, SpO_2 = 84.3, U-score = 60.5.

Cluster 2 included 7 patients with mean NEWS = 8.3, SpO_2 = 78.0, U-score = 49.8.

A diagram of standardized values of SpO_2 and U-score is shown in Figure 2. As we can see, at the beginning of treatment, patients from

TABLE 1 Patient characteristics at ICU admission.

Parameter	Value
Age (years)	63 ± 14.1
Sex (female)	10 (38.47%)
Sex (male)	16 (61.53%)
Weight (kg)	80.4 ± 16.65
Length (cm)	166.5 ± 9.09
BMI (kg/m^2)	28.48 ± 6.5
Hypertension	22 (84%)
Diabetes mellitus	9 (34%)
Ischemic heart disease	18 (69%)
Heart failure	11 (42%)
Pulmonary embolism	11 (42%)
Pulmonary hypertension	26 (100%)
Malignant disease	3 (11.5%)
Liver failure	2 (7.6%)
Vascular disease	4 (15.3%)
Days with symptoms at ICU admission	11.5 (5–27)
SAPS II	25 (9–35)
NEWS	7.34 (3–10)

Data are presented as (mean ± standard deviation), median (interquartile range) and numbers (percentages). ICU: Intensive care unit. BMI, body mass index; SAPS II, Simplified Acute Physiology Score II; NEWS, National Early Warning Score.

TABLE 2 Mean values and Spearman's correlation coefficients for vital signs and U-score composite ECG index.

Variables	Mean ± SD	HR	SpO_2	*t*	NEWS	SAPS II	U-score
HR (heart rate)	80.05 ± 11.98	1.00	−0.24*	−0.10	0.10	0.06	−0.11
SpO_2 (blood oxygen saturation)	86.46 ± 4.76	−0.24*	1.00	−0.24*	−0.64*	−0.12	0.28*
Body temperature (t°C)	37.17 ± 0.66	−0.10	−0.24*	1.00	0.46*	−0.52	−0.22*
NEWS score	6.14 ± 2.24	0.10	−0.64*	0.46*	1.00	−0.25	−0.25*
SAPS II	23.5 ± 6.14	0.06	−0.12	−0.52*	−0.25	1.00	−0.17
U-score	59.49 ± 10.66	−0.11	0.28*	−0.22*	−0.25*	−0.17	1.00

SD, standard deviation, *$p < 0.05$.

cluster 1 have higher levels of oxygen and an integral ECG index compared to cluster 2. Cluster 2 (subgroup 2) is characterized by a combination of greater severity with low oxygen level and lower U-score (integral ECG index level).

The average values of vital signs in the identified subgroups are shown in Table 4.

In addition, the subgroup with a more severe course of the disease (cluster 2) significantly differs in the age of patients. In subgroup 2, patients are older but do not differ in physiological severity. They do not have a significant difference in the indicator of physiological severity SAPS II.

We studied the dynamics of the abovementioned main parameters (SpO_2, NEWS, U-score) in two clusters throughout treatment using a repeated measures analysis of variance (RepANOVA). The changes of these parameters at the beginning (1) and by the end (2) of treatment are shown in Figures 3–5.

As shown from the figures above, both subgroups show a decrease in NEWS severity score and an increase in SpO_2 as a result of treatment, and these changes are statistically significant. The impact of therapy on main parameters in subgroups can be assessed by partial effect sizes (partial eta-squared, η_p^2). The effect of increasing oxygen in each cluster is significant: cluster 1 R1 SpO_2 $\eta_p^2 = 0.73$, $p = 0.00001$; cluster 2 $\eta_p^2 = 0.72$, $p = 0.007$.

In addition, the dynamics of severity reduction in the 1st subgroup is more pronounced. Sub NEWS partial eta-squared = $\eta_p^2 = 0.7$, $p = 0.0001$. Note that a decrease in the mean NEWS is not statistically significant in subgroup 2 (cluster 2 with severe baseline).

In addition, the dynamics of severity reduction in the 1st subgroup is more pronounced. Sub NEWS partial eta-squared = $\eta_p^2 = 0.7$, $p = 0.0001$. Note that a decrease in the mean NEWS is not statistically significant in subgroup 2 (cluster 2 with severe baseline).

The U-score composite index has a positive tendency to increase. However, the wide dispersion observed indicates a heterogeneity of U-score changes. Note that in subgroup 2, with low initial levels of integral indicator and oxygen, the part of unfavorable outcomes (ratio deceased/survivors) is 3 out of 7, >3 out of 19 in subgroup 1.

TABLE 3 The distribution of patient severity score at the beginning of treatment.

NEWS score	Number of cases	%
3	2	6.25
5	2	6.25
6	2	6.25
7	8	25.00
8	13	40.63
9	3	9.38
10	2	6.25

FIGURE 2
Normalized means SpO_2 and U-score in cluster 1 and cluster 2.

3.2 Comparison of vital signs and ECG indicators in two groups by the outcome of treatment

Next, we studied the differences between patient groups formed according to treatment outcomes. The study group of 26 patients consisted of two classes according to the treatment outcome: 20 survivors and 6 non-survivors. Clinical data and ECG parameters at the beginning and end of treatment were compared between survivors (S) and non-survivors (D).

At the beginning of treatment, there were no significant differences between groups S and D in the vital signs (SAPS II, SpO_2, NEWS score and U-score), except for the patient's age and body temperature (Table 5).

Then, an analysis was performed to identify any statistically significant differences among all partial ECG indicators constituting the U-score. The results of comparisons of ECG indicators in these

TABLE 4 Comparison of vital signs in cluster 1 and cluster 2 (*t*-test).

Vital signs	Cluster 1 (n = 19) Mean ± SD	Cluster 2 (n = 7) Mean ± SD	p-value
Age	59.63 ± 10.23	71.43 ± 18.48	0.047
BR (breathing rate)	23.84 ± 3.20	24.86 ± 2.91	0.470
HR	82.05 ± 15.06	88.29 ± 11.35	0.331
SpO_2	84.37 ± 2.48	78.00 ± 4.47	0.0001
Body temperature	37.46 ± 0.81	37.79 ± 0.73	0.358
NEWS	7.16 ± 1.77	8.29 ± 1.25	0.136
SAPS II	23.19 ± 5.75	25.67 ± 7.37	0.413
U-score	60.05 ± 11.87	49.86 ± 7.90	0.046

SD, standard deviation; n, number of patients.

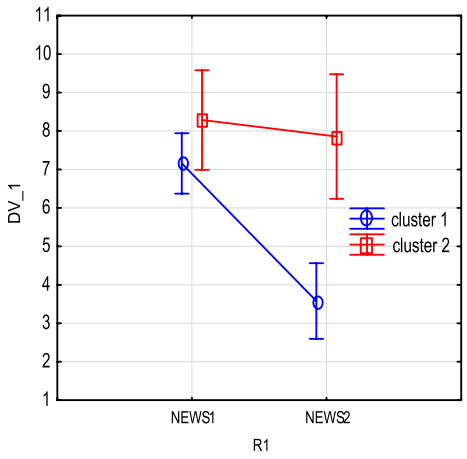

FIGURE 3
NEWS severity score changes due to treatment in clusters 1 and 2.

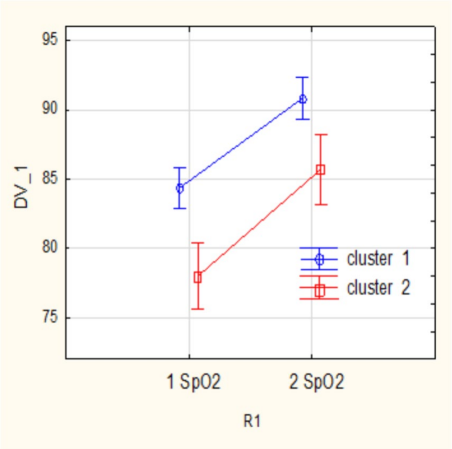

FIGURE 4
SpO_2 changes due to treatment in clusters 1 and 2.

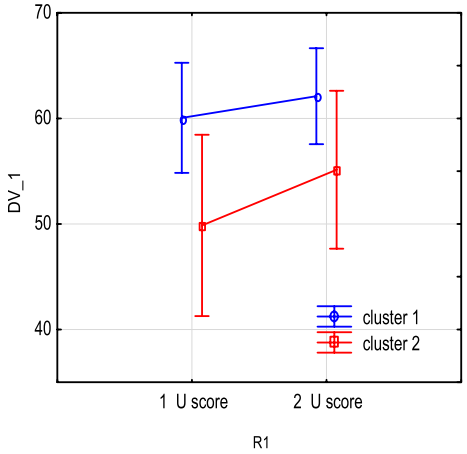

FIGURE 5
U-score integral ECG index changes due to treatment in clusters 1 and 2.

groups at the beginning of treatment according to the *t*-test are shown in Table 6 and according to the Mann–Whitney *U* test in Table 7.

At the beginning of the treatment, statistically significant differences were observed among the following parameter values:

1 T-wave morphology (SVD);
2 P/QRS integrals ratio;
3 U-score based composite index of cardiovascular events risk;
4 R-wave amplitude (μV) (lead I);
5 Integral of QRS complex (lead I).

By the end of the treatment, statistically significant differences were observed, in addition to the severity of the disease, among the same ECG parameters and also in one of the fundamental heart rate variability indicators, RMSSD, which represents parasympathetic nervous system activity.

3.3 Vital signs and ECG indicators as treatment outcome predictors

In the last step of statistical treatment, we explored which parameters had the highest correlation with treatment outcomes using vital signs and the entire ECG and HRV dataset. We used univariate feature selection for classification based on the Chi-square statistical test (Tables 7, 8). Variables with higher Chi-square statistical values are then selected as predictors for classification.

As follows from Table 8, vital signs at the beginning of treatment cannot be selected as treatment outcome predictors.

ECG indicators that significantly correlate with the treatment outcome class attribute and previously identified features (independent-group *t*-test) listed in Table 9 ranged *p*-value.

After testing the features with several statistics, we applied the wrapping method to the extracted features. With this approach, we evaluate the effectiveness of a subset of features, considering the final result of the applied learning algorithm (increase in accuracy when solving the classification problem).

We used the CART machine learning algorithm (decision tree method) with 10-fold cross-validation to divide patients into two classes (S, D) according to the tested feature sets. As a result, a set of features with the highest classification accuracy (one classification error) was obtained. Figure 6 shows the optimal classification tree built based on three features.

Contribution of 3 ECG parameters to the resulting rules listed in Table 9. Classification accuracy—96%. One of the recovered patients was erroneously classified as deceased.

If one builds a tree using both parameters above and an additional attribute, NEWS score value, the result will be as shown in Figure 7. This tree has the same structure as the previous one. Still, it is right branch has an additional split determined by the NEWS score condition. In this case, the classification accuracy on the training set was 100%.

4 Discussion

This study and our previous works showed that the combination of ECG and HRV parameters has the best diagnostic value

TABLE 5 Comparison of mean vital signs between survivors and non-survivors (*t*-test).

ECG indicators	Survivors (*n* = 20) Mean ± SD	Non-survivors (*n* = 6) Mean ± SD	*p*-value
SAPS II	23.61 ± 5.91	25.00 ±	0.693
Age	59.30 ± 11.06	74.50 ±	0.013
SpO_2	82.95 ± 4.16	81.67 ±	0.521
Body temperature	37.47 ± 0.83	37.80 ± 0.63	0.023
NEWS	7.30 ± 1.89	8.00 ± 0.63	0.388
HR	77.25 ± 20.67	78.50 ± 13.91	0.891
U-score	58.10 ± 10.83	54.67 ± 15.23	0.541

TABLE 6 Mean ECG indicators for survivors and non-survivors at the beginning of treatment (*t*-test).

ECG indicators	Survivors (*n* = 20) Mean ± SD	Non-survivors (*n* = 6) Mean ± SD	*p*-value
T-wave morphology (SVD)	28.95 ± 49.16	98.67 ± 84.44	0.017
P/QRS integrals ratio	0.151 ± 0.09	0.062 ± 0.05	0.039
Integral of QRS complex (I)	0.023 ± 0.01	0.037 ± 0.02	0.025
U-score based composite index of cardiovascular events risk assessment	60.4 ± 21.30	38.3 ± 22.40	0.037
R-wave amplitude (μV) (lead I)	593.1 ± 231.00	861.8 ± 356.08	0.036
Q-wave amplitude (μV) (lead I)	−18.40 ± 29.19	−32.67 ± 43.83	0.358
R-wave amplitude (μV) (lead II)	412.85 ± 210.75	313.00 ± 263.49	0.345

n, number of patients.

TABLE 7 Chi-square statistics: outcomes—vital signs.

Vital signs (at the beginning of treatment)	Chi-square	*p*-value
Age	13.09	0.109
SAPS II	11.02	0.201
Body temperature	5.95	0.428
HR	5.15	0.741
BR	2.90	0.714
SpO_2	5.11	0.647
NEWS	3.211	0.782

TABLE 8 ECG indicators—treatment outcome predictors.

Treatment outcome predictors (at the beginning of treatment)	Chi-square	*p*-value
R-wave amplitude (mV) (lead II)	18.01	0.011
T-wave morphology (SVD)	9.56	0.043
Q-wave amplitude (μV) (lead I)	12.26	0.052
Shift of the ST segment after 0.08 s after point J (mV) (lead I)	13.23	0.104
R-wave amplitude (μV) (lead I)	11.17	0.135
U-score based composite index of cardiovascular events risk	8.29	0.141

TABLE 9 The contribution (rank) of parameters to the survival classification model.

Parameters	Variable rank	Importance
T-wave morphology (SVD)	100	1.000
R-wave amplitude (μV) (lead II)	98	0.977
Q-wave amplitude (μV) (lead I)	28	0.278

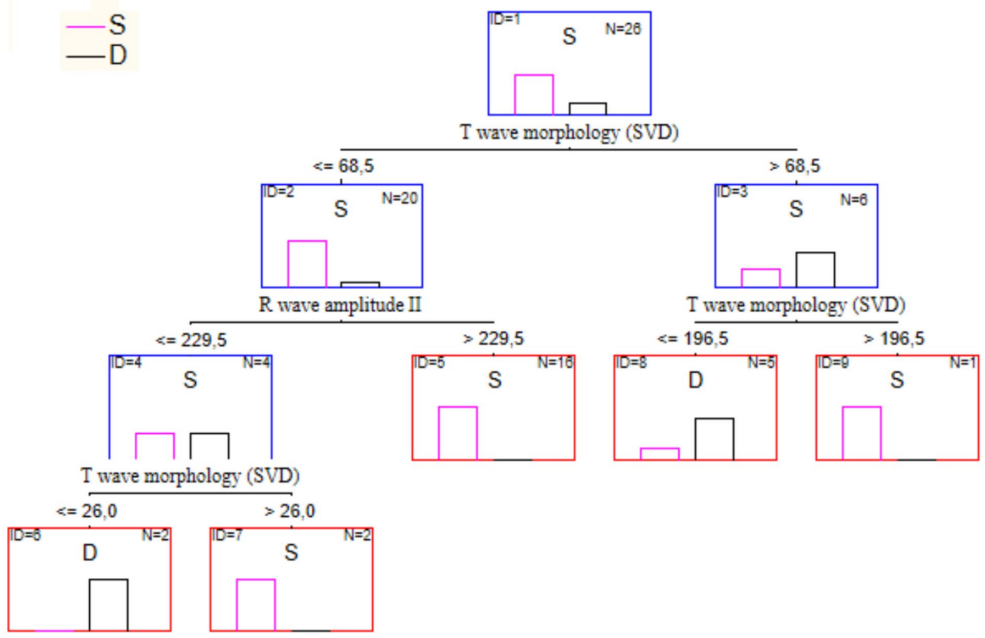

FIGURE 6
The optimal decision tree for the treatment outcome prognosis.

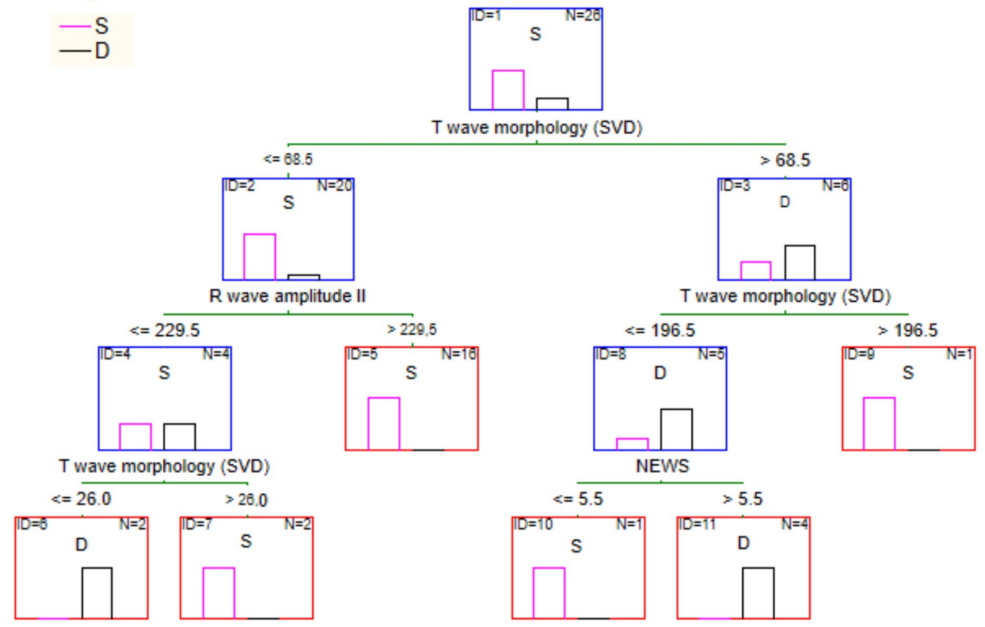

FIGURE 7
Full tree for the survival classification according to ECG parameters and NEWS score.

(Chaikovsky et al., 2019; Chaikovsky et al., 2020; Apykhtin et al., 2018; Patrick Neary et al., 2014; Clarke et al., 2020; Chaikovsky et al., 2020).

Changes in individual ECG and HRV parameters demonstrate only certain aspects of the examined phenomenon. Moreover, they can occur in opposite directions. Therefore, to reach a particular conclusion, in our case concerning the degree of subtle myocardial injury, a specific summarizing index that would synthesize the effects of individual components is necessary. The calculation method of such an index can be implemented in different ways. However, it must always include such sequential steps as theoretical justification of the composite index for the particular task selection of data adequate to the problem at hand analysis of this data, including its normalization, using methods of multivariate statistics selection of the informative private indicators (including the exclusion of correlated parameters)

and finally an actual construction of the composite index through the aggregation of private indicators. As was shown above, we have fulfilled all those steps. At the same time, it is crucial to consider that in addition to the composite index, various partial indicators are the most informative for detecting subtle changes in various clinical scenarios.

Usually, those are modern electrocardiographic indexes with a common pathophysiological basis. All of them assess the electrical homogeneity of the myocardium through different means, as the more heterogeneous the myocardium is from an electrical point of view (the higher the dispersion of the generated transmembrane action potentials in amplitude and length), the higher the likelihood of serious cardiovascular events. In our study, such a highly informative modern electrocardiographic index was the T-wave SVD.

The SVD of T-wave represents the complexity of ventricular repolarization. One major spatial component (eigenvector) can be identified when repolarization is uniform, as in normal individuals. Conversely, when the repolarization pattern becomes fragmented, the relative value of the smaller vectors increases proportionally. Such an approach allows a comparison between the morphology of the T-wave across the 12 leads and the quantification of T-wave abnormalities in an observer-independent way (Clarke et al., 2020).

A key element in conducting our original research was the integration of an advanced Hybrid Cloud Environment for Telerehabilitation (HCET) (Malakhov, 2024; Malakhov and Semykopna, 2024), developed at the Glushkov Institute of Cybernetics of the National Academy of Sciences of Ukraine. Building on the concept of Research and Development Workstation Environments (RDWEs) (Palagin et al., 2018), HCET provides a robust, ontology-driven, service-oriented architecture that supports each stage of clinical research: from data collection and secure storage to real-time analysis and interpretation of results.

The HCET infrastructure (Malakhov, 2024) merges distributed hardware resources, specialized software platforms, and dedicated telemedicine modules, enabling seamless collaboration among multidisciplinary teams. Its three-layer model – Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS), and Software-as-a-Service (SaaS) – ensures the flexible allocation and on-demand scaling of computational resources. At the core of the infrastructure is a high-performance server (HP ProLiant DL380p Gen8) running Ubuntu 22.04.3 LTS, configured with virtualization tools (KVM, LibVirt) to host multiple virtualized operating systems. This setup supports data-intensive applications and fosters reliable data processing pipelines.

Specifically, HCET facilitated several key steps in this study of ECG changes in severe COVID-19 patients:

Secure Data Acquisition and Management. Patient ECG data, along with relevant clinical parameters (e.g., NEWS scores), were collected in the ICU setting and securely transmitted to the HCET server. The platform's integrated Electronic Health Record (EHR) managing module ensured proper storage, encryption, and retrieval of sensitive information, adhering to ethical and legal requirements for patient confidentiality.

Advanced Analytical Tool Integration. The HCET environment included support for Statistica 12.0 software, which was pivotal for our repeated-measures ANOVA, cluster analysis, and decision tree modeling. By leveraging the platform's PaaS capabilities, researchers could seamlessly install, configure, and run Statistica within multiple virtual operating systems, optimizing both performance and collaborative workflows.

Interactive remote collaboration. Telemedicine components within HCET, such as the "Physician's Digital Workplace" and "Patient's Digital Cabinet," enabled continuous interaction between clinicians, biomedical engineers, data scientists, and other stakeholders. This not only streamlined the process of updating patient records but also allowed real-time discussion of ECG findings, NEWS scores, and relevant biomarkers across geographically dispersed locations.

Machine Learning and Decision Support. Building on the RDWE principles, the cognitive subsystem of HCET (referred to as the information-analytical subsystem) supported iterative model development. Automated interactive systems—such as OntoChatGPT (Palagin et al., 2023)—facilitated the semantic and contextual analysis of text documents. Researchers also used the specialized MedRehabBot (Palagin et al., 2024) services to manage domain-specific ontologies, aiding in the consistent labeling and categorization of ECG parameters. This approach improved the accuracy of clustering methods (Kaverinskiy et al., 2024a; Kaverinskiy et al., 2024b) used to pinpoint subtle myocardial injury and strengthened the decision-tree models predicting COVID-19 mortality.

Scalability and future directions. As additional clinical evidence accumulates, HCET's hybrid cloud setup ensures that resources can be dynamically scaled to accommodate larger patient cohorts, additional ECG recordings, and refined machine learning algorithms. Such scalability supports ongoing research into minor ECG changes, fostering deeper insights into myocardial injury, especially in post-COVID-19 care.

In essence, integrating the HCET platform into our experimental design bolstered every phase of this original research, from data stewardship to advanced statistical analyses. By uniting telemedicine, digital health, and cloud-based data processing services, HCET exemplifies how innovative hybrid cloud solutions can effectively support cutting-edge studies in physical medicine, rehabilitation, and beyond.

This work is the first study to assess minor electrocardiogram changes using the original scaling method in patients with COVID-19.

We reiterate that patients with unstable comorbid conditions and/or significant resting ECG abnormalities, as defined by Minnesota coding, were excluded from the study. This step was taken to ensure that any dynamic ECG changes observed would be attributed to COVID-19 rather than other underlying diseases. These findings underscore that our results may not be fully generalizable to individuals entirely free of comorbidities. Future research should expand recruitment to patients without these comorbidities to verify whether U-score based ECG changes exhibit similar predictive value.

Limitations of the study are the following: Firstly, the number of patients is relatively small. Secondly, no comparison of minor ECG changes with the levels of biomarkers of myocardial damage and inflammation was performed. Finally, the prognostic value of detected ECG changes regarding long-term COVID complications has yet to be analyzed.

Further larger-scale studies are planned to confirm and clarify the results.

5 Conclusion

The suggested ECG and HRV scaling method allow for registering and analyzing minor electrocardiogram changes during treatment. Modern ECG parameters used for advanced ECG analysis were the most informative. Contrary to this outline, the ECG analysis must be more informative for this task.

Two subgroups were identified that differed significantly in the severity of COVID-19 and the integral indicator of the cardiovascular system at the beginning of treatment. At the end of treatment, differences between subgroups remained. In the severe subgroup, there were 50 percent of deaths.

A comparison of potential predictors of mortality showed that at the beginning of treatment, there were no significant differences in vital signs between those who survived and those who died. In our study, the average age in the group of deceased patients was slightly higher, and the SAPS II score was not associated with the treatment outcome. A set of ECG parameters significantly associated with treatment outcome and may be predictors of treatment outcome were identified.

In addition to the U-score based composite index, partial indicators are the most informative for detecting subtle changes in various clinical scenarios, such as treatment outcome prediction. A decision tree for the survival classification of patients with COVID-19 was built based on the partial ECG parameters and NEWS score.

The obtained results allow us to create a combined decision rule that includes both the well-known NEWS-score and AI-based analysis of subtle ECG changes. Such a combined decision rule should have the highest possible predictive accuracy for COVID-19 Treatment Outcome. Further large-scale studies are needed to confirm these findings.

An important future direction involves utilizing innovative informatics infrastructures, such as the Hybrid Cloud Environment for Telerehabilitation (Malakhov, 2024; Malakhov and Semykopna, 2024), to manage and analyze large ECG datasets alongside clinical metrics. By delivering on-demand computational resources within an integrated platform, HCET enhances the scalability of machine learning pipelines for real-time ECG data processing.

Data availability statement

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

Ethics statement

The study was conducted according to the guidelines of the Declaration of Helsinki and approved by the Ethics Committee of Shupik National Healthcare University of Ukraine, Kyiv (protocol code no. 5 and date of approval is April 4, 2021). Informed consent was obtained from all subjects involved in the study. 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

IC: Conceptualization, Formal analysis, Methodology, Writing – original draft, Writing – review & editing, Project administration. DD: Writing – review & editing, Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Validation, Writing – original draft. OK: Software, Visualization, Writing – original draft. KaM: Investigation, Writing – original draft. JV: Resources, Validation, Writing – original draft. KyM: Software, Writing – review & editing. OL: Conceptualization, Methodology, Resources, Supervision, Validation, Writing – original draft.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Acknowledgments

The authors are sincerely grateful to Anna Starynska (Cardiolyse Oy, Finland) for her long-term support concerning data management.

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.

Generative AI statement

The author(s) declare that no Gen AI was used in the creation of this manuscript.

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.

References

Akhmerov, A., and Marbán, E. (2020). COVID-19 and the heart. *Circ. Res.* 126, 1443–1455. doi: 10.1161/circresaha.120.317055

Allyn, J., Ferdynus, C., Bohrer, M., Dalban, C., Valance, D., and Allou, N. (2016). Simplified acute physiology score II as predictor of mortality in intensive care units: a decision curve analysis. *PLoS One* 11:e0164828. doi: 10.1371/journal.pone.0164828

Apykhtin, K., Chaikovsky, I., Yaroslavska, S., Starynska, A., and Stadnyk, L. (2018). Adaptation of cardiovascular system to work in the night shifts of doctors and nurses. *J. Am. Coll. Cardiol.* 72:C243. doi: 10.1016/j.jacc.2018.08.1056

Baker, K. F., Hanrath, A. T., Schim Van Der Loeff, I., Kay, L. J., Back, J., and Duncan, C. J. (2021). National Early Warning Score 2 (NEWS2) to identify inpatient COVID-19 deterioration: a retrospective analysis. *Clin. Med.* 21, 84–89. doi: 10.7861/clinmed.2020-0688

Bandyopadhyay, D., Akhtar, T., Hajra, A., Gupta, M., Das, A., Chakraborty, S., et al. (2020). COVID-19 pandemic: cardiovascular complications and future implications. *Am. J. Cardiovasc. Drugs* 20, 311–324. doi: 10.1007/s40256-020-00420-2

Battaglini, D., Lopes-Pacheco, M., Castro-Faria-Neto, H. C., Pelosi, P., and Rocco, P. R. M. (2022). Laboratory biomarkers for diagnosis and prognosis in COVID-19. *Front. Immunol.* 13:857573. doi: 10.3389/fimmu.2022.857573

Bergamaschi, L., D'Angelo, E. C., Paolisso, P., Toniolo, S., Fabrizio, M., Angeli, F., et al. (2021). The value of ECG changes in risk stratification of COVID-19 patients. *Noninvasive Electrocardiol* 26:e12815. doi: 10.1111/anec.12815

Bertsimas, D., Lukin, G., Mingardi, L., Nohadani, O., Orfanoudaki, A., Stellato, B., et al. (2020). COVID-19 mortality risk assessment: an international multi-center study. *PLoS One* 15:e0243262. doi: 10.1371/journal.pone.0243262

Cameli, M., Pastore, M. C., Soliman Aboumarie, H., Mandoli, G. E., D'Ascenzi, F., Cameli, P., et al. (2020). Usefulness of echocardiography to detect cardiac involvement in COVID-19 patients. *Echocardiography* 37, 1278–1286. doi: 10.1111/echo.14779

Chaikovsky, I. (2020). Electrocardiogram scoring beyond the routine analysis: subtle changes matters. *Expert Rev. Med. Devices* 17, 379–382. doi: 10.1080/17434440.2020.1754795

Chaikovsky, I., Kryvova, O., Kazmirchyk, A., Mjasnikov, G., Sofienko, S., Bugay, A., et al. (2019). "Assessment of the post-traumatic damage of myocardium in patients with combat trauma using a data mining analysis of an electrocardiogram," in *2019 Signal Processing Symposium (SPSympo)*, Krakow, Poland: IEEE, p. 34–38.

Chaikovsky, I., Lebedev, E., Ponomarev, V., and Necheporuk, A. (2020). The relationship between ECG/HRV variables and socio-economic factors: results of mass screening in the rural region of Ukraine. *Eur. J. Prev. Cardiolog.* 27, S1–S119. doi: 10.1177/2047487320935268

Chaikovsky, I., Oshlianska, O., Artsymovych, A., Kryvova, O., Kovalenko, O., and Stadniuk, L. (2020). "Using of data mining methods to evaluate the myocardial damage in children with juvenile idiopathic arthritis," in *2020 IEEE 40th International Conference on Electronics and Nanotechnology* (ELNANO), Kyiv, Ukraine: IEEE, p. 391–395.

Chaykovskyy, I. A., Budnyk, M. M., and Starynska, G. A. (2019). Method of ECG evaluation based on universal scoring system. Washington, DC: U.S. Patent and Trademark Office.

Clarke, R., Chaikovsky, I., Wright, N., Du, H., Chen, Y., Guo, Y., et al. (2020). Independent relevance of left ventricular hypertrophy for risk of Ischaemic heart disease in 25, 000 Chinese adults. *Eur. Heart J.* 41:2938. doi: 10.1093/ehjci/ehaa946.2938

Colan, S. D. (2013). The why and how of Z scores. *J. Am. Soc. Echocardiogr.* 26, 38–40. doi: 10.1016/j.echo.2012.11.005

Crook, H., Raza, S., Nowell, J., Young, M., and Edison, P. (2021). Long COVID—mechanisms, risk factors, and management. *BMJ* 374:n1648. doi: 10.1136/bmj.n1648

Fathi, M., Vakili, K., Sayehmiri, F., Mohamadkhani, A., Hajiesmaeili, M., Rezaei-Tavirani, M., et al. (2021). The prognostic value of comorbidity for the severity of COVID-19: a systematic review and meta-analysis study. *PLoS One* 16:e0246190. doi: 10.1371/journal.pone.0246190

Garg, M., Maralakunte, M., Garg, S., Dhooria, S., Sehgal, I., Bhalla, A. S., et al. (2021). The conundrum of 'Long-COVID-19': a narrative review. *IJGM* 14, 2491–2506. doi: 10.2147/ijgm.s316708

Guzik, T. J., Mohiddin, S. A., Dimarco, A., Patel, V., Savvatis, K., Marelli-Berg, F. M., et al. (2020). COVID-19 and the cardiovascular system: implications for risk assessment, diagnosis, and treatment options. *Cardiovasc. Res.* 116, 1666–1687. doi: 10.1093/cvr/cvaa106

Italia, L., Tomasoni, D., Bisegna, S., Pancaldi, E., Stretti, L., Adamo, M., et al. (2021). COVID-19 and heart failure: from epidemiology during the pandemic to myocardial injury, myocarditis, and heart failure sequelae. *Front Cardiovasc Med* 8:713560. doi: 10.3389/fcvm.2021.713560

Izcovich, A., Ragusa, M. A., Tortosa, F., Lavena Marzio, M. A., Agnoletti, C., Bengolea, A., et al. (2020). Prognostic factors for severity and mortality in patients infected with COVID-19: a systematic review. *PLoS One* 15:e0241955. doi: 10.1371/journal.pone.0241955

Kang, J., Chen, T., Luo, H., Luo, Y., Du, G., and Jiming-Yang, M. (2021). Machine learning predictive model for severe COVID-19. *Infect. Genet. Evol.* 90:104737. doi: 10.1016/j.meegid.2021.104737

Kaverinskiy, V. V., Palagin, O. V., and Malakhov, K. S. (2024a). A distributed classification method for real-time healthcare data processing. *CEUR Workshop Proc.* 3777, 162–174.

Kaverinskiy, V. V., Vakulenko, D., Vakulenko, L., and Malakhov, K. S. (2024b). Machine learning analysis of arterial Oscillograms for depression level diagnosis in cardiovascular health. *CSIMQ* 2024, 94–110. doi: 10.7250/csimq.2024-40.04

Liang, W., Liang, H., Ou, L., Chen, B., Chen, A., Li, C., et al. (2020). Development and validation of a clinical risk score to predict the occurrence of critical illness in hospitalized patients with COVID-19. *JAMA Intern. Med.* 180, 1081–1089. doi: 10.1001/jamainternmed.2020.2033

Long, B., Brady, W. J., Bridwell, R. E., Ramzy, M., Montrief, T., Singh, M., et al. (2021). Electrocardiographic manifestations of COVID-19. *Am. J. Emerg. Med.* 41, 96–103. doi: 10.1016/j.ajem.2020.12.060

Lovell, J. P., Čiháková, D., and Gilotra, N. A. (2022). COVID-19 and myocarditis: review of clinical presentations, pathogenesis management. *Heart Int.* 16, 20–27. doi: 10.17925/hi.2022.16.1.20

Malakhov, K. S. (2024). Innovative hybrid cloud solutions for physical medicine and telerehabilitation research. *Int. J. Telerehab.* 16, e6635–e6619. doi: 10.5195/ijt.2024.6635

Malakhov, K. S., and Semykopna, T. V. (2024). Integrating hybrid cloud solutions in telerehabilitation. *CEUR Workshop Proc* 3806, 434–448.

Matsumori, A., Auda, M. E., Bruno, K. A., Shapiro, K. A., Kato, T., Nakamura, T., et al. (2022). Cardiovascular factors associated with COVID-19 from an international registry of primarily Japanese patients. *Diagnostics* 12:2350. doi: 10.3390/diagnostics12102350

Mueller, C., Giannitsis, E., Jaffe, A. S., Huber, K., Mair, J., Cullen, L., et al. (2021). Cardiovascular biomarkers in patients with COVID-19. *Eur. Heart J. Acute Cardiovasc. Care* 10, 310–319. doi: 10.1093/ehjacc/zuab009

Ojha, V., Verma, M., Pandey, N. N., Mani, A., Malhi, A. S., Kumar, S., et al. (2021). Cardiac magnetic resonance imaging in coronavirus disease 2019 (COVID-19): a systematic review of cardiac magnetic resonance imaging findings in 199 patients. *J. Thorac. Imaging* 36, 73–83. doi: 10.1097/rti.0000000000000574

Palagin, O. V., Kaverinskiy, V. V., Litvin, A., and Malakhov, K. S. (2023). OntoChatGPT information system: ontology-driven structured prompts for chat GPT Meta-learning. *IJC* 22, 170–183. doi: 10.47839/ijc.22.2.3086

Palagin, O. V., Kaverinskiy, V. V., Malakhov, K. S., and Petrenko, M. G. (2024). Fundamentals of the integrated use of neural network and Ontolinguistic paradigms: a comprehensive approach. *Cybern. Syst. Anal.* 60, 111–123. doi: 10.1007/s10559-024-00652-z

Palagin, O. V., Velychko, V. Y., Malakhov, K. S., and Shchurov, O. S. (2018). Research and Development workstation environment: the new class of current research information systems. *CEUR Workshop Proc.* 2139, 255–269.

Patrick Neary, J., Baker, T. P., Jamnik, V., Gledhill, N., Chaikovsky, I., Frolov, Y. A., et al. (2014). Multimodal approach to cardiac screening of elite ice hockey players during the NHL scouting combine. *Med. Sci. Sports Exerc.* 46:742. doi: 10.1249/01.mss.0000495720.24160.ee

Pourhomayoun, M., and Shakibi, M. (2021). Predicting mortality risk in patients with COVID-19 using machine learning to help medical decision-making. *Smart Health* 20:100178. doi: 10.1016/j.smhl.2020.100178

Wang, K., Zuo, P., Liu, Y., Zhang, M., Zhao, X., Xie, S., et al. (2020). Clinical and laboratory predictors of in-hospital mortality in patients with coronavirus Disease-2019: a cohort study in Wuhan, China. *Clin. Inf. Dis.* 71, 2079–2088. doi: 10.1093/cid/ciaa538

Xie, Y., Xu, E., Bowe, B., and Al-Aly, Z. (2022). Long-term cardiovascular outcomes of COVID-19. *Nat. Med.* 28, 583–590. doi: 10.1038/s41591-022-01689-3

Yan, L., Zhang, H.-T., Goncalves, J., Xiao, Y., Wang, M., Guo, Y., et al. (2020). An interpretable mortality prediction model for COVID-19 patients. *Nat. Mach. Intell.* 2, 283–288. doi: 10.1038/s42256-020-0180-7

Zhou, K., Sun, Y., Li, L., Zang, Z., Wang, J., Li, J., et al. (2021). Eleven routine clinical features predict COVID-19 severity uncovered by machine learning of longitudinal measurements. *Comput. Struct. Biotechnol. J.* 19, 3640–3649. doi: 10.1016/j.csbj.2021.06.022

frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 04 June 2025
DOI 10.3389/frai.2025.1597727

OPEN ACCESS

EDITED BY
Dmytro Chumachenko,
National Aerospace University – Kharkiv Aviation Institute, Ukraine

REVIEWED BY
Ramkumar Hariharan,
Northeastern University, United States
Yan Wang,
University of California, Los Angeles, United States

*CORRESPONDENCE
Neema Nicodemus Lyimo
✉ neemanico@sua.ac.tz

RECEIVED 25 March 2025
ACCEPTED 12 May 2025
PUBLISHED 04 June 2025

CITATION
Lyimo NN, Fue KG, Materu SF, Kilatu NC and Telemala JP (2025) Assessing the potential for application of machine learning in predicting weather-sensitive waterborne diseases in selected districts of Tanzania. *Front. Artif. Intell.* 8:1597727. doi: 10.3389/frai.2025.1597727

Assessing the potential for application of machine learning in predicting weather-sensitive waterborne diseases in selected districts of Tanzania

Neema Nicodemus Lyimo[1]*, Kadeghe Goodluck Fue[2], Silvia Francis Materu[3], Ndimile Charles Kilatu[4] and Joseph Philipo Telemala[1]

[1]Department of Informatics and Information Technology, Sokoine University of Agriculture, Morogoro, Tanzania, [2]Department of Engineering Sciences and Technology, Sokoine University of Agriculture, Morogoro, Tanzania, [3]Department of Biosciences, Sokoine University of Agriculture, Morogoro, Tanzania, [4]Department of Health, Morogoro Municipal Council, Morogoro, Tanzania

Introduction: This study evaluates the potential of machine learning (ML) to predict and manage weather-sensitive waterborne diseases (WSWDs) in selected Tanzanian districts, focusing on environmental health officers' (EHOs) knowledge and perceptions. It explores EHOs' familiarity with information and communication technology (ICT) and artificial intelligence (AI)/ML, alongside challenges and opportunities for integrating AI-driven public health solutions.

Methods: A census-style survey was conducted among EHOs in three district councils. A structured questionnaire, piloted in one district, was administered to 76 EHOs, achieving a 66% response rate. Data were analyzed using descriptive and inferential statistics to assess knowledge levels, perceptions, and gender-related differences.

Results: Most EHOs were moderately familiar with ICT; however, only 54% had prior exposure to AI/ML concepts, and 64% reported limited AI familiarity. Among the variables examined, only prior exposure to AI/ML concepts and self-reported familiarity with AI demonstrated statistically significant associations with gender. Despite this, the majority recognized AI/ML's potential to improve disease prediction accuracy. Key barriers to ML adoption include inadequate technical infrastructure, data quality issues, and a shortage of expertise. Opportunities identified included utilizing historical disease data, integrating AI with meteorological information, and using satellite imagery for surveillance.

Discussion: The study highlights frontline health workers' perceived barriers to ML adoption and suggests that gender influences awareness and engagement with AI and ML technologies. Strengthening technical capacity, improving data quality, and fostering cross-sector collaboration are critical for successful AI/ML integration. These insights offer a roadmap for resilience to WSWDs in developing countries like Tanzania through data-driven technologies.

KEYWORDS

climate-sensitive, artificial intelligence, Tanzania, predictive ML, waterborne diseases, developing countries

1 Introduction

The intersection of climate change and public health presents significant challenges, particularly in low-income countries like Tanzania, where water insecurity and waterborne diseases are prevalent (Mutono et al., 2021). Changes in rainfall patterns, rising temperatures, and extreme weather events driven by climate change exacerbate water quality issues, increasing the incidence of diseases such as cholera, diarrhea, and typhoid. In sub-Saharan Africa (SSA), including Tanzania, frequent extreme hydrological events—such as floods and droughts—result in water contamination and shortages, further compromising public health (Mutono et al., 2021; World-Bank, 2024). During flooding, runoff from pit latrines contaminates open-dug wells and rivers, heightening the risk of disease transmission (Mtwangi Limbumba et al., 2019; Mguni et al., 2016; Pantaleo et al., 2018). Conversely, droughts reduce access to clean water, creating hygiene challenges that foster environments conducive to disease outbreaks (Hunter et al., 2010; Howard et al., 2016).

The urban population in Tanzania is particularly vulnerable to these climate-sensitive diseases due to rapid urbanization and inadequate infrastructure development (Mshida et al., 2020). According to a report from WHO/UNICEF Joint Monitoring Programme for Water Supply, Sanitation, and Hygiene (2018/2019), only 47% of the urban population in Tanzania has access to basic sanitation, and merely 23.5% has access to basic hygiene facilities (Mshida et al., 2020). This lack of access to safe water increases the risk of diseases such as amoebiasis, typhoid fever, and dysentery among the population, particularly affecting children under 5 years (WHO, 2024). Nearly 1.7 billion cases of childhood diarrheal disease occur globally each year, ranking as the third leading cause of death among children aged 1–59 months (Adedokun and Yaya, 2020). In Tanzania, inadequate water, sanitation, and hygiene (WASH) services contribute to over 31,000 annual deaths and result in economic losses exceeding USD 2.4 billion due to increased medical expenses and reduced productivity (World-Bank, 2024). Studies have shown a direct correlation between rising temperatures and increased disease risk. For instance, for every 1°C increase in temperature, the likelihood of cholera outbreaks can rise by 15%–29% (Trærup et al., 2011). Additionally, erratic rainfall patterns elevate the risk of water contamination, further endangering public health (Mboera et al., 2011). Children from poorer households in SSA, particularly those living in areas with unimproved sanitation, face a higher risk of contracting waterborne diseases (Flückiger and Ludwig, 2022; Demissie et al., 2021; Adedokun and Yaya, 2020).

In response to these challenges, technological advancements, particularly in artificial intelligence (AI) and machine learning (ML), offer promising solutions for predicting and managing outbreaks of waterborne diseases. Research highlights various applications of AI in health sectors of low- and middle-income countries, including epidemiological surveillance and disease prediction (Adigwe et al., 2024). ML algorithms have shown remarkable potential in analyzing large datasets to identify patterns that can forecast disease outbreaks, often detecting anomalies that traditional methods might overlook. Recent studies highlight the effectiveness of ML models, such as Random Forest and Support Vector Machines (SVMs), in predicting disease cases based on historical patient data, achieving up to 77% accuracy for typhoid prediction when key input features like age and medical history are incorporated (Hussain et al., 2023; Flückiger and Ludwig, 2022). Additionally, IoT-based systems for real-time water quality monitoring have demonstrated their utility in early detection of potential health risks (Nemade et al., 2024). ML techniques, including deep learning models such as Long Short-Term Memory Networks (LSTMs) (Hochreiter and Schmidhuber, 1997) and Convolutional Neural Networks (CNNs) (LeCun et al., 1998), have been widely applied in healthcare research for developing predictive and diagnostic models. For example, Jia et al. (2019) used LSTM models to predict typhoid outbreaks, while Guo et al. (2020) employed SVM models for hepatitis prediction.

Familiarity with AI/ML among EHOs is a critical factor in the practical implementation of these technologies for disease monitoring and prediction. In many low-resource settings, including regions in South Asia and sub-Saharan Africa, studies have emphasized the importance of local health professionals' digital literacy in influencing the success of AI-driven interventions (Rajkomar et al., 2019; Topol, 2019). For instance, a review by Rasheed et al. (2020) found that AI approaches significantly supported frontline workers and decision-makers during the COVID-19 pandemic by enhancing real-time diagnostics, optimizing resource allocation, and improving situational awareness, thereby reducing the burden on healthcare systems and improving response effectiveness. A study by Chettri et al. (2025) highlights that in the Indian healthcare sector, the integration of trustworthy AI systems faces challenges due to ethical and regulatory constraints, emphasizing the need for capacity building among healthcare workers to bridge this gap. Similarly, a scoping review by Botha et al. (2024) underscores the perceived threats posed by AI tools in healthcare, such as unpredictable errors and data privacy concerns, which can be mitigated through proper training and involvement of healthcare professionals in the deployment process. These insights underscore that the familiarity and acceptance of AI/ML among EHOs not only facilitates smooth adoption but also enhances the relevance and contextual accuracy of such tools in mitigating climate-sensitive health risks.

Despite the potential of these technologies, their application in Tanzania remains limited, with significant gaps in research, practical knowledge, and infrastructure. The Tanzanian government recognizes the importance of adopting digital health technologies, including AI and ML, as outlined in the Tanzania National Digital Health Strategy 2019–2021. However, more work is needed to build capacity and develop locally relevant AI-driven solutions for public health.

Given the high burden of climate-sensitive waterborne diseases and the growing impact of climate change on public health, this study aims to assess the knowledge, perceptions, challenges, and opportunities related to using ML in predicting these diseases among public health officers in Tanzania. The study contributes to providing insights for designing AI-based and data-driven approaches to strengthen public health resilience to climate-sensitive waterborne diseases in Tanzania. Specifically, the objectives are:

1. To assess the knowledge of information and communication technology (ICT) and AI/ML among environmental health officers in selected study areas in Tanzania.
2. To evaluate the perceptions of environmental health officers regarding the application of AI/ML in predicting climate-sensitive waterborne diseases.
3. To identify the perceived challenges in integrating AI/ML into public health practices in Tanzania.

The rest of the paper is organized as follows: Section 2 outlines the data collection approach and analysis techniques. The Section 3 presents the key findings, followed by the Section 4, which interprets and compares these results. Finally, the Conclusions summarize key insights and propose future research directions.

2 Methods

2.1 Study area

The research was conducted across three district councils in Tanzania: Morogoro Municipal Council (MC), Ilala City Council (CC) (in Dar es Salaam), and Dodoma CC. These districts were selected as part of the larger ACHE project (AI for Climate and Health), which aimed to collect a comprehensive dataset on climate-sensitive waterborne diseases to support predictive machine learning models in public health in Tanzania. The ACHE project encompassed five district councils. Of the five councils engaged in the ACHE project, three were included in this study; Temeke was used for piloting the data collection tool, while Singida was excluded due to unavailability of EHOs during the data collection period, as many were on leave.

Morogoro MC is an urban district located in eastern Tanzania, with a population of ∼471,000. Ilala District in Dar es Salaam is a highly urbanized and densely populated area, serving as a commercial and administrative hub, with a population exceeding 1.6 million. Dodoma City Council is situated in the central part of the country and serves as the national capital, has a population of around 765,000. These population statistics are according to the National Census report of 2022 [The United Republic of Tanzania (URT) et al., 2024]. Despite variations in their geographic locations and population sizes, EHOs across these districts operate under uniform national policy frameworks. They are recruited based on standardized qualifications in Environmental Health and share similar job descriptions and professional responsibilities across the country.

The survey reported in this paper was conducted concurrently with the ACHE project, not after its completion, meaning that prior experience in the project did not influence participant selection. Tanzania has relatively consistent climate conditions across districts, and the roles and responsibilities of EHOs are the similar nationally. For these reasons, the study findings remain broadly applicable to other Tanzanian districts beyond those included in this sample.

The dataset collected through the ACHE project is openly accessible via the Zenodo repository (Fue et al., 2025), to allow AI/ML researchers and data scientists to use it for training predictive ML models on climate-sensitive waterborne disease outbreaks. A detailed description of the dataset will be presented in a separate paper, while the initial baseline models developed using the data are also covered in another companion paper. This baseline survey focused on assessing the readiness and familiarity of EHOs with AI/ML, a crucial step before deploying ML-supported public health interventions.

2.2 Participants

The study focused on Environmental Health Officers (EHOs), for two reasons: their involvement in the projects data collection for open datasets for predictive ML; and their formal and daily duties, which include monitoring and controlling environmental hazards, enforcing public health regulations, conducting disease surveillance, and promoting community health initiatives to prevent and manage diseases. The selected districts comprised a total of 106 wards (as of December 2024), each ideally assigned one EHO. However, due to personnel shortages, some EHOs were responsible for multiple wards, resulting in a total of 76 EHOs available across the three councils. While participation in the study was voluntary, selection bias is unlikely to have been a major concern. The survey targeted all the available 76 EHOs who were actively serving across the three councils.

2.3 Sampling method

Due to the specialized expertise needed for this baseline study and the relatively small population size, the study did not employ a sample size calculation in its traditional sense; it aimed to include all the 76 EHOs who were actively on duty at the time of the study. Since this was a small, well-defined population, we aimed to include all EHOs rather than rely on sampling calculations. As such, clustering by study site was not considered relevant for sample size determination. This census-style approach was appropriate for capturing baseline knowledge and perceptions among the full workforce available in the selected areas. Since participation was voluntary in accordance with research ethics, which requires informed consent, only a subset of EHOs ultimately took part in the survey.

2.4 Data collection

Data was gathered using a structured questionnaire, which was administered in two formats: a printed version and an online version via Google Forms. The questionnaire was organized into five thematic sections. The first section collected demographic information, mainly sex, age group, and education level. The second section assessed the respondents' familiarity with and use of ICT tools in public health. The third section explored methods EHOs currently use to predict and manage weather-sensitive waterborne diseases. The fourth section assessed their knowledge of AI and ML, along with their percepectives on the applicability of these technologies in public health. The final section explored the perceived challenges of implementing AI and ML in public health in this context. To capture perceptions effectively, the questionnaire

used Likert scale questions, ranging between 4-point and 5-point scales depending on the nature of the question.

The survey instrument was developed by the project team based on a thorough literature review and insights from preliminary fieldwork in Morogoro, where researchers are based. Although not adapted from an existing validated tool, the questionnaire underwent pilot testing in Temeke District with nine EHOs. Their feedback was used to refine the language, sequence, and content of several questions, including the addition of items targeting potential causative factors behind responses, particularly regarding experiences and attitudes toward AI/ML. These refinements enhanced relevance of the tool, comprehensiveness, and overall effectiveness in capturing the study's intended objectives. The survey was administered in English, as this is the official language of communication among EHOs in Tanzania.

2.5 Data analysis

The collected data was analyzed using both descriptive and inferential statistical methods, with Python scripts. Descriptive statistics were used to summarize demographic information and identify overall trends. For inferential statistics, the Chi-square test was applied to assess associations between nominal and ordinal variables, while Kendall's Tau was utilized to evaluate relationships among ordinal variables. Kendall's Tau was chosen due to the relatively small sample size.

2.6 Ethical clearance and research permit

The study adhered to all required ethical guidelines, obtaining official research clearance and permits before data collection commenced. Ethical approval was secured from Sokoine University of Agriculture (Research Clearance No. SUA/DPRTC/R/126/CoNAS/2/2023/3), while government research permits (Permit No. JC.156/254/01) were granted by the President's Office, Regional Administration, and Local Government Authorities (PO-RALG). In addition to these central approvals, further permissions were obtained at the regional and district levels. Specifically, at each region, authorization was sought from the Regional Administrative Secretary's (RAS) office, which formally introduced the research team to the respective district authorities. Following this, each district issued research clearance and introduction letters to the Ward Executive Officers (WEOs) and EHOs, requesting their cooperation and support during data collection.

3 Results

3.1 Demographics information

Questionnaires were distributed to all the 76 EHOs across the three councils. Of these, 50 were completed and submitted successfully, a response rate of 66%. Table 1 shows that the proportion of female and male EHOs responses to the questionnaire is 48 and 52% respectively. The distribution of respondents across the three councils is further illustrated in Table 2. Morogoro MC represented 50% of the total respondents, Dodoma CC represented 26%, and Ilala CC represented 24%. In Morogoro MC, the demographic breakdown revealed that 52% of respondents were male, while 48% were female. In Dodoma CC, the male population constituted 61%, while the female population accounted for 39%. In Ilala CC, the gender distribution among respondents showed that 33% were male, while 67% were female.

Additionally, Table 1 shows the largest proportion of the respondents were in the age range of 31–40 (46%) whereas the smallest proportion was in the age range of 18–30 (12%). With regard to their education level, Table 1 shows that 76% of the EHOs had Diploma, 18% had Bachelor degree and 6% had Certificate. The education levels of respondents across the three councils are further elaborated in Table 3. Within the Morogoro CC, the data indicates that 4% of participants had a Certificate, 80% held a Diploma, and 16% attained a Bachelor's degree. In Dodoma CC, the educational qualifications were distributed as follows: 15% of individuals had a Certificate, 70% possessed a Diploma, and 15% attained a Bachelor's degree. Within the Ilala CC, it was observed that none of the participants possessed a Certificate, whereas 75% had attained a Diploma and 25% had completed a Bachelor's degree.

TABLE 1 Demographic characteristics of the respondents by frequency distribution.

Category	Group	Count	Percentage (%)
Sex	Female	24	48.0
	Male	26	52.0
Education	Diploma	38	76.0
	Bachelor	9	18.0
	Certificate	3	6.0
Age group	18–30	6	12.0
	31–40	23	46.0
	41–50	11	22.0
	51–60	10	20.0

TABLE 2 Distribution of the respondents' gender per council.

Council	Total count	Male%	Female%	% Distribution of the respondents per council
Morogoro MC	25	52	48	50
Dodoma CC	13	61	39	26
Ilala CC	12	33	67	24

3.2 The familiarity and usage of ICT tools and facilities among EHOs for public health

When asked to rate their familiarity with information and communication technology (ICT) tools, 46% of participants reported being moderately familiar, 32% felt very familiar (Table 4). Table 4 also shows a smaller proportion of respondents (14%),

TABLE 3 Distribution of respondents' education levels per council.

Council	Total Count	Certificate (%)	Diploma (%)	Bachelor (%)
Morogoro MC	25	4	80	16
Dodoma CC	13	15	70	15
Ilala CC	12	0	75	25

TABLE 4 ICT familiarity and digital tools usage among EHOs by frequency distribution.

Category	Group	Count	Percentage (%)
Rate your familiarity with using information and communication technology (ICT) tools	Extremely familiar	2	4.0
	Very familiar	16	32.0
	Moderately familiar	23	46.0
	Somewhat familiar	7	14.0
	Not familiar at all	2	4.0
How frequently do you use digital devices (computers, smartphones, tablets) for work-related tasks in the public health sector?	Daily	38	76.0
	Sometimes	8	16.0
	At least once in a week	2	4.0
	Never	2	4.0

considered themselves somewhat familiar, while 4% each rated themselves as extremely familiar and not familiar at all. Regarding the frequency of using digital devices (computers, smartphones, tablets) for work-related tasks in the public health sector, Table 4 illustrates that 76% of respondents use these devices daily. Table 4 also shows that 16% use them occassionally in a week, 4% use them at least once a week, and the remaining 4% never use digital devices for public health activities. The study also shows that EHOs in the selected districts employ various ICT tools in public health initiatives. As presented in Figure 1, email and communication applications, such as WhatsApp, accounted for the highest usage at 92%. Statistical analysis software was used by 22% of respondents, health information systems by 12%, and mobile health applications by 10%.

3.3 Knowledge of AI and ML among EHOs

Table 5 reveals that 54% of the respondents had encountered the terms AI and ML before participating in the study, while 46% had not. On the other hand, Table 5 indicates that 64% were somewhat familiar with the concept of Artificial Intelligence, a notable 30% were not familiar at all, and only 6% were very familiar with the AI concept.

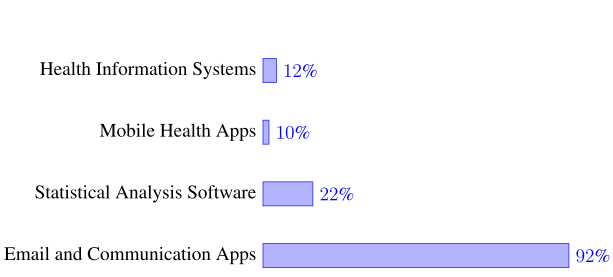

FIGURE 1
ICT tools usage in public health activities.

TABLE 5 AI/ML familiarity among EHOs by frequency distribution.

Category	Group	Count	Percentage
Had you encountered the term "Machine Learning (ML) or Artificial Intelligence (AI)" before this survey?	Yes	27	54.0
	No	23	46.0
How familiar are you with the concept of Artificial Intelligence (such as Machine Learning)?	Somewhat familiar	32	64.0
	Not familiar	15	30.0
	Very familiar	3	6.0

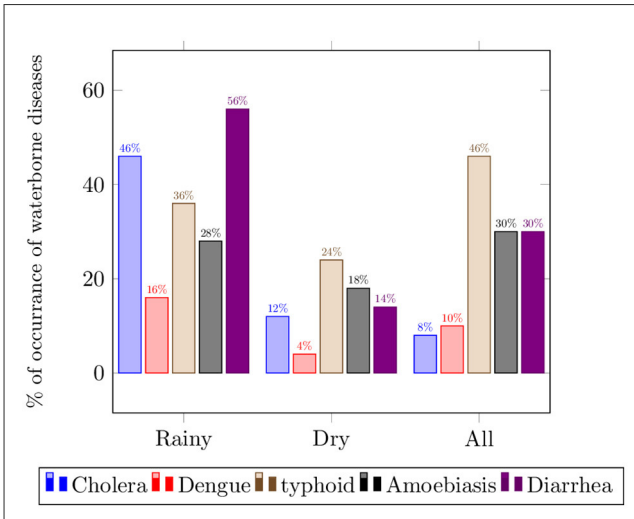

FIGURE 2
Perceived prevalence of waterborne diseases.

3.4 EHOs observations on waterborne disease trends

To understand the relationship between weather-sensitive waterborne diseases and weather conditions, the EHOs were asked to identify prevalent waterborne diseases in their area and the associated weather patterns. Figure 2 shows that diarrhea and cholera are reported to be frequently observed during the wet/rainy seasons, with cholera (56%) having more peak occurrence, followed by diarrhea (46%). Cholera was reported to have a substantial

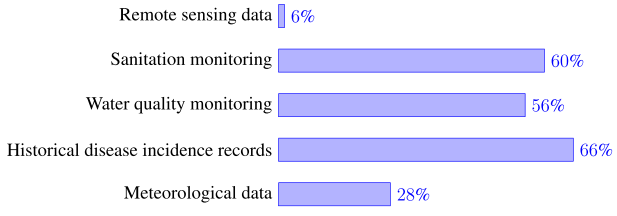

FIGURE 3
Key data sources for monitoring and prediction of waterborne diseases.

decrease in occurrences during the dry season. Typhoid fever, categorized as a seasonal disease, has been reported in both wet (24%) and dry (46%) seasons, but it exhibits higher prevalence during the dry season.

3.5 Key data sources for monitoring and predicting WSWB diseases

Understanding the data sources used by health professionals is essential for designing effective monitoring systems for weather-related health risks. In this study, respondents were asked to identify the primary data sources they use for tracking and predicting waterborne disease outbreaks in the areas of their jurisdictions in Tanzania. The findings reveal that historical disease incidence records are the most frequently used source, with 66% of respondents depending on them for monitoring and predicting weather-sensitive waterborne (WSWB) disease outbreaks (Figure 3). These records are primarily obtained from the DHIS2, a national health data reporting and monitoring system, as well as news reports on disease outbreaks. Additionally, 60% of respondents reported using sanitation data, while 56% relied on water quality monitoring data. Meteorological data are utilized by 14% of the respondents who reported relying on them for predicting WSWB diseases. However, despite the valuable environmental insights offered by remote sensing data, only 6% of respondents reported using this source.

3.6 The potential of integrating AI and ML into existing WSWB disease prediction and management strategies

Regarding perceptions of AI/ML in enhancing the accuracy and timeliness of disease outbreak predictions amid changing climate patterns, 50% of respondents agreed, while 20% strongly agreed. Meanwhile, 26% were uncertain, and 4% disagreed, as shown in Table 6. When asked whether AI/ML could replace or complement traditional methods for predicting and managing waterborne diseases, the same Table 6 reveals that 40% believed AI/ML could serve as a full replacement. Additionally, 26% felt it could fully complement traditional methods, 22% believed it would provide partial support, and 12% thought it would not replace traditional approaches at all.

TABLE 6 Perceptions of EHOs on AI/ML by frequency distribution.

Category	Group	Count	Percentage
ML/AI contributes to more accurate disease outbreak predictions	Strongly agree	10	20.0
	Agree	25	50.0
	No idea	13	26.0
	Disagree	2	4.0
	Strongly disagree	0	0.0
Can ML/AI replace or complement traditional methods?	Replace	20	40.0
	Fully complement	13	26.0
	Partially complement	11	22.0
	Not at all	6	12.0
Confidence in AI/ML for waterborne disease management	Very confident	9	18.0
	Confident	29	58.0
	Not sure	12	24.0
	Not confident	0	0.0
Healthcare professionals' trust in AI/ML for disease prediction	Strongly agree	10	20.0
	Agree	27	54.0
	No idea	12	24.0
	Disagree	1	2.0
	Strongly disagree	0	0.0

Furthermore, Table 6 indicates that 58% of respondents expressed confidence in the reliability and accuracy of machine learning models for predicting waterborne diseases, while 18% were very confident, and 24% remained uncertain. In terms of healthcare professionals' trust in machine learning algorithms for disease prediction and management, 54% of respondents agreed, 20% strongly agreed, 24% were unsure, and 2% disagreed (Table 6).

EHOs identified multiple ways in which AI/ML could be integrated into existing disease prediction and management strategies (Figure 4). The most commonly selected approach, chosen by 54.3% of respondents, was utilizing AI/ML to analyze historical disease data. Similarly, 52.2% envisioned developing predictive models to enhance outbreak forecasting. Around 45.7% suggested integrating meteorological data, water quality data, and disease incidence records to provide a more comprehensive understanding of disease trends. Additionally, 41.3% saw AI-driven early warning systems as a valuable tool for proactive disease management. Approximately 30.4% supported using AI to detect sudden spikes in disease incidence in real time by analyzing data streams from various sources. Meanwhile, 28.3% highlighted the potential of incorporating satellite imagery with disease data to assess the relationship between land cover changes and outbreaks. Lastly, 19.6% envisioned AI automating routine tasks to improve efficiency in disease monitoring and response efforts.

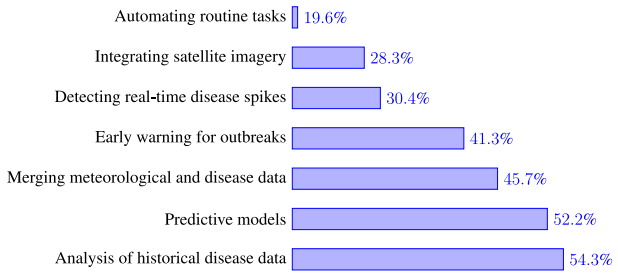

FIGURE 4
Different ways of integrating AI/ML in waterborne diseases prediction and management.

TABLE 7 Relationship between sex and other variables.

Variable 1	Variable 2	Statistic (X^2)	p-value
Sex	ICT familiarity	2.708	0.608
	ICT usage	2.872	0.412
	AI/ML encounter	3.862	0.049
	AI familiarity	10.337	0.006
	AI/ML prediction	5.566	0.135
	AI replaces traditional	3.688	0.297
	ML trust	3.474	0.324

3.7 Demographic factors and their influence on AI perceptions and ICT usage

Table 7 explores how respondents' gender relates to various factors, including familiarity with ICT (ict_familiarity), usage of digital tools in public health (ict_usage), prior exposure to AI/ML terms (ai_ml_encounter), familiarity with AI (ai_familiarity), perceptions of AI/ML in predicting weather-sensitive waterborne diseases (ai_ml_prediction), belief in AI replacing traditional methods (ai_replace_traditional), and trust in ML decisions (ml_trust). Among these associations, only the relationships between sex and both prior exposure to AI/ML terms and AI familiarity were statistically significant ($p = 0.049$) and ($p = 0.006$), respectively. On the other hand, Table 8 examines whether age group (age_group) is associated with the same set of variables. The results indicate no statistically significant relationships between age and any of the tested factors. The low correlation coefficients suggest minimal association, while the high p-values indicate that any observed trends are likely due to chance. Since none of the associations reach statistical significance ($p < 0.05$), we cannot conclude that age influences AI/ML awareness, ICT familiarity and usage, or related perceptions in a meaningful way.

Table 9 investigates how education level (education) correlates with ICT and AI-related variables. The findings reveal that education level is significantly associated with the use of ICT tools in public health ($p = 0.012$), indicating that individuals with higher education levels are more likely to use such tools. However, all other relationships show weak or negligible correlations and are not statistically significant.

TABLE 8 Relationship between age group and other variables.

Variable 1	Variable 2	Kendall's T	p-value
Age group	ICT familiarity	0.055	0.659
	ICT usage	0.049	0.701
	AI/ML encounter	−0.100	0.450
	AI familiarity	0.045	0.726
	AI/ML prediction	−0.200	0.107
	AI replaces traditional	−0.182	0.135
	ML trust	−0.070	0.574

TABLE 9 Relationship between education and other variables.

Variable 1	Variable 2	Kendall's T	p-value
Education	ICT familiarity	0.144	0.271
	ICT usage	0.341	0.012
	AI/ML encounter	−0.027	0.845
	AI familiarity	0.029	0.831
	AI/ML prediction	0.138	0.293
	AI replaces traditional	0.112	0.387
	ML trust	0.128	0.332

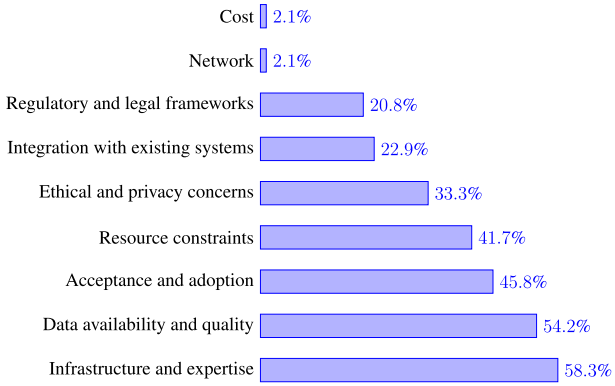

FIGURE 5
Challenges for integration of AI/ML in waterborne disease prediction and management.

3.8 Challenges and opportunities for AI and ML adoption in public health

To evaluate the challenges and opportunities associated with integrating machine learning (ML) into Tanzania's public health sector, respondents highlighted several key barriers (see Figure 5). The most pressing challenge, cited by 58.3% of the participants, was the lack of technical infrastructure and expertise necessary for effective implementation. Data availability and quality were also major concerns, with 54.2% identifying these as critical obstacles.

Other notable challenges include the acceptance and adoption of AI/ML technologies (45.8%) and resource constraints, such as funding and skilled personnel shortages (41.7%). Ethical and

privacy concerns were raised by 33.3% of the respondents, while 22.9% pointed to difficulties in integrating AI/ML with existing public health systems. Regulatory and legal frameworks were also seen as potential barriers, with 20.8% recognizing the need for clearer guidelines and policies.

Although less frequently mentioned, network connectivity issues and financial costs were identified as potential hindrances by a small percentage (2.1%) of respondents. These findings highlight both the technical and socio-economic challenges that must be addressed to fully leverage AI/ML for public health improvements in Tanzania.

4 Discussion

The findings of this study offer important reflections on the feasibility of integrating machine learning (ML) into public health practices for managing weather-sensitive waterborne diseases (WSWDs) in Tanzania. The discussion focuses on the key thematic areas of ICT and AI/ML knowledge among environmental health officers (EHOs), their perceptions of AI/ML applications, and the critical challenges and opportunities that influence the adoption of these technologies.

4.1 Knowledge of ICT and AI/ML among EHOs

The study indicates that while EHOs have a reasonable level of familiarity with general ICT tools, their understanding of AI and ML remains limited. This knowledge AI and ML gap is a potential barrier to the successful adoption of AI-driven solutions in public health. The findings suggest that despite a moderate engagement with digital tools in their daily activities, many EHOs have not had significant exposure to AI/ML concepts, which may limit their ability to evaluate and implement such technologies effectively. One approach to bridging this gap is targeted training programs and continuous professional development opportunities to equip EHOs with the necessary competencies in AI/ML applications.

Targeted training programs have proven effective in enhancing digital competencies among health professionals, thereby facilitating the adoption of AI and ML technologies in public health. For instance, the Sustainable Healthcare with Digital Health Data Competence (Susa) project in Europe, funded by the EU's Digital Europe programme, aims to equip over 6,500 health professionals with advanced digital skills, including AI integration, through structured bachelor's and master's programs, as well as lifelong learning modules. This initiative illustrates the importance of involving healthcare workers early in the technology development process to ensure successful implementation (Cookson, 2025). Similarly, in African contexts, studies have highlighted that AI can significantly improve public health surveillance and disease monitoring. However, the lack of skilled health professionals remains a barrier, emphasizing the need for targeted training to harness AI's full potential (Tshimula et al., 2024).

4.2 Utilization of key data sources for efficient monitoring and prediction of WSWB diseases

EHOs rely on traditional methods to predict waterborne diseases, primarily through manual reviews of historical disease outbreak records and observable weather patterns, without the support of dedicated tools or systems. This approach limits the effective utilization of available data. EHOs specifically reported underutilization of key sources such as meteorological and remote sensing data. Lack of integrated systems or analytical tools that can process and interpret such data effectively, hindering the potential for more accurate and timely disease prediction. Meteorological data such as precipitation, temperature, and humidity are essential for identifying environmental variables that affect disease outbreaks (Chowdhury et al., 2018). Likewise, remote sensing data offers significant insights into changes in land cover, variations in water quality, and environmental risk factors, which are essential for early warning systems (Mashala et al., 2023; Deng et al., 2024). The limited utilization of these data sources could be due to a lack of tools or systems that utilize these data, which hinders the ccuracy of disease prediction and hinders the effectiveness of proactive interventions. Strengthening capacity-building initiatives is essential to promote the adoption of advanced analytical methods, such as ML algorithms, enabling the effective integration of diverse data types for comprehensive analysis and more accurate predictive modeling.

4.3 Perceptions of AI/ML in managing weather-sensitive waterborne diseases

EHOs generally express optimism about the role of AI and ML in enhancing disease prediction and management. However, this optimism is accompanied by uncertainties, particularly regarding the efficacy of AI/ML models in real-world settings. While some respondents view AI/ML as a potential replacement for traditional methods, others see it as a complementary tool that should be integrated cautiously. This divergence in perception underscores the need for clear implementation frameworks and capacity-building initiatives that illustrate how AI/ML can augment rather than replace existing public health strategies. Ensuring that these technologies are positioned as decision-support tools rather than automated replacements for human expertise will be crucial for fostering trust and facilitating adoption. This divergence in perception is not unique to our study area; similar sentiments have also been reported across low- and middle-income countries.

There are significant advancements in AI-driven health initiatives within low- and middle-income countries, with African nations such as Kenya, Nigeria, and Ghana leading the way. Nonetheless, most of these AI systems have remained confined to academic environments, mainly concentrating on model training, validation, and testing, without advancing to practical implementation (Ndagire et al., 2022; Baik et al., 2020). This underscores a notable disparity between the advancement of AI technologies and their limited uptake in the healthcare

sector. The slow uptake of AI and other new technologies in healthcare, as expressed in literature, is largely due to varying perceptions among stakeholders. Some healthcare professionals worry that technology could threaten their independence, patient care, or be used as a means of administrative control. However, some view it as a means to improve patient engagement. Patients generally embrace technology because it gives them more choice in treatment (Safi et al., 2018; Mirugwe, 2024). Effective change management is crucial for technology adoption since insufficient inclusive engagement in technology creation sometimes causes resistance.

4.4 Gender and AI awareness

Among the various variables examined in relation to gender, only two, prior exposure to AI/ML terminology and self-reported familiarity with AI, demonstrated statistically significant associations. These findings suggest that gender plays a notable role in shaping awareness and engagement with artificial intelligence and machine learning. Specifically, one gender group appears more likely to have encountered or interacted with AI/ML concepts, which may reflect differences in access to relevant educational resources, workplace roles, or broader socio-cultural dynamics that influence exposure to digital technologies.

This observation aligns with existing literature highlighting persistent gender gaps in digital literacy and attitudes toward AI. For instance, Russo et al. (2025) reported that women tend to exhibit greater anxiety toward AI, hold less favorable attitudes, engage with the technology less frequently, and perceive themselves as having lower knowledge compared to men. These patterns are reflected in our sample, where female respondents reported lower familiarity and exposure. Moreover, as noted by Armutat et al. (2024), structural barriers, including gender stereotypes, unequal access to educational opportunities, and systemic discrimination, continue to limit women's participation in AI and other technology-driven fields.

The implications of these findings are significant for efforts to promote inclusive AI literacy and workforce development. Addressing gender disparities in AI knowledge and exposure requires targeted interventions, such as gender-sensitive training programs, outreach initiatives in underrepresented communities, and policies aimed at dismantling structural barriers in education and employment. Without deliberate efforts to close these gaps, the underrepresentation of women in AI-related domains may persist, ultimately limiting the diversity of perspectives and innovations in this rapidly evolving field.

4.5 Challenges and opportunities

Despite the potential of AI and ML, several significant challenges hinder their adoption in Tanzania. These challenges, if unaddressed, could limit the applicability, scalability, and sustainability of AI-driven public health interventions in Tanzania.

One of the most pressing challenges is the lack of infrastructure. Deploying AI/ML models effectively requires robust digital infrastructure, including reliable internet access, sufficient computational power, and a stable electricity supply—resources that remain insufficient in many parts of Tanzania. The absence of these foundational elements restricts the ability to run complex AI models in real time and limits opportunities for cloud-based computing solutions. Without substantial investments in digital infrastructure, AI/ML applications risk remaining confined to research settings without meaningful real-world impact. This also limits affect the deployment of Electronic Health Record Systems (EHRS) in Tanzanian primary healthcare facilities, according to Mwogosi (2024). Furthermore, the limited availability of AI/ML specialists in the public health sector means that even where infrastructure is available, there are few personnel with the necessary skills to develop, implement, and maintain AI-driven solutions. Addressing this issue requires incorporating AI/ML education into public health curricula and establishing continuous professional training programs tailored for EHOs and other key stakeholders. The second issue is data availability and quality, which also present significant barriers. Many public health decisions rely on historical disease records, which are often incomplete, inconsistent, or outdated. Additionally, the integration of diverse data sources—such as meteorological data, water quality indicators, and satellite imagery—remains underutilized. The absence of standardized data collection protocols and insufficient data-sharing mechanisms among institutions further exacerbate the challenge. Without high-quality, well-structured datasets with supporting infrastructure in place AI/ML models cannot be effectively trained to generate reliable predictions.

According to Yonazi (2012) and Muhunzi et al. (2024), poor data governance and insufficient investment in digital systems infrastructures hinder the use of big data analytics in healthcare in many developing nations. However, high-income countries have strong digital infrastructures to facilitate AI and machine learning in healthcare. Electronic health record systems are widely used in developed countries due to its standardization, interoperability, and immediate data availability (Slawomirski et al., 2023). The European Health Data Space's open data regulations and platforms have improved data accessibility and innovation (Hussein et al., 2023). The disparities underscore a notable digital divide, with high-income nations more equipped to utilize data-driven health solutions, whereas developing nations face the challenge of addressing essential infrastructure and data quality issues first. Despite these challenges, Arinze (2024) suggests that East African Open Data Initiatives may pave the way for more advanced digital health applications. Strengthening data governance frameworks and fostering collaborations between public health agencies and research organizations could improve data accessibility and the development of data-driven solutions in the health sector.

Beyond technical and data-related constraints, there are significant sociocultural and institutional challenges related to the acceptance and adoption of AI/ML in public health. Resistance to new technologies is common when professionals perceive them as disruptive or when their benefits are not clearly demonstrated. In many cases, EHOs may be skeptical of AI/ML-driven decision-making, particularly if these technologies are introduced without adequate stakeholder engagement. Ethical concerns, such as transparency in AI-driven decision-making

and the risk of algorithmic biases, also contribute to hesitancy. To overcome these barriers, there is a need for structured implementation strategies that involve EHOs in the co-design of AI/ML systems, ensuring that these technologies align with their workflows and address real-world challenges in disease management. Therefore, a collective effort is needed to advance research on AI/ML applications for predicting and managing diseases such as WSWD, particularly in low- and middle-income countries (LMICs) such as Tanzania. Research should focus on collecting high quality data and developing AI/ML models that are optimized for resource-constrained environments, ensuring that these technologies are both accessible and applicable. Additionally, studies exploring strategies for capacity-building, infrastructure development, and ethical AI adoption will be essential for facilitating the successful integration of AI-driven solutions in public health.

4.6 Limitations of the study

While the findings of this study provide valuable insights into the readness and perceptions of EHOs ragrding AI/ML in public health, seral limitations must be acknowledged. First, the sample consisted of EHOs who participated in a specific data collection project (ACHE), which may introduce selection bias. Although their participation was not based on prior experience with AI/ML, their engagement in a broader health data initiative may reflect a more proactive or accessible segment of EHOs, potentially limiting generalizability to all EHOs across Tanzania.

Second, the survey instrument, while piloted and refined through feedback from EHOs in Temeke district, was not adapted from pre-existing validated tools. Although this iterative process improved clarity and relevance, the absence of a formally validated instrument introduces a degree of measurement bias.

Third, the survey was administered in Kiswahili, the national language, to ensure accessibility and reduce language-related misunderstandings. However, minor interpretation differences across regions may have influenced how some questions were perceived or answered.

Lastly, the study relied on self-reported data, which is subject to recall and social desirability biases. Future research should aim to expand the sample across more diverse districts and incorporate mixed-methods approaches—such as in-depth interviews and focus group discussions—to enrich and triangulate the findings.

5 Conclusion and future work

This study explored the knowledge, perceptions, and preparedness of Environmental Health Officers (EHOs) in selected Tanzanian districts regarding the application of machine learning (ML) in predicting weather-sensitive WSWDs. The findings reveal that while general ICT familiarity among EHOs is reasonably strong, there remains a substantial gap in awareness and understanding of AI and ML concepts. This gap presents a critical challenge for integrating ML-based systems into public health decision-making processes.

Importantly, the study highlights several perceived barriers to ML adoption from the perspective of frontline health workers, including limited technical skills, data quality concerns, and institutional readiness. While the study did not directly assess existing infrastructure or data systems, the perceptions gathered from EHOs offer valuable insights into the enabling conditions required for future ML applications.

Rather than presenting ML as an immediately deployable solution, this study emphasizes the importance of targeted capacity-building and policy support to prepare the public health workforce for digital transformation. Future research should complement these findings by examining actual data readiness, infrastructure capabilities, and the development of contextually appropriate ML models for low-resource settings. Pilot interventions, co-designed with health professionals, may also help bridge the gap between potential and practical implementation of ML for disease prediction and control.

Data availability statement

All relevant data are included within the article and its supplementary materials. Further inquiries regarding the generated and analyzed data from this study can be obtained from the corresponding author upon request.

Ethics statement

The studies involving humans were approved by (i) President's Office, Regional Administration, and Local Government Authorities and (ii) Sokoine University of Agriculture. 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.

Author contributions

NL: Conceptualization, Writing – original draft, Data curation, Formal analysis, Visualization, Writing – review & editing, Supervision, Investigation. KF: Writing – review & editing. SM: Methodology, Writing – review & editing, Investigation. NK: Resources, Writing – review & editing, Project administration, Investigation. JT: Writing – original draft, Data curation, Formal analysis, Funding acquisition, Visualization, Supervision.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported by a grant from Lacuna Fund, an initiative co-founded by The Rockefeller Foundation, Google.org, Canada's International Development Research Centre (IDRC), and GIZ on behalf of the German Federal Ministry for Economic Cooperation and Development (BMZ); and by Wellcome.

Conflict of interest

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

Generative AI statement

The author(s) declare that Gen AI was used in the creation of this manuscript. The author(s) declare that generative AI tools, including ChatGPT, were used to assist with paraphrasing, improving clarity, and enhancing the flow of certain sections of this paper. All AI-assisted content was reviewed and refined by the author(s) to ensure accuracy and originality.

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/frai.2025.1597727/full#supplementary-material

References

Adedokun, S. T., and Yaya, S. (2020). Childhood morbidity and its determinants: evidence from 31 countries in Sub-Saharan Africa. *BMJ Glob. Health* 5:e003109. doi: 10.1136/bmjgh-2020-003109

Adigwe, O. P., Onavbavba, G., and Sanyaolu, S. E. (2024). Exploring the matrix: knowledge, perceptions and prospects of artificial intelligence and machine learning in Nigerian healthcare. *Front. Artif. Intell.* 6:1293297. doi: 10.3389/frai.2023.1293297

Arinze E. D. (2024). A thorough examination of Open Data Initiatives in East Africa, focusing on how they improve the accessibility of data. *Res. Invent. J. Sci. Exp. Sci.* 3, 30–37.

Armutat, S., Wattenberg, M., and Mauritz, N. (2024). "Artificial intelligence: gender-specific differences in perception, understanding, and training interest," in *International Conference on Gender Research* (Reading: Academic Conferences International Limited), 36–43. doi: 10.34190/icgr.7.1.2163

Baik, Y., Rickman, H. M., Hanrahan, C. F., Mmolawa, L., Kitonsa, P. J., Sewelana, T., et al. (2020). A clinical score for identifying active tuberculosis while awaiting microbiological results: development and validation of a multivariable prediction model in Sub-Saharan Africa. *PLoS Med.* 17:e1003420. doi: 10.1371/journal.pmed.1003420

Botha, N. N., Segbedzi, C. E., Dumahasi, V. K., Maneen, S., Kodom, R. V., Tsedze, I. S., et al. (2024). Artificial intelligence in healthcare: a scoping review of perceived threats to patient rights and safety. *Arch. Public Health* 82:188. doi: 10.1186/s13690-024-01414-1

Chettri, S. K., Deka, R. K., and Saikia, M. J. (2025). Bridging the gap in the adoption of trustworthy ai in Indian healthcare: challenges and opportunities. *AI* 6:10. doi: 10.3390/ai6010010

Chowdhury, F. R., Ibrahim, Q. S. U., Bari, M. S., Alam, M. J., Dunachie, S. J., Rodriguez-Morales, A. J., et al. (2018). The association between temperature, rainfall and humidity with common climate-sensitive infectious diseases in Bangladesh. *PLoS ONE* 13:e0199579. doi: 10.1371/journal.pone.0199579

Cookson, C. (2025). *EU project launched to prepare health workers for a digital future. Financial Times*. Available online at: https://www.ft.com/content/a56ef5a3-f5d8-446d-ae9b-f503cce20de7

Demissie, G. D., Yeshaw, Y., Aleminew, W., and Akalu, Y. (2021). Diarrhea and associated factors among under five children in sub-saharan africa: evidence from demographic and health surveys of 34 sub-Saharan countries. *PLoS ONE* 16:e0257522. doi: 10.1371/journal.pone.0257522

Deng, Y., Zhang, Y., Pan, D., Yang, S. X., and Gharabaghi, B. (2024). Review of recent advances in remote sensing and machine learning methods for lake water quality management. *Remote Sens.* 16:4196. doi: 10.3390/rs16224196

Flückiger, M., and Ludwig, M. (2022). Temperature and risk of diarrhoea among children in Sub-Saharan Africa. *World Dev.* 160:106070. doi: 10.1016/j.worlddev.2022.106070

Fue, K., Nicodemus Lyimo, N., Telemala, J. P., Materu, S. F., and Kilatu, N. (2025). *Tanzania Climate-sensitive Waterborne Diseases Dataset for Predictive Machine Learning: A Comprehensive Five-year Analysis of Health, Infrastructure, and Weather Data (2019–2023)*.

Guo, Y., Feng, Y., Qu, F., Zhang, L., Yan, B., Lv, J., et al. (2020). Prediction of hepatitis e using machine learning models. *PLoS ONE* 15:e0237750. doi: 10.1371/journal.pone.0237750

Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.* 9, 1735–1780. doi: 10.1162/neco.1997.9.8.1735

Howard, G., Calow, R., Macdonald, A., and Bartram, J. (2016). Climate change and water and sanitation: likely impacts and emerging trends for action. *Annu. Rev. Environ. Resour.* 41, 253–276. doi: 10.1146/annurev-environ-110615-085856

Hunter, P. R., MacDonald, A. M., and Carter, R. C. (2010). Water supply and health. *PLoS Med.* 7:e1000361. doi: 10.1371/journal.pmed.1000361

Hussain, M., Cifci, M. A., Sehar, T., Nabi, S., Cheikhrouhou, O., Maqsood, H., et al. (2023). Machine learning based efficient prediction of positive cases of waterborne diseases. *BMC Med. Inform. Decis. Mak.* 23:11. doi: 10.1186/s12911-022-02092-1

Hussein, R., Scherdel, L., Nicolet, F., and Martin-Sanchez, F. (2023). Towards the European health data space (EHDS) ecosystem: a survey research on future health data scenarios. *Int. J. Med. Inform.* 170:104949. doi: 10.1016/j.ijmedinf.2022.104949

Jia, W., Wan, Y., Li, Y., Tan, K., Lei, W., Hu, Y., et al. (2019). Integrating multiple data sources and learning models to predict infectious diseases in China. *AMIA Summits Transl. Sci. Proc.* 2019, 680–685.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 2278–2324. doi: 10.1109/5.726791

Mashala, M. J., Dube, T., Mudereri, B. T., Ayisi, K. K., and Ramudzuli, M. R. (2023). A systematic review on advancements in remote sensing for assessing and monitoring land use and land cover changes impacts on surface water resources in semi-arid tropical environments. *Remote Sens.* 15:3926. doi: 10.3390/rs15163926

Mboera, L. E., Mayala, B. K., Kweka, E. J., and Mazigo, H. D. (2011). Impact of climate change on human health and health systems in Tanzania: a review. *Tanzan. J. Health Res.* 13, 407–426. doi: 10.4314/thrb.v13i5.10

Mguni, P., Herslund, L., and Jensen, M. B. (2016). Sustainable urban drainage systems: examining the potential for green infrastructure-based stormwater management for Sub-Saharan cities. *Nat. Hazards* 82, 241–257. doi: 10.1007/s11069-016-2309-x

Mirugwe, A. (2024). *Adoption of Artificial Intelligence in the Ugandan Health Sector: A Review of Literature*. Available at SSRN 4735326. doi: 10.2139/ssrn.4735326

Mshida, H., Malima, G., Machunda, R., Muzuka, A. N., Banzi, J., Gautam, O. P., et al. (2020). Sanitation and hygiene practices in small towns in Tanzania: the case of Babati district, Manyara region. *Am. J. Trop. Med. Hyg.* 103:1726. doi: 10.4269/ajtmh.19-0551

Mtwangi Limbumba, T., Herslund, L. B., and Kombe, W. J. (2019). The institutional challenges and opportunities for adopting landscape-based storm water management options in informal Settlements-Dar Es Salaam City. *J. Sustain. Dev.* 12, 46–55. doi: 10.5539/jsd.v12n2p46

Muhunzi, D., Kitambala, L., and Mashauri, H. L. (2024). Big data analytics in the healthcare sector: opportunities and challenges in developing countries. a literature review. *Health Informatics J.* 30:14604582241294217. doi: 10.1177/14604582241294217

Mutono, N., Wright, J. A., Mutembei, H., Muema, J., Thomas, M. L., Mutunga, M., et al. (2021). The nexus between improved water supply and water-borne diseases in urban areas in Africa: a scoping review. *AAS Open Res.* 4:27. doi: 10.12688/aasopenres.13225.1

Mwogosi, A. (2024). Revolutionizing primary health care in Tanzania: unravelling the contextual factors on electronic health record systems implementation for effective

decision support. *J. Sci. Technol. Policy Manag*. doi: 10.1108/JSTPM-11-2023-0205. [Epub ahead of print].

Ndagire, E., Ollberding, N., Sarnacki, R., Meghna, M., Pulle, J., Atala, J., et al. (2022). Modelling study of the ability to diagnose acute rheumatic fever at different levels of the ugandan healthcare system. *BMJ Open* 12:e050478. doi: 10.1136/bmjopen-2021-050478

Nemade, B., Maharana, K. K., Kulkarni, V., Mondal, S., Ghantasala, G. P., Al-Rasheed, A., et al. (2024). IOT-based automated system for water-related disease prediction. *Sci. Rep*. 14:29483. doi: 10.1038/s41598-024-79989-6

Pantaleo, P. A., Komakech, H. C., Mtei, K. M., and Njau, K. N. (2018). Contamination of groundwater sources in emerging African towns: the case of Babati town, Tanzania. *Water Pract. Technol*. 13, 980–990. doi: 10.2166/wpt.2018.104

Rajkomar, A., Dean, J., and Kohane, I. (2019). Machine learning in medicine. *N. Engl. J. Med*. 380, 1347–1358. doi: 10.1056/NEJMra1814259

Rasheed, J., Jamil, A., Hameed, A. A., Aftab, U., Aftab, J., Shah, S. A., et al. (2020). A survey on artificial intelligence approaches in supporting frontline workers and decision makers for the COVID-19 pandemic. *Chaos Solitons Fract*. 141:110337. doi: 10.1016/j.chaos.2020.110337

Russo, C., Romano, L., Clemente, D., Iacovone, L., Gladwin, T. E., Panno, A., et al. (2025). Gender differences and in artificial intelligence: the moderating role of artificial intelligence anxiety. *Front. Psychol*. 16:1559457. doi: 10.3389/fpsyg.2025.1559457

Safi, S., Thiessen, T., Schmailzl, K. J., et al. (2018). Acceptance and resistance of new digital technologies in medicine: qualitative study. *JMIR Res. Protoc*. 7:e11072. doi: 10.2196/11072

Slawomirski, L., Lindner, L., de Bienassis, K., Haywood, P., Hashiguchi, T. C. O., Steentjes, M., et al. (2023). *Progress on Implementing and Using Electronic Health Record Systems: Developments in OECD Countries as of 2021*. Organisation for Economic Co-operation and Development (OECD). doi: 10.1787/4f4ce846-en

The United Republic of Tanzania (URT), Ministry of Finance, Tanzania National Bureau of Statistics, President's Office - Finance and Planning, and Office of the Chief Government Statistician, Zanzibar (2024). *The 2022 Population and Housing Census Basic Demographic and Socio-economic Profile Report*. Technical report, The United Republic of Tanzania, Tanzania.

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

Trærup, S. L., Ortiz, R. A., Markandya, A. (2011). The costs of climate change: a study of cholera in tanzania. *Int. J. Environ. Res. Public Health* 8, 4386–4405. doi: 10.3390/ijerph8124386

Tshimula, J. M., Kalengayi, M., Makenga, D., Lilonge, D., Asumani, M., Madiya, D., et al. (2024). Artificial intelligence for public health surveillance in Africa: applications and opportunities. *arXiv* [Preprint]. arXiv:2408.02575. doi: 10.48850/arXiv.2408.02575

WHO (2024). *WHO Bacterial priority Pathogens list, 2024: Bacterial Pathogens of Public Health Importance, to Guide Research, Development, and Strategies to Prevent and Control antimicrobial Resistance*. Geneva: World Health Organization.

World-Bank (2024). *Global Economic Prospects, January 2024*. Geneva: World Bank Publications.

Yonazi, J. J. (2012). Exploring facilitators and challenges facing ICT4D in Tanzania. *J. e-Govern. Stud. Best Pract*. 2012, 1–16. doi: 10.5171/2012.703053

frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 05 June 2025
DOI 10.3389/frai.2025.1388180

OPEN ACCESS

EDITED BY
Dmytro Chumachenko,
National Aerospace University, Ukraine

REVIEWED BY
Shailesh Tripathi,
University of Applied Sciences Upper Austria, Austria
Sergiy Yakovlev,
Lodz University of Technology, Poland

*CORRESPONDENCE
Nataliia Melnykova
✉ melnykovanatalia@gmail.com

RECEIVED 19 February 2024
ACCEPTED 24 February 2025
PUBLISHED 05 June 2025

CITATION
Melnykova N, Pavlyk B, Basystiuk O and Skopivskyi S (2025) Analysis and correcting pronunciation disorders based on artificial intelligence approach.
Front. Artif. Intell. 8:1388180.
doi: 10.3389/frai.2025.1388180

Analysis and correcting pronunciation disorders based on artificial intelligence approach

Nataliia Melnykova*, Bohdan Pavlyk, Oleh Basystiuk and Stepan Skopivskyi

Department of Artificial Intelligence, Lviv Polytechnic National University, Lviv, Ukraine

The main aim of this study is to employ artificial intelligence and machine learning methods to assess and correct pronunciation disorders in post-traumatic military patients, acknowledging the critical need for effective communication rehabilitation in individuals who have experienced trauma, such as head injuries or war-related incidents. Tasks include reviewing existing research, selecting appropriate machine learning methods, generating relevant training data, and implementing a software architecture tailored to analyze and correct pronunciation defects in this specific population. The analysis of machine learning methods led to the selection of two experimental models: a Convolutional Neural Network (CNN) utilizing mel-spectrograms for image-based sound representation and a Long Short-Term Memory (LSTM) network combined with mel-frequency cepstral coefficients, aiming to explore the effectiveness of sequential data processing in the context of pronunciation disorder classification in post-traumatic military patients. The results of the two models were compared based on the loss and accuracy functions of the training and validation data, error matrices, and such key metrics as precision, recall, and F1-score. Both models showed promising results in classifying dysarthria stages, but the CNN model performed slightly better in predicting all classes than the LSTM.

KEYWORDS

pronunciation, unclear pronunciation, dysarthria, classification, CNN, LSTM, artificial intelligence tools

1 Introduction

Technological evolution is relatively rapid, and now, many processes and duties performed by humans can be replaced by computers; in many professions, artificial intelligence can easily replace humans, including in medicine. In these difficult times, our country has faced many challenges that must be addressed immediately. One of the most critical tasks is to save the lives of every Ukrainian, for which the military suffers injuries on the battlefield that can last a lifetime. One of these problems can be a speech impediment. In such cases, automatic voice analysis could help the military rehabilitate faster. It will also be helpful for post-traumatic patients who have problems with speech for some reason.

The relevance of this topic is due to the problematic situation in the country, as pronunciation disorders can occur in military personnel who have suffered trauma, such as head injuries or war. This also supports the practical value, as research in this area can help military personnel who have suffered trauma to improve their pronunciation and make their speech more understandable to others.

This research aims to find practical solutions for developing a software module to provide analysis, selection of artificial intelligence tools, and verification of the effectiveness of machine learning methods in classifying stages of disease for post-traumatic patients.

The object of study is the process of analysis and correction of pronunciation in post-traumatic patients, and the subject is the voice parameters of post-traumatic military patients for speech defects and the artificial intelligence method for analysis and correction of pronunciation in post-traumatic patients.

2 Current research analysis

Artificial intelligence in medicine is already commonplace today, and this trend will continue to grow. The relevance of this topic can also be confirmed by the situation in our country, where every second counts.

The importance of this issue is associated with the following key factors:

- Expansion of rehabilitation capabilities: Post-traumatic speech disorders are a significant problem that arises after various types of brain injuries (e.g., stroke or traumatic brain injury). Patients facing these problems often require long-term rehabilitation, and artificial intelligence can provide additional tools to support this process. AI-driven speech recognition systems can analyze patients' speech patterns in real-time, identifying pronunciation errors and providing feedback (Abdulmajeed et al., 2022).
- Accessibility and convenience: Artificial intelligence technologies can allow patients to engage in rehabilitation at a convenient time and place, improving their ability to recover (Lu et al., 2020).
- Adaptability: Artificial intelligence can adapt to the individual needs of each patient, providing more personalized and effective interventions.
- Monitoring and tracking progress: By using artificial intelligence methods, medical professionals can better track patients' progress in real time, allowing them to adjust treatment plans as needed (Rastogi, 2020).
- Resources: In a world with a constantly increasing number of patients and limited medical resources, artificial intelligence can be an effective means of maximizing the use of these resources.
- Improvement of treatment outcomes: Artificial intelligence can help find new approaches to speech exercises that can accelerate recovery (Duda, 2025).
- Classification of speech problems: Recognition and classification of different types of speech disorders are crucial elements in this field. Speech involves various aspects, including articulation, tempo, pitch, and rhythm. Artificial intelligence can assist in analyzing and classifying these complex speech characteristics. Such classification will enable medical professionals better to understand the specifics of the patient's problem and tailor rehabilitation programs for maximum effectiveness (Deruty, 2025).

The PRISMA scheme was employed to systematically analyze the relevant literature, ensuring a comprehensive and transparent approach to source selection. This process provided a foundation for identifying key studies that address pathology detection through advanced analytical methods.

In article by Panek et al. (2015), issues related to pathology detection are addressed by creating a feature vector consisting of 28 different parameters obtained through voice signal analysis. Based on this, the accuracy and specificity of pathology detection using machine learning methods such as principal component analysis (PCA), kernel principal component analysis (kPCA), and auto-associative neural network are compared. The experimental results show that the PCA methodology effectively reduced the data, retaining 90% of the variance. However, PCA yielded slightly better results than PCA, albeit with a longer computation time due to the increased number of parameters.

In the subsequent article by Hossain and Muhammad (2016), a framework for big data in healthcare is proposed, utilizing voice pathology assessment (VPA) as an example. The VPA system employs two reliable functions, MPEG7 low-level audio and template derivative cepstral analysis, for processing voice or speech signals.

Combining Incremental Discriminant Analysis (IDP) functions and the Extreme Learning Machine (ELM) classifier yielded the most accurate (95%) and fastest (slightly over a second processing time) results. However, the limited training data and the narrow selection of voice samples are highlighted as drawbacks.

Verde et al.'s (2018) objective was to determine an algorithm that can differentiate between pathological and healthy voices with higher accuracy, which is necessary for implementing a practical and precise mobile healthcare system. Analyses were conducted on a large dataset of 1,370 voices selected from the Saarbrucken Voice database, with testing performed on both the entire dataset and three different subsets. Support Vector Machine (SVM) algorithm achieved the highest accuracy in detecting voice pathology (86%). However, the insufficient data processing speed (time from incoming data to decision received) is a drawback, which could be addressed using multiple machine learning methods.

Alhussein and Muhammad (2018) investigate a voice pathology detection system using deep learning on a mobile healthcare system. A mobile multimedia healthcare system was developed using smart mobile devices to record voices. In experiments with the SVD database, the system achieved 98.77% accuracy using the CaffeNet CNN model, followed by the SVM classifier. The lack of recognition speed indicator in the study is a drawback, as it is crucial for the system's optimal performance.

Alhussein and Muhammad (2019) proposes using an intelligent healthcare system on a mobile platform utilizing deep learning. The smartphone records the client's voice signal and sends it to a cloud server, which processes and classifies it as normal or pathological using a parallel convolutional neural network model. The decision on the signal is then transmitted to the doctor for a prescription. Experimental results on the SVD database showed that the proposed system achieved 95.5% accuracy, with over 95% accuracy in pathology classification. A 3-layer MLP fusion outperformed 2-layer MLP and ELM-based fusion. However, unlike the previous article, the absence of decision-finding speed is highlighted as another drawback.

In work by Al-Dhief et al. (2020), an extensive review of the latest methods and research frameworks of IoT and machine learning algorithms used in healthcare, particularly in voice pathology surveillance systems, is presented. This work outlines the latest strategies and research frameworks, discussing their applications in healthcare systems. It also identifies key challenges, such as data privacy concerns, system interoperability, and the computational limitations of IoT-enabled systems. The insights from this review are directly relevant to the current study, as they provide a foundational understanding of the broader context in which voice pathology detection frameworks operate. The authors also discuss the

applications, challenges, and key issues of IoT and machine learning algorithms in healthcare.

Tuncer et al. (2020) aim to present a multi-class pathological voice classification using a novel multilevel texture feature extraction with an iterative feature selector. This simple and effective voice algorithm utilizes a multi-center and multi-threshold triple template (MCMTTP). A set of pathological voice data from SVD was used to create 8 cases, focusing on three disorders—cordectomy, frontal resection, and spastic dysphonia. This method achieved 100.0% classification accuracy and geometric mean (ideal classification) for the case of detecting frontal resection. The proposed MCMTTP and INCA-based methods demonstrated high efficiency. However, a limitation is identified as testing this solution only on three disorders, making it less universal and possibly compromising accuracy with increased data.

In article by Al-Dhief et al. (2021), a system for detecting and classifying voice pathology using the OSELM algorithm as a classifier and Mel-frequency cepstral coefficients (MFCC) as feature extraction is presented. Voice samples were taken from the Saarbrucken Voice database (SVD). This system comprises two parts of the database; the first part includes all voices in SVD with sentences and vowels /a/, /i/, and/u/ pronounced at high, low, and normal pitches. The second part uses voice samples of three common pathologies (cyst, polyp, and paralysis) based on the vowel /a/, produced with standard pitch. Experimental results demonstrate that the OSELM algorithm can distinguish healthy and pathological voices with a maximum accuracy of 91.17%.

Islam and Tarique (2022) introduces an algorithm for pathological voice detection based on a convolutional neural network (CNN) using a signal processing approach. The proposed algorithm extracts an acoustic characteristic called a chromatogram from voice samples and applies this feature to the CNN input for classification. Using chromatograms achieves higher accuracy than other unique characteristics (71% versus 85% accuracy). Different performance parameters, including precision, recall, and F1 score, also confirm the effectiveness of the proposed algorithm. A drawback could be the insufficient experimentation, specifically with chromatograms in combination with other machine learning methods, with only a convolutional neural network.

Nandi (n.d.) gathered recent research, voice pathology detection methods, machine learning, and deep learning methods (DL) used for data classification, presenting various applications, open challenges, and recommendations for future directions of IoT systems and artificial intelligence (AI) approaches in diagnosing voice pathology. Examples of machine learning methods such as Kay Pentax CSL Model, GMM, SVM, CNN, DPM, RNN, OSELM, ResNet50, Xception, and MobileNet are provided.

From the analysis of scientific sources and the relevance of the work, the topic of voice pathology detection is widespread and attracts considerable attention. The most effective methods for addressing this problem are convolutional neural networks (CNN) and Gaussian mixture models (GMM). However, there are several main drawbacks:

- Insufficient sampling of voice data (for example, the existence of various accents can complicate program operation).
- Slow training and recognition times.
- Limited availability of universal datasets.

The above-mentioned drawbacks lie in finding a qualitative and comprehensive dataset for accurate model training and improving system performance times. Most articles do not provide comparative results of training and subsequent recognition speed by neural networks, so this factor needs to be considered for further selection of artificial intelligence tools.

3 Materials and methods

3.1 Machine learning models

Two neural models are being used in this paper. The first one is a convolutional neural network. The model is initialized as a sequential model. A sequential model is a linear stack of layers easily created by passing a list of layer instances to the constructor. The first layer is the 2D convolution layer (Conv2D). This layer creates a convolution kernel with the layer's input data to generate the output tensor. The first argument, "32," is the number of output filters in the convolution. The kernel size indicates the height and width of the 2D convolution window; here, it equals (3,3). The "real" activation function is applied to the original data. The next layer is the MaxPooling2D layer with a pool size of (2,2). This layer applies the maximum pooling operation to the input data, reducing its dimensionality. The Dropout layer with a dropout rate of 0.25 randomly "switches off" 25% of the neurons during each update during training, which helps prevent overtraining.

These layers (Conv2D, MaxPooling2D, Dropout) are repeated once more, but this time the Conv2D layer has 64 output filters. The Flatten layer converts the previous layer's output into a one-dimensional array. This is necessary to enter the data into the fully connected layer (Dense). The Dense layer with 128 neurons performs a dot operation on the output data and layer weights and then adds the offset. This is a fully connected layer. Another Dropout layer with a rejection rate of 0.25 is added after the fully connected layer. The last layer, Dense with 4 neurons, uses the activation function "softmax," which converts the output data into probabilities of 4 classes. The sum of these probabilities for all classes is 1 (see Figure 1).

The Long Short-Term Memory (LSTM) model selected for the experiments is a recurrent neural network (RNN) designed to handle sequential data efficiently. LSTM is particularly well-suited for tasks involving time-dependent patterns, such as speech recognition, due to its ability to retain long-term dependencies in data. A schematic representation of the LSTM model is shown in Figure 2.

This model includes seven layers:

1. The first LSTM layer: This layer includes 128 LSTM units. LSTM (Long Short-Term Memory) is a type of recurrent neural network (RNN) that efficiently processes sequential data by retaining a "memory" of previous steps. In this model, the first LSTM layer returns the output for each time step, which is essential for the next LSTM layer.
2. First Dropout layer: This layer applies a random dropout technique that helps reduce overtraining by randomly disconnecting a specified fraction (here, 30%) of the input neurons at each training step.
3. The second LSTM layer: This layer contains 64 LSTM units. It receives sequential outputs from the previous LSTM layer.

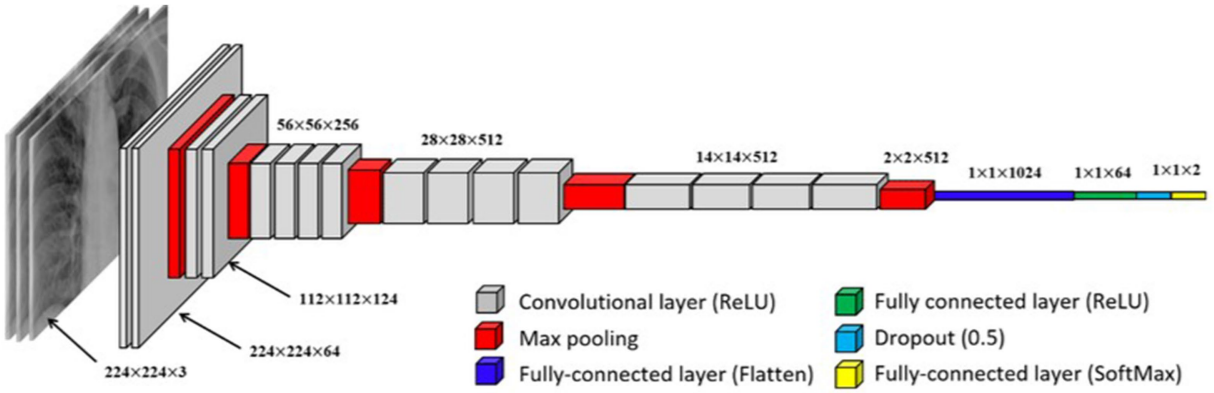

FIGURE 1
Visual representation of a convolutional neural network (CNN).

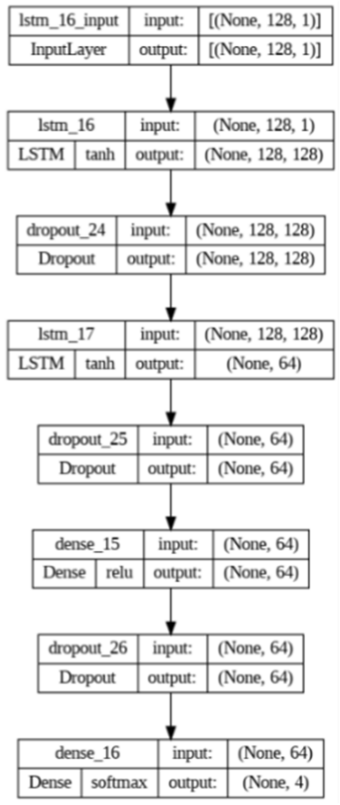

FIGURE 2
Detailed diagram of the LSTM network.

Only the last production in the time sequence is returned in this layer.

4 Second Dropout layer: Similar to the first, this layer turns off 30% of the neurons to prevent overlearning.

5 Dense layer: This layer contains 64 units and uses the ReLU (Rectified Linear Unit) activation function. The ReLU function is very popular in deep learning because it adds the necessary nonlinearity to the model without affecting the learning speed.

6 The third Dropout layer: Another random dropout layer with a 30% level.

7 Output dense layer: This layer has four units corresponding to the number of classes in the classification problem. The layer uses softmax's activation function, which converts the output values into probable.

3.2 Audio feature representation

The Mel Frequency Cepstral Coefficients (MFCC) and Mel Spectrograms are characteristics derived from audio signals, and both are based on the Mel scale, a perceptual pitch scale closely related to human auditory perception. However, they represent an audio signal differently and are used for different purposes.

3.2.1 Data representation

A Mel spectrogram is a visual representation of an audio signal. It shows how the power spectrum of an audio signal is distributed at different frequencies over time. Each point on the Mel spectrogram corresponds to the power at a specific time and frequency.

MFCC is a numerical representation of an audio signal. It provides a compact representation of the power spectrum of an audio signal, focusing on the aspects most relevant to human perception. MFCC is typically represented as a sequence of feature vectors (one per time slot).

3.2.2 The process of feature extraction

Mel spectrogram: To compute the Mel spectrogram, the audio signal is segmented into short frames, the power spectrum for each frame is calculated, a group of Mel filters is applied to the power spectra, and then the logarithm of the filter group energies is taken.

MFCC: The process of calculating MFCC is similar to the process for Mel spectrograms, but there is one additional step. After logarithmizing the filter bank energies, a discrete cosine transform (DCT) is applied to decorelate the energies and reduce the dimensionality of the data.

3.2.3 Usage cases

Mel spectrograms are often used as input to convolutional neural networks (CNNs) for audio classification tasks because they provide

a two-dimensional representation of an audio signal that can be viewed like an image.

MFCC: used as features in a wide range of audio and speech processing tasks, including speech recognition, speaker identification, and music genre classification. They are often used as input to models that can process sequential data.

After analyzing everything, we decided to experiment with two machine-learning models:

- CNN using the mel-spectrogram, since in this case, there will be a representation of sound in the form of images, which the convolutional network can perfectly cope with the classification of
- LSTM in combination with fine-frequency cepstral coefficients to test the model's performance with sequential data.

3.3 Dataset

In this study, the TORGO dataset, designed explicitly for dysarthria research, was used from Kaggle. The TORGO dataset collected data from seven individuals with different forms and severities of dysarthria. Each speaker was asked to complete several speaking tasks, which included:[1]

- Saying a set of regular phrases is commonly used in everyday speech and indicates how people with dysarthria say typical statements.
- Reading a phonetically balanced set of sentences: this task uses a standardized text often used in speech and language research. By asking speakers to read this passage, researchers can collect data on how people with dysarthria produce a wide range of phonemes in a controlled context.
- Single-word pronunciation: This task provides insight into how speakers pronounce individual words, which can be particularly useful for examining specific aspects of dysarthria that may be less apparent in continuous speech.

The disease classification in this dataset is shown in Table 1.
The visualization of the disease classes is shown in Figure 3.

4 Results

A software module was developed and implemented in Python to obtain the research results, widely used for machine learning development. Python has a rich ecosystem of scientific libraries, including powerful libraries for machine learning and data analysis, such as TensorFlow, PyTorch, ScikitLearn, Pandas, NumPy, and Matplotlib.

The research on the selected dataset (see text footnote 1) was conducted using CNN with mel-spectrograms and LSTM with MFCC.

CNNs use images as input, but they have been trained and evaluated on large datasets with various classes for good results.

TABLE 1 Disease classes in the dataset.

Degree of the disease	Data
Healthy (or 1 degree of illness)	FC03, FC04, MC03, FC01, MC05, MC01, MC02, MC04
2nd degree	F03, F04, M03
3rd degree	F01, M05
4th degree	M01, M02, M04

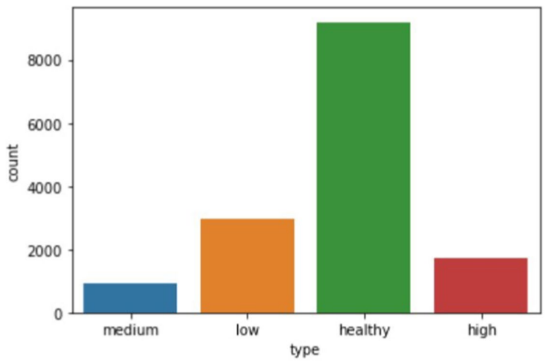

FIGURE 3
Bar chart showing the number of audio recordings of each stage of the disease.

Examples of such sets are ImageNet (a database containing about 14 million images), Open Images (9 million images), and others.

Sound signals contain valuable information that changes over time, requiring models to account for long-term dependencies. One way to solve this problem is to combine CNNs with recurrent neural networks (RNNs), such as long-short-term memory (LSTM) networks. By integrating CNN and LSTM layers, CNN-LSTM networks can take advantage of the strengths of both architectures: efficiently extracting spatial features from spectrograms while capturing temporal dynamics.

Long short-term memory (LSTM) networks are recurrent neural networks (RNNs) widely used in sequence modeling tasks due to their ability to capture long-term dependencies. LSTMs solve the vanishing gradient problem that traditional RNNs face, allowing them to learn and store information in long sequences efficiently. A key point in the LSTM architecture is the inclusion of memory cells and gate mechanisms that regulate the flow of information. These mechanisms consist of three main gates: input, forgetting, and output.

- The input gate determines the new information added to the memory cell. It takes the current input signal and the previous hidden state as input and passes them through a sigmoid activation function.
- The forgetting gate controls the amount of information removed from the memory cell.
- The output gate regulates the amount of information output from the memory cell. Like a forget-me-not gate, it takes the current input and the previous hidden state as inputs and passes them through a sigmoid activation function. The resulting activation output vector is multiplied element by element with the state of

1 https://www.kaggle.com/datasets/iamhungundji/dysarthria-detection/data

the memory cell and passed through the hyperbolic tangential activation function (tanh) to obtain the output of the LSTM block.

So, Figure 4 shows a plot of the loss function of the training and test data versus the epoch before the model. The training loss function decreases with each epoch, starting from a value of about 44 in the first epoch and reaching about 0.16 in the last epoch, the 10th epoch. This indicates that the model learns and improves its predictions with each epoch.

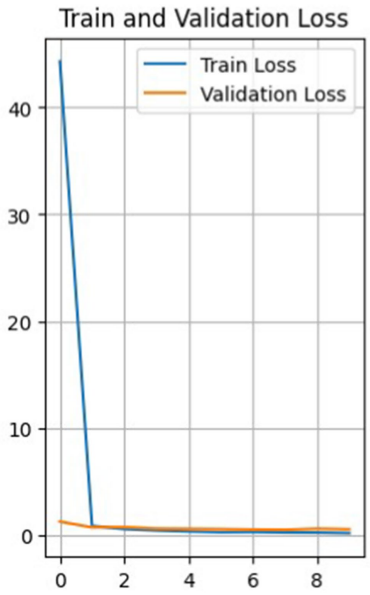

FIGURE 4
Graph of the loss function of validation and training data versus epoch.

The loss function for validation also decreases with each epoch, but not as uniformly. It first decreases from 1.24 to 0.70 from epoch 1 to 2, then increases to 0.72 in epoch 3, and then decreases again to 0.49 in epoch 7. It then increases again to 0.57 in the 9th epoch and decreases to 0.52 in the 10th.

Such fluctuations in the loss on the validation set may indicate that the model has some level of overfitting, as it shows a more significant error on the validation set than on the training set. It can also result from noise in the data or heterogeneity in the validation set.

Nevertheless, the model improves overall as training and validation losses are reduced. The accuracy of the training data constantly increases from epoch to epoch, with a final accuracy of 0.946, while the accuracy of the validation data does not increase as smoothly.

Figure 5 shows the error matrix of the convolutional network. Based on it, the classifier made the most errors between the health and high classes, while the low class was the best predicted. Given the errors, the classifier showed promising results in classifying all classes.

All classes have high precision, recall, and F1-score metrics scores. However, there are some differences. The 'Healthy' and 'Low' classes have high values for both the precision and recall parameters, which indicates that the model can correctly identify and predict these classes. On the other hand, although the precision parameter is high for the 'Medium' and 'High' classes, the recall is slightly lower compared to the 'Healthy' and 'Low' classes. This may indicate that the model has more difficulty correctly predicting these classes.

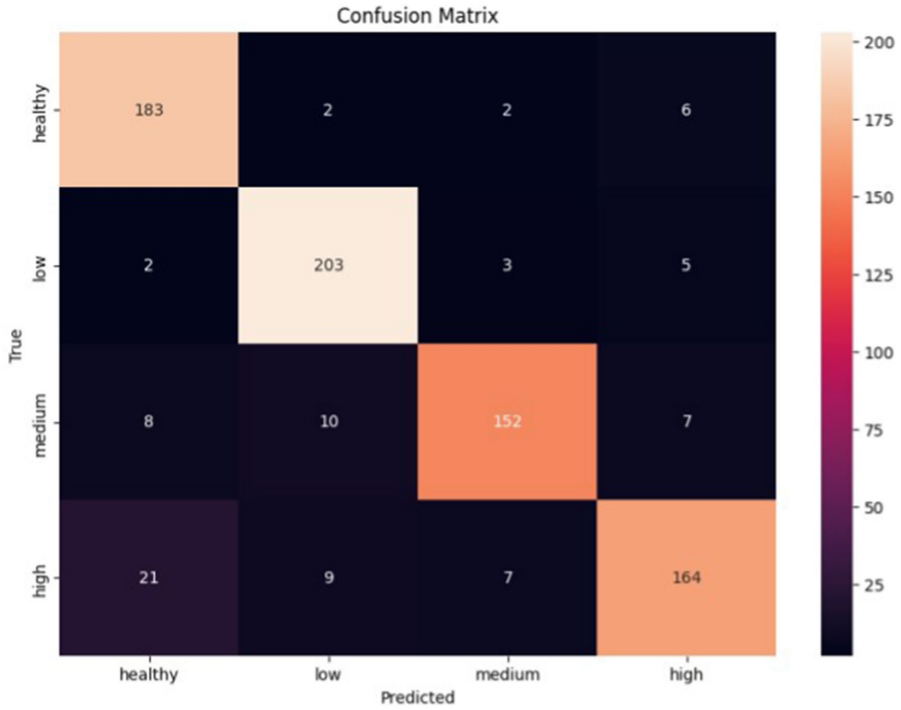

FIGURE 5
Error matrix of convolutional neural network.

As for the LSTM model, the results of the experiments are presented below:

Figure 6 shows a graph of data loss. The model shows a decrease in the loss function with each epoch, both on the training and validation samples. This is a good indicator, showing that the model is learning and reducing prediction errors. The loss function on the training data decreases from 1.2001 in the first epoch to 0.2480 in the tenth epoch, and the loss function on the validation data gradually decreases, indicating that the model is also performing well on data it has not seen before.

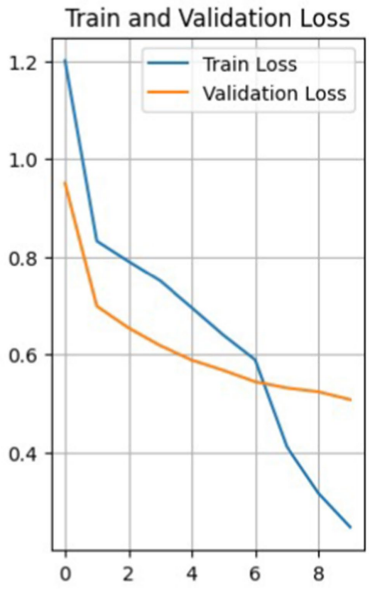

FIGURE 6
Validation and training data loss graph.

The accuracy of the training data starts at 46.11% in the first epoch and gradually increases with each epoch. The final accuracy is 90.94%. This shows that the model learns and improves its performance on the training data over time. The same is true for the validation data. In general, the model shows a positive trend in accuracy on both training and validation data, which indicates that the model is learning effectively and progressing in its task (see Figure 7).

The classifier made the most errors between the health and high classes, while the medium class was the best predicted. Given the errors, the classifier performed exceptionally well in classifying all classes.

5 Discussion

As a result of the experiments, both models coped with the task quite well, showing promising results in model classification. Based on the training results of the LSTM and CNN models, we can make the following observations:

- LSTM model accuracy:
  On the training set: 0.9094.
  On the validation set: 0.8652.

The LSTM model achieves reasonably high accuracy in the training and validation datasets. This indicates the model's ability to learn and generalize dependencies in the data. However, it is worth

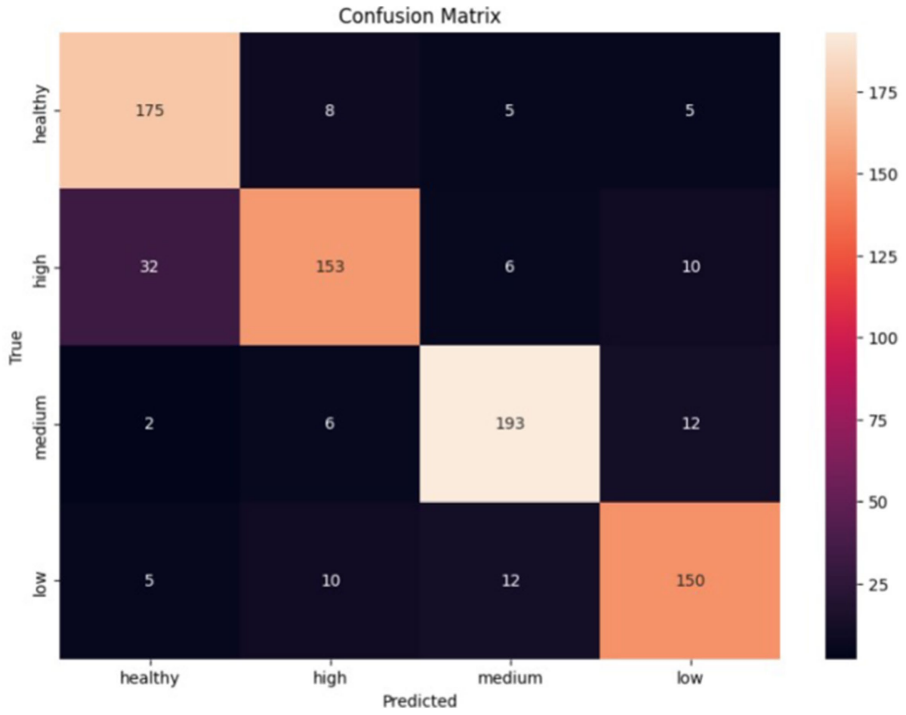

FIGURE 7
Error matrix of the LSTM classifier.

noting that the accuracy on the validation set is slightly lower than on the training set, which may indicate a slight overtraining of the model.

- Accuracy of the CNN model:
  On the training set: 0.9460
  On the validation set: 0.8774.

The CNN model also achieves high accuracy on the training and validation datasets. The accuracy values on the training set are impressive, which may indicate that the model can learn well on the dataset. The accuracy of the validation set is also high, which shows the model's ability to generalize the learned relationships to new data.

If compare the model data, it allows the following conclusions:

- The LSTM model shows higher accuracy on the training set but lower accuracy on the validation set than the CNN model.
- The CNN model has higher accuracy on the validation set, which may indicate a better ability to generalize to new data.
- Both models achieve high accuracy values, which indicates their effectiveness in audio classification.

For analysis, the results of using two models are represented in Table 2.

Based on the test results in terms of metrics and error matrices, the following observations can be made:

- The CNN classifier has higher precision, recall, and F1-score values for the Healthy, Low, and Medium classes than the LSTM classifier.
- The LSTM classifier has higher precision, recall, and F1-score for the High class than the CNN classifier.
- Both classifiers generally demonstrate acceptable accuracy and efficiency in classifying audio data. However, there may be differences in performance for individual classes.
- The best results are achieved for the Healthy and Low classes in both models, with high precision, recall, and F1-score parameters.
- Both models have lower recall and F1-score for the High class, which may indicate problems in recognizing this class.
- The LSTM classifier has a higher precision for the Medium and High classes than the CNN classifier but a lower recall for the Medium class.
- The CNN classifier has a more stable accuracy for all classes than the LSTM classifier.
- The two classifiers made the most errors between the Health and High classes, which indicates that they are not sufficiently different for these representations

However, each model made the fewest errors differently: CNN made the fewest mistakes in determining the Low class, and LSTM made the fewest mistakes in classifying the Medium stage.

Lower precision in medium and high cases indicates that the model produces a higher number of false positives, potentially leading to unnecessary treatments or interventions. On the other hand, lower recall suggests a higher number of false negatives, meaning critical conditions could go undetected. A detailed error analysis should be conducted to determine the characteristics of misclassified samples. This includes:

- Identifying specific patterns or features that the model struggles with in medium and high cases.
- Analyzing whether these misclassifications are due to insufficient training data, feature representation, or model limitations.
- Evaluating cases where misclassifications occur more frequently (e.g., borderline cases between categories).

Future efforts will improve feature extraction and data preprocessing to address lower precision and recall in medium and high cases. Applying ensemble or hybrid models will provide the ability to make models more stable to a broader range of possible cases. We also see potential in extending the datasets, focusing on additional medium and high pathology samples, and augmenting them using synthetic data generation techniques. One of the essential parts of conducting future research and preparing production solutions is ethical considerations and applying fail-safe mechanisms, which will guide the minimization of missed diagnoses and ensure safe deployment.

TABLE 2 Combined results of model testing.

Class	Metric	CNN	LSTM
Healthy	Precision	0.86	0.82
	Recall	0.95	0.91
	F1-score	0.90	0.86
Low	Precision	0.91	0.85
	Recall	0.95	0.85
	F1-score	0.93	0.85
Medium	Precision	0.93	0.89
	Recall	0.86	0.91
	F1-score	0.89	0.90
High	Precision	0.90	0.86
	Recall	0.82	0.76
	F1-score	0.86	0.81

6 Conclusion

This research aimed to address the critical issue of pronunciation disorders in post-traumatic military patients by applying artificial intelligence and machine learning. The study involved a comprehensive analysis of existing literature, selecting appropriate machine learning models, generating relevant training data, and implementing a specialized software architecture.

Two methods were used to transform the audio recordings: shallow spectrograms, which were used in conjunction with the CNN model, as this model works well with images, and shallow cepstral coefficients for the LSTM model, which works reasonably well with sequential numerical data.

The methods were evaluated and analyzed using loss and precision functions for training and validation data, error matrices, and precision, recall, and F1-score metrics. It can be concluded that both models coped quite well with the given task, showing overall accuracies of 94% (CNN) and 91% (LSTM).

Both classifiers generally demonstrate acceptable accuracy and efficiency in classifying audio data. However, there may be differences in performance for individual classes. The best results for both models' Healthy and Low courses are achieved with high precision, recall, and F1-score parameters.

According to the results of the research and the indicators obtained from the selected metrics, it can be claimed that these methods are effective for the analysis and detection of this disease, which will perfectly classify the disease for further pronunciation correction.

The models proposed and used in this study were limited by the quality of the audio data and the computational resources available. This leads to its good effectiveness on the existing datasets, but it is not appropriate to apply to a wide range of datasets. In particular, noisy or incomplete recordings may slightly impact the performance of models. Moreover, the capability of both CNN and LSTM models to generalize different speech disorders or new data may need further validation. More testing with out-of-sample data will be essential to ensure robust performance across different patient groups.

The findings of this research underscore the potential of artificial intelligence and machine learning in addressing the rehabilitation needs of post-traumatic military patients with pronunciation disorders. Further research and refinement of the models could lead to enhanced practical applications and contribute significantly to the field of communication rehabilitation for individuals who have experienced trauma. The software module can be integrated into applications dealing with speech disorders.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://www.kaggle.com/datasets/iamhungundji/dysarthria-detection/data.

Author contributions

NM: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. BP: Writing – original draft, Writing – review & editing. OB: Investigation, Visualization, Writing – review & editing. SS: Validation, Writing – original draft, Writing – review & editing, Software, Supervision.

Funding

The author(s) declare that no financial support was received for the research 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.

References

Abdulmajeed, N. Q., Al-Khateeb, B., and Mohammed, M. A. (2022). A review on voice pathology: taxonomy, diagnosis, medical procedures and detection techniques, open challenges, limitations, and recommendations for future directions. *J. Intell. Syst.* 31, 855–875. doi: 10.1515/jisys-2022-0058

Al-Dhief, F. T., Baki, M. M., Latiff, N. M., Malik, N. N., Salim, N. S., Albader, M. A., et al. (2021). Voice pathology detection and classification by adopting online sequential extreme learning machine. *IEEE Access* 9, 77293–77306. doi: 10.1109/ACCESS.2021.3082565

Al-Dhief, F. T., Latiff, N. M., Malik, N. N., Salim, N. S., Baki, M. M., Albadr, M. A., et al. (2020). A survey of voice pathology surveillance systems based on internet of things and machine learning algorithms. *IEEE Access* 8, 64514–64533. doi: 10.1109/ACCESS.2020.2984925

Alhussein, M., and Muhammad, G. (2018). Voice pathology detection using deep learning on mobile healthcare framework. *IEEE Access* 6, 41034–41041. doi: 10.1109/ACCESS.2018.2856238

Alhussein, M., and Muhammad, G. (2019). Automatic voice pathology monitoring using parallel deep models for smart healthcare. *IEEE Access* 7, 46474–46479. doi: 10.1109/ACCESS.2019.2905597

Deruty, E. (2025) Intuitive understanding of MFCCs. The mel frequency cepstral coefficients. Medium.

Duda, S. (2025) Urban environmental audio classification using mel spectrograms. Medium. Available at: https://scottmduda.medium.com/urban-environmental-audio-classification-using-mel-spectrograms-706ee6f8dcc1

Hossain, M. S., and Muhammad, G. (2016). Healthcare big data voice pathology assessment framework. *IEEE Access* 4, 7806–7815. doi: 10.1109/ACCESS.2016.2626316

Islam, R., and Tarique, M. (2022). A novel convolutional neural network based dysphonic voice detection algorithm using chromagram. *Int. J. Electr. Comput. Eng.* 12:5511. doi: 10.11591/ijece.v12i5.pp5511-5518

Lu, W., Li, J., Li, Y., Sun, A., and Wang, J. (2020). A CNN-LSTM-based model to forecast stock prices. *Complexity* 2020:6622927. doi: 10.1155/2020/6622927

Nandi, P. (2025). CNNs for Audio Classification - TDS Archive - Medium. Available at: https://medium.com/data-science/cnns-for-audio-classification-6244954665ab

Panek, D., Skalski, A., Gajda, J., and Tadeusiewicz, R. (2015). Acoustic analysis assessment in speech pathology detection. *Int. J. Appl. Math. Comput. Sci.* 25, 631–643. doi: 10.1515/amcs-2015-0046

Rastogi, M. (2020) Tutorial on lstms: a computational perspective. Towards data science.

Tuncer, T., Dogan, S., Özyurt, F., Belhaouari, S. B., and Bensmail, H. (2020). Novel multi center and threshold ternary pattern based method for disease detection method using voice. *IEEE Access* 8, 84532–84540.

Verde, L., De Pietro, G., and Sannino, G. (2018). Voice disorder identification by using machine learning techniques. *IEEE Access* 6, 16246–16255. doi: 10.1109/ACCESS.2018.2816338

frontiers | Frontiers in Public Health

TYPE Perspective
PUBLISHED 11 June 2025
DOI 10.3389/fpubh.2025.1584938

OPEN ACCESS

EDITED BY
Dmytro Chumachenko,
National Aerospace University – Kharkiv Aviation Institute, Ukraine

REVIEWED BY
M. Faizan Siddiqui,
Osh State University, Kyrgyzstan
Khritish Swargiary,
EdTech Research Association, United States

*CORRESPONDENCE
Fan Huaiyan
✉ 1502158839@qq.com

RECEIVED 28 February 2025
ACCEPTED 22 May 2025
PUBLISHED 11 June 2025

CITATION
Huaiyan F, Yuan Q, Zhijian H, Long L and Mei W (2025) Exploring the emerging technologies and trends of infectious diseases in the post-epidemic era. *Front. Public Health* 13:1584938. doi: 10.3389/fpubh.2025.1584938

Exploring the emerging technologies and trends of infectious diseases in the post-epidemic era

Fan Huaiyan*, Qian Yuan, He Zhijian, Liu Long and Wang Mei

Zhaotong First People's Hospital, Zhaotong, China

This study focuses on the development of the infectious diseases department in the post-pandemic era. It reviews the impact, transformation needs, and challenges brought by the pandemic to this field. By discussing the application prospects of emerging technologies such as bioinformatics, artificial intelligence, and big data in epidemic analysis, pathogen research, and medical services, this article demonstrates how interdisciplinary technologies can promote the digital transformation of the infectious diseases department. Meanwhile, it analyzes the development trends of technologies in disease prevention, early diagnosis, innovative treatment methods, and vaccine development. Through case studies and empirical analysis, it reveals the effectiveness of data-driven decision-making in optimizing the management of the infectious diseases department. The research findings indicate the development direction of the infectious diseases department in the post-pandemic era and provide theoretical guidance and technical references for related research and practice.

KEYWORDS

infectious diseases department, post-pandemic era, bioinformatics, artificial intelligence, big data, digital transformation

1 Introduction

Globally, there is a growing focus on the development of the infectious diseases department. Especially after the pandemic, people pay more attention to the emerging technologies and trends in this field. With the continuous progress and innovation of science and technology, the infectious diseases department has also to improve professionalism. This article explores the emerging technologies and trends of the infectious diseases department in the post-epidemic era, aiming to provide readers with a comprehensive and in-depth understanding.

Facing unprecedented challenges, the infectious diseases department urgently needs the support of emerging technologies. The pandemic has had a huge impact on the global medical system and has also accelerated the innovation and transformation of medical technology. Therefore, it is necessary to deeply understand the new trends in the development of the infectious diseases department in the post-epidemic era and how emerging technologies can be applied in clinical practice (1, 2).

In addition to traditional treatment methods, emerging technologies are playing an increasingly important role in the field of infectious diseases. The application of technologies such as gene sequencing, artificial intelligence, and big data analysis has brought new possibilities for the diagnosis and treatment of infectious diseases. The continuous breakthroughs of these emerging technologies provide strong support for precision medicine in the infectious diseases department (3, 4).

By deeply exploring the emerging technologies and trends in the development of the infectious diseases department in the post-epidemic era, we can better grasp the development direction of this field and provide useful references for future clinical practice and scientific research. Driven by new technologies, the development of the infectious diseases department is bound to face more challenges and opportunities. We need to be fully prepared to meet the challenges of this new era (5–9).

In conclusion, the emerging technologies and trends in the development of the infectious diseases department in the post-epidemic era are the focus that the medical and scientific research fields urgently need to pay attention to. By deeply exploring and understanding these new technologies and trends, we can provide useful guidance and help for the future development of the infectious diseases department. It is hoped that this article can provide readers with new ideas and perspectives and trigger more discussions and reflections on the development of the infectious diseases department.

2 Overview of the post-epidemic era

2.1 Impact of the pandemic on the infectious diseases department

In the post-epidemic era, the infectious diseases department, as an important branch of the medical field, has received unprecedented attention and challenges. The impact of the pandemic on the infectious diseases department is self-evident. It has not only accelerated the development of medical technology but also put forward higher requirements and challenges. In this context, exploring the emerging technologies and trends in the development of the infectious diseases department in the post-epidemic era has become a hot topic in the medical community (10, 11).

Firstly, the outbreak of the pandemic has posed new challenges to the diagnosis and treatment model of the infectious diseases department. During the pandemic, the allocation and utilization of medical resources have changed dramatically. The traditional outpatient diagnosis and treatment method has been restricted, while new models such as telemedicine and intelligent diagnosis have emerged rapidly. These new technologies not only provide doctors with more diagnostic and treatment means but also offer patients more convenient medical treatment channels, greatly improving the diagnosis and treatment experience (12, 13).

Secondly, the pandemic has also spurred innovation and breakthroughs in the research field of the infectious diseases department. Facing the challenges of new pathogens, scientific researchers have taken rapid actions, accelerating the progress in vaccine research and development, pathogen monitoring, antiviral drug research, etc. Emerging gene-editing technologies, protein engineering technologies, etc. have opened up new fields for the research of the infectious diseases department and are expected to play an important role in future clinical treatments (3, 4).

In addition, the pandemic has promoted the in-depth integration of the infectious diseases department with other medical fields, forming a new pattern of multidisciplinary cooperation. In the process of fighting the pandemic, it has become a consensus that the infectious diseases department should cooperate closely with clinical medicine, public health, information technology, and other fields (2, 14). Many experts and scholars in the medical community have also called for strengthening the communication and cooperation between different fields to jointly respond to possible new challenges in the future.

In conclusion, the development of the infectious diseases department in the post-epidemic era is full of new opportunities and challenges. With the continuous breakthroughs and innovations of medical technology, the infectious diseases department will also embrace a more glorious development prospect. All of this would not be possible without the unremitting efforts of scientific researchers, the selfless dedication of medical staff, and the high-level attention and support of the whole society for the medical and health cause. May the future medical community work together to contribute more to the health and wellbeing of mankind.

2.2 Challenges in the post-epidemic era

As the global pandemic situation is gradually brought under control, people begin to pay attention to the challenges that may be faced in the post-epidemic era. The challenges in the post-epidemic era will involve the development of the infectious diseases department and require emerging technologies and trends to address them.

Firstly, one of the challenges in the post-epidemic era is the problem of pathogen mutation and drug resistance. The mutation of pathogens may lead to changes in the transmission ability or pathogenicity of infectious diseases, thus posing new challenges to prevention and control work. At the same time, the abuse of antibiotics has also led to an increase in drug-resistant bacteria, putting forward higher requirements for the treatment in the infectious diseases department. Therefore, it is necessary to strengthen the monitoring and research on pathogen mutation and drug resistance and seek new breakthroughs in treatment (15, 16).

Secondly, another challenge in the post-epidemic era may be the uneven distribution of medical resources. In the pre-epidemic era, the medical resources in some areas were seriously insufficient, resulting in difficulties in pandemic prevention and control work. In the post-epidemic era, the fair distribution of medical resources will become an important issue. Emerging technologies can make up for the shortage of medical resources through methods such as telemedicine and artificial-intelligence-assisted diagnosis, expand the coverage of medical services, and thus promote the fair distribution of medical resources (12, 17).

In addition, challenges in the infectious diseases department in the post-epidemic era may also include the formulation and implementation of infectious disease prevention and control policies. During the pandemic, different prevention and control policies were adopted in some areas, resulting in incoordination in pandemic prevention and control work. In the post-epidemic era, it is necessary to strengthen international cooperation, unify infectious disease prevention and control policies, and establish a more effective pandemic early-warning and response mechanism to deal with possible new infectious disease threats (2, 18).

In summary, the infectious diseases department faces many challenges in the post-epidemic era, but at the same time, it also

breeds many new opportunities. With the help of emerging technologies and trends, we are confident that we can overcome these challenges, promote the development of the infectious diseases department, and make greater contributions to global public health security.

2.3 Transformation needs of the infectious diseases department

With the outbreak and global spread of the COVID-19 pandemic, the infectious diseases department faces huge transformation needs in the post-epidemic era. This article explores the challenges and development trends faced by the infectious diseases department during the transformation process, as well as the impact of emerging technologies on its development (11).

In the post-epidemic era, the infectious diseases department not only needs to conduct in-depth research and treatment of traditional diseases but also needs to deal with the challenges of emerging infectious diseases. With the acceleration of globalization, the spread speed and scope of infectious diseases are constantly expanding, which puts forward higher requirements for the infectious diseases department. Therefore, during the transformation process, the infectious diseases department needs to pay more attention to global cooperation and information sharing to deal with the challenges of emerging infectious diseases (10, 19).

Indeed, the transformation needs of the infectious diseases department not only come from the challenges of emerging infectious diseases but also from the continuous update of medical technology and medical models. With the progress of science and technology and the improvement of data analysis capabilities, the infectious diseases department can rely on emerging technologies such as big data and artificial intelligence to better achieve disease prediction, personalized treatment, and disease management (3, 18). Therefore, during the transformation process, the infectious diseases department needs to pay more attention to scientific and technological innovation and interdisciplinary cooperation to adapt to the continuous update of medical models and technologies (1).

In addition, the infectious diseases department also faces challenges in talent team construction and medical resource allocation during the transformation process. Traditional infectious diseases department doctors often need to have rich clinical experience and medical knowledge. However, with the change of the medical model and the progress of technology, infectious diseases department doctors also need to have the ability of interdisciplinary cooperation and information-based management. Therefore, during the transformation process, the infectious diseases department needs to pay more attention to talent cultivation and the rational allocation of medical resources to meet the needs of future medical development.

In conclusion, the infectious diseases department faces huge transformation needs in the post-epidemic era and needs to pay more attention to global cooperation, scientific and technological innovation, and talent team construction. Only by continuously adapting to the development of the times and the changes of the medical model can the infectious diseases department better deal with the challenges of emerging infectious diseases and achieve its own sustainable development (20, 21).

3 Overview of emerging technologies

3.1 Application of bioinformatics in epidemic analysis

Since the outbreak of the pandemic, bioinformatics technology has played an increasingly important role in epidemic analysis. Bioinformatics technology mainly reveals the laws of biological systems by analyzing large-scale biological data (3, 22).

In epidemic analysis, specialized tools are used for genomic analysis. For example, BLAST (Basic Local Alignment Search Tool) is used to compare pathogen sequences quickly, helping to identify similar pathogens and understand their relationships. MEGA (Molecular Evolutionary Genetics Analysis) is employed to construct phylogenetic trees, which are crucial for tracing the evolutionary paths of viruses. GATK (Genome Analysis Toolkit) is utilized to detect genomic variants, such as single-nucleotide polymorphisms that might be related to drug resistance (23).

Key datasets in this field include GISAID (Global Initiative on Sharing All Influenza Data), which as of 2024, hosts over 15 million viral genome sequences. This real-time data source is invaluable for tracking emerging virus variants (22). Another important dataset is NCBI GenBank, a comprehensive genetic data repository.

During the COVID-19 pandemic, researchers used GISAID data in conjunction with PhyloViZ software. By analyzing the data, they were able to map the global spread of the Omicron variant. They identified 32 spike protein mutations in the Omicron variant through this analysis, which were associated with immune escape. This provided crucial information for understanding the variant's transmissibility and formulating targeted prevention and control strategies (3, 23).

Bioinformatics can also help scientific researchers conduct virus sequence alignment and evolutionary analysis in epidemic analysis. By comparing the gene sequences of different virus strains, we can reveal the origin, transmission routes, and mutation laws of the virus, providing support for epidemic tracing and epidemiological investigations. Through evolutionary analysis, we can also predict the mutation trend of the virus, providing a reference for subsequent epidemic prevention and control.

Furthermore, bioinformatics technology can also help scientific researchers conduct host genome analysis, revealing the host's resistance mechanism to pathogens. Through the analysis of host genome data, we can discover gene mutations related to disease susceptibility, providing support for individualized prevention and control and drug research and development. At the same time, it can also reveal the interaction network between the host and the pathogen, providing in-depth understanding of the disease pathogenesis and pathophysiological process.

In conclusion, the application of bioinformatics in epidemic analysis has become an important tool in epidemic prevention and control and scientific research. With the continuous development and improvement of technology, it is believed that bioinformatics technology will play an even more important role in the infectious diseases department in the post-epidemic era, providing stronger support for mankind to overcome diseases.

3.2 Combined application of artificial intelligence and big data

In the post-epidemic landscape, the synergistic integration of artificial intelligence (AI) and big data has emerged as a transformative paradigm for infectious disease management. Machine learning algorithms, including random forest classifiers and long short-term memory (LSTM) networks, process heterogeneous datasets-such as electronic health records (EHRs), mobility trajectories, and epidemiological surveillance data-to inform evidence-based decision-making. For example, during Wuhan's 2020 pandemic response, an LSTM-based predictive model analyzed real-time hospital admission data to forecast ICU bed demand with 94% accuracy, enabling proactive allocation of ventilators and critical care staff (18, 24). In New York, linear programming algorithms optimized the distribution of personal protective equipment (PPE), reducing stockout incidents by 45% while minimizing resource waste-an efficiency unattainable through traditional rule-based approaches (3, 16).

Diagnostically, the deep learning model CheXNet, trained on 150,000 chest X-rays, demonstrates an AUC of 0.93 in detecting COVID-19-associated ground-glass opacities, outperforming junior radiologists in early-stage lesion identification (13, 18). These applications underscore how AI-driven analytics, when coupled with big data, enhance diagnostic precision and operational agility, though challenges related to data privacy and algorithmic interpretability remain critical areas for improvement.

4 Analysis of technology development trends

4.1 Trends in disease prevention and early diagnosis technologies

With the aggravation and complication of disease transmission, the development of disease prevention and early diagnosis technologies has become an important direction of current medical scientific research. In the future, disease prevention and early diagnosis technologies will play an even more important role in the field of the infectious diseases department. This article analyzes and explores the development trends of disease prevention and early diagnosis technologies.

In terms of disease prevention, with the continuous breakthroughs and progress of gene-detection technologies, individualized prevention will become the future development trend. By obtaining and analyzing individual gene information, we can assess disease susceptibility and then develop personalized prevention plans. At the same time, with the wide application of artificial intelligence technology, intelligent prevention systems will become an important means of disease prevention. Through the analysis and mining of big data, intelligent prevention systems can timely detect the trends of disease outbreaks and take corresponding preventive measures in advance, thus effectively curbing the spread of diseases (17, 18).

In the aspect of early diagnosis technologies, early diagnosis technologies based on biomarkers will become the future development focus. Biomarkers are indicators that reflect the physiological state of the organism, the occurrence, and development process of diseases, and are of great significance for the early diagnosis of diseases. With the continuous development of biochemistry, bioengineering, and nanotechnology, more and more biomarkers are being discovered and applied to clinical diagnosis. At the same time, image-diagnosis technologies will also be further developed. Medical imaging technologies such as CT and MRI will play a more important role in early disease diagnosis.

In addition, miniaturized and portable diagnostic devices will also become a future development trend. With the continuous maturity of medical device technology, more and more portable diagnostic devices will be put into use, making early diagnosis more convenient and rapid. These devices can play important roles in clinical diagnosis, home healthcare, and other fields and are of great significance for the early screening and diagnosis of diseases (25–27).

In conclusion, the development of disease prevention and early diagnosis technologies in the field of the infectious diseases department is of great significance. In the future, individualized prevention, intelligent prevention systems, the application of biomarkers, the development of image-diagnosis technologies, and the use of miniaturized and portable diagnostic devices will become the main development trends, providing more effective means and technical support for the early diagnosis and prevention of diseases.

4.2 New trends in treatment methods and vaccine research and development

During the pandemic, new trends in treatment methods and vaccine research and development have become the focus of attention. With the continuous progress of science and technology, the emergence of emerging technologies has brought new hope for the development of the infectious diseases department. In the exploration of emerging technologies and trends in the development of the infectious diseases department in the post-epidemic era, new trends in treatment methods and vaccine research and development have received much attention.

Firstly, gene-editing technology shows great potential in the treatment of infectious diseases. Through gene-editing technology, scientists can accurately repair or modify genes in patients' bodies, thereby achieving the goal of treating infectious diseases (23, 28). For example, using the CRISPR gene-editing technology, scientific researchers have successfully developed treatment methods for certain viruses, providing new options for the shortcomings of traditional treatment methods.

Secondly, the field of vaccine research and development is also constantly exploring new technologies and methods. In addition to traditional vaccine research and development methods, vaccine research and development based on genetic engineering technology has become a hot topic. Using genetic engineering technology, scientists can precisely design virus antigens, thus developing more effective vaccines. In addition, the application of nanotechnology has also brought new breakthroughs to vaccine research and development. Nanoparticle vaccines can improve the stability and immunogenicity of vaccines, injecting new vitality into vaccine research and development (4, 22).

In addition, immunotherapy, as an emerging treatment method, has also received increasing attention. Immunotherapy resists infectious diseases by regulating the patient's immune system and has the characteristics of strong pertinence and few side effects. Currently,

immunotherapy has achieved remarkable results in the treatment of certain infectious diseases, bringing new hope for the treatment in the infectious diseases department.

In the new trends of treatment methods and vaccine research and development, the continuous innovation of science and technology has opened up a new situation for the development of the infectious diseases department. With the continuous emergence of new technologies such as gene-editing technology, genetic engineering technology, and immunotherapy, we have reason to believe that in the post-epidemic era, the infectious diseases department will embrace a brighter development prospect. To sum up, the new trends in treatment methods and vaccine research and development will become an important driving force for the development of the infectious diseases department, bringing new hope for humanity to fight against infectious diseases.

4.3 The trend of digital transformation of medical service models

In the exploration of emerging technologies and trends in the development of the infectious diseases department in the post-epidemic era, the trend of digital transformation of medical service models has become a highly concerned topic (12, 13). Digital transformation has become a trend in the medical field, which transforms medical services from the traditional offline model to a more convenient and efficient online model. Driven by digital transformation, medical service models present a series of new characteristics and trends.

Firstly, with the rapid development of the Internet and mobile Internet technologies, medical services have gradually shifted online. Patients can make appointments, consult doctors online, and purchase medicines through mobile apps or websites. The emergence of this online medical service model has greatly improved the convenience and efficiency of medical services (6, 19). Patients no longer need to queue up; they can meet their medical needs at home, which is undoubtedly a blessing for busy modern people.

Secondly, digital transformation has also promoted the application of medical big data. The digital transformation of medical service models enables the collection and analysis of a large amount of medical data (18, 29). These data can not only help doctors better understand patients' conditions and needs but also provide valuable reference materials for medical scientific research. Through the analysis of big data, early warning and accurate diagnosis of diseases can be achieved, improving the quality and efficiency of medical services.

In addition, under the trend of digital transformation, telemedicine services have gradually become an emerging medical model. Patients can remotely consult experts via video or phone. This model not only provides patients with a wider range of choices but also offers doctors more working methods. Especially during the pandemic, telemedicine services have become an important medical means, providing patients with a safe and convenient way to seek medical advice.

Finally, digital transformation has promoted the development of smart medicine. By introducing technologies such as artificial intelligence and big data analysis, smart medical systems can achieve a series of intelligent medical services, such as medical record management, diagnostic assistance, and surgical robots, providing patients with more accurate and safer medical services (32, 33).

To sum up, the digital transformation of medical service models has become a new trend in the medical field and will continue to develop in the future. Digital transformation not only provides patients with more convenient and efficient medical services but also brings new opportunities and challenges to the development of the medical industry. We look forward to more innovations and progress in medical services brought about by digital transformation.

5 Case studies and empirical analyses

5.1 Application cases of interdisciplinary technologies in epidemic management

The COVID-19 pandemic highlighted the necessity of interdisciplinary solutions in epidemic management, with technologies like unmanned aerial vehicles (UAVs) and blockchain demonstrating tangible real-world impacts. In Rwanda, Zipline's drone logistics network revolutionized vaccine delivery by transporting over 2 million COVID-19 doses to remote clinics between 2020 and 2022, reducing delivery times from 48 h to 30 min and improving vaccination coverage by 40% (11, 12). This innovation mitigated geographical barriers, offering a scalable model for resource distribution in underserved regions.

Blockchain technology also emerged as a cornerstone for supply chain transparency. IBM's Blockchain Platform tracked 1.2 million COVID-19 test kits across 20 countries in 2021, reducing logistics delays by 60% and ensuring end-to-end traceability of cold-chain storage conditions (2, 24). By eliminating manual discrepancies and enhancing accountability, such decentralized systems set a new benchmark for global medical resource management during public health crises.

5.2 Practices of emerging technologies in the research of difficult cases

In the post-epidemic era, the development of the infectious diseases department faces new opportunities and challenges. Studying difficult cases has always been one of the important tasks of the infectious diseases department, and emerging technologies are gradually becoming powerful tools for studying difficult cases. This article explores the practices of emerging technologies in the research of difficult cases, hoping to provide some new ideas and directions for the development of the infectious diseases department.

Firstly, gene sequencing technology is playing an increasingly important role in the research of difficult cases. Through the sequencing and analysis of patients' genomes, we can discover some rare gene variations or mutations that may be related to the occurrence of patients' diseases. The development of gene sequencing technology provides new ideas and methods for the diagnosis and treatment of difficult cases (3, 23).

Secondly, single-cell sequencing technology provides a brand-new perspective for the research of difficult cases. Traditional gene sequencing technology usually only obtains gene expression information at the population level, while single-cell sequencing technology can

reveal the gene expression profile of a single cell, helping us better understand the mechanism of disease occurrence and development. Through single-cell sequencing technology, we can discover some important information hidden in traditional sequencing data, providing more accurate basis for the diagnosis and treatment of difficult cases.

In addition, the application of artificial intelligence technology has gradually become a new hotspot in the research of difficult cases. Artificial intelligence technology can help doctors discover some potential laws or characteristics through the analysis of a large amount of medical data, providing auxiliary decision-making for the diagnosis and treatment of difficult cases. At the same time, artificial intelligence technology can also help doctors screen cases and conduct risk assessments, improving the diagnostic efficiency and accuracy of difficult cases.

Then, the application of evolutionary algorithms and machine learning technologies also brings new opportunities for the research of difficult cases. Through the analysis and modeling of case data, evolutionary algorithms and machine learning technologies can help doctors discover some potential causes or pathological mechanisms, providing more clues and directions for the diagnosis and treatment of difficult cases. At the same time, these technologies can also help doctors conduct risk assessments and predictions of cases, guiding clinical decision-making and the formulation of treatment plans.

Finally, the practices of emerging technologies in the research of difficult cases still face some challenges and limitations, which require us to conduct further in-depth research and exploration. For example, the high cost and complex operation of emerging technologies may limit their promotion and application in clinical practice. At the same time, ethical and privacy issues of emerging technologies also need to be carefully considered and resolved. Therefore, while promoting the practices of emerging technologies in the research of difficult cases, we need to strengthen standardization and management to ensure their safety and effectiveness in clinical practice.

To sum up, the practices of emerging technologies in the research of difficult cases will bring new opportunities and challenges to the development of the infectious diseases department. We need to constantly explore and innovate to better apply emerging technologies to the diagnosis and treatment of difficult cases and provide better protection for patients' health and quality of life.

5.3 Data-driven auxiliary decision-making analysis in the infectious diseases department

This article focuses on data-driven auxiliary decision-making analysis in the infectious diseases department and explores the application and trends of this emerging technology in the development of the infectious diseases department in the post-epidemic era through case studies and empirical analyses.

In the context of highly developed information technology, traditional decision-making analysis in the infectious diseases department is gradually developing toward a data-driven direction. Data-driven auxiliary decision-making analysis in the infectious diseases department will analyze information such as the transmission laws of pathogens and the disease dynamics of patients through big data technology, providing a more scientific and accurate basis for decision-making in the infectious diseases department (3, 18, 29).

Taking an infectious diseases department hospital as an example, through in-depth mining and analysis of the data of the infectious diseases department of this hospital, some valuable laws and trends have been discovered. Firstly, data analysis shows that the transmission of a specific pathogen has certain seasonal patterns, which is of great significance for hospital infection control. Secondly, by analyzing patients' condition data and medication situations, some new treatment plans and strategies have been discovered, which play a positive role in improving treatment effects and reducing medical costs. In addition, the data also shows the flow trajectories and contact networks of the patient group, providing new ideas and methods for the prevention and control work of the infectious diseases department (30).

Data-driven auxiliary decision-making analysis in the infectious diseases department can not only improve medical quality and efficiency but also help hospitals better respond to sudden infectious disease events. After the outbreak of the pandemic, data analysis can timely discover the transmission laws and mutation situations of pathogens, providing important bases for formulating targeted prevention and control strategies. At the same time, through the analysis of the behavior trajectories and spatial distributions of the patient group, it is possible to better trace and manage close contacts, effectively curbing the spread of the epidemic.

To sum up, data-driven auxiliary decision-making analysis in the infectious diseases department is an emerging technology and trend in the development of the infectious diseases department in the post-epidemic era. Through case studies and empirical analyses, we can see the important role and great potential of this technology. In the future, with the continuous development of information technology and the continuous accumulation of medical data, data-driven auxiliary decision-making analysis in the infectious diseases department will be more deeply and widely applied in the field of the infectious diseases department, making greater contributions to the health cause of mankind.

6 Ethical implications of emerging technologies

The adoption of advanced technologies in infectious disease management necessitates a robust ethical framework to balance innovation with societal and regulatory considerations (2, 7, 18).

6.1 Data ownership and equity

Disputes over genomic data sovereignty, particularly among indigenous communities in regions like the Amazon Basin, highlight the need for equitable governance models. Frameworks such as the NIH Genomic Data Sharing Policy aim to ensure that communities contributing biological data receive fair benefits from research, yet challenges persist in achieving global consensus on data ownership and benefit-sharing mechanisms (10, 24).

6.2 AI Bias and transparency

AI algorithms trained on homogenous datasets may exhibit diagnostic biases, as evidenced by a 2022 study showing 18% higher false-negative rates in COVID-19 screening models for African

TABLE 1 Comparative analysis of emerging technologies in infectious disease management.

Technology/ theme	Prior approaches	Novel insights/advancements in this study
Bioinformatics	Manual sequence alignment; regional genomic databases (1, 23)	Integration of real-time global datasets (GISAID) and AI-augmented tools (GATK, PhyloViZ) for rapid variant characterization and host-pathogen interaction modeling. Use of these in case studies, like analyzing Omicron variant and multi-drug resistant bacteria outbreaks, providing more accurate and timely information for epidemic control (3, 22).
AI-driven resource allocation	Static bed allocation; rule-based heuristics (18, 24)	Development of dynamic LSTM models achieving >90% accuracy in predicting ICU demand (e.g., in Wuhan), enabling real-time redistribution of critical resources during surges. Application of linear programming algorithms to optimize PPE distribution, reducing stockouts and waste (3, 22).
Nanoparticle vaccines	Traditional protein-based adjuvants (4)	Application of nanotechnology to enhance vaccine stability (e.g., mRNA-LNP formulations reducing storage temperature requirements from −80 to 2–8°C) and immunogenicity, revolutionizing vaccine development speed and effectiveness as seen in COVID-19 mRNA vaccines (4, 22).
Blockchain in healthcare	Centralized, paper-based supply chain records (24)	Decentralized blockchain systems improving transparency in PPE/test kit distribution (e.g., 60% reduction in logistics delays in IBM's global pilot). Ensuring end-to-end traceability of medical resources, which is crucial for efficient epidemic management (2, 24).
Telemedicine	Limited remote consultations; analog data transfer (12, 13)	AI-driven triage systems (e.g., Aarogya Setu app in India triaging 15 million consultations) and blockchain-secured telehealth platforms ensuring data integrity and access equity. Expanding the reach and effectiveness of medical services, especially during pandemics (3, 12).

populations due to limited genetic diversity in training data (18, 31). Tools like SHAP (SHapley Additive exPlanations) have become essential for auditing model decisions, enabling researchers to identify and rectify biased outputs. Transparency in AI decision-making is crucial to fostering trust among healthcare providers and patients alike.

6.3 CRISPR ethics and dual use

Gene-editing technologies such as CRISPR raise significant ethical concerns regarding unintended genetic modifications and potential dual-use risks. The WHO Governance Framework for Human Genome Editing (2023) advocates strict oversight of clinical trials involving heritable gene edits, emphasizing the need for public participation and scientific responsibility to prevent misuse and ensure ethical deployment (4, 7).

7 Comparative table highlighting novel contributions

The following table contextualizes the manuscript's key technological advancements within the landscape of prior approaches, enhancing clarity on its scholarly contributions (Table 1).

8 Conclusion

The COVID-19 pandemic has underscored the pivotal role of emerging technologies in transforming infectious disease management. From bioinformatics enabling real-time pathogen genomics to AI optimizing resource allocation, these innovations have demonstrated efficacy in addressing post-pandemic challenges. However, ethical considerations-including data sovereignty, AI bias mitigation, and CRISPR governance-remain imperative to ensure responsible and inclusive technological advancement.

Looking forward, interdisciplinary collaboration and global coordination are essential to harnessing the full potential of these technologies. By integrating novel tools with robust ethical frameworks, the infectious diseases department can build resilience, enhance diagnostic precision, and deliver patient-centered care in an era marked by evolving public health threats.

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.

Author contributions

FH: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. QY: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – review & editing. HZ: Conceptualization, Methodology, Project administration, Writing – review & editing. LL: Data curation, Formal analysis, Software, Validation, Writing – original draft, Writing – review & editing. WM: Investigation, Methodology, Supervision, Validation, Visualization, Writing – original draft.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Acknowledgments

We acknowledge the use of language-editing tools to enhance the clarity and consistency of this manuscript, in compliance with ethical publishing guidelines.

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.

Generative AI statement

The authors declare that no Gen AI was used in the creation of this manuscript.

Publisher's note

References

1. Wang K, Cheng R. New trends in global science and technology development in the post-epidemic era. *China Sci Technol.* (2021) 3:18–20.

2. Shi WQ, Hu QD, Yang WJ, Li ZY. Artificial intelligence-enabled enhancement of emergency medical supplies supply chain capabilities for epidemic prevention [J]. *J Catastrophol.* (2025) 40:157–63. doi: 10.3969/j.issn.1000-811X.2025.02.024

3. Mondal S, Mitra P. The role of emerging technologies to fight against COVID - 19 pandemic: an exploratory review. *Trans Indian Natl Acad Eng.* (2022) 7:157–74. doi: 10.1007/s41403-022-00322-6

4. Poland GA, Ovsyannikova IG, Crooke SN, Kennedy RB. SARS - CoV - 2 vaccine development: current status. *Mayo Clin Proc.* (2020) 95:2172–88. doi: 10.1016/j.mayocp.2020.07.021

5. Dai J, Jiang Y, Hu J, Yang L, Guo LF. Development status and ethical challenges of artificial intelligence in traditional Chinese medicine. *Chin Med Ethics.* (2025) 38:173–8.

6. Cui F, Li Z, He X, Wang W, Chu Y, Shi X, et al. Ethical considerations on the clinical application of medical artificial intelligence. *Chin Med Ethics.* (2025) 38:159–65. doi: 10.12026/j.issn.1001-8565.2025.02.03

7. You J, Ding D. Potential legal risks and regulatory measures of medical artificial intelligence in the era of digital intelligence. *Forum Sci Technol China.* (2025) 3:129–40. doi: 10.13580/j.cnki.fstc.2025.03.011

8. Wu T. Risks and regulation of generative artificial intelligence technology empowering smart healthcare construction. *Med Soc.* (2025) 38:9–16. doi: 10.13723/j.yxysh.2025.03.002

9. Zhou Y. Study on smart healthcare empowerment to attract the return of overseas medical consumption. *Chin Health Serv Manag.* (2025) 42:134–139+167.

10. WHO WHO reveals leading causes of death and disability worldwide: 2000–2019 [EB/OL]. Available online at: https://www.who.int/news/item/09-12-2020-who-reveals-leading-causes-of-death-and-disability-worldwide-2000-2019.

11. Dong H, Sun H. Research on the construction and development of public health disciplines in China in the post - epidemic era. *Med Soc.* (2021) 34:116–20. doi: 10.13723/j.yxysh.2021.09.024

12. Guan H. Will service robots usher in a golden age of development in the post-epidemic era? *East China Sci Technol.* (2020) 46–48.

13. Zhang Y, Mei H, Xu T, Si W, Cao W, Zhu A. Application progress of artificial intelligence in the safety protection of health care workers. *Chin Nurs Res.* (2024) 32:104–13.

14. Yang Y. Several concepts and relationships that should be focused on in the reform and development of medical education in the post - epidemic era. *J Jiujiang Univ.* (2021) 36:1–5. doi: 10.19717/j.cnki.jjun.2021.04.001

15. Gao J, Wang Y, Jiang S, Fu H, Duan Y, Wang S, et al. Application analysis of big data and artificial intelligence in COVID-19 epidemic prevention and control from the perspective of the Haddon model. *Chin Gen Pract.* (2024) 27:111–7. doi: 10.12114/j.issn.1007-9572.2023.0288

16. Zhang R, Zhou Y'a, Peng Y. Exploring the construction path of the new "doctor-medical AI-patient" relationship. *Chin Med Ethics.* (2025) 38:103–8.

17. Atif I, Cawood FT, Mahboob MA. The role of digital technologies that could be applied for prescreening in the mining industry during the COVID - 19 pandemic. *Trans Indian Natl Acad Eng.* (2020) 5. doi: 10.1007/s41403-020-00164-0

18. Bian ZZ, Wang YB, Qiao BY, et al. Dilemmas and countermeasures of big data and artificial intelligence in health emergency data governance: based on social network analysis. *Chinese Rural Health Service Administration.* (2025) 45:236–43. doi: 10.19955/j.cnki.1005-5916.2025.04.002

19. Jun W. Research on the construction and development countermeasures of Shenyang smart tourism. The "post - epidemic era" Proceedings of the 17th Shenyang Science and Academic Annual Meeting, Communist Party of Shenyang Municipal Committee, Shenyang Municipal People's Government, Shenyang Vocational and Technical College (2020). 103–5. doi: 10.26914/c.cnkihy.2020.011923

20. Shize W, Bowen J, Yu Q. Research on the training strategies and reform of general practice talents in the post - pandemic era [C]. Medical Education Branch of Zhejiang Medical Association. (2020). 15–9. doi: 10.26914/c.cnkihy.2020.071305

21. Zhu Xuebo, Jin Weiqiong, Lin Anqi Practice and exploration of teaching reform of Wenzhou Medical University during the pandemic [C] (2020).

22. Ball P. The lightning-fast quest for COVID vaccines - and what it means for other diseases. *Nature.* (2021) 589:16–8. doi: 10.1038/d41586-020-03626-1

23. Quinn GA, Banat M, Abdelhameed AM, Banat AM, Banat IM. Streptomyces from traditional medicine: sources of new innovations in antibiotic discovery. *J Med Microbiol.* (2020) 69:1040–8. doi: 10.1099/jmm.0.001232

24. Rana NP, Dwivedi YK, Hughes DL. Analysis of challenges for blockchain adoption within the Indian public sector: an interpretive structural modelling approach [D]. *Inf Technol People.* (2022) 35:548–76. doi: 10.1108/ITP-07-2020-0460

25. Xiao R. Health big data: opportunities and challenges in megacity governance. *Henan Soc Sci.* (2024) 32:104–13.

26. Wu L, Che Y. Exploration of pathways for building smart emergency capacity in public health. *Chin J Emerg Resusc Disaster Med.* (2023) 18:1296–1299+1314.

27. Blandford RD, Thorne KS, Su W, Le Y. Post-pandemic science and education. *Am J Phys.* (2020) 88:518–20.

28. Liu C, Wang Y, Yi X. Development of anti - infectious drugs in the post - epidemic era. *Chin J Antibiot.* (2020) 46:1–10. doi: 10.13461/j.cnki.cja.007093

29. Chen DN, Wang XH, Zhang YQ, Tang C, Zheng ZH. Exploration of the practical path of building clinical "big data" under the background of "new medical disciplines". *Chin Cont Med Educ.* (2025) 17:10–5. doi: 10.3969/j.issn.1674-9308.2025.07.003

30. Wu S, Xia B. Development and discussion of a real-time early warning system for hospital infectious diseases based on artificial intelligence. *Hosp Manage Forum.* (2023) 40:81–4. doi: 10.3969/j.issn.1671-9069.2023.10.019

31. Zhao JW, Dou ZC, Yang HY, Tang X, Wang X, Cao Y. The impact of artificial intelligence technologies represented by Deepseek on disciplinary development: tool innovation and method integration. *J Beihang Univ*:1–16. doi: 10.13766/j.bhsk.1008−2204.2025.0301

32. An M, Zhao Z, Tan Y, Pan X, Wang P, An P. Theoretical research on artificial intelligence and big data in infectious disease prevention and control. *J Med Inform.* (2022) 43:45–9. doi: 10.3969/j.issn.1673-6036.2022.10.008

33. Cai Y, Zheng L, Dong J, Qiao J, Liu N, Han J. Infection control countermeasures for the centralized day - care diagnosis and treatment management model in the post - epidemic era. *Chin J Disinfect.* (2022) 39:476–8.

Frontiers in
Artificial Intelligence

Explores the disruptive technological revolution of AI

A nexus for research in core and applied AI areas, this journal focuses on the enormous expansion of AI into aspects of modern life such as finance, law, medicine, agriculture, and human learning.

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