frontiers | Research Topics

Internet of medical things and computational intelligence in healthcare 4.0

Edited by
Sujata Dash, Wellington Pinheiro dos Santos and
Subhendu Kumar Pani

Published in
Frontiers in Big Data
Frontiers in Artificial Intelligence

ISSN 1664-8714
ISBN 978-2-8325-5595-8
DOI 10.3389/978-2-8325-5595-8

About Frontiers

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

Frontiers journal series

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

Dedication to quality

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

What are Frontiers Research Topics?

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

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

Internet of medical things and computational intelligence in healthcare 4.0

Topic editors

Sujata Dash — Nagaland University, India

Wellington Pinheiro dos Santos — Federal University of Pernambuco, Brazil

Subhendu Kumar Pani — Krupajal Engineering College, India

Citation

Dash, S., Santos, W. P. d., Pani, S. K., eds. (2024). *Internet of medical things and computational intelligence in healthcare 4.0*. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-5595-8

Table of contents

TYPE Editorial
PUBLISHED 21 February 2024
DOI 10.3389/fdata.2024.1368581

OPEN ACCESS

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

*CORRESPONDENCE
Sujata Dash
✉ sujata238dash@gmail.com

RECEIVED 10 January 2024
ACCEPTED 30 January 2024
PUBLISHED 21 February 2024

CITATION
Dash S, Pani SK and Santos WPd (2024) Editorial: Internet of Medical Things and computational intelligence in healthcare 4.0. *Front. Big Data* 7:1368581. doi: 10.3389/fdata.2024.1368581

Editorial: Internet of Medical Things and computational intelligence in healthcare 4.0

Sujata Dash[1]*, Subhendu Kumar Pani[2] and Wellington Pinheiro dos Santos[3]

[1]Department of Information Technology, School of Engineering and Technology, Nagaland University, Dimapur, India, [2]Department of Computer Science and Engineering, Krupajal Engineering College, Bhubaneswar, India, [3]Department of Biomedical Engineering, Federal University of Pernambuco, Recife, Brazil

KEYWORDS

computational intelligence, Internet of Medical Things (IoMT), healthcare 4.0 applications, biomedical informatics, machine learning, deep learning

Editorial on the Research Topic
Internet of Medical Things and computational intelligence in healthcare 4.0

We are delighted to present this special editorial, exploring the dynamic Research Topic of "*Internet of Medical Things and computational intelligence in healthcare 4.0.*" As we navigate the intricate realms of healthcare's digital transformation, the synergy between the Internet of Things (IoT) and computational intelligence stands as a beacon of innovation, promising a paradigm shift in the way we perceive, deliver, and experience healthcare. Healthcare 4.0 encapsulates a vision where interconnected devices, advanced analytics, and artificial intelligence converge to create a holistic, patient-centric ecosystem. At the heart of this transformative journey lies the intersection of the Internet of Medical Things (IoMT) and Computational Intelligence, propelling healthcare into an era marked by unprecedented efficiency, personalized care, and empowered patients.

The Internet of Medical Things (IoMT) has seen substantial growth, with an increasing number of connected healthcare devices. According to a report by Grand View Research, the global IoMT market size was valued at over USD 44 billion in 2020 and is expected to exhibit a compound annual growth rate (CAGR) of around 19.2% from 2021 to 2028. This underscores the rapid adoption of IoMT technologies in healthcare systems worldwide. Additionally, a study by Allied Market Research highlighted that the computational intelligence market in healthcare is also on the rise.

There was a large diversity of submissions covering different aspects, from IoMT, Computational Intelligence, Patient-Centric Care in the Digital Age, Ethical Considerations and Security Challenges, and Future Horizons and Collaborative Innovation. The articles and insights shared herein reflect the dedication and innovation of researchers who are at the forefront of this transformative journey. Moreno et al. employ a pattern-based classification method on the African-American Study of Chronic Kidney Disease with Hypertension dataset, revealing 15 distinct clinical features and SNP patterns. Notably, four clinical features and two SNPs show high predictive accuracy for CKD progression. These findings promise to inform future research and advance therapeutic interventions for individuals with chronic kidney disease.

Sharafutdinov et al. propose a new method to assess the generalization of ML models across hospitals. The study shows how the method works using patient data from different hospitals. The results highlight the importance of evaluating model transferability and creating diverse datasets.

Liu et al. analyzed SEER database data from 2004 to 2015 to explore prognostic factors in pancreatic cancer metastasis to the liver across different age groups. They found gender-specific primary sites and age-dependent prognostic factors such as tumor grade, histology, treatment, AJCC N stage, and race. The study underscores the importance of age-specific treatment strategies for pancreatic cancer metastasis to the liver.

Toubiana et al. published an editorial on the use of blockchain for Electronic Vaccine Certificates (EVCs) for COVID-19 vaccination. Blockchain may not be the best solution for EVCs, and the authors suggest exploring alternative cryptographic methods that involve centralized authorities for practical use.

Merhbene et al. used NLP on Reddit data to detect burnout. Their ensemble classifier achieved 0.93 balanced accuracy, outperforming single classifiers. NLP is a highly effective tool for identifying burnout indicators and improves standard classifiers.

Gao et al. reviewed the oral microbiome's relationship with systemic autoimmune diseases (SADs) like SLE, RA, and SS. The review highlights the importance of multiomics data and emphasizes the need for standardized methodologies to improve the understanding of SADs' etiology and potential therapies.

Rath et al. used imbalanced ECG samples to train ML models for detecting HD. AdaBoost and LR outperformed other classifiers. The ensemble model achieved the best HD detection performance. The methodology is versatile and applicable to various disease detection scenarios.

Mishra et al. developed a novel technique for detecting COVID-19 using phoneme analysis and audio signal smearing. They proposed a classification system based on phoneme grouping and achieved 97.22% accuracy for specific phoneme grouping using machine learning classifiers. This technique shows promise for quick and effective early-stage disease detection, with potential for application in other speech-related diseases.

Chicco and Jurman stress the importance of validating supervised machine learning results in biomedical informatics. The challenge is to achieve reliable results in the face of over-optimistic findings. Past guidelines have been too complex, especially for beginners. In response, Walsh et al. (2021) proposed ABC tips to simplify validation. The tips are meant to provide an effective tool for practitioners of all levels to enhance the reliability of scientific results in biomedical sciences.

The market for IoT healthcare is projected to grow at a CAGR of 21.2% from 2024 to 2030, with a value of USD 44.21 billion expected in 2023. This growth is driven by several factors, including the use of wearables and smartphones for patient monitoring, an increased adoption of remote patient monitoring during the COVID-19 pandemic, and investments in digital health infrastructure. The prevalence of chronic conditions and investments in digital healthcare technologies have also led to a significant rise in telemedicine adoption. IoT technology is expanding in healthcare due to several reasons, including advancements in smartphone technology, improved data security measures, and the growing accessibility of wearable sensors and connected health monitors. This demand is further fueled by emerging economies such as India, China, Indonesia, Bangladesh, and some African and Latin American countries, which are contributing to this growth through improvements in network infrastructure and growing network coverage. Physicians are also increasingly using mobile devices, creating further demand for IoT solutions in healthcare.

Author contributions

SD: Writing – original draft, Writing – review & editing. SP: Writing – review & editing. WS: Writing – review & editing.

Funding

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

Acknowledgments

We would like to express our gratitude to all of the authors and editors of Frontiers for their hard work and dedication in making this Research Topic a success.

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

Walsh, I., Fishman, D., Garcia-Gasulla, D., Titma, T., Pollastri, G., Capriotti, E., et al. (2021). DOME: recommendations for supervised machine learning validation in biology. *Nat. Methods* 5, 1122–1127. doi: 10.1038/s41592-021-01205-4

frontiers in Big Data

ORIGINAL RESEARCH
published: 14 January 2021
doi: 10.3389/fdata.2020.528828

Identifying Clinical and Genomic Features Associated With Chronic Kidney Disease

M. Megan Moreno[1], Travaughn C. Bain[1], Melissa S. Moreno[1], Katherine C. Carroll[1,2,3], Emily R. Cunningham[1,2,4], Zoe Ashton[1], Roby Poteau[1], Ersoy Subasi[5], Michael Lipkowitz[6] and Munevver Mine Subasi[1]*

[1]Department of Mathematical Sciences, Florida Institute of Technology, Melbourne, FL, United States, [2]Department of Biomedical and Chemical Engineering and Sciences, Melbourne, FL, United States, [3]Department of Biology, University of Florida, Gainesville, FL, United States, [4]Department of Mathematics, SUNY Potsdam, Potsdam, NY, United States, [5]Department of Computer Engineering and Sciences, Florida Institute of Technology, Melbourne, FL, United States, [6]Department of Medicine, Georgetown University Medical Center, Washington, DC, United States

We apply a pattern-based classification method to identify clinical and genomic features associated with the progression of Chronic Kidney disease (CKD). We analyze the African-American Study of Chronic Kidney disease with Hypertension dataset and construct a decision-tree classification model, consisting 15 combinatorial patterns of clinical features and single nucleotide polymorphisms (SNPs), seven of which are associated with slow progression and eight with rapid progression of renal disease among African-American Study of Chronic Kidney patients. We identify four clinical features and two SNPs that can accurately predict CKD progression. Clinical and genomic features identified in our experiments may be used in a future study to develop new therapeutic interventions for CKD patients.

Keywords: classification, genomic analysis, AASK, chronic kidney disease, decision trees

OPEN ACCESS

Edited by:
Tuan D. Pham,
Prince Mohammad bin Fahd University, Saudi Arabia

Reviewed by:
Dinh Tuan Phan Le,
New York City Health and Hospitals Corporation, United States
Lin Liu,
Tsinghua University, China

****Correspondence:***
Munevver Mine Subasi
msubasi@fit.edu

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

Received: *22 January 2020*
Accepted: *30 October 2020*
Published: *14 January 2021*

Citation:
Moreno MM, Bain TC, Moreno MS, Carroll KC, Cunningham ER, Ashton Z, Poteau R, Subasi E, Lipkowitz M and Subasi MM (2021) Identifying Clinical and Genomic Features Associated With Chronic Kidney Disease.
Front. Big Data 3:528828.
doi: 10.3389/fdata.2020.528828

1 INTRODUCTION

The main function of kidney is to remove excess water and waste products from blood. It also helps to regulate the levels of minerals such as sodium, calcium, and potassium in blood. One suffers from chronic kidney disease (CKD), also known as renal disease, when kidney losses its function gradually and usually permanently. CKD, defined by reduced glomerular filtration rate (GFR), proteinuria, or structural kidney disease, is a worldwide growing public health problem[1]. Many subjects with renal disease of most etiologies progress to severe renal failure and/or end stage renal disease (ESRD), requiring renal replacement therapy, which may involve a form of dialysis or renal transplantation (Lewis et al., 1993; Klahr et al., 1994; DCCT, 1995; Brenner et al., 2001; Lewis et al., 2001; Wright et al., 2002; Niki et al., 2015). However, progression rate of CKD is very heterogeneous (Lindeman et al., 1985; Lindeman, 1990; Hallan et al., 2006). While a few predictive factors for progression such as proteinuria have been detected, identification of those at risk to progress remains a significant problem. It has also been established that there are several therapies that can ameliorate the progression of renal disease including ACE inhibitors, blood pressure control, tight diabetes control and perhaps low protein diets; however, in trials examining these therapeutic modalities there

[1] Chronic Kidney disease Surveillance Project, Center for disease Control and Prevention—http://nccd.cdc.gov/ckd/

remains a very significant risk of progression of renal disease in the subjects receiving optimal therapy (Lewis et al., 1993; Klahr et al., 1994; DCCT, 1995; Brenner et al., 2001; Lewis et al., 2001; Wright et al., 2002; Niki et al., 2015).

African-American Study of Chronic Kidney disease with Hypertension (AASK) was motivated by the high rate of hypertension-related chronic kidney disease in the African-American population and the scarcity of effective therapies. The study involved 21-center randomized double-blinded treatment trial of 1,094 African-American patients with hypertension at ages ranging from 18 to 70 years. Patients had renal failure with GFR between 20 and 65 ml/min/1.73m^2. Patients were randomized to the angiotensinogen converting enzyme inhibitor (ACEi) ramipril, the β-blocker (BB) metoprolol or the dihydropyridine calcium channel blocker (CCB) amlodipine, and to usual (mean arterial pressure (MAP 102–107) or low (MAP <92) blood pressure (BP) goals. The rationale for the treatment arms was that there was human and animal data suggesting that ACEi and CCB might slow progression of renal disease independent of their BP effects (Lewis et al., 1993; Hallan, 1998), and there was data from observational and treatment studies that a lower BP might have beneficial effects (Klahr et al., 1994; Klag et al., 1997). Although other studies had attempted to achieve a 10 $mmHg$ MAP separation (Hansson et al., 1998; Lewis et al., 2001), AASK is the first major trial to actually achieve this goal. The primary outcome was rate of decline of GFR (GFR slope) based on iothalamate GFR studies at 6 months intervals, with a secondary clinical composite outcome of end stage renal disease (ESRD), a 25 ml/min or 50% drop in GFR from baseline (GFR event), or death (Subasi et al., 2017).

The initial AASK results were not conclusive (Wright et al., 2002). While the adopted therapy was shown to slow the progression of renal disease, there was still high rate of progression to renal failure. The CCB arm of the study was stopped early when interim analysis indicated that CCB was inferior to both BB and ACEi in patients with >0.22 urine protein/creatinine ratio (about 300 mg proteinuria/24 h) (Agodoa et al., 2001). The low BP goal of the study did not improve outcomes: there was no beneficial effect of low MAP on rate of progression of renal disease as defined by GFR slope or clinical composite outcomes (GFR events, end stage renal disease (ESRD) or death). Subsequently, a similar result was found in the REIN trial (Ruggenenti et al., 1999). Studies in Type 2 diabetes have demonstrated a linear relation of achieved BP to renal outcomes (Bakris et al., 2003; Pohl et al., 2005); however, it should be noted that all the patients in these studies were treated to the same goal BP, so that rather than low BP being protective, the ability to achieve lower BPs may have defined a sub-population in these studies with low risks of disease progression. Despite the lack of effect on renal outcomes in AASK, proteinuria was diminished by the lower BP goal. This finding is similar to that previously reported for diabetics (Lewis et al., 2001). Finally, a subgroup analysis in AASK did suggest that patients on a non-protective regimen (CCB) may have benefited from the low BP goal (Contreras et al., 2005). Most importantly in AASK, ACEi decreased the number of events as compared to both BB and CCB (Wright et al., 2002). These data for ACEi vs. CCB are tabulated in **Table 1** (risk reduction adjusted for baseline covariates) and were most dramatic for the hard outcomes, especially ESRD.

TABLE 1 | Analysis of clinical composite outcomes - 95% confidence interval (CI).

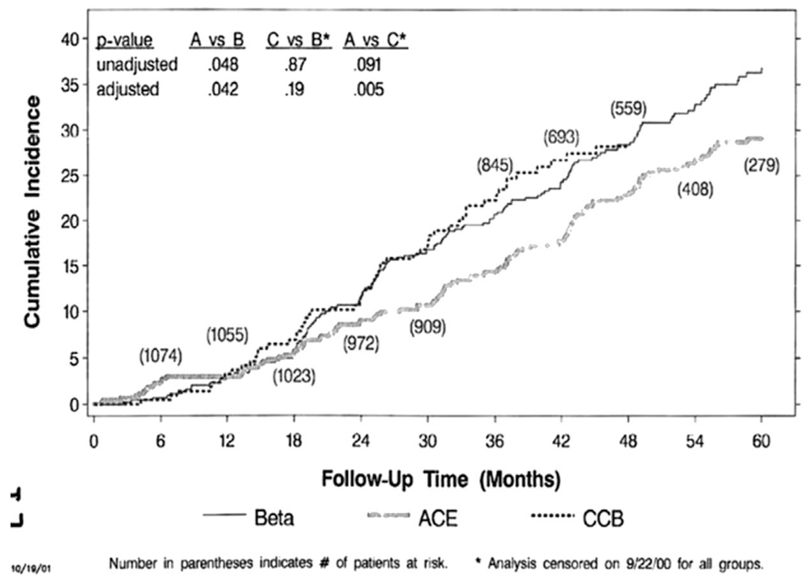
ramipril vs. Amlodipine	% Risk Reduction	95% CI	*p*-value
GFR event, ESRD or death	38%	14% – 56%	0.004
GFR event or ESRD	40%	14% – 59%	0.006
ESRD or death	49%	26% – 65%	< 0.001
ESRD alone	59%	36% – 74%	< 0.001

Several possible interventions such as blood pressure control (Wright et al., 2002), diabetes treatment (DCCT, 1995), controlling dietary protein intake (Klahr et al., 1994) and medications with possible renoprotective effects (Ruggenenti et al., 1999; Agodoa et al., 2001; Wright et al., 2002) have been tested in clinical trials. In all cases, the residual rate of progression of chronic kidney disease has remained significant. To date, there are few prediction models to identify which patients are likely to progress significantly. Subasi et al. (2017) (Subasi et al., 2017) identified serum proteomic patterns that can accurately distinguish rapid progression and slow progression among AASK patients. Recently, Lipkowitz et al. (2013) (Parsa et al., 2013) examined effects of variants in gene encoding apolipoprotein L1 (APOL1) on the disease progression and observed that renal risk variants in APOL1 were associated with the higher rates of ESRD and progression of chronic kidney disease in African-American patients as compared to white patients. Other recent studies include Rahman et al. (2013), where the effects of two antihypertensive drug dose (PM dose and add-on dose) schedules on nocturnal blood pressure vs. usual therapy (AM dose) in former participants were determined and Chen et al. (2016), where the longitudinal changes in hematocrit in hypertensive renal disease were studied.

The goal of our current study is to apply a pattern-based classification method to identify clinical and genomic features that may serve as prognostic markers for the progression of renal disease among AASK patients. Clinical and genomic features identified in our analysis shall be used in a future study to obtain comparison of the disease progression in white patients and African-American patients, both of those with and those without apolipoprotein L1 (APOL1) high-risk variants. The ultimate goal of our AASK data analysis, started in (Subasi et al., 2017) and continued in this current work, is to identify new targets and provide basis for new therapeutic interventions for chronic kidney disease.

2 STUDY SUBJECTS

Closer inspection of the data highlights the current dilemma: although there is a $30-60\%$ decrease in the number of events with ACEi still a residual event rate of $>6\%/yr$ in the trial as a whole and $>11\%/yr$ in subjects with urine protein/creatinine >0.22, a mild degree of proteinuria of $200-300mg/day$ (**Figures 1** and **2**). In addition it can be seen that the event rate is essentially constant throughout the 5 years of the trial, indicating that remaining patients are still at risk to progress. This finding is



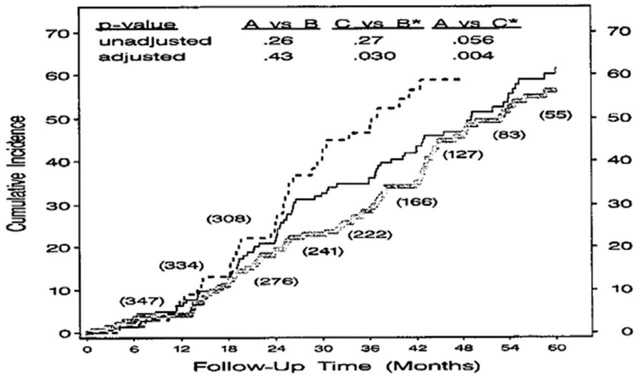
Number in parentheses indicates # of patients at risk. * Analysis censored on 9/22/00 for all groups.

FIGURE 1 | AASK clinical composite events–all patients.

similar to that of other trials such as MDRD (Klahr et al., 1994; Hebert et al., 1997), the Collaborative Study Group Trial (Lewis et al., 1993), RENAAL (Brenner et al., 2001) and IDNT (Lewis et al., 2001).

Figure 3 indicates the significant heterogeneity of progression rate of renal disease in the AASK Trial, where the rate of decline of GFR after 6 months in the trial (chronic GFR slope) is depicted in blue for each patient from most rapid decline (negative slope) on the left, to the least rapid decline (positive slope) on the right. The expected rate of decline of GFR with aging is generally assumed to be $-1ml/min/yr$ (Berg, 2006; Murussi et al., 2006), although longitudinal studies have raised questions about this assumption (Lindeman et al., 1985; Lindeman, 1990). Based on this estimate, approximately 30% of the AASK patients in **Figure 3** did not progress (right side, slope $> -1ml/min/yr$) while approximately 30% progressed rapidly (left side, slope $< -3ml/min/yr$). The figure also shows that proteinuria, the strongest predictor of progression rate reported in literature, is not an ideal predictor in that there are a number of slow progressors with significant proteinuria (red spikes, right), while a significant number of rapid progressors had no or minimal proteinuria (absence of red bars, left) (Subasi et al., 2017). This data is supported by the observation in genetics studies that proteinuria and progression of renal disease may be disparate phenotypes (Fogarty et al., 2000; Krolewski et al., 2006).

Pts with Baseline UP/Cr > 0.22



Number in parentheses indicates # of patients at risk.
* Analysis censored on 9/22/00 for all groups.

FIGURE 2 | AASK clinical composite events–proteinuria.

2.1 Pre-processing of AASK Data to Predict Progression of Renal Disease

An avenue that has not been carefully explored is a data mining approach to detect the combinations of clinical features and/or single nucleotide polymorphisms (SNPs) that better determine the population at risk for progression of CKD. The goal of this section is to identify combinatorial patterns of clinical features and SNPs that can accurately predict progression of the renal disease among AASK patients. In order to achieve this, we perform a study on a selected subset of subjects from the AASK Clinical Trial based on the glomerular filtration slope (GFR) of all AASK patients presented in **Figure 3**. The original AASK data contains 1,094 African-American patients with 88 clinical features and 130 SNPs. Before we start our analysis, we remove features with more than 80% missing values in the dataset. We then remove AASK patients with missing GFR values and more than 10% missing values. This results in 800 AASK patients with 77 clinical features and 113 SNPs. In order to

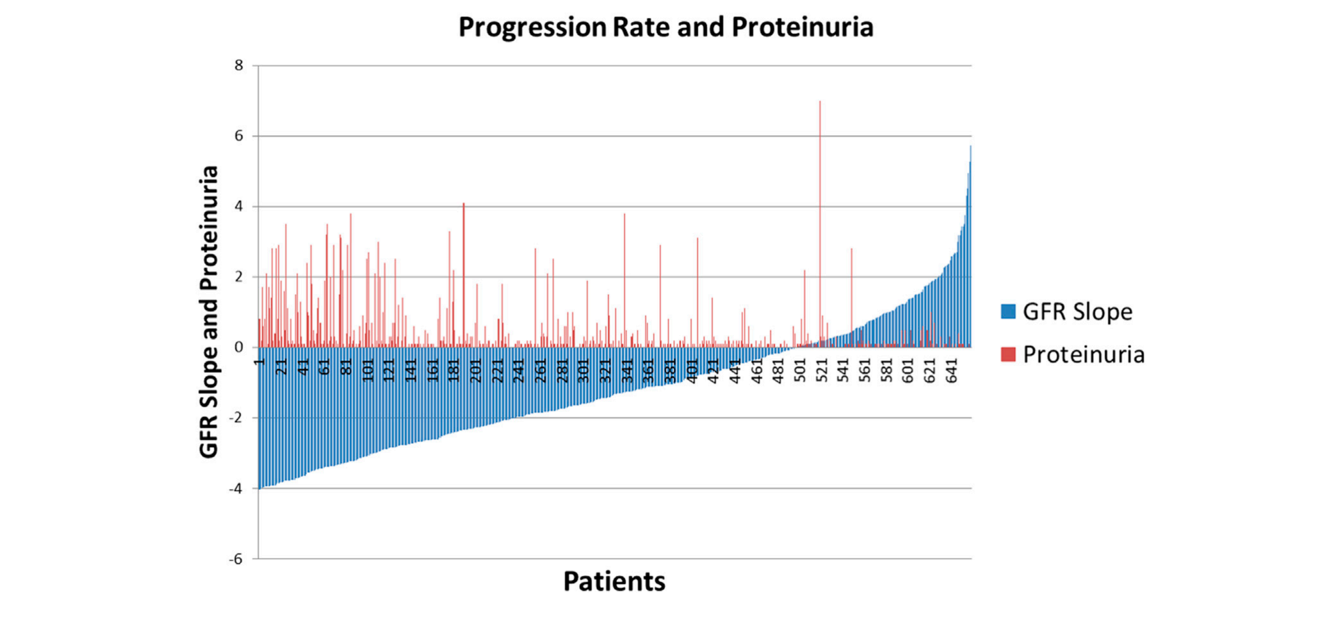


FIGURE 3 | AASK Patients stratified by GFR slope with degree of proteinuria superimposed.

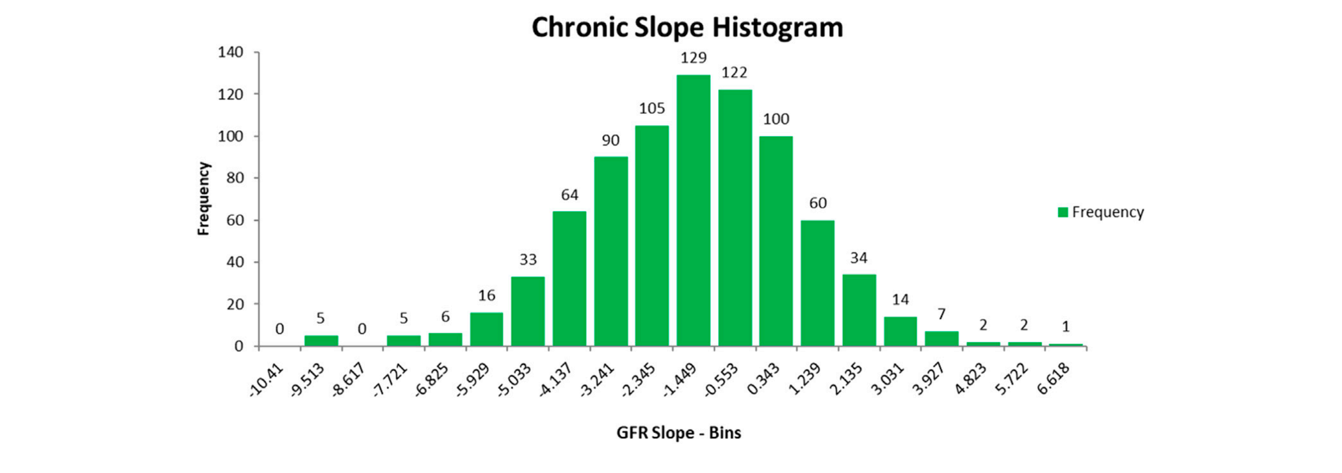


FIGURE 4 | Chronic GFR slope of AASK patients in the reduced data.

develop a classification model that can predict the rate of decline of kidney function, we identify two "extreme" groups of patients whose disease progression is "slow" (GFR chronic slope $> 1 ml/min/yr$) or "rapid" (GFR chronic slope $< -4 ml/min/yr$). The two subsets of patients, referred to as slow progressors and rapid progressors are selected from the AASK study based on the chronic GFR slope histogram presented in **Figure 4**. The resulting reduced dataset contains 138 AASK patients identified as rapid progressors and 75 AASK patients as slow progressors.

Figure 5 shows the PCA plot of the AASK patients in the reduced dataset. **Table 2** describes the patient population for this study. As can be seen from the table, proteinuria is very different between the two groups of disease progression, which supports the previous studies showing that proteinuria is the strongest predictor of GFR slope progression in AASK (Wang et al., 2006).

2.2 Identification of Significant Clinical and Genomic Features

The resulting AASK dataset consisting of 138 rapid progressors, 75 slow progressors, 77 clinical features, and 113 SNPs, is further investigated to remove any features irrelevant for the recognition of a rapid progressor as opposed to a slow progressor. In order to obtain a classification model effectively and efficiently, we first apply a correlation-based feature selection procedure (Hall and Smith, 1998) to retain only those relevant features successfully distinguishing between rapid progressors and slow progressors in

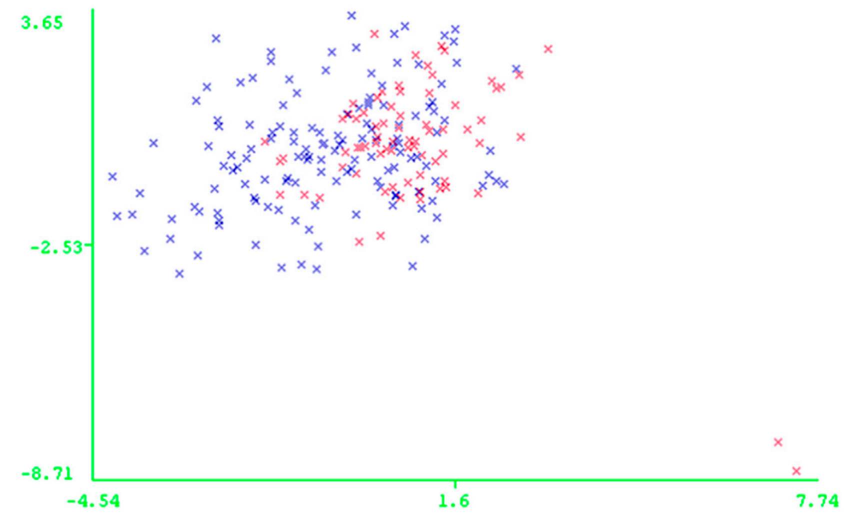


FIGURE 5 | PCA plot of AASK patients in the reduced data: * Rapid Progressors and * Slow Progressors

AASK data. Correlation-based feature selection method evaluates the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them. Subsets of features that are highly correlated with the outcome (rapid/slow progression) while having low intercorrelation are preferred. AASK data is randomly partitioned into ten approximately equal parts; one of these subsets is designated as “test set”, correlation based feature selection is built on the remaining nine subsets which form the “training dataset”, and then evaluated on the cases in the test set. This procedure is repeated ten times, always taking another one of the ten parts in the role of the test set (re-randomizing the patients into ten new subsets and repeating the procedure nine additional times for a total of 100 tests).

Table 3 shows the features selected from ten times 10-folding cross-validation of the correlation-based feature subset selection procedure in WEKA, a commonly used open source data mining software (Hall et al., 2009). The rationale for using small numbers of features is both for ease in collecting the relevant data for prediction on patients from different sources (health systems) and the possibility that finding a small number of novel predictors may help inform studies into the mechanisms and treatment of CKD progression if they suggest new and unexplored pathways. The SNPs and the fact that the alpha-2 agonist antihypertensive medicine use are predictors may help in this manner.

3 PATTERN-BASED CLASSIFICATION MODEL TO PREDICT PROGRESSION OF RENAL DISEASE

3.1 Identification of Combinatorial Patterns of Significant Clinical Features and SNPs

Study Subjects analysis provides us with a reduced AASK data, containing 138 rapid progressors and 75 slow progressor with.

TABLE 2 | Baseline characteristics of study population.

Basic Clinical Features	Rapid Progressors	Slow Progressors
Chronic slope	−5.41 ± 1.36	2.11 ± 1.03
GFR	42.83 ± 13.25	52.30 ± 10.55
Proteinuria	1.12 ± 1.40	0.13 ± 0.20
Age	50.22 ± 11.94	52.52 ± 9.52
Weight (kg)	96.42 ± 22.42	87.52 ± 19.65
(cm)	171.69 ± 10.56	169.21 ± 10.80
BMI	32.69 ± 7.06	30.57 ± 6.09

TABLE 3 | Feature Selection - 10 fold stratified cross validation.

% Absolute Frequency	Feature
90%	α-agonist
100%	Proteinuria
100%	U.Protein/U.Creatinine
70%	GFR value at G1 visit
100%	CHGB-1
90%	PLCG2 rs4399527

- four clinical features: *α-agonist (peripherol base), proteinuria, urine-protein/urine-creatinine, GFR value at G1 visit*, where α-agonist represents the use of peripheral alpha-2 agonist blood pressure medication
- two SNPs: CHGB-1, PLCG2 rs4399527.

These six features were validated using 10 × 10-folding cross-validation experiments on seven commonly used and well-known classification methods, including Random Forest, Decision Trees, Nearest Neighbor, Support Vector Machines, Neural Networks, Logistic Regression, and Naïve Bayes (Hall et al., 2009). In this step the AASK data is randomly partitioned into ten approximately equal parts; one of these subsets is designated as “test set”, a model is built on the remaining nine subsets which

TABLE 4 | Cross-validation of classification methods for AASK samples.

Classification Method	Accuracy	Sensitivity	Specificity	Precision	Recall	F-Measure	ROC Area
Random forest	78.33%	83.63%	68.79%	0.71	0.69	0.68	0.86
C4.5 decision tree	76.77%	80.53%	70.18%	0.68	0.70	0.67	0.78
Nearest neighbor	70.21%	76.97%	58.02%	0.59	0.58	0.57	0.68
Support vector machines	72.70%	77.91%	63.34%	0.62	0.63	0.61	0.71
Neural networks	73.07%	78.19%	63.79%	0.63	0.64	0.62	0.81
Logistic regression	75.88%	81.70%	65.39%	0.68	0.65	0.65	0.85
Naïve bayes	70.20%	57.90%	93.02%	0.56	0.93	0.69	0.85

TABLE 5 | C4.5 classification model for AASK samples.

Patterns	C4.5 Classification Model for Renal disease Progression
S1	U. Protein ≤ 0 and PLCG2 rs4399527=*GC* and CHGB 1=*TT*
S2	U. Protein ≤ 0 and PLCG2 rs4399527=*GC* and CHGB 1=*CT* and α-agonist ≤ 0 and Pro./Creat.Ratio > 0.01706
S3	U. Protein ≤ 0 and PLCG2 rs4399527=*GC* and CHGB 1=*CC*
S4	U. Protein ≤ 0.5 and PLCG2 rs4399527=*CC* and Pro./Creat.Ratio ≤ 0.15714
S5	U. Protein ≤ 0.5 and PLCG2 rs4399527=*GG* and CHGB 1=*TT* and 41.4 < GFR G1 ≤ 59.5816
S6	U. Protein ≤ 0.5 and PLCG2 rs4399527=*GG* and CHGB 1=*CT* and Pro./Creat.Ratio > 0.02177
S7	U. Protein ≤ 0.5 and PLCG2 rs4399527=*GG* and CHGB 1=*CC*
R1	U. Protein ≤ 0 and PLCG2 rs4399527=*GC* and CHGB 1=*CT* and α-agonist ≤0 and Pro./Creat.Ratio ≤ 0.01706
R2	U. Protein ≤ 0 and PLCG2 rs4399527=*GC* and CHGB 1=*CT* and α-agonist >0
R3	0 < U. Protein ≤ 0.5 and PLCG2 rs4399527=*GC*
R4	U. Protein ≤ 0.5 and PLCG2 rs4399527=*CC* and Pro./Creat.Ratio > 0.15714
R5	U. Protein ≤ 0.5 and PLCG2 rs4399527=*GG* and CHGB 1=*TT* and GFR G1 ≤ 41.4
R6	U. Protein ≤ 0.5 and PLCG2 rs4399527=*GG* and CHGB 1=*TT* and GFR G1 > 59.5816
R7	U. Protein ≤ 0.5 and PLCG2 rs4399527=*GG* and CHGB 1=*CT* and Pro./Creat.Ratio ≤ 0.02177
R8	U. Protein > 0.5

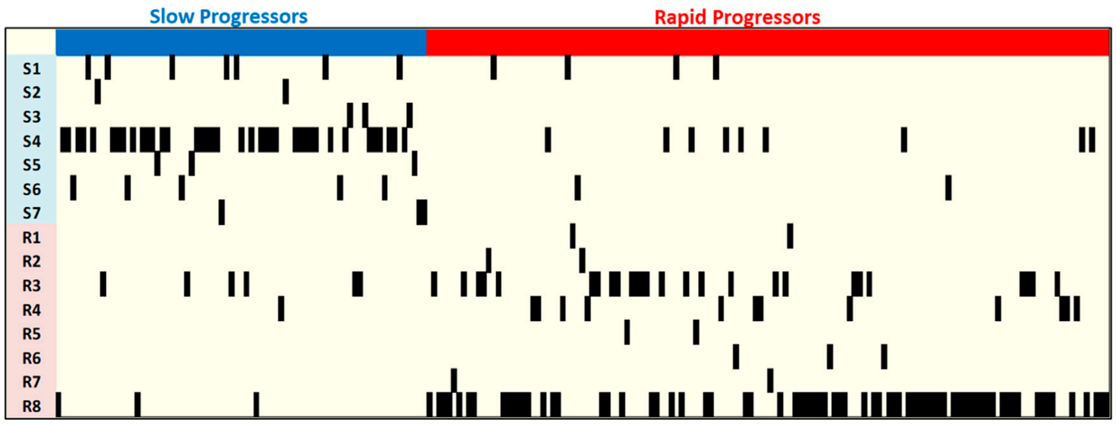


FIGURE 6 | C4.5 decision tree for AASK samples.

form the "training dataset", and then tested by predicting the classes of patients in the test set using a classification method. This procedure is repeated 10 times, always taking another one of the ten parts in the role of the test set (re-randomizing the patients into 10 new subsets and repeat the procedure nine additional times) for a total of 100 tests for each of the seven classification methods. **Table 4** shows average accuracy, sensitivity (proportion of correctly classified rapid progressors), specificity (proportion of correctly classified slow progressors) as well as average precision, recall, F-measure, and area under Receiver Operating Characteristic (ROC) curve.

As can be seen in **Table 4**, while Random Forest provides us with highest accuracy, C4.5 Decision Tree (Quinlan, 1993), a non-parametric supervised learning method used for classification and regression, provides the best sensitivity and specificity, i.e., the best prediction for rapid and slow prediction. C4.5 classification model consisting of seven patterns, S1-S7, for slow progressors and eight patterns, R1-R8, for rapid progressors

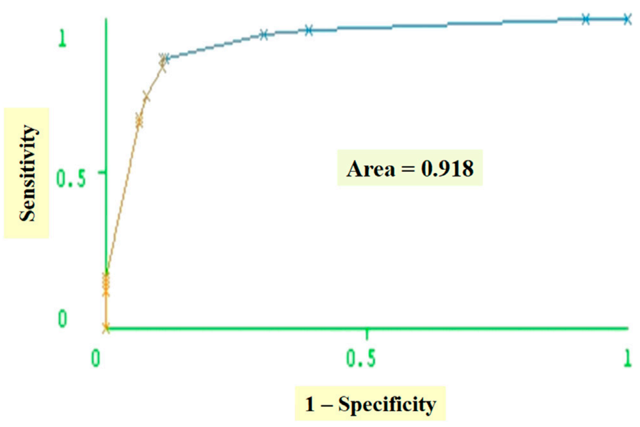


FIGURE 7 | Heatmap of the C4.5 patterns for AASK samples.

is presented in **Table 5** as combinatorial patterns of clinical features and SNPs associated with slow and rapid progression in the AASK dataset. **Figures 6** and **8** show the C4.5 decision tree and heatmap corresponding to the combinatorial patterns presented in **Table 5**, respectively.

The pattern characteristics including

- *rapid prevalence*: proportion of rapid progressors covered by a pattern to the total number of rapid progressors,
- *slow prevalence*: proportion of slow progressors covered by a pattern to the total number of slow progressors,
- *rapid homogeneity*: proportion of rapid progressors covered by the pattern,
- *slow homogeneity*: proportion of slow progressors covered by the pattern,
- *degree*: number of conditions appear in the description of the pattern of the C4.5 classification model are given in **Table 6**.

3.2 Validation of Combinatorial Patterns

We remark that the C4.5 classification model given in **Table 5** consists of explicit patterns, where the four clinical features and two SNPs selected in **Identification of Significant Clinical and Genomic Features** are assigned threshold values. Note that patterns S1-S7 exhibit high homogeneity for the slow progressors and R1-R8 exhibit high homogeneity for the rapid progressors in AASK data. For example, patterns S2, S3, S5, S7 have 100% homogeneity, meaning that all patients covered by each of these patterns are slow progressors. Similarly, the homogeneity of patterns R1, R2, R5, R6, R7 is also 100%, i.e., all patients covered by each of these patterns are rapid progressors. We refer to such patterns as pure patterns associated with the respective subgroups of AASK patients. We also remark that the classification model contains fuzzy patterns, S1, S4, S6, R3, R4, R8, i.e., patterns with homogeneity < 100%. For example, the homogeneity of pattern S4 is 81%, meaning that 81% of the patients covered by pattern S4 are slow progressors and the remaining 19% of the patients covered by this pattern are rapid progressors in AASK Clinical Trial.

As for the prevalence, patterns S4 and R8 are significant patterns, S4 covering 51% of all slow progressors, but only 12% of the rapid progressors and R8 covering 54% of all rapid progressors, but only 2% of the slow progressors in the data. While the other patterns in the classification model does not exhibit high prevalence in the associated subgroups within the data, they are still required to predict the progression of all AASK patients in the study. Finally, we observe that these patterns use

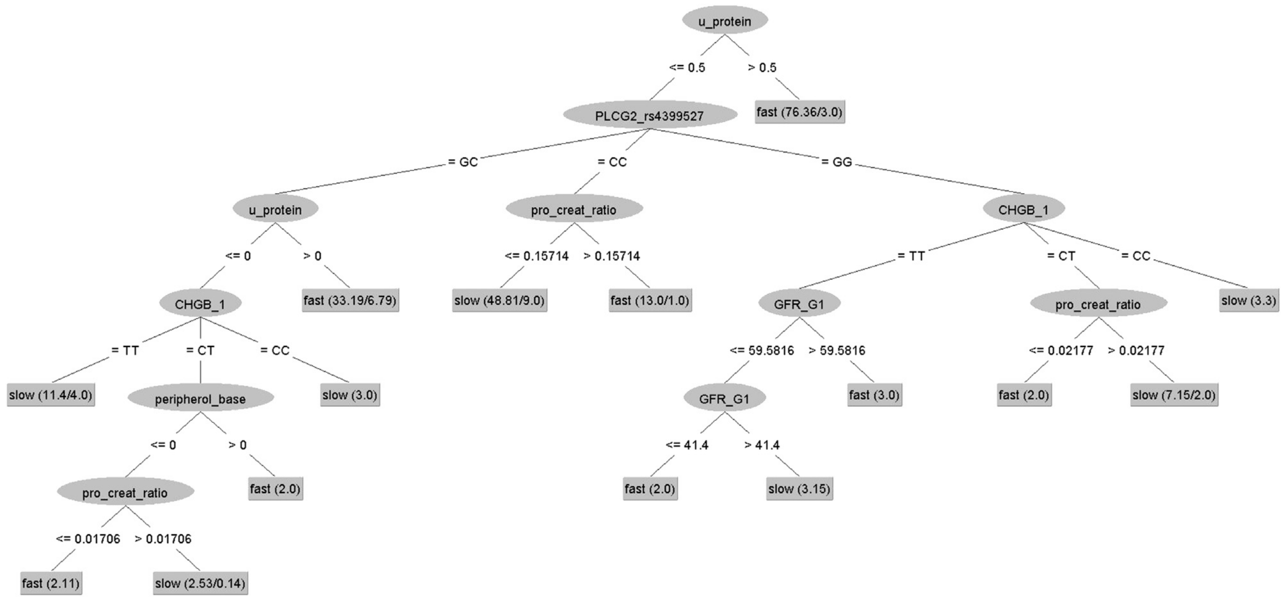


FIGURE 8 | Receiver operating curves (ROC).

TABLE 6 | C4.5 decision tree pattern characteristics.

Pattern	Homogeneity (%)	Slow prevalence	Rapid prevalence	Degree
S1	63.64%	9.33%	5.33%	3
S2	100%	2.67%	0%	5
S3	100%	4%	0%	3
S4	80.85%	50.67%	12%	3
S5	100%	4%	0%	4
S6	71.43%	6.67%	2.67%	4
S7	100%	4%	0%	3
R1	100%	1.45%	0%	5
R2	100	1.45%	0%	4
R3	81.25%	18.84%	4.35%	2
R4	92.31%	8.70%	0.72%	3
R5	100%	1.45%	0%	4
R6	100%	2.17%	0%	4
R7	100%	1.45%	0%	4
R8	96.10%	53.62%	2.17%	1

small number of features of AASK patients. The degrees of the patterns (number of features used in pattern description) range from one to 5. Note that according to pattern R8, the U. Protein levels of 54% of rapid progressors exceeds 0.5 and 96% of the patients covered by this pattern are rapid progressors. Similar observations can be done for other patterns forming the classification model in **Table 5**.

Based on the 10 × 10-folding cross-validation experiments, the classification model correctly classifies 80.53% of rapid progressors and 70.18% of slow progressors and exhibits an average accuracy of 76.77% with 0.68 precisiom, 0.70 recall, and 0.67 F-measure, validating the distinguishing power of the classification model for the AASK patients in our study. As another measure of the effectiveness of the classification model at predicting rapid or slow progressors, we generate receiver operating characteristic (ROC) curve that shows how much the classification model is capable of distinguishing between the rapid progressors and slow progressors in AASK Clinical Trial. ROC curve is obtained by plotting *sensitivity* (true positive rate) against 1 − *specificity* (false positive rate). Based on 10 × 10-folding cross-validation experiments, the area under the ROC curve is 0.78. ROC curve corresponding to the C4.5 classification model (built on entire dataset) in **Table 5** is shown in **Figure 8**.

Thus, we can conclude that the combinatorial patterns forming the classification model in **Table 5** are high quality decision rules that can be easily interpreted by medical experts, allowing them to target the clinical features and SNPs associated with the progression of the renal disease to develop new therapies.

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found in the African American Study of Kidney Disease and Hypertension Study (Clinical Trial) (AASK Trial) https://repository.niddk.nih.gov/studies/aask-trial/.

AUTHOR CONTRIBUTIONS

ES, ML, and MMS are senior co-authors who designed and supervised the entire project and participated in writing the manuscript. MMM, TB, and MSM participated in the study design and performed the combinatorial analysis and participated in writing the manuscript. KC, EC, ZA, and RP were involved in various steps of the combinatorial analysis.

FUNDING

ML, ES, and MMS's work was supported by National Institutes of Health—Grant number: 5R21DK67468. KC and EC's work was supported by National Science Foundation (NSF) Research Experience for Undergraduates (REU) Grant—Award number: 1,359,341.

ACKNOWLEDGMENTS

Special thanks to the AASK Investigators.

REFERENCES

Agodoa, L. Y., Appel, L., Bakris, G. L., Beck, G., Bourgoignie, J., Briggs, J. P., et al. (2001). Effect of ramipril vs amlodipine on renal outcomes in hypertensive nephrosclerosis: a randomized controlled trial. *Jama* 285, 2719–2728. doi:10.1001/jama.285.21.2719

Bakris, G. L., Weir, M. R., Shanifar, S., Zhang, Z., Douglas, J., van Dijk, D. J., et al. (2003). Effects of blood pressure level on progression of diabetic nephropathy: results from the RENAAL studyEffects of intensive blood-pressure lowering and low-dose aspirin in patients with hypertension: principal results of the Hypertension Optimal Treatment (HOT) randomised trial. *Arch. Intern. Med.* 163, 1555–1565. doi:10.1001/archinte.163.13.1555

Berg, U. (2006). Differences in decline in gfr with age between males and females. reference data on clearances of inulin and pah in potential kidney donors. *Nephrol. Dial. Transplant.* 21, 2577–2582. doi:10.1093/ndt/gfl227

Brenner, B. M., Cooper, M. E., de Zeeuw, D., Keane, W. F., Mitch, W. E., Parving, H. H., et al. (2001). Effects of losartan on renal and cardiovascular outcomes in patients with type 2 diabetes and nephropathy. *N. Engl. J. Med.* 345, 861–869. doi:10.1056/NEJMoa011161

Chen, E., Miller, G. E., Yu, T., and Brody, G. H. (2016). The Great Recession and health risks in African American youth. *Brain Behav. Immun.* 53, 234–241. doi:10.1016/j.bbi.2015.12.015

Contreras, G., Greene, T., Agodoa, L. Y., Cheek, D., Junco, G., Dowie, D., et al. (2005). Blood pressure control, drug therapy, and kidney disease. *Hypertension*. 46, 44–50. doi:10.1161/01.HYP.0000166746.04472.60

DCCT (1995). Effect of intensive therapy on the development and progression of diabetic nephropathy in the diabetes control and complications trial. The Diabetes Control and Complications (DCCT) Research Group. *Kidney Int* 47, 1703–1720.

Fogarty, D. G., Hanna, L. S., Wantman, M., Warram, J. H., Krolewski, A. S., and Rich, S. S. (2000). Segregation analysis of urinary albumin excretion in families with type 2 diabetes. *Diabetes* 49, 1057–1063. doi:10.2337/diabetes.49.6.1057

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutremann, P., and Witten, I. (2009). The WEKA data mining software: an update. *SIGKDD Explorations* 11 (1), 10–18. doi:10.1145/1656274.1656278

Hall, M. A., and Smith, L. A. (1998). *Practical feature subset selection for machine learning*. Springer.

Hallan, M. (1998). Calcium antagonists and renal disease. *Kidney Int*. 54, 1771–1784. doi:10.1046/j.1523-1755.1998.00168.x

Hallan, S. I., Coresh, J., Astor, B. C., Asberg, A., Powe, N. R., Romundstad, S., et al. (2006). International comparison of the relationship of chronic kidney disease prevalence and esrd risk. *J. Am. Soc. Nephrol*. 17, 2275–2284. doi:10.1681/ASN.2005121273

Hansson, L., Zanchetti, A., Carruthers, S. G., Dahlöf, B., Elmfeldt, D., Julius, S., et al. (1998). Effects of intensive blood-pressure lowering and low-dose aspirin in patients with hypertension: principal results of the Hypertension Optimal Treatment (HOT) randomised trial. HOT Study Group. *Lancet* 351, 1755–1762. doi:10.1016/s0140-6736(98)04311-6

Hebert, L. A., Kusek, J. W., Greene, T., Agodoa, L. Y., Jones, C. A., Levey, A. S., et al. (1997). Effects of blood pressure control on progressive renal disease in blacks and whites. modification of diet in renal disease study group. *Hypertension* 30, 428–435. doi:10.1161/01.hyp.30.3.428

Klag, M. J., Whelton, P. K., Randall, B. L., Neaton, J. D., Brancati, F. L., and Stamler, J. (1997). End-stage renal disease in African-American and white men. 16-year MRFIT findings. *Jama* 277, 1293–1298.

Klahr, S., Levey, A. S., Beck, G. J., Caggiula, A. W., Hunsicker, L., Kusek, J. W., et al. (1994). The effects of dietary protein restriction and blood-pressure control on the progression of chronic renal disease. Modification of Diet in Renal Disease Study Group. *N. Engl. J. Med*. 330, 877–884. doi:10.1056/NEJM199403313301301

Krolewski, A. S., Poznik, G. D., Placha, G., Canani, L., Dunn, J., Walker, W., et al. (2006). A genome-wide linkage scan for genes controlling variation in urinary albumin excretion in type II diabetes. *Kidney Int*. 69, 129–136. doi:10.1038/sj.ki.5000023

Lewis, E. J., Hunsicker, L. G., Bain, R. P., and Rohde, R. D. (1993). The effect of angiotensin-converting-enzyme inhibition on diabetic nephropathy. The Collaborative Study Group. *N. Engl. J. Med*. 329, 1456–1462. doi:10.1056/NEJM199311113292004

Lewis, E. J., Hunsicker, L. G., Clarke, W. R., Berl, T., Pohl, M. A., Lewis, J. B., et al. (2001). Renoprotective effect of the angiotensin-receptor antagonist irbesartan in patients with nephropathy due to type 2 diabetes. *N. Engl. J. Med*. 345, 851–860. doi:10.1056/NEJMoa011303

Lindeman, R. D., Tobin, J., and Shock, N. W. (1985). Longitudinal studies on the rate of decline in renal function with age. *J. Am. Geriatr. Soc*. 33, 278–285. doi:10.1111/j.1532-5415.1985.tb07117.x

Lindeman, R. (1990). Overview: renal physiology and pathophysiology of aging. *Am. J. Kidney Dis*. 16, 275–282. doi:10.1016/s0272-6386(12)80002-3

Murussi, M., Gross, J. L., and Silveiro, S. P. (2006). Glomerular filtration rate changes in normoalbuminuric and microalbuminuric Type 2 diabetic patients and normal individuals A 10-year follow-up. *J. Diabetes Complicat*. 20, 210–215. doi:10.1016/j.jdiacomp.2005.07.002

Niki, P., Panos, K., and Christos, C. (2015). New targets for end-stage chronic kidney disease therapy. *J. Crit. Care Med*. 1, 92–95. doi:10.1515/jccm-2015-0015

Parsa, A., Kao, W. H., Xie, D., Astor, B. C., Li, M., Hsu, C. Y., et al. (2013). APOL1 risk variants, race, and progression of chronic kidney disease. *N. Engl. J. Med*. 369, 2183–2196. doi:10.1056/NEJMoa1310345

Pohl, M. A., Blumenthal, S., Cordonnier, D. J., De Alvaro, F., Deferrari, G., Eisner, G., et al. (2005). Independent and additive impact of blood pressure control and angiotensin II receptor blockade on renal outcomes in the irbesartan diabetic nephropathy trial: clinical implications and limitations. *J. Am. Soc. Nephrol*. 16, 3027–3037. doi:10.1681/ASN.2004110919

Quinlan, J. (1993). *C4.5: programs for machine learning*. Morgan Kaufmann Publishers.

Rahman, M., Greene, T., Phillips, R. A., Agodoa, L. Y., Bakris, G. L., Charleston, J., et al. (2013). A trial of 2 strategies to reduce nocturnal blood pressure in blacks with chronic kidney disease. *Hypertension* 61, 82–88. doi:10.1161/HYPERTENSIONAHA.112.200477

Ruggenenti, P., Perna, A., Gherardi, G., Garini, G., Zoccali, C., Salvadori, M., et al. (1999). Renoprotective properties of ace-inhibition in non-diabetic nephropathies with non-nephrotic proteinuria. *Lancet* 354, 359–364. doi:10.1016/S0140-6736(98)10363-X

Subasi, E., Subasi, M. M., Hammer, P. L., Roboz, J., Anbalagan, V., and Lipkowitz, M. S. (2017). A classification model to predict the rate of decline of kidney function. *Front. Med*. 4, 97. doi:10.3389/fmed.2017.00097

Wang, X., Lewis, J., Appel, L., Cheek, D., Contreras, G., Faulkner, M., et al. (2006). Validation of creatinine-based estimates of gfr when evaluating risk factors in longitudinal studies of kidney disease. *J. Am. Soc. Nephrol*. 17, 2900–2909. doi:10.1681/ASN.2005101106

Wright, J. T., Bakris, G., Greene, T., Agodoa, L. Y., Appel, L. J., Charleston, J., et al. (2002). Effect of blood pressure lowering and antihypertensive drug class on progression of hypertensive kidney disease: results from the aask trial. *Jama* 288, 2421–2431. doi:10.1001/jama.288.19.2421

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

frontiers in Big Data

ORIGINAL RESEARCH
published: 27 September 2021
doi: 10.3389/fdata.2021.654972

Prognostic Factors of Survival in Pancreatic Cancer Metastasis to Liver at Different Ages of Diagnosis: A SEER Population-Based Cohort Study

Meiqi Liu[1,2†], *Moran Wang*[3†] *and Sheng Li*[3*‡]

[1]Department of Infectious Disease, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China, [2]Department and Institute of Infectious Disease, Xi'an Children's Hospital, Xi'an Jiaotong University, Xi'an, China, [3]Department of Cardiology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

Background: Liver is a common metastatic organ for most malignancies, especially the pancreas. However, evidence for prognostic factors of pancreatic cancer metastasis to the liver at different ages is lacking. Thus, we aimed to evaluate the predictors of patients with pancreatic cancer metastasis to liver grouped by age of diagnosis.

Methods: We chose the patients diagnosed between 2004 and 2015 from the SEER database. The primary lesions of metastatic liver cancer between sexes were compared using the Pearson's chi-square test for categorical variables. The overall survival (OS) and cancer-specific survival (CSS) were the endpoint of the study. The prognostic factors were analyzed with the Kaplan-Meier method and log-rank test, and Cox proportional-hazards regression model.

Results: The main primary sites of metastatic liver cancer for our patients are lung and brunchu, sigmoid colon, pancreas, which in males are lung and bronchu, sigmoid colon and pancreas, while breast, lung and bronchu, sigmoid colon in females. Furthermore, we explored the prognostic factors of pancreatic cancer metastasis to liver grouped by age at diagnosis. Tumor grade, histology and treatment are valid prognostic factors in all age groups. Additionally, gender and AJCC N stage in age<52 years old, while race and AJCC N stage in age >69 years old were predictors. Surgery alone was the optimal treatment in group age>69 years old, whereas surgery combined with chemotherapy was the best option in the other groups.

Conclusion: Our study evaluated the predictors of patients with pancreatic cancer metastasis to liver at various ages of diagnosis.

Keywords: pancreatic cancer, liver, prognosis, metastasis, treatment

Abbreviations: AJCC, American Joint Committee on Cancer (seventh); CI, coincidence intervals; CSS, cancer-specific survival; HR, hazard ratio; OS, overall survival; SEER, Surveillance Epidemiology and End Results.

OPEN ACCESS

Edited by:
Shi-Cong Tao,
Shanghai Jiao Tong University, China

Reviewed by:
Raghvendra Mall,
St. Jude Children's Research Hospital, United States
Zhan Wang,
Zhejiang University, China
Shengming Jin,
Fudan University, China

****Correspondence:***
Sheng Li
shengli410@126.com

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

[‡]These authors share last authorship

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

Received: *18 January 2021*
Accepted: *16 August 2021*
Published: *27 September 2021*

Citation:
Liu M, Wang M and Li S (2021) Prognostic Factors of Survival in Pancreatic Cancer Metastasis to Liver at Different Ages of Diagnosis: A SEER Population-Based Cohort Study. Front. Big Data 4:654972. doi: 10.3389/fdata.2021.654972

INTRODUCTION

The liver is the most frequently afflicted metastatic organ second to the lymph nodes for most malignancies (Jaques et al., 1995; Hess et al., 2006; Amankwah et al., 2013; Ryu et al., 2013). The most common tumors with liver metastases arise from the portal venous drainage system, which provides about two-thirds of the liver's blood supply. Because lesions are usually asymptomatic, liver involvement in metastasis is often neglected and poorly studied, and even extensive infiltration of metastatic tumor may not alter its function or homeostasis until late in the disease (Clark et al., 2016). There are few epidemiological studies on metastatic liver cancer, but 30–70% of patients die of liver metastasis (Pickren J and Lane, 1982) and most patients with liver metastases will die of the primary disease (Gilbert Ha et al., 1982).

As one of the deadliest malignant tumors in the world (Ferlay et al., 2015; Schild and Vokes, 2016), pancreatic cancer is the eighth most common cause of cancer in males and the sixth most common cause of cancer in females. In decades, a large number of studies have shown that the development of pancreatic cancer was closely related to age. The aging trend of the population in the world is challenging the current treatments and caring for patients with pancreatic cancer (Bray et al., 2018; Ferlay et al., 2018). The underlying mechanisms of pancreatic cancer is complicated and uncertain, accompanied with poor prognosis (Maisonneuve, 2019). According to the original site in pancreas, pancreatic cancer is classified as endocrine and exocrine pancreatic cancer, and the latter is more common and has a higher risk of mortality in both females and males (Fesinmeyer et al., 2005). Additionally, the majority of exocrine pancreatic cancer is adenocarcinoma (Li, 2001; Cowgill and Muscarella, 2003). Approximately 50% of pancreatic cancer patients are diagnosed with distant metastases (Mayo et al., 2012), and the most common site of distal metastases found at autopsy was the liver, followed by the peritoneum, lungs and pleura, bones, and adrenal glands (Kamisawa et al., 1995; Mao et al., 1995; Embuscado et al., 2005; Disibio and French, 2008). Previous studies suggested risk factors of pancreatic cancer involving smoking, positive family history and genetics, diabetes, obesity, dietary factors, alcohol consumption, and physical inactivity (Yadav and Lowenfels, 2013; Ilic and Ilic, 2016). Age, race, tumor size, grade, lymph node metastasis (Mayo et al., 2012), AJCC stage (Kamarajah et al., 2017) and treatment (Ansari et al., 2019) are also reported associated with the survival of pancreatic cancer patients. However, evidence for prognostic factors in pancreatic cancer with distant metastasis is rare. However, evidence for prognostic factors in pancreatic cancer with distant metastasis is rare. Moreover, Andrew A et al. and previous studies reported that treatment strategies for pancreatic cancer differentiate in diverse range of ages (Wheeler and Nicholl, 2014). Thus, the objective of this study is to determine the differences in primary sites of metastatic liver cancer between males and females. Furthermore, we evaluated the prognostic risk factors of pancreatic cancer metastasis to liver at different ages of diagnosis through the Cox regression model.

MATERIALS AND METHODS

Data Source

The data was from the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) program between 2004 and 2015. The program contains the population-based central cancer registries of 18 geographically defined regions. Because all the data used in the study was retrieved from the SEER database with publicly available methods, the study did not require local moral approval or a declaration.

Patient Selection

The inclusion criteria included: 1) The disease was diagnosed between 2004 and 2015; 2) metastases of the primary tumor were at the liver; 3) there was only one primary tumor; 4) the diagnosis of the disease was histologically positive; 5) there were more than 0 days of survival.

The exclusion criteria included: 1) age≥85 years old; 2) the demographics of patients were incomplete, including race and marital status; 3) the clinicopathological characteristics of patients were incomplete, including grade, AJCC seventh stage (TNM), tumor size, laterality, causes of death and treatment methods; 4) patients treated with radiotherapy; 5) the type of reporting source was autopsy only or death certificate only. (**Figure 1**).

We used the histopathology codes from the International Classification of Disease for Oncology third edition (ICD-O-3) to define the primary sites of patients with hepatic metastatic carcinoma. In the ICD-O-3, the codes were defined as follows: code 19-29 (tongue), code 50-69 (gum and other mouth), code 70-89 (salivary gland), code 90-99 (tonsil), code 110-119 (nasopharynx), code 129-139 (hypopharynx), code 150-159 (Esophagus), code 160-169 (stomach), code 170-179 (small intestine), code 180 (cecum), code 181 (appendix), code 182 (ascending colon), code 183 (hepatic flexure), code 184 (transverse colon), code 185 (splenic flexure), code 186 (descending colon), code 187 (sigmoid colon), code 199 (rectosigmoid junction), code 209 (rectum), code 210-218 (Anus, Anal Canal and Anorectum), code 220 (liver), code 221 (intrahepatic bile duct), code 239 (gallbladder), code 240-241 (other biliary), code 250-259 (pancreas), code 300-319 (nose, nasal cavity and middle ear), code 320-329 (larynx), code 340-349 (lung and bronchu), code 380, 472-479, 490-499 (soft tissue including heart), code 381-383 (trachea, mediastinum and other respiratory organs), code 384 (pleura), code 400-419 (bones and joints), code 440-449 (skin excluding basal and squamous), code 480 (retroperitoneum), code 481-482 (peritoneum, omentum and mesentery), code 500-509 (breast), code 510-519 (vulva), code 529 (vagina), code 530-539 (cervix uteri), code 540-549 (corpus uteri), code 569 (ovary), code 570 (other female genital organs), code 601 (penis), code 619 (prostate), code 620-629 (testis), code 649-659 (kidney and renal pelvis), code 669 (ureter), 670-679 (urinary bladder), code 739 (thyroid) and code 740-755 (other endocrine including thymus).

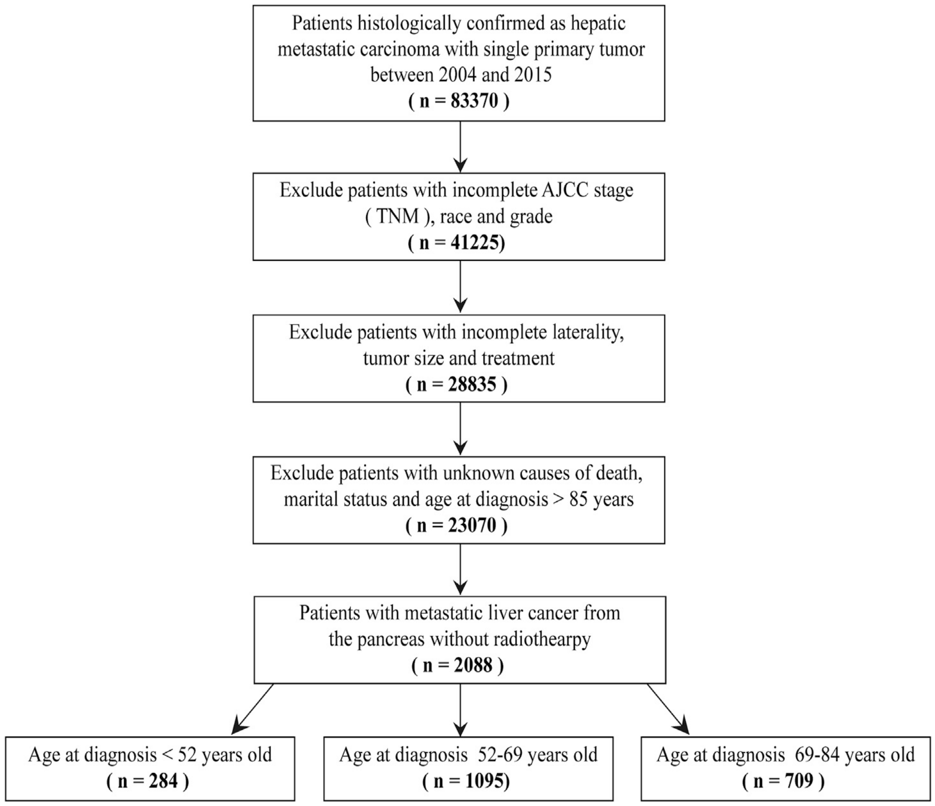

Patients histologically confirmed as hepatic metastatic carcinoma with single primary tumor between 2004 and 2015
(n = 83370)
Exclude patients with incomplete AJCC stage (TNM), race and grade
(n = 41225)
Exclude patients with incomplete laterality, tumor size and treatment
(n = 28835)
Exclude patients with unknown causes of death, marital status and age at diagnosis > 85 years
(n = 23070)
Patients with metastatic liver cancer from the pancreas without radiothearpy
(n = 2088)
Age at diagnosis < 52 years old
(n = 284)
Age at diagnosis 52-69 years old
(n = 1095)
Age at diagnosis 69-84 years old
(n = 709)

FIGURE 1 | Study cohort.

According to the age at diagnosis of patients, we divided them into three groups, including age at diagnosis <52 years old, age at diagnosis 52–69 years old and age at diagnosis 69–84 years old.

Clinical Variables of Patients

Information on demographic factors (age, race, sex and marital status), tumor-related factors (tumor size, grade, histology and AJCC TNM staging system), therapeutic factors (surgery and chemotherapy) and follow-up were collected from the SEER database. And follow-up period ended in 2015. Based on the Surgery Codes of the SEER program and information about other treatments, we divided the treatment options into categories: no treatment (N), surgery alone (S), chemotherapy alone (C), surgery combined with chemotherapy (SC).

OS and CSS were the interesting endpoint, and the cancer-specific death was based on the code of “SEER cause-specific death classification” in the SEER database. OS was measured from the date on which the first-time definite diagnosis was made until the date of death caused by any cause or the most recent follow-up.

Statistical Analysis

Age and tumor size are categorized according to the best cut-off value produced by the x-tile software version 3.6.1 (Yale University School of Medicine, US). (S2) The incidence rates were calculated by using R software. And baseline patients' demographics and clinicopathological characteristics were compared using the Pearson's chi-square test for categorical variables. The independent risk factors were identified by univariate and multivariate Cox proportional-hazards regression analyses for OS. R software version 4.0.2 (R Project, Vienna, Austria) was used for all analysis. Statistically significant cutoff value was set up as $p < 0.05$, two-sided. $p < 0.2$ was selected as filter value for univariate to multivariate analysis.

RESULTS

The Frequency Distribution of Primary Lesions of Metastatic Liver Cancer

Regardless of gender, the most common primary site of hepatic metastatic carcinoma was lung and brunchu that accounted for 15.18% of all primary lesions, followed by sigmoid colon (11.11%), pancreas (9.15%), breast (8.92%), cecum (8.18%) and rectum (7.81%). The result of Pearson's chi-square test showed that the primary sites of hepatic metastatic carcinoma were significantly different between males and females including anus, anal canal and anorectum ($p < 0.001$), ascending colon ($p < 0.05$), breast ($p < 0.001$), cervix uteri ($p < 0.001$), corpus colon ($p < 0.001$), descending colon ($p < 0.05$), esophagus ($p < 0.001$), gallbladder ($p < 0.001$), hepatic flexure ($p < 0.05$), kidney and renal pelvis ($p < 0.001$), larynx ($p < 0.05$), liver ($p < 0.001$), lung

TABLE 1 | The frequency distribution of primary lesions of metastatic liver cancer.

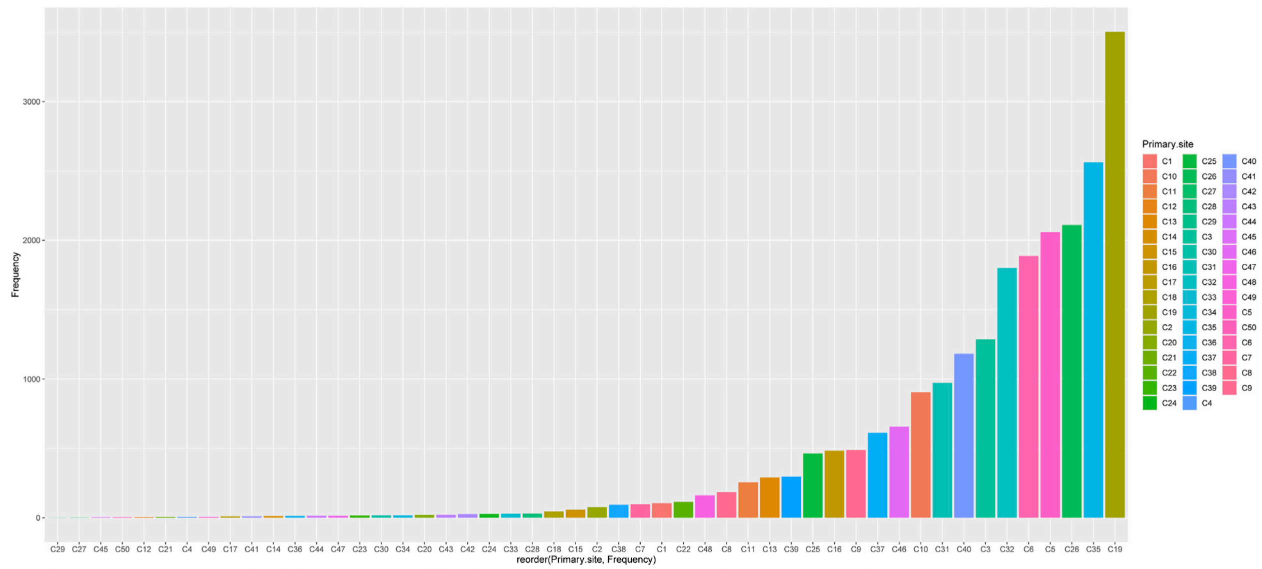
Primary.site, n (%)	Total (*n* = 23,070)	Female (*n* = 11,139)	Male (*n* = 11,931)	*p* Value
Anus, Anal Canal and Anorectum	105 (0.46)	72 (0.65)	33 (0.28)	<0.001***
Appendix	77 (0.33)	42 (0.38)	35 (0.29)	>0.05
Ascending Colon	1,286 (5.57)	656 (5.89)	630 (5.28)	<0.05*
Bones and Joints	6 (0.03)	1 (0.01)	5 (0.04)	>0.05
Breast	2058 (8.92)	2048 (18.39)	10 (0.08)	<0.001***
Cecum	1887 (8.18)	948 (8.51)	939 (7.87)	>0.05
Cervix Uteri	97 (0.42)	97 (0.87)	0 (0)	<0.001***
Corpus Uteri	185 (0.80)	185 (1.66)	0 (0)	<0.001***
Descending Colon	488 (2.12)	210 (1.89)	278 (2.33)	<0.05*
Esophagus	905 (3.92)	115 (1.03)	790 (6.62)	<0.001***
Gallbladder	256 (1.11)	183 (1.64)	73 (0.61)	<0.001***
Gum and Other Mouth	5 (0.02)	2 (0.02)	3 (0.03)	>0.05
Hepatic Flexure	290 (1.26)	120 (1.08)	170 (1.42)	<0.05*
Hypopharynx	12 (0.05)	2 (0.02)	10 (0.08)	>0.05
Intrahepatic Bile Duct	58 (0.25)	28 (0.25)	30 (0.25)	>0.05
Kidney and Renal Pelvis	483 (2.09)	180 (1.62)	303 (2.54)	<0.001***
Larynx	10 (0.04)	1 (0.01)	9 (0.08)	<0.05*
Liver	46 (0.20)	6 (0.05)	40 (0.34)	<0.001***
Lung and Bronchu	3,503 (15.18)	1,505 (13.51)	1998 (16.75)	<0.001***
Nasopharynx	21 (0.09)	3 (0.03)	18 (0.15)	<0.01**
Nose, Nasal Cavity and Middle Ear	6 (0.03)	4 (0.04)	2 (0.02)	>0.05
Other Biliary	115 (0.50)	49 (0.44)	66 (0.55)	>0.05
Other Endocrine including Thymus	16 (0.07)	49 (0.05)	10 (0.08)	>0.05
Other Female Genital Organs	27 (0.12)	27 (0.24)	0 (0)	<0.001***
Ovary	464 (2.01)	464 (4.17)	0 (0)	<0.001***
Pancreas	2,110 (9.15)	910 (8.17)	1,200 (10.06)	<0.001***
Penis	2 (0.01)	0 (0)	2 (0.02)	>0.05
Peritoneum, Omentum and Mesentery	30 (0.13)	27 (0.1)	3 (0.15)	<0.001***
Pleura	1 (0.00)	0 (0)	1 (0.01)	>0.05
Prostate	18 (0.08)	0 (0)	18 (0.15)	<0.001***
Rectosigmoid Junction	973 (4.22)	384 (3.45)	589 (4.94)	<0.001***
Rectum	1801 (7.81)	605 (5.43)	1,196 (10.02)	<0.001***
Retroperitoneum	29 (0.13)	11 (0.10)	18 (0.15)	>0.05
Salivary Gland	18 (0.08)	7 (0.06)	11 (0.09)	>0.05
Sigmoid Colon	2,563 (11.11)	1,064 (9.55)	1,499 (12.56)	<0.001***
Skin excluding Basal and Squamous	13 (0.06)	5 (0.04)	8 (0.07)	>0.05
Small Intestine	613 (2.66)	284 (2.55)	329 (2.76)	>0.05
Soft Tissue including Heart	94 (0.41)	54 (0.48)	40 (0.34)	>0.05
Splenic Flexure	297 (1.29)	113 (1.01)	184 (1.54)	<0.001***
Stomach	1,182 (5.12)	329 (2.95)	853 (7.15)	<0.001***
Testis	10 (0.04)	0 (0)	10 (0.08)	<0.01**
Thyroid	26 (0.11)	14 (0.13)	12 (0.10)	>0.05
Tongue	22 (0.10)	6 (0.05)	16 (0.13)	>0.05
Tonsil	14 (0)	4 (0)	10 (0)	>0.05
Trachea, Mediastinum and Other Respiratory Organs	4 (0.02)	0 (0)	4 (0.03)	>0.05
Transverse Colon	658 (2.85)	302 (2.71)	356 (2.98)	>0.05
Ureter	14 (0.06)	7 (0.06)	7 (0.06)	>0.05
Urinary Bladder	162 (0.70)	49 (0.44)	113 (0.95)	<0.001***
Vagina	6 (0.03)	6 (0.05)	0 (0)	<0.05*
Vulva	4 (0.02)	4 (0.04)	0 (0)	>0.05

, two-sided* p *values <0.05; **, two-sided* p *values <0.01; *, two-sided* p *values <0.001.*

and bronchus ($p < 0.001$), nasopharynx ($p < 0.01$), other female genital organs ($p < 0.001$), ovary ($p < 0.001$), pancreas ($p < 0.001$), peritoneum, omentum and mesentery ($p < 0.001$), prostate ($p < 0.001$), rectosigmoid junction ($p < 0.001$), rectum ($p < 0.001$), sigmoid colon ($p < 0.001$), splenic flexure ($p < 0.001$), stomach ($p < 0.001$), testis ($p < 0.01$), urinary bladder ($p < 0.001$) and vulva ($p < 0.05$). (**Table 1**). In females, the top five most common primary lesions of hepatic metastases were breast (18.39%), lung and bronchu (13.51%), sigmoid colon (9.55%), cecum (8.51%) and pancreas (8.17%), while in males were lung and bronchu (16.75%), sigmoid colon (12.56%), pancreas (10.06%), rectum (10.02%) and cecum (7.87%). (**Figures 2,3**).

The Effect of Age at Diagnosis With Pancreatic Cancer Metastasis to Liver

The Kaplan Meier survival curve showed significant difference in overall survival for patients diagnosed at different age groups ($p < 0.001$). The overall survival time was negatively correlated



C1 Anus, Anal Canal and Anorectum;
C2 Appendix ;
C3 Ascending Colon;
C4 Bones and Joints;
C5 Breast;
C6 Cecum;
C7 Cervix Uteri;
C8 Corpus Uteri;
C9 Descending Colon;
C10 Esophagus;

C11 Gallbladder;
C12 Gum and Other Mouth;
C13 Hepatic Flexure ;
C14 Hypopharynx;
C15 Intrahepatic Bile Duct;
C16 Kidney and Renal Pelvis;
C17 Larynx;
C18 Liver;
C19 Lung and Bronchu ;
C20 Nasopharynx;

C21 Nose, Nasal Cavity and Middle Ear ;
C22 Other Biliary;
C23 Other Endocrine including Thymus;
C24 Other Female Genital Organs;
C25 Ovary ;
C26 Pancreas;
C27 Penis;
C28 Peritoneum, Omentum and Mesentery;
C29 Pleura;
C30 Prostate;

C31 Rectosigmoid Junction;
C32 Rectum;
C33 Retroperitoneum;
C34 Salivary Gland;
C35 Sigmoid Colon;
C36 Skin excluding Basal and Squamous;
C37 Small Intestine;
C38 Soft Tissue including Heart;
C39 Splenic Flexure;
C40 Stomach;

C41 Testis;
C42 Thyroid;
C43 Tongue;
C44 Tonsil;
C45 Trachea, Mediastinum and Other Respiratory Organs;
C46 Transverse Colon;
C47 Ureter;
C48 Urinary Bladder;
C49 Vagina;
C50 Vulva

FIGURE 2 | Frequency Distribution of primary tumour sources of hepatic metastatic carcinoma.

with the age at diagnosis. Among the three groups, the prognosis of patients diagnosed at age less than 52 years old was the best, and of which the median survival time was 1 year. (**Figure 4**).

The Effect of Treatment with Pancreatic Cancer Metastasis to Liver

Regardless of age at diagnosis, surgery alone (S) was the optimal treatment option for patients with pancreatic cancer metastasis to liver, followed by surgery combined with chemotherapy (SC), chemotherapy alone (C) and no treatment (N) ($p < 0.001$). And the median survival time of patients with surgery alone was approximately 3.5–4 years. (**Figure 5**).

The Relative Hazard Ratio of Treatment and Age at Diagnosis

As the multivariable hazard ratio of in prognosis displayed in **Figure 6**, with the increase of the age of diagnosis, treatment showed significantly protective effect, while grade had a significant effect on prognosis only in younger age. And other prognostic factors had almost no significant change. (**Figure 6A**). Thus, we further analyzed the relative hazard ratio of diverse treatment options and age of diagnosis in patients, we found that when patients were diagnosed at a younger age, chemotherapy alone was the most adverse risk factor, while when diagnosed at an older age, age at diagnosis was the most adverse risk factor for the outcome. What's more, for patients diagnosed at all ages, chemotherapy alone was the treatment with the worst effect on prognosis, while for patients diagnosed at age more than 69 years old, surgery was better than combined with chemotherapy. (**Figure 6B**).

Clinical Characteristics of the Patients With Pancreatic Cancer Metastasis to Liver

Demographic characteristics of 2088 patients with pancreatic cancer metastasis to liver grouped by age at diagnosed during the 12-years study period (between 2004 and 2015) in the SEER database are shown in **Table 2**. In this study, sex ($p = 0.002$), race ($p = 0.031$), marital status ($p < 0.001$), tumor grade ($p < 0.001$), AJCC N stage ($p = 0.005$), treatment ($p < 0.001$), median survival time ($p < 0.001$) and vital status ($p < 0.001$) were the parameters with significant difference among different groups. On the whole, most patients were married white males whose tumors were poorly differentiated and less than 4.9 cm in size, treated with chemotherapy alone (C). The most common histological type of tumors was adenomas and adenocarcinomas. Compared with the other groups, well differentiated tumors (25%), surgery alone (S, 14.44%) or surgery combined with chemotherapy (SC, 11.27%) for treatment strategies and longer survival time (12 months) would more likely to occur in age <52 years old group.

Univariate and Multivariate of OS in the Patients with Pancreatic Cancer Metastasis to Liver

As illustrated in **Table 3**, on the basis of the overall survival (OS), univariate analysis showed that the significant indicators were sex, grade, tumor size, AJCC N stage, histology and treatment in group age <52 years old; marital status, grade, tumor size, histology and

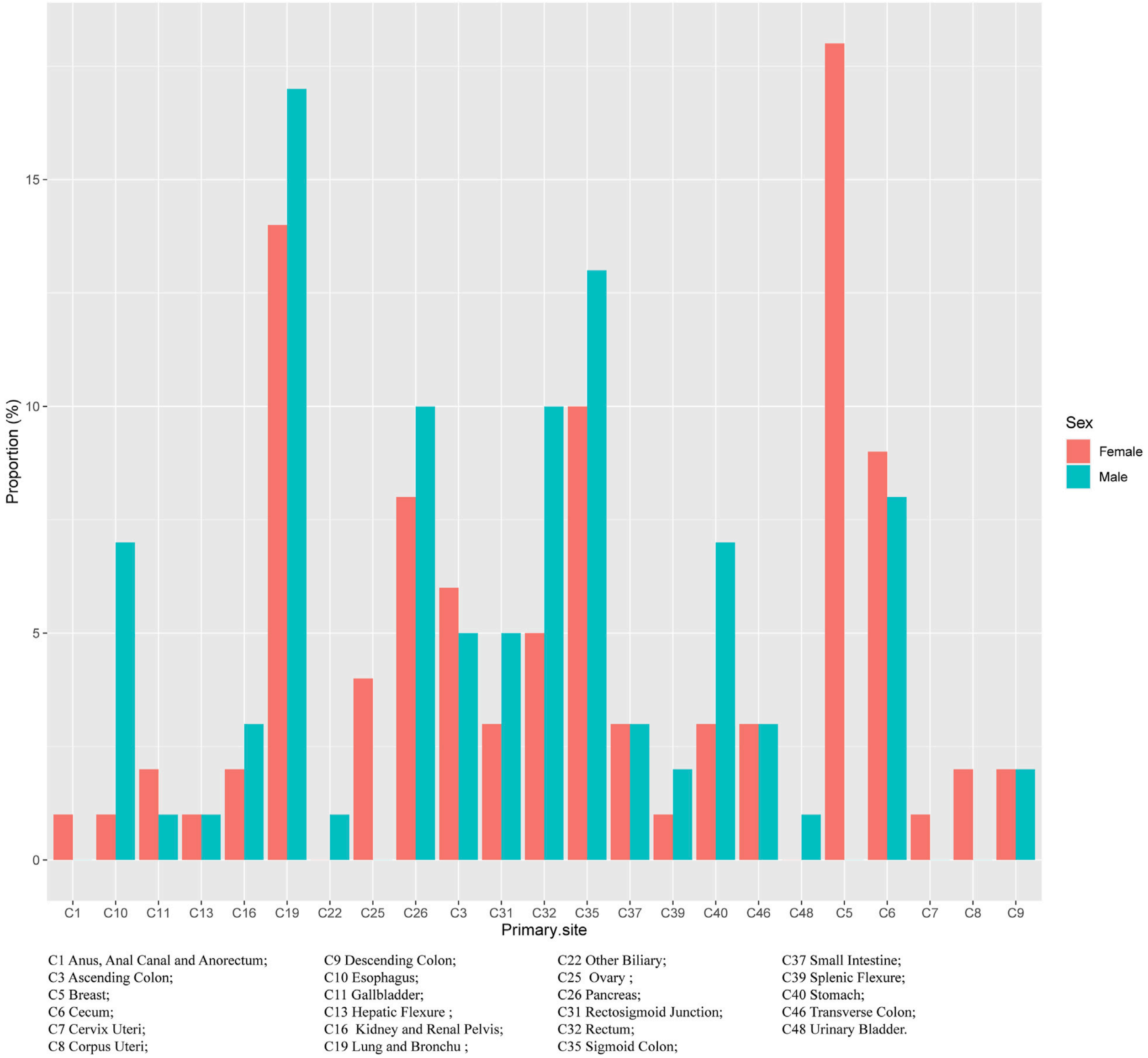


FIGURE 3 | The purpose of primary tumour sources of liver metastatic carcinoma in both sexes.

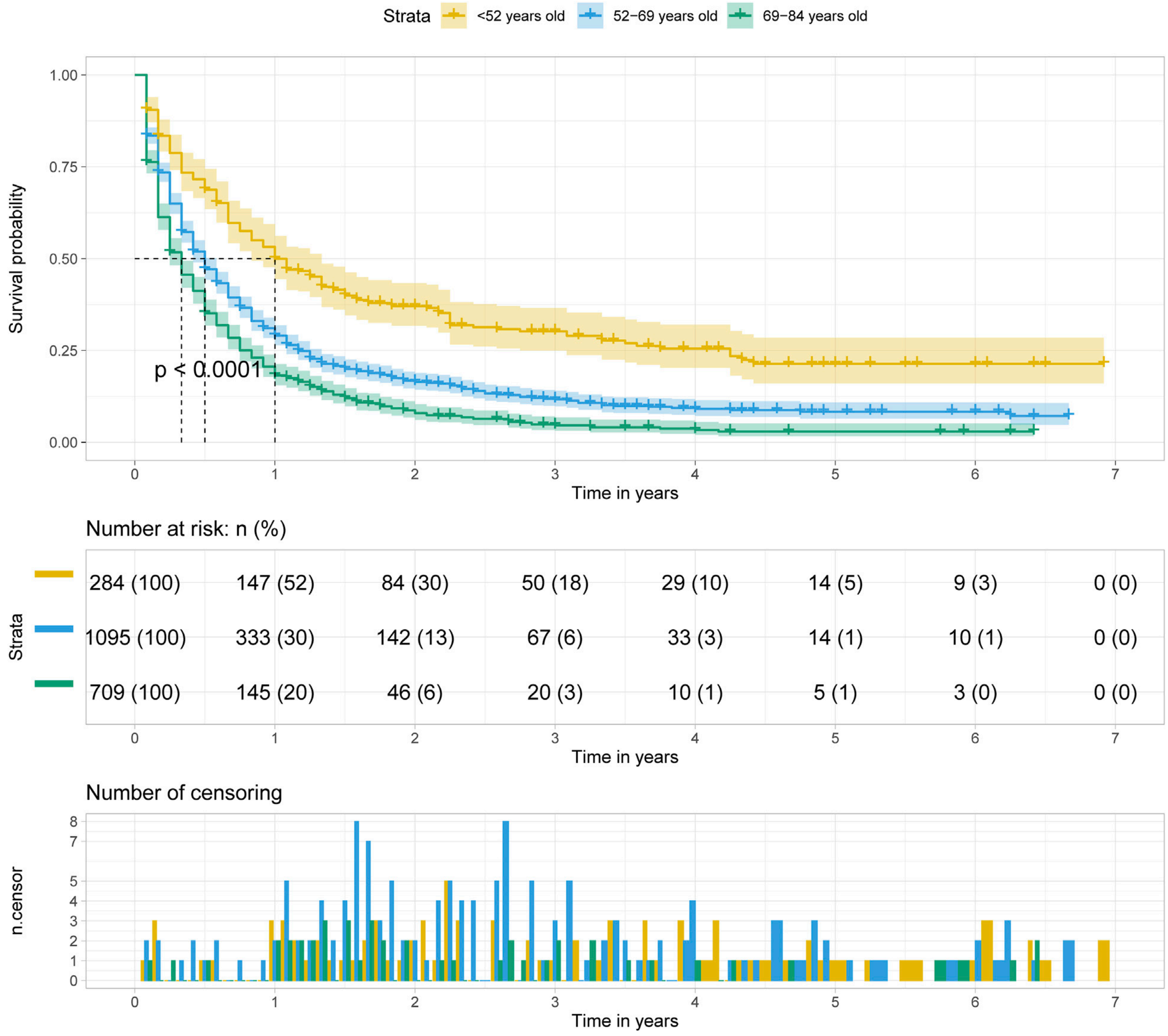
treatment in group age 52–69 years old; and race, grade, AJCC N stage, histology and treatment in group age 69–84 years old.

In multivariate analysis, we further observed the variables selected from univariate analysis (p < 0.2). Cox regression analysis was performed to compete hazard ratios and 95% confidence intervals. In the three groups, tumor grade was all associated with poor overall survival, and surgery alone (S) was the best treatment option for the overall survival of patients.

Using AJCC N0 stage as reference, AJCC N1 stage (p = 0.020, HR = 1.18, 95%CI, 1.03–1.36) in group age 52–69 years and AJCC NX stage (p = 0.039, HR = 1.33, 95%CI, 1.01–1.74) in group age 69–84 years old were indicated to be associated with poor overall survival, while in group age <52 years old, AJCC N stage was not correlated with the prognosis. Choosing adenomas and adenocarcinomas as reference in histological types, in addition to ductal and lobular neoplasms (age <52 years, p = 0.027, HR = 1.69, 95%CI, 1.06-2.69; age 52–69 years old, p < 0.001, HR = 1.59, 95%CI, 1.24-2.04; age 69–84 years old, p = 0.045, HR = 1.34, 95%CI, 1.01-1.79) in the three groups, other histological types (p = 0.016, HR = 1.90, 95%CI, 1.13-3.21) in group age <52 years old, and epithelial neoplasms (age 52–69 years old, p = 0.007, HR = 1.48, 95%CI, 1.11-1.96; age 69–84 years old, p = 0.004, HR = 1.65, 95%CI, 1.18-2.31) in the other two groups (age >52 years old) were associated with a poor overall survival.

Univariate and Multivariate of CSS in the Patients with Pancreatic Cancer Metastasis to Liver

As illustrated in **Table 4**, on the basis of the cancer-specific survival (CSS), univariate analysis showed that the significant



Number at risk: n (%)

Strata	0	1	2	3	4	5	6	7
🟨	284 (100)	147 (52)	84 (30)	50 (18)	29 (10)	14 (5)	9 (3)	0 (0)
🟦	1095 (100)	333 (30)	142 (13)	67 (6)	33 (3)	14 (1)	10 (1)	0 (0)
🟩	709 (100)	145 (20)	46 (6)	20 (3)	10 (1)	5 (1)	3 (0)	0 (0)

Time in years

FIGURE 4 | Kalpan Meier survival curve showing the effect of age at diagnosis with pancreatic cancer metastasis to liver.

indicators were sex, grade, histology and treatment methods in group age <52 years old; grade, histology and treatment methods in group age 52–69 years old; and race, grade, AJCC T stage and treatment methods in group age 69–84 years old.

Using well differentiated grade as reference, multivariate analysis in **Table 4** indicated tumor grade was associated with poor overall survival at different ages. In addition, treatment (S, C, SC) was associated with better cancer-specific survival in all three groups compared with no treatment. Notably, in group age 69–84 years old, surgery alone (S, $p < 0.001$, HR = 0.40, 95%CI, 0.26-0.60) was the optimal treatment, whereas surgery combined with chemotherapy (SC, group age <52 years old, $p < 0.01$, HR = 0.17, 95%CI, 0.08-0.33; group age 52–69 years old, $p < 0.001$, HR = 0.22, 95%CI, 0.16-0.30) was the best option in the other groups.

When using AJCC N0 as reference, patients with AJCC N1 stage ($p < 0.001$, HR = 1.82, 95%CI, 1.29-2.56) had a poor prognosis only in group age <52 years old. And epithelial neoplasms (age 52-69, $p = 0.026$, HR = 1.38, 95%CI, 1.04-1.84; age 69–84 years old, $p = 0.042$ HR = 1.43, 95%CI, 1.01-2.02) were associated with a poor cancer-specific survival only in group age >52 years old when using adenomas and adenocarcinomas as reference. Additionally, in group age 69–84 years old, other racial patients ($p = 0.017$, HR = 1.42, 95%CI, 1.07-1.90) had a worse prognosis.

DISCUSSION

It was reported that 90% cancer-related deaths resulted from metastasis of the primary tumor. The formation of local infiltrates and metastases are clinically most relevant to the progression of cancer (Christofori, 2006). Organ damage due to growth-related

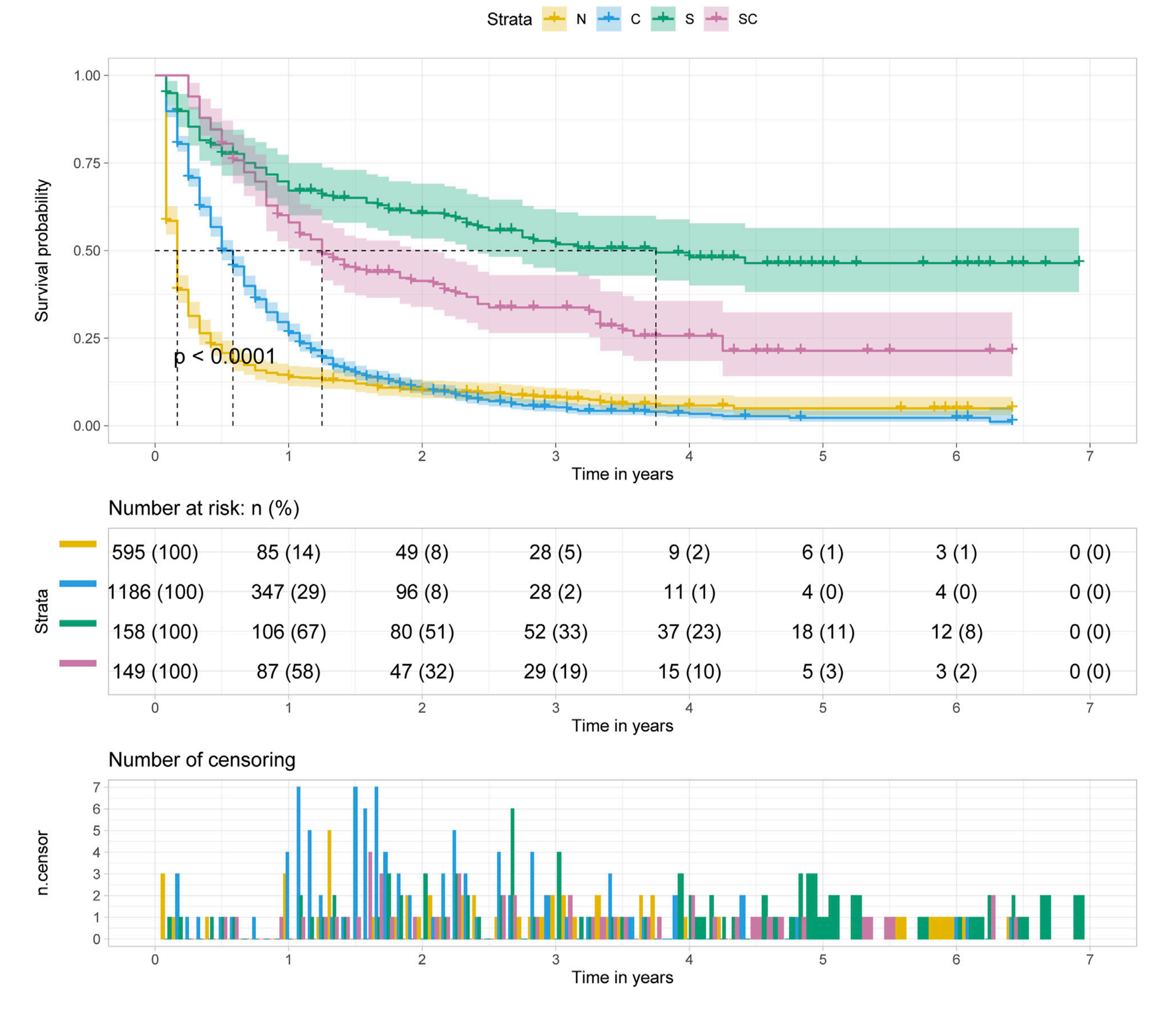
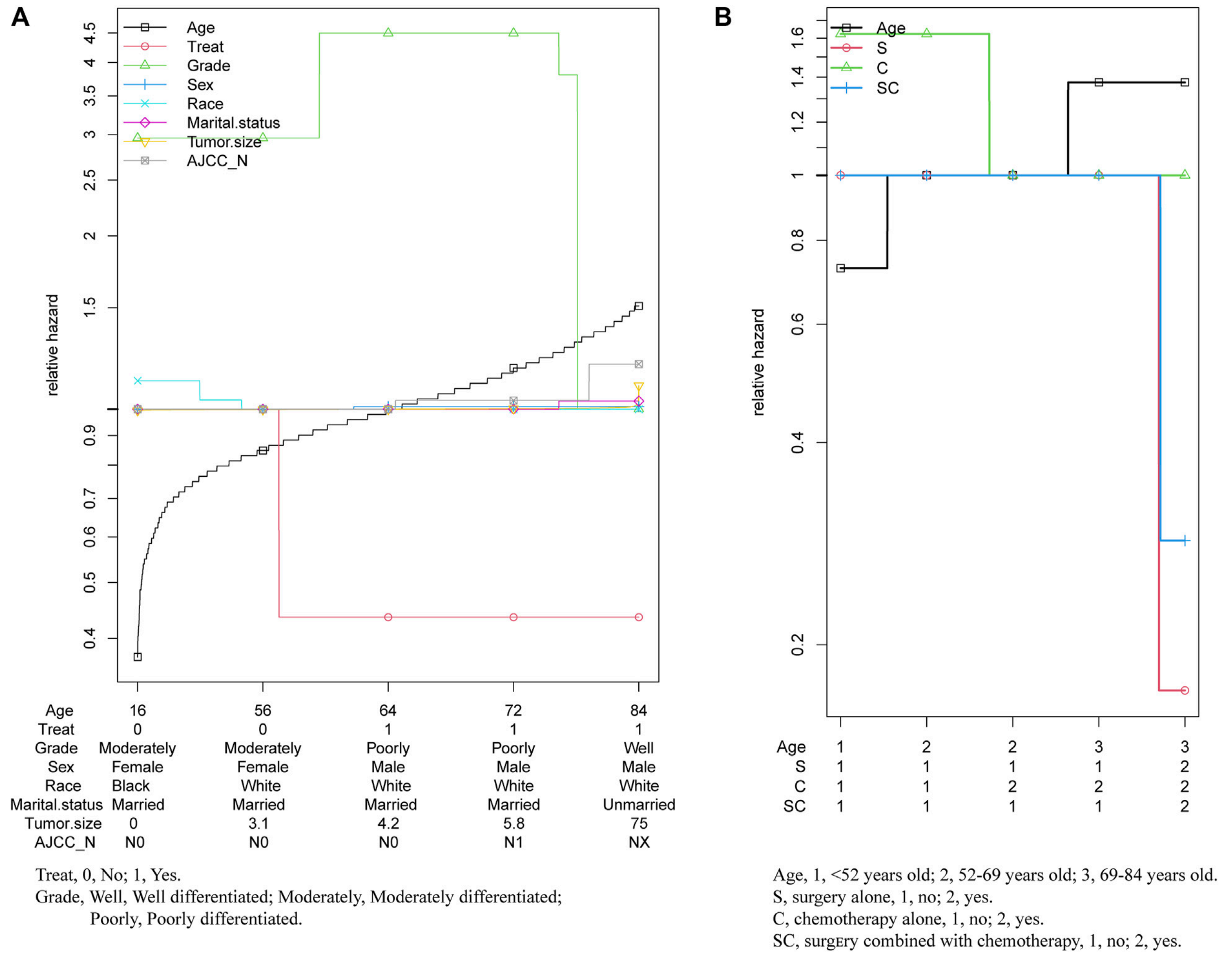


FIGURE 5 | Kalpan Meier survival curve showing the effect of treatment with pancreatic cancer metastasis to liver.

lesions, paraneoplastic syndromes, or treatment complications was significantly associated with morbidity and mortality of metastatic disease (Steeg, 2006). In general, cancer metastasis can be divided into different stages from local invasion, intravasation, survival in circulation, extravasation, finally to colonization and metastasis (Hanahan and Weinberg, 2011). The unique biological characteristics of the liver make it a vulnerable site for tumor metastasis: 1) structural and hemodynamic features - characteristic microcirculation in the liver makes it easier for diffuse tumor cells carried in the blood to enter. In addition, molecules on the surface of hepatic nonparenchymal cells (NPCs) lining the hepatic capillaries contribute to the adhesion and retention of circulating tumor cells. The pore on the hepatic sinusoidal endothelial cell (LSECs) facilitates the tumor cells to enter the basement membrane directly; 2) regenerative capabilities—the cellular tissue remodeling mechanism involved in self-renewal and reconstruction that promotes intratumoral stroma and blood vessel formation through signals generated by tumor cells, creating an enabling environment for survival and growth; 3) regional immunosuppression—the general foreign body reaction is reduced to limit potential damage to the liver, resulting in a relatively tolerant microenvironment that allows for the survival and growth of foreign tumor cells (Vidal-Vanaclocha, 2011; Clark et al., 2016).

Pancreatic cancer is the fourth leading cause of cancer-related death worldwide, and its main metastatic site is liver (Stott et al., 2010). Studies have shown that, in addition to smoking, a family history of pancreatic cancer, black race, diabetes, and increased body mass index were also predictors of pancreatic cancer mortality (Coughlin et al., 2000). A lack of early signs and symptoms, as well as high aggressiveness, leads to a low



Treat, 0, No; 1, Yes.
Grade, Well, Well differentiated; Moderately, Moderately differentiated; Poorly, Poorly differentiated.

Age, 1, <52 years old; 2, 52-69 years old; 3, 69-84 years old.
S, surgery alone, 1, no; 2, yes.
C, chemotherapy alone, 1, no; 2, yes.
SC, surgery combined with chemotherapy, 1, no; 2, yes.

FIGURE 6 | Relative hazard ratio of multivariables in patients with pancreatic cancer metastasis to liver.

survival rate. The prognosis of patients with pancreatic cancer is closely related to tumor stage and tumor grade/aggressiveness (Bolm et al., 2015) that can only be evaluated by biopsy or surgery. Our present data showed tumor grade was also a significant predictor of overall survival and cancer-specific survival in patients with liver metastasis, independent of age at diagnosis (**Table 3**, **Table 4**). In addition, 85% of the histology types of pancreatic cancer are ductal adenocarcinoma of the pancreas (PDAC) (Ryan et al., 2014; Hogendorf et al., 2018). For patients with PDAC, younger age, male sex, larger tumor size, low ALT level and high CA 19-9 level could predict unexpected distant metastasis (Liu et al., 2018). Histologically, pancreatic adenocarcinoma accounts for the largest proportion in pancreatic cancer (Simard et al., 2012), accompanied with the worst prognosis, and the most common site of metastasis is liver (Lemke et al., 2013; Deeb et al., 2015; Kumar et al., 2015), which is consistent with our results (**Table 2**). Our data suggested that most histologic types of pancreatic metastases to liver were adenocarcinomas. The prognosis of patients with pancreatic cancer with liver metastasis was poorer than that of patients with distant lymph node metastasis or lung metastasis. The factors predicting the better prognosis included age<65 years, white race, being married, female sex and surgery treatment (Oweira et al., 2017). Furthermore, our study showed that the younger the age, the higher the overall survival rate of patients with pancreatic cancer with liver metastasis (**Figure 4**). In addition, we found differences in prognostic factors among the groups after grouping by age at diagnosis. Histologically, compared with pancreatic adenocarcinoma, ductal and lobular neoplasms and epithelial neoplasms were associated with poor overall survival in the group age >52 years old, while the latter were not correlated with the prognosis in group age <52 years old. In the multivariate regression analysis, histological type was a significant predictor for cancer-specific survival only for patients diagnosed at age >52 years old. AJCC N1 stage with significance in predicting poor overall survival only in group age 52–69 years old, and predicting poor cancer-specific survival only in group age <52 years old. (**Table 3**, **Table 4**).

At present, the only treatment for pancreatic cancer is surgery, and adjuvant therapy based on chemotherapy can improve the

TABLE 2 | Clinical characteristics of the patients with pancreatic cancer metastasis to liver grouped by age at diagnosis.

Variables	Total (*n* = 2088)	<52 years old (*n* = 284)	52–69 years old (*n* = 1095)	69-84 years old (*n* = 709)	*p* value
Sex, n (%)	—	—	—	—	0.002**
Female	900 (43.10)	119 (41.90)	438 (40.00)	343 (48.38)	—
Male	1188 (56.90)	165 (58.10)	657 (60.00)	366 (51.62)	—
Race, n (%)	—	—	—	—	0.031*
White	1,651 (79.07)	207 (73.89)	865 (79.00)	579 (81.66)	—
Black	261 (12.50)	44 (15.49)	143 (13.06)	74 (13.06)	—
Other	176 (8.43)	33 (11.62)	87 (7.95)	56 (7.90)	—
Marital status, n (%)	—	—	—	—	<0.001***
Unmarried	333 (15.95)	87 (30.63)	186 (16.99)	60 (8.46)	
Married	1755 (84.05)	197 (69.37)	909 (83.01)	649 (91.54)	
Grade, n (%)	—	—	—	—	<0.001***
Well differentiated	257 (12.31)	73 (25.70)	108 (9.86)	76 (10.72)	—
Moderately differentiated	756 (36.21)	82 (28.87)	414 (37.81)	260 (36.67)	—
Poorly differentiated	1,000 (47.89)	116 (40.85)	539 (49.22)	345 (48.66)	—
Undifferentiated	75 (3.59)	15 (4.58)	34 (3.95)	28 (3.95)	—
Tumor size, n (%)	—	—	—	—	0.325
<4.9 cm	1,285 (61.54)	164 (57.75)	669 (61.10)	452 (63.75)	
4.9–7.4 cm	565 (27.06)	83 (29.33)	294 (26.85)	188 (26.52)	
>7.4 cm	238 (11.40)	37 (13.03)	132 (12.05)	69 (9.73)	
AJCC N, n (%)	—	—	—	—	<0.005**
N0	1,074 (51.44)	129 (45.42)	544 (49.68)	401 (56.56)	
N1	804 (38.51)	129 (45.42)	432 (39.45)	243 (34.27)	
NX	210 (10.06)	26 (9.15)	119 (10.87)	65 (9.17)	
Histology, n (%)	—	—	—	—	0.935
Adenomas and adenocarcinomas	1,688 (80.84)	225 (79.23)	896 (81.83)	567 (79.97)	—
Ductal and lobular neoplasms	166 (7.95)	25 (8.80)	82 (7.49)	59 (8.32)	—
Epithelial neoplasms	116 (5.56)	16 (5.63)	60 (5.48)	40 (5.64)	—
Others	118 (5.65)	18 (6.34)	57 (5.21)	43 (6.06)	—
Treat n (%)	—	—	—	—	<0.001***
N	595 (28.50)	51 (17.96)	273 (24.93)	271 (38.22)	—
C	1,186 (56.80)	160 (56.34)	659 (60.18)	367 (51.76)	—
S	158 (7.57)	41 (14.44)	75 (6.85)	42 (5.92)	—
SC	149 (7.14)	32 (11.27)	75 (6.85)	42 (5.92)	—
Survival time, Median (IQR)	6.00 (2.00,13.00)	12.00 (4.00, 27.00)	6.00 (2.00, 14.00)	4.00 (2.00, 9.00)	<0.001***
Vital status, n (%)	—	—	—	—	<0.001***
Alive	280 (13.41)	86 (30.28)	148 (13.52)	46 (6.49)	—
Cancer-specific death	1773 (84.91)	194 (68.31)	933 (85.21)	646 (91.11)	—
Other causes-specific death	35 (1.68)	4 (1.41)	14 (1.28)	17 (2.40)	—

, two-sided p values <0.05; **, two-sided p values <0.01; *, two-sided p values <0.001. AJCC, American Joint Committee on Cancer (seventh).*
Treat, N, no treatment; C, chemotherapy alone; S, surgery alone; SC, surgery combined with chemotherapy.

survival rate (McGuigan et al., 2018). For elderly patients (age>80 years old), postoperative adjuvant chemotherapy is critical to the prognosis (Sho et al., 2016). Surgery is limited to patients with localized disease, and metastatic spread is often considered a contraindication to resection, regardless of whether it is observed synchronously or ectopic (Seufferlein et al., 2012). However, metastatic excision or local treatment is occasionally performed in centers around the world based on individual clinical experience, and there is no objective evidence to guide treatment methods taking into account patient choice or metastatic spread (Gleisner et al., 2007; Shrikhande et al., 2007; De Jong et al., 2010; Nentwich et al., 2012; Edwards et al., 2013). In general, palliative chemotherapy with FOLFIRONOX (mFOLFIRINOX with 5-fluorouracil) is the preferred chemotherapy regimen for metastatic pancreatic cancer (McGuigan et al., 2018). Despite this, T. Hackert (Hackert et al., 2017) proved that resection of liver or interaortocaval lymph nodes (ILN) metastases could be superior to palliative treatment for pancreatic cancer patients with metastasis. Mitsuka et al. (2020) found that the median survival time was significantly improved for patients diagnosed between 44 and 83 years old who underwent liver resection or pancreatectomy. Other study (Warschkow et al., 2020) showed that lymphadenectomy had only 18% direct effect on improved overall survival, while 82% of its effect were mediated by other factors like treatment at high-volume hospitals and adjuvant chemotherapy for patients whose median age were 66 years. However, the analysis on differences among different ages of patients is scarce. As we know, there is insufficient evidence that the efficacy of different therapies in patients with metastatic pancreatic cancer is age-related. Our data showed chemotherapy alone was the most important prognostic factor for patients who diagnosed at younger age, and age of diagnosis was the most prognostic factor for patients diagnosed at an older age. For the diagnosis of pancreatic cancer at all ages, surgery was the best treatment method to improve the overall survival rate of

TABLE 3 | Univariate and multivariate of OS in the patients with pancreatic cancer metastasis to liver grouped by age at diagnosis.

Variables	<52 years old				52–69 years old				69–84 years old			
	Univariate analysis		Multivaraiate analysis		Univariate analysis		Multivaraiate analysis		Univariate analysis		Multivaraiate analysis	
	N	P Value	HR (95%CI)	P Value	N	P Value	HR (95%CI)	P Value	N	P Value	HR (95%CI)	P Value
Sex, n (%)	—	—	—	—	—	—	—	—	—	—	—	—
Female	119	—	—	—	438	—	—	—	343	—	—	—
Male	165	0.004**	1.28 (0.94–1.75)	0.112	657	0.126	1.10 (0.97–1.26)	0.15	366	0.113	0.87 (0.74–1.01)	0.075
Race, n (%)	—	—	—	—	—	—	—	—	—	—	—	—
White	207	—	—	—	865	—	—	—	579	—	—	—
Black	44	0.913	—	—	143	0.156	1.10 (0.90–1.33)	0.349	74	0.0645	1.00 (0.78–1.01)	0.976
Other	33	0.250	—	—	87	0.914	—	0.940	56	0.037*	1.22 (0.92–1.62)	0.176
Marital status, n (%)	—	—	—	—		—	—	—	—	—	—	—
Unmarried	87	—	—	—	186	—	—	—	60	—	—	—
Married	197	0.425	—	—	909	0.017*	0.85 (0.71–1.01)	0.063	0.940	0.0674	—	—
Grade, n (%)	—	—	—	—	—	—	—	—	—	—	—	—
Well differentiated	73	—	—	—	108	—	—	—	76	—	—	—
Moderately differentiated	82	<0.001***	2.52 (1.53–4.13)	<0.001***	414	<0.001***	3.16 (2.35–4,25)	<0.001***	260	<0.001***	2.07 (1.5–2.76)	<0.001***
Poorly differentiated	116	<0.001***	5.17 (3.14–8.52)	<0.001***	539	<0.001***	4.61 (3.43–6.19)	<0.001***	345	<0.001***	2.68 (2.02–3.57)	<0.001***
Undifferentiated	13	<0.001***	5.08 (2.24–11.53)	<0.001***	34	<0.001***	1.10 (0.82–1.47)	<0.001***	—	<0.001***	2.28 (1.43–3.65)	<0.001***
Tumor size, n (%)	—	—	—	—	—	—	—	—	—	—	—	—
<4.9 cm	164	—	—	—	544	—	—	—	401	—	—	—
4.9–7.4 cm	83	0.322	1.03 (0.95–1.81)	0.091	432	0.516	1.18 (1.03–1.36)	0.20*	69	0.744	—	—
>7.4 cm	37	0.008**	0.69 (0.41–1.16)	0.130	119	0.075	0.97 (0.78–1.20)	0.750	188	0.641	—	—
AJCC N, n (%)	—	—	—	—	—	—	—	—	—	—	—	—
N0	129	—	—	—	669	—	—	—	452	—	—	—
N1	1299	0.887	1.31 (0.75–1.42)	0.850	294	0.057	1.09 (0.94–1.27)	0.243	69	0.744	—	—
NX	26	<0.001***	0.69 (0.41–1.16)	0.162	132	0.003*	0.82 (0.66–1.03)	0.082	188	0.641	—	—
Histology, n (%)	—	—	—	—	—	—	—	—	—	—	—	—
Adenomas and adenocarcinomas	225	—	—	—	896	—	—	—	567	—	—	—
Ductal and lobular neoplasms	25	0.001**	1.69 (1.06–2.69)	0.027*	82	0.438	1.59 (1.24–2.04)	<0.001***	59	0.832	1.34 (1.01–1.79)	0.045*
Epithelial neoplasms	16	<0.001***	1.68 (0.93–3.01)	0.083	60	0.008*	1.48 (1.11–1.96)	0.007**	40	<0.001***	1.65 (1.18–2.31)	0.004*
Others	18	<0.001***	1.90 (1.13–3.21)	0.016	57	0.146	1.10 (O.82–1.47)	0.521	43	0.929	0.79 (0.57–1.10)	0.166
Treat n (%)	—	—	—	—	—	—	—	—	—	—	—	—
N	51	—	—	—	273	—	—	—	271	—	—	—
C	160	0.533	0.62 (0.41–0.94)	0.024*	659	<0.001***	0.43 (0.37–0.50)	<0.001***	367	<0.001***	0.54 (0.45–0.63)	<0.001***
S	41	<0.001***	0.16 (0.08–0.32)	<0.001***	75	<0.001***	0.15 (0.10–0.21)	<0.001***	42	<0.001***	1.65 (1.18–0.40)	<0.001***
SC	32	0.005**	0.35 (0.19–0.64)	<0.001***	88	<0.001***	0.16 (0.12–0.21)	<0.001***	29	<0.001***	0.33 (0.21–0.51)	<0.001***

, two-sided* p *values <0.05; **, two-sided* p *values <0.01; *, two-sided* p *values <0.001. AJCC, American Joint Committee on Cancer (seventh).*
HR, hazard ratio.
CI, coincidence intervals. OS, overall survival.
Treat, N, no treatment; C, chemotherapy alone; S, surgery alone; SC, surgery combined with chemotherapy.

patients with pancreatic cancer with liver metastasis (**Table 3**). Considering the tumor-specific survival rate, surgery combined with chemotherapy is the best choice for patients under 69 years of age at the time of diagnosis, while surgery alone is the best choice for patients aged 69–84 years at the time of diagnosis. In addition, surgery alone and combined chemotherapy were

TABLE 4 | Univariate and multivariate of CSS in the patients with pancreatic cancer metastasis to liver grouped by age at diagnosis.

Variables	<52 years old				52–69 years old				69–84 years old			
	Univariate		Multivariate		Univariate		Multivariate		Univariate		Multivariate	
	N	p value	HR (95%CI)	p value	N	p value	HR (95%CI)	p value	N	p value	HR (95%CI)	p value
Sex, n (%)												
Female	73	—	—	—	365	—	—	—	317	—	—	—
Male	121	0.045*	1.08 (0.78–1.48)	0.645	568	0.274	—	—	329	0.428	—	—
Race, n (%)												
White	142	—	—	—	735	—	—	—	527	—	—	—
Black	31	0.308	—	—	127	0.814	—	—	65	0.208	1.08 (0.83–1.40)	0.576
Other	21	0.796	—	—	71	0.253	—	—	54	0.028*	1.42 (1.07–1.90)	0.017*
Marital status, n (%)												
Unmarried	56	—	—	—	158	—	—	—	55	—	—	—
Married	138	0.706	—	—	775	0.073	0.88 (0.74–1.04)	0.14	591	0.975	—	—
Grade, n (%)												
Well differentiated	25	—	—	—	54	—	—	—	58	—	—	—
Moderately differentiated	53	<0.001***	4.05 (2.25–7.29)	<0.001***	353	0.024*	1.61 (1.20–2.16)	0.002**	237	<0.001***	1.99 (1.48–2.68)	<0.001***
Poorly differentiated	106	<0.001***	5.26 (2.97–9.33)	<0.001***	497	<0.001***	2.29 (1.71–3.06)	<0.001***	325	<0.001***	2.50 (1.86–3.36)	<0.001***
Undifferentiated	10	<0.001***	9.43 (3.76–23.67)	<0.001***	29	0.003**	2.00 (1.25–3.20)	0.004**	26	0.002**	2.14 (1.33–3.46)	0.002**
Tumor size, n (%)												
<4.9 cm	113	—	—	—	576	—	—	—	414	—	—	—
4.9–7.4 cm	64	0.601	—	—	260	0.514	—	—	170	0.651	—	—
>7.4 cm	17	0.854	—	—	97	0.928	—	—	62	0.255	—	—
AJCC N, n (%)												
N0	86	—	—	—	468	—	—	—	362	—	—	—
N1	85	0.601	1.82 (1.29–2.56)	<0.001***	359	0.491	—	—	223	0.041*	0.94 (0.79–1.12)	0.47
NX	23	0.077	1.17 (0.72–1.92)	0.523	106	0.338	—	—	61	0.055	1.13 (0.86–1.50)	0.382
Histology, n (%)												
Adenomas and adenocarcinomas	139	—	—	—	753	—	—	—	513	—	—	
Ductal and lobular neoplasms	23	0.764	1.05 (0.66–1.68)	0.832	75	0.309	1.19 (0.92–1.54)	0.178	55	0.639	1.15 (0.85–1.56)	0.353
Epithelial neoplasms	14	0.024*	1.17 (0.63–2.17)	0.622	55	0.009**	1.38 (1.04–1.84)	0.026*	38	0.002**	1.43 (1.01–2.02)	0.042*
Others	18	0.016*	1.35 (0.80–2.27)	0.264	50	0.136	1.18 (0.88–1.58)	0.268	40	0.834	0.84 (0.60–1.17)	0.293
Treat, n (%)												
N	36	—	—	—	245	—	—	—	252	—	—	—
C	131	0.036*	0.44 (0.29–0.69)	<0.001***	601	<0.001***	0.40 (0.35–0.47)	<0.001***	340	<0.001***	0.48 (0.40–0.57)	<0.001***
S	9	0.008**	0.22 (0.09–0.52)	<0.001***	33	<0.001***	0.30 (0.21–0.44)	<0.001***	31	<0.001***	0.40 (0.26–0.60)	<0.001***
SC	18	<0.001***	0.17 (0.08–0.33)	<0.001***	54	<0.001***	0.22 (0.16–0.30)	<0.001***	23	<0.001***	0.44 (0.28–0.69)	<0.001***

, two-sided* p *values <0.05; **, two-sided* p *values <0.01; *, two-sided* p *values <0.001. AJCC, American Joint Committee on Cancer (seventh).*
HR, hazard ratio.
CI, coincidence intervals. CSS, cancer-specific survival.
Treat, N, no treatment; C, chemotherapy alone; S, surgery alone; SC, surgery combined with chemotherapy.

significantly superior to chemotherapy alone in terms of overall survival and tumor-specific survival (**Table 3**, **Table 4**). Interestingly, in a case report (Katsura et al., 2019), after the combination therapy of pancreatoduodenectomy and chemotherapy, a 66-year-old patient with pancreatic ductal carcinoma metastasis to liver showed the disappearance of liver metastasis and without other new metastasis. This case report partially confirms the conclusion from our analysis that surgery alone was the optimal treatment in group age>69 years old, while surgery combined with chemotherapy was the best option in the other groups. Both surgical treatment and chemotherapy cause damage to human bodies. Especially, the elderly can hardly bear the double blow, as surgical treatment and chemotherapy both exerting in therapy. In addition, chemotherapy is often accompanied with many side effects. The analyzed data in this manuscript demonstrates that surgical treatment alone is superior to surgery plus chemotherapy in patients older than 69 years of age. It suggests that surgery should be a priority for the older population (age>69) with pancreatic cancer metastasis to liver. Certainly, clinical treatment selection depends on the multiple assessment of patient, and this manuscript provides an epidemiological reference for the selection of clinical treatment.

Although the SEER database provides a large amount of clinical data, there are still many limitations in our research. First, we need to further conduct a follow-up clinical trial to verify this result. Second, we did not include patients undergoing radiotherapy because of the small number of cases, and we need to compare the effects of radiotherapy, chemotherapy and surgery on prognosis. Finally, the sequence of chemotherapy and surgery, the diverse methods of surgery and chemotherapy can be further studied.

CONCLUSION

In this population-based analysis, we found the main primary sites of metastatic liver cancer are lung and brunchu, sigmoid colon and pancreas. Furthermore, we explored the prognostic factors of pancreatic cancer metastasis to liver grouped by age at diagnosis. Tumor grade, histology and treatment are valid prognostic factors in all age groups. Additionally, gender and AJCC N stage in age<52 years old, while race and AJCC N stage in age>69 years old were predictors. Surgery alone was the optimal treatment in group age>69 years old, whereas surgery combined with chemotherapy was the best option in the other groups. In conclusion, these findings would help to choose better treatment for patients with metastatic liver cancer.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://seer.cancer.gov/.The data was from the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) program between 2004 and 2015.

AUTHOR CONTRIBUTIONS

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

FUNDING

This work was founded by National Natural Science Foundation of China (81974032, 81570337 to SL), Natural Science Foundation of Hubei Province (2015CFB455 to SL).

ACKNOWLEDGMENTS

We wish to thank the help from the Division of Infectious Disease as well as Division of Cardiology, Department of Internal Medicine in Tongji Hospital, and the Department and Institute of Infectious Disease, Xi'an Children's Hospital.

REFERENCES

Ansari, D., Althini, C., Ohlsson, H., and Andersson, R. (2019). Early-onset pancreatic cancer: a population-based study using the SEER registry. *Langenbecks Arch. Surg.* 404 (5), 565–571. doi:10.1007/s00423-019-01810-0

Bolm, L., Janssen, S., Käsmann, L., Wellner, U., Bartscht, T., Schild, S. E., et al. (2015). Predicting Survival after Irradiation of Metastases from Pancreatic Cancer. *Anticancer Res.* 35 (7), 4105–4108.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer J. Clinicians* 68 (6), 394–424. doi:10.3322/caac.21492

Christofori, G. (2006). New signals from the invasive front. *Nature* 441 (7092), 444–450. doi:10.1038/nature04872

Clark, A. M., Ma, B., Taylor, D. L., Griffith, L., and Wells, A. (2016). Liver metastases: Microenvironments and *ex-vivo* models. *Exp. Biol. Med. (Maywood)* 241 (15), 1639–1652. doi:10.1177/1535370216658144

Coughlin, S. S., Calle, E. E., Patel, A. V., and Thun, M. J. (2000). Predictors of pancreatic cancer mortality among a large cohort of United States adults. *Cancer Causes Control* 11 (10), 915–923. doi:10.1023/a:1026580131793

Cowgill, S. M., and Muscarella, P. (2003). The genetics of pancreatic cancer. *Am. J. Surg.* 186 (3), 279–286. doi:10.1016/s0002-9610(03)00226-5

De Jong, M. C., Farnell, M. B., Sclabas, G., Cunningham, S. C., Cameron, J. L., Geschwind, J.-F., et al. (2010). Liver-Directed Therapy for Hepatic Metastases in Patients Undergoing Pancreaticoduodenectomy. *Ann. Surg.* 252 (1), 142–148. doi:10.1097/SLA.0b013e3181dbb7a7

Deeb, A., Haque, S. U., and Olowokure, O. (2015). Pulmonary metastases in pancreatic cancer, is there a survival influence? *J. Gastrointest. Oncol.* 6 (3), E48–E51. doi:10.3978/j.issn.2078-6891.2014.114

Disibio, G., and French, S. W. (2008). Metastatic patterns of cancers: results from a large autopsy study. *Arch. Pathol. Lab. Med.* 132 (6), 931–939. doi:10.5858/2008-132-931-mpocrf

Edwards, J., Scoggins, C., McMasters, K., and Martin, R. (2013). Combined pancreas and liver therapies: resection and ablation in hepato-pancreatico-biliary malignancies. *J. Surg. Oncol.* 107 (7), 709–712. doi:10.1002/jso.23318

Embuscado, E. E., Laheru, D., Ricci, F., Yun, K. J., de Boom Witzel, S., Seigel, A., et al. (2005). Immortalizing the complexity of cancer metastasis: genetic features of lethal metastatic pancreatic cancer obtained from rapid autopsy. *Cancer Biol. Ther.* 4 (5), 548–554. doi:10.4161/cbt.4.5.1663

Ferlay, J., Colombet, M., Soerjomataram, I., Dyba, T., Randi, G., Bettio, M., et al. (2018). Cancer incidence and mortality patterns in Europe: Estimates for 40 countries and 25 major cancers in 2018. *Eur. J. Cancer* 103, 356–387. doi:10.1016/j.ejca.2018.07.005

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., et al. (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* 136 (5), E359–E386. doi:10.1002/ijc.29210

Fesinmeyer, M. D., Austin, M. A., Li, C. I., De Roos, A. J., and Bowen, D. J. (2005). Differences in Survival by Histologic Type of Pancreatic Cancer. *Cancer Epidemiol. Biomarkers Prev.* 14 (7), 1766–1773. doi:10.1158/1055-9965.epi-05-0120

Gilbert HA, K. A., Hintz, B. L., Rao, A. R., and Nussbaum, H. (1982). "Patterns of metastases," in *Liver Metastases*. Editor W. LGH (Boston: GK Hall Medical Publishers), 19–39.

Gleisner, A. L., Assumpcao, L., Cameron, J. L., Wolfgang, C. L., Choti, M. A., Herman, J. M., et al. (2007). Is resection of periampullary or pancreatic adenocarcinoma with synchronous hepatic metastasis justified? *Cancer* 110 (11), 2484–2492. doi:10.1002/cncr.23074

Hackert, T., Niesen, W., Hinz, U., Tjaden, C., Strobel, O., Ulrich, A., et al. (2017). Radical surgery of oligometastatic pancreatic cancer. *Eur. J. Surg. Oncol. (Ejso)* 43 (2), 358–363. doi:10.1016/j.ejso.2016.10.023

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell* 144 (5), 646–674. doi:10.1016/j.cell.2011.02.013

Hess, K. R., Varadhachary, G. R., Taylor, S. H., Wei, W., Raber, M. N., Lenzi, R., et al. (2006). Metastatic patterns in adenocarcinoma. *Cancer* 106 (7), 1624–1633. doi:10.1002/cncr.21778

Hogendorf, P., Durczyński, A., and Strzelczyk, J. (2018). Metastatic Pancreatic Cancer. *J. Invest. Surg.* 31 (2), 151–152. doi:10.1080/08941939.2017.1291774

Ilic, M., and Ilic, I. (2016). Epidemiology of pancreatic cancer. *Wjg* 22 (44), 9694–9705. doi:10.3748/wjg.v22.i44.9694

Jaques, D. P., Coit, D. G., Casper, E. S., and Brennan, M. F. (1995). Hepatic metastases from soft-tissue sarcoma. *Ann. Surg.* 221 (4), 392–397. doi:10.1097/00000658-199504000-00010

Kamarajah, S. K., Burns, W. R., Frankel, T. L., Cho, C. S., and Nathan, H. (2017). Validation of the American Joint Commission on Cancer (AJCC) 8th Edition Staging System for Patients with Pancreatic Adenocarcinoma: A Surveillance, Epidemiology and End Results (SEER) Analysis. *Ann. Surg. Oncol.* 24 (7), 2023–2030. doi:10.1245/s10434-017-5810-x

Kamisawa, T., Isawa, T., Koike, M., Tsuruta, K., and Okamoto, A. (1995). Hematogenous metastases of pancreatic ductal carcinoma. *Pancreas* 11 (4), 345–349. doi:10.1097/00006676-199511000-00005

Katsura, Y., Takeda, Y., Ohmura, Y., Sakamoto, T., Kawai, K., Inatome, J., et al. (2019). A Case Report of Conversion Surgery for Pancreatic Ductal Adenocarcinoma with Liver Metastasis after Chemotherapy. *Gan To Kagaku Ryoho* 46 (4), 802–804.

Kumar, A., Dagar, M., Herman, J., Iacobuzio-Donahue, C., and Laheru, D. (2015). CNS involvement in pancreatic adenocarcinoma: a report of eight cases from the Johns Hopkins Hospital and review of literature. *J. Gastrointest. Canc* 46 (1), 5–8. doi:10.1007/s12029-014-9667-y

Lemke, J., Scheele, J., Kapapa, T., Wirtz, C., Henne-Bruns, D., and Kornmann, M. (2013). Brain metastasis in pancreatic cancer. *Ijms* 14 (2), 4163–4173. doi:10.3390/ijms14024163

Li, D. (2001). Molecular epidemiology of pancreatic cancer. *Cancer J.* 7 (4), 259–265.

Liu, X., Fu, Y., Chen, Q., Wu, J., Gao, W., Jiang, K., et al. (2018). Predictors of distant metastasis on exploration in patients with potentially resectable pancreatic cancer. *BMC Gastroenterol.* 18 (1), 168. doi:10.1186/s12876-018-0891-y

Maisonneuve, P. (2019). Epidemiology and burden of pancreatic cancer. *La Presse Médicale* 48 (3 Pt 2), e113–e123. doi:10.1016/j.lpm.2019.02.030

Mao, C., Domenico, D. R., Kim, K., Hanson, D. J., and Howard, J. M. (1995). Observations on the Developmental Patterns and the Consequences of Pancreatic Exocrine Adenocarcinoma. *Arch. Surg.* 130 (2), 125–134. doi:10.1001/archsurg.1995.01430020015001

Mayo, S. C., Nathan, H., Cameron, J. L., Olino, K., Edil, B. H., Herman, J. M., et al. (2012). Conditional survival in patients with pancreatic ductal adenocarcinoma resected with curative intent. *Cancer* 118 (10), 2674–2681. doi:10.1002/cncr.26553

McGuigan, A., Kelly, P., Turkington, R. C., Jones, C., Coleman, H. G., and McCain, R. S. (2018). Pancreatic cancer: A review of clinical diagnosis, epidemiology, treatment and outcomes. *Wjg* 24 (43), 4846–4861. doi:10.3748/wjg.v24.i43.4846

Mitsuka, Y., Yamazaki, S., Yoshida, N., Yan, M., Higaki, T., and Takayama, T. (2020). Time interval-based indication for liver resection of metastasis from pancreatic cancer. *World J. Surg. Onc* 18 (1), 294. doi:10.1186/s12957-020-02058-5

Nentwich, M. F., Bockhorn, M., König, A., Izbicki, J. R., and Cataldegirmen, G. (2012). Surgery for advanced and metastatic pancreatic cancer--current state and trends. *Anticancer Res.* 32 (5), 1999–2002.

Oweira, H., Petrausch, U., Helbling, D., Schmidt, J., Mannhart, M., Mehrabi, A., et al. (2017). Prognostic value of site-specific metastases in pancreatic adenocarcinoma: A Surveillance Epidemiology and End Results database analysis. *Wjg* 23 (10), 1872–1880. doi:10.3748/wjg.v23.i10.1872

Pickren J, T. Y., and Lane, W. (1982). "Liver metastases: Analysis of autopsy data," in *Liver Metastases*. Editor G. H. L. Weiss (Boston: GK Hall Medical Publishers), 2–18.

Reed, D., Amankwah, A. P., and Conley, D. R. (2013). Epidemiology and therapies for metastatic sarcoma. *Clep* 5, 147–162. doi:10.2147/clep.s28390

Ryan, D. P., Hong, T. S., and Bardeesy, N. (2014). Pancreatic adenocarcinoma. *N. Engl. J. Med.* 371 (11), 1039–1049. doi:10.1056/NEJMra1404198

Ryu, S. W., Saw, R., Scolyer, R. A., Crawford, M., Thompson, J. F., and Sandroussi, C. (2013). Liver resection for metastatic melanoma: equivalent survival for cutaneous and ocular primaries. *J. Surg. Oncol.* 108 (2), 129–135. doi:10.1002/jso.23361

Schild, S. E., and Vokes, E. E. (2016). Pathways to improving combined modality therapy for stage III nonsmall-cell lung cancer. *Ann. Oncol.* 27 (4), 590–599. doi:10.1093/annonc/mdv621

Seufferlein, T., Bachet, J. B., Van Cutsem, E., and Rougier, P. (2012). Pancreatic adenocarcinoma: ESMO-ESDO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann. Oncol.* 23 (Suppl. 7), vii33–vii40. doi:10.1093/annonc/mds224

Sho, M., Murakami, Y., Kawai, M., Motoi, F., Satoi, S., Matsumoto, I., et al. (2016). Prognosis after surgical treatment for pancreatic cancer in patients aged 80 years or older: a multicenter study. *J. Hepatobiliary Pancreat. Sci.* 23 (3), 188–197. doi:10.1002/jhbp.320

Shrikhande, S. V., Kleeff, J., Reiser, C., Weitz, J., Hinz, U., Esposito, I., et al. (2006). Pancreatic resection for M1 pancreatic ductal adenocarcinoma. *Ann. Surg. Oncol.* 14 (1), 118–127. doi:10.1245/s10434-006-9131-8

Simard, E. P., Ward, E. M., Siegel, R., and Jemal, A. (2012). Cancers with increasing incidence trends in the United States: 1999 through 2008. *CA: A Cancer J. Clinicians* 62 (2), 118–128. doi:10.3322/caac.20141

Steeg, P. S. (2006). Tumor metastasis: mechanistic insights and clinical challenges. *Nat. Med.* 12 (8), 895–904. doi:10.1038/nm1469

Stott, S. L., Hsu, C.-H., Tsukrov, D. I., Yu, M., Miyamoto, D. T., Waltman, B. A., et al. (2010). Isolation of circulating tumor cells using a microvortex-generating herringbone-chip. *Proc. Natl. Acad. Sci.* 107 (43), 18392–18397. doi:10.1073/pnas.1012539107

Vidal-Vanaclocha, F. (2011). "Architectural and Functional Aspects of the Liver with Implications for Cancer Metastasis," in *Liver Metastasis: Biology and Clinical Management*. Editor P. Brodt (Dordrecht: Springer Netherlands), 9–42. doi:10.1007/978-94-007-0292-9_2

Warschkow, R., Tsai, C., Köhn, N., Erdem, S., Schmied, B., Nussbaum, D. P., et al. (2020). Role of lymphadenectomy, adjuvant chemotherapy, and treatment at high-volume centers in patients with resected pancreatic cancer-a distinct view on lymph node yield. *Langenbecks Arch. Surg.* 405 (1), 43–54. doi:10.1007/s00423-020-01859-2

Wheeler, A. A., and Nicholl, M. B. (2014). Age Influences Likelihood of Pancreatic Cancer Treatment, but not Outcome. *World J. Oncol.* 5 (1), 7–13. doi:10.14740/wjon789w

Yadav, D., and Lowenfels, A. B. (2013). The epidemiology of pancreatitis and pancreatic cancer. *Gastroenterology* 144 (6), 1252–1261. doi:10.1053/j.gastro.2013.01.068

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

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

frontiers | Frontiers in Big Data

ORIGINAL RESEARCH
published: 05 April 2022
doi: 10.3389/fdata.2022.863100

BurnoutEnsemble: Augmented Intelligence to Detect Indications for Burnout in Clinical Psychology

Ghofrane Merhbene[1], *Sukanya Nath*[2], *Alexandre R. Puttick*[1] *and Mascha Kurpicz-Briki*[1]*

[1] Applied Machine Intelligence Research Group, Department of Engineering and Information Technology, Bern University of Applied Sciences, Bern, Switzerland, [2] Institute for Research in Open, Distance and eLearning (IFeL), Swiss Distance University of Applied Sciences, Brig, Switzerland

Burnout, a state of emotional, physical, and mental exhaustion caused by excessive and prolonged stress, is a growing concern. It is known to occur when an individual feels overwhelmed, emotionally exhausted, and unable to meet the constant demands imposed upon them. Detecting burnout is not an easy task, in large part because symptoms can overlap with those of other illnesses or syndromes. The use of natural language processing (NLP) methods has the potential to mitigate the limitations of typical burnout detection *via* inventories. In this article, the performance of NLP methods on anonymized free text data samples collected from the online forum/social media platform Reddit was analyzed. A dataset consisting of 13,568 samples describing first-hand experiences, of which 352 are related to burnout and 979 to depression, was compiled. This work demonstrates the effectiveness of NLP and machine learning methods in detecting indicators for burnout. Finally, it improves upon standard baseline classifiers by building and training an ensemble classifier using two methods (subreddit and random batching). The best ensemble models attain a balanced accuracy of 0.93, test F1 score of 0.43, and test recall of 0.93. Both the subreddit and random batching ensembles outperform the single classifier baselines in the experimental setup.

Keywords: burnout, natural language processing, machine learning, augmented intelligence, ensemble classifier, psychology

1. INTRODUCTION

Stress at the workplace is an increasingly relevant topic. In a study involving almost 10,000 working adults in eight territories throughout Europe, it was found that 18% of the respondents feel stressed daily, and three out of ten participants feel so stressed that they have considered finding a new job (ADP, 2018). A Swiss study (SECO, 2015) estimates that 24.2% of employees feel often or always stressed at their workplace, while 35.2% feel mostly (22.2%) or always (13%) exhausted at the end of the working day. In the latter group, 25.5% still feel exhausted the next morning, a circumstance which, if prolonged, can lead to various health hazards. Studies from the United States give the same indication. The Stress in America's Report of 2019 by the American Psychological Association shows that Americans consider a healthy stress level at an average of 3.8 (scale ranging from 1 to 10, where 10 is "a great deal of stress" and 1 is "little or no stress;") however, they report having experienced an average stress level of 4.9 (American Psychological Association, 2019).

OPEN ACCESS

Edited by:
Kathiravan Srinivasan,
Vellore Institute of Technology, India

Reviewed by:
Juan Pedro Martínez-Ramón,
University of Murcia, Spain
Francisco Manuel Morales Rodríguez,
University of Granada, Spain

***Correspondence:**
Mascha Kurpicz-Briki
mascha.kurpicz@bfh.ch

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

Received: *26 January 2022*
Accepted: *25 February 2022*
Published: *05 April 2022*

Citation:
Merhbene G, Nath S, Puttick AR and Kurpicz-Briki M (2022) BurnoutEnsemble: Augmented Intelligence to Detect Indications for Burnout in Clinical Psychology. Front. Big Data 5:863100. doi: 10.3389/fdata.2022.863100

This stress can lead to workplace burnout. In 2019, the WHO included burnout in the 11th Revision of the International Classification of Diseases (ICD-11) as a syndrome.[1] In particular, during the pandemic crisis, burnout in the healthcare sector was an important issue: it has been shown, for instance, that the COVID-19 crisis has had an overwhelming psychological impact on intensive care workers (Azoulay et al., 2020).

Identifying burnout syndrome is complex because symptoms can overlap with other diseases or syndromes (Jaggi, 2019). In particular, the overlap between depression and burnout is an important subject of scientific discussion, e.g., (Schonfeld and Bianchi, 2016). In clinical intervention and field research, burnout is typically detected *via* the use of *inventories*. Potential burnout patients fill out a psychological test, usually in the form of a questionnaire with scaled-response answers (e.g., not at all, sometimes, often, very often). Although such inventories are used in most studies and are well-established in the clinical environment, major limitations have been identified. For example, in personality inventories, participants are liable to fake their results, e.g., (Holden, 2007). They may adapt their responses in high-stake situations in order to increase their chances for the desired outcome (Lambert, 2013). A further issue with inventories is known as *extreme response bias* (ERB); some respondents will tend to choose (or avoid) only the highest or the lowest options in such tests (Greenleaf, 1992; Brulé and Veenhoven, 2017). It has also been shown that on self-reported tests for subjective well-being, the respondent's mood during testing sometimes contributes as a predictor (Diener et al., 1991). Furthermore, defensiveness (the denial of symptoms) and social bias can influence the outcome of inventories (Williams et al., 2019).

A potential way to mitigate the existing and well-known problems with inventories is to explore the use of free text questions or transcribed interviews. Previous studies have demonstrated promise in such methods (Burisch, 2014), but, in practice, the manual effort of analyzing the resulting data often results in untenable overhead costs. Fortunately, recent developments in the field of natural language processing (NLP) make approaches using such unstructured textual data feasible. It has been shown that computational linguistic markers can be used to predict depressivity of the writer (Havigerová et al., 2019).

Existing work applying NLP to psychology focuses on the identification of indicators for different types of mental health disorders by using data obtained from social media, comprising the majority of available research in this area. For example, such work concentrates on suicide risk assessment (Morales et al., 2019), (Just et al., 2017), depression (Moreno et al., 2011), (Schwartz et al., 2014), post-partum depression (De Choudhury et al., 2013), (De Choudhury et al., 2014), or different mental health signals (Coppersmith et al., 2014). In some cases, data from Reddit online forums have been used, for example, to detect mental health disorders (Thorstad and Wolff, 2019), anxiety (Shen and Rudzicz, 2017), or depression (Tadesse et al., 2019).

However, very little work exists in the field of burnout detection. Burnout detection in data extracted from issues and comments posted within software development tools have been studied (Mäntylä et al., 2016). The authors used the valence-arousal-dominance (VAD) model to study burnout risk in a corporate setting. This model distinguishes three emotions: *valence* ("the pleasantness of a stimulus,") *arousal* ("the intensity of emotion provoked by a stimulus,") and *dominance* ("the degree of control exerted by a stimulus") (Warriner et al., 2013). To measure burnout risk, the metric is based on low valence and dominance and high arousal (Mäntylä et al., 2016). In other work, a first attempt to detect burnout based on patient and expert interviews in the German language were done; it was found that a combination of NLP and machine learning techniques in this field leads to promising results (Nath and Kurpicz-Briki, 2021).

In the context of earlier work focused on gathering data from social media websites and the study of mental health conditions, this work extends state-of-the-art predictive models in the field while focusing specifically on detecting indicators for burnout in data collected from Reddit. It aims to develop the base technology for potential new directions in tool development for clinical psychology. Herein, the authors emphasize that this work is oriented toward the approach of augmented intelligence rather than artificial intelligence (Rui, 2017); instead of replacing clinical professionals, it strives toward technology that empowers humans in the decision-making process, providing input to be considered in human decision-making.

The work in this article addresses the following objectives:

- It evaluates whether NLP methods applied to free text are an effective means to detect indicators for burnout, compared to a control group using general text samples, and a control group with depression-related texts.
- In particular, it investigates how the use of an ensemble classifier can leverage the accuracy of such methods.
- Furthermore, the approach is compared to single machine learning classifiers such as logistic regression.

This article is structured as follows: first, the materials and methods used in this work are discussed. In particular, this includes data collection, the characteristics of the datasets used in the experiments, and the experimental setup. Then, the results are presented, first for single classifier models and then for the ensemble models. Finally, the results are discussed and an outlook on potential future work is provided.

2. MATERIALS AND METHODS

2.1. Reddit Data Collection

On Reddit, users can organize posts based on a subject, so-called *subreddits*, which are online micro communities dedicated to a particular topic. Reddit has the advantage of allowing the possibility to create micro communities *via* subreddits. As a result, in addition to topics such as gaming and music, there are thriving communities dedicated to various mental health topics, such as depression, anxiety, and bipolar disorder. In particular, there is a subreddit dedicated to burnout; unfortunately, the number of entries was too low at the time of our data collection to provide a sufficiently large dataset. However, users discuss the subject of burnout in various other subreddit threads. One can

[1] https://icd.who.int/browse11/l-m/en#/http://id.who.int/icd/entity/129180281

thus collect textual data related to burnout by scraping Reddit for burnout-related posts. In this work, `praw` (Boe, 2011), a Python Reddit API Wrapper, was used to extract submissions with the keyword "burnout" and its different variations, such as "burnout," "burn out," "burned out," "burning out," "burnt out," "burn-out," etc., 1,536 such submissions were found.

However, the word *burnout* also widely occurs in other contexts, such as "The tires are burnt out." It is also frequently used in informal discussions, such as having *game burnout* or *music burnout*. It was therefore necessary to isolate submissions describing burnout in the professional or educational context. A total of 677 submissions satisfying these conditions were manually identified. The replies to the selected submissions were also collected, as they were likely to contain posts by other users describing their experiences with burnout. This increased the size of the dataset to 23,371 posts. However, not all of the posts and replies were relevant to professional or educational experiences with burnout. Therefore, 352 instances were extracted manually that describe burnout experiences from a first-person perspective. This formed the test group for the data classified as *burnout*.

To create the first control group, the *no burnout* dataset, our method employed the strategy described in Shen and Rudzicz (2017). Namely, 17,025 posts from a variety of subreddits were collected: "askscience", "relationships", "writingprompts", "teaching", "writing", "parenting", "atheism", "christianity", "showerthoughts", "jokes", "lifeprotips", "writing", "personalfinance", "talesfromretail", "theoryofreddit", "talesfromtechsupport", "randomkindness", "talesfromcallcenters", "books", "fitness", "askdocs", "frugal", "legaladvice", "youshouldknow", and "nostupidquestions", Since a number of these collected posts consisted of empty or very little text, all posts consisting of fewer than 100 characters were dropped, resulting in a final *no burnout* dataset consisting of 13,216 posts.

The second control group, the *depression* dataset, was collected from the subreddit for depression and contains 979 posts. As for burnout, only entries using the first-person perspective were selected.

The authors emphasize that no information concerning user identity (e.g., username or age) was collected.

2.2. Datasets for Experiments

Using the raw data consisting of 13,216 posts labeled *no burnout* (control group), 352 labeled *burnout*, and 979 labeled *depression*, four datasets for use in the experiments were compiled. Dataset statistics are presented in **Table 1**.

Dataset 1: Burnout vs. No Burnout (BNB): It combines the 13,216 *no burnout* posts with the 352 *burnout* posts, resulting in a highly unbalanced dataset of size 13,568.

Dataset 2: Burnout vs. No Burnout (Balanced) (BNB-Balanced): Balanced dataset of 704 posts, of which, 352 posts are selected from the *no burnout* dataset through random sampling (without replacement). Additionally, an equal number of 352 posts are added from the *burnout* data.

Dataset 3: Burnout vs. No Burnout (No Keywords) (BNB-No-Keywords): It is obtained from Dataset 2 by removing the keywords from the *burnout* dataset that were used during data collection to search for burnout-related posts: "burnout," "burn-out," "burning out," etc.

Dataset 4: Burnout vs. Depression (BD): Balanced dataset of 704 entries, of which 352 posts are selected from the *depression* dataset through random sampling (without replacement). Further, an equal number of 352 posts are added from the *burnout* data.

2.3. Vectorization

The `spacy`[2] Python NLP-library was used in order to vectorize text data for use in our NLP models. Each Reddit post was tokenized using the pre-trained `en_core_web_sm` English language pipeline and converted into a 500-dimensional bag-of-words vector, which simply counts the occurrences of each of the 500 most commonly appearing words in the text corpus.

2.4. Experimental Setup

2.4.1. Single Classifier Models

The following experiment was repeated on Datasets 1–4. The feature set consisted of the vectorized Reddit posts, each labeled with either 1 (burnout) or 0 (no burnout/depression). Using a 70-30% training-test split[3] and 10-fold cross-validation (CV), a variety of classifier models was trained: logistic regression, Support Vector Machine (SVM) (with linear, RBF, degree 3 polynomial and sigmoid kernels), and random forest. Each model's performance was measured by using the mean CV accuracy and F1 scores averaged across all folds, as well as the (balanced) accuracy, F1, and recall scores on the test data. It was chosen to specifically include recall as a metric because, in a real-world setting, it would be important to capture all possible *burnout* samples (recall $= 1$), even at the expense of a larger number of false positives (see Section 4.5 for further discussion).

2.4.2. Ensemble Classifier Models

Ensemble classifiers allow aggregating the decisions of several single classifier models. The ensemble methods presented in this work closely resemble a method known as UnderBagging (Barandela et al., 2003). Each ensemble is built according to the template below.

Ensemble model template:

- The ensemble consists of n submodels.[4]
- Each submodel is trained with 10-folds CV on a balanced dataset of 492 posts.
- These datasets share the same 246 *burnout* samples but contain pairwise disjoint sets of *no burnout* samples.
- The prediction of the whole ensemble is determined by voting, i.e., for a given test sample, a label of *burnout* is predicted if the

[2] https://spacy.io/

[3] For the unbalanced dataset, a 70–30% split was taken on each class separately in order to ensure that the training and test sets had roughly the same class distribution.

[4] In our experiments, $10 \leq n \leq 20$.

TABLE 1 | Dataset statistics.

Dataset Name	No. of Samples	Mean Text Length (chars)	Std. Dev of Text Length	Test Group %age	Control Group %age
1. Burnout vs. No Burnout(BNB)	13,568	1158	1451	2.6%	97.4%
2. Burnout vs. No Burnout (Bal.) (BNB-balanced)	704	867	850	50%	50%
3. Burnout vs. No Burnout (No KWs)(BNB-no-keywords)	704	863	846	50%	50%
4. Burnout vs. Depression (BD)	704	1009	905	50%	50%

"Control" refers to either no burnout or depression, while test refers to burnout.

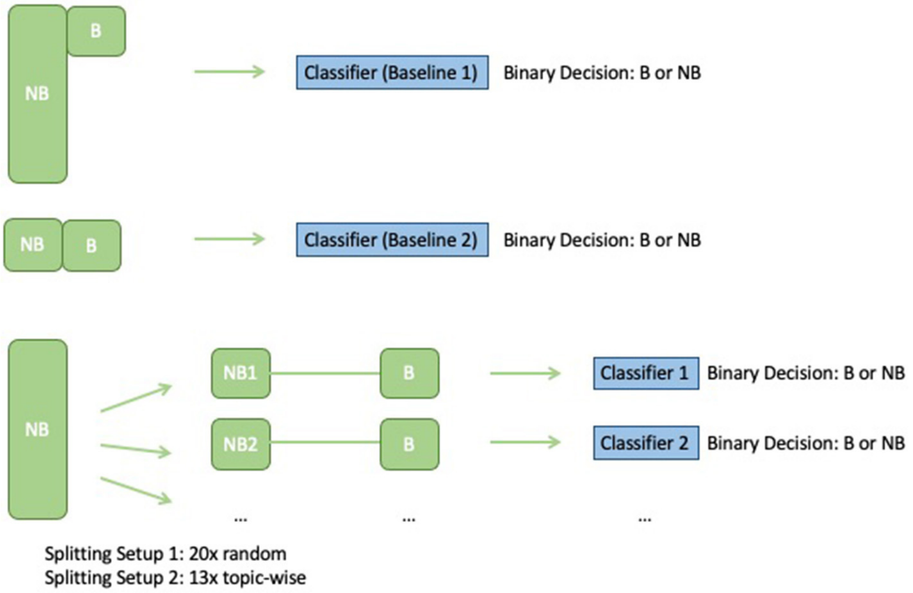


FIGURE 1 | Training the baselines vs. training ensembles on balanced batches.

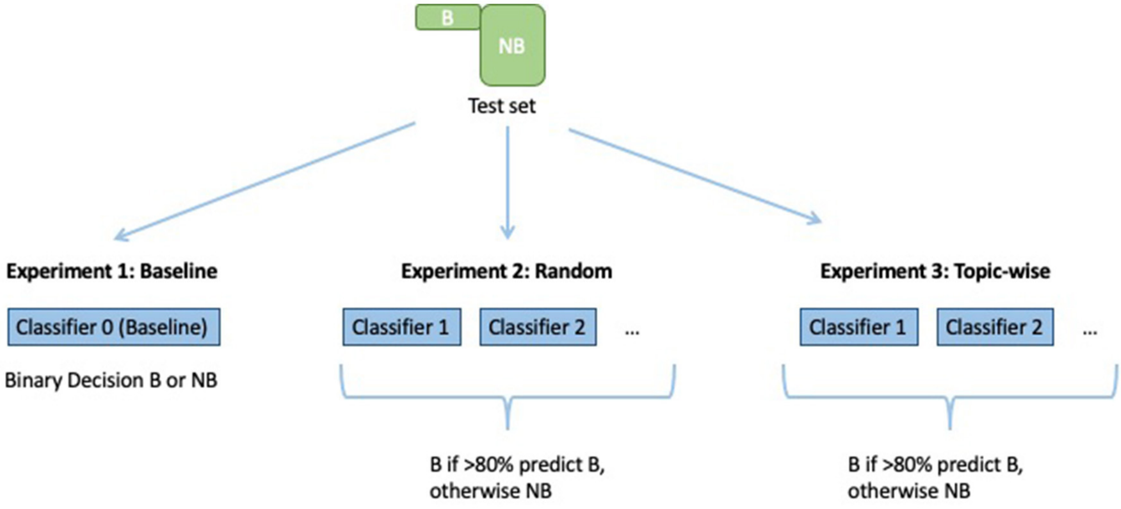


FIGURE 2 | Computing predictions on the test set, with a voting threshold of $p = 0.8$ for example purposes.

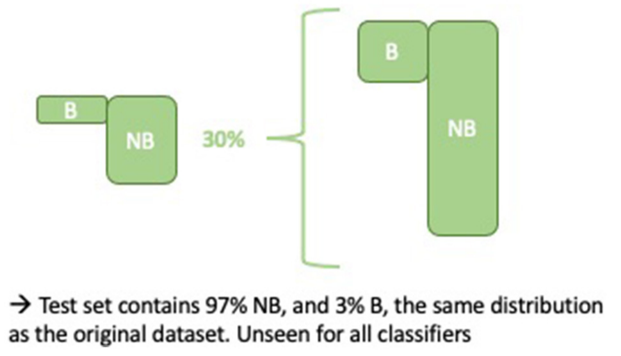

B
NB
30%
B
NB
→ Test set contains 97% NB, and 3% B, the same distribution as the original dataset. Unseen for all classifiers

FIGURE 3 | Constructing the unbalanced test set of 4,071 samples.

voting $>$ p% of the submodels classify the sample as *burnout*.[5] Otherwise, the sample is classified as *no burnout*.

The classifier type of the submodels was restricted to logistic regression, which demonstrated the most consistent performance in our initial experiments, although RBF, linear SVMs, and random forests also showed promise. **Figure 1** shows a depiction of our ensemble setup, along with a comparison to the two baseline models to which the ensemble results were compared:

- **Baseline 1:** Logistic regression classifier trained *via* 70–30% train-test split on the unbalanced Dataset 1 (BNB).
- **Baseline 2:** Logistic regression classifier trained on a balanced dataset obtained by randomly sampling 246 *no burnout* samples and combining them with the 246 *burnout* samples used for training.

Figure 1 depicts the setup for training the baseline classifier (logistic regression on the full unbalanced training set) and the ensemble classifiers, and **Figure 2** depicts how each model makes predictions on the test data.

Training and test data were allocated according to a 70-30% split. This was done in a stratified manner, i.e., the *no burnout* and *burnout* class distribution in the training and test data were approximately equal to the distribution in Dataset 1 (BNB) (as shown in **Figure 3**). Note that the same test data were used for both baselines and ensembles.

Two types of data batching were tested in our ensembles (Ensemble 1 and Ensemble 2, see description below) and the following metrics were measured:

- *Mean CV accuracy:*[6] Computed by first taking the mean CV accuracies for each submodel over the 10 folds, followed by averaging over the n submodels.
- *Mean CV F1 (macro):* Identical with F1 in place of accuracy.
- *Mean test balanced accuracy*: The balanced accuracy on test data averaged across the n submodels.
- *Mean test F1 (macro)*: Identical for F1.
- *Mean test recall*: Identical for recall.
- The corresponding SDs of the above three test metrics.

Ensemble 1: Random sample batching:

The *random sample batching* ensemble was trained using $n = 20$ batches, each consisting of 246 randomly sampled (without replacement) posts from the *no burnout* training samples concatenated with 246 burnout training samples to create BNB-balanced datasets.

Ensemble 2: Batching by subreddit:

The *subreddit batching* ensemble was trained by creating a balanced dataset corresponding to each of the subreddits appearing in the *no burnout* training data for which at least 246 samples had been collected. There were $n = 17$ such subreddits in total.

The effect of changing the voting threshold p on the ensemble performance was also tested. Values of $p = 0.4$, 0.5, 0.6, 0.7, 0.8, and 0.9 were evaluated.

3. RESULTS

3.1. Single Classifier Models

3.1.1. Burnout vs. No Burnout

The results of the single classifier experiments on Dataset 1 (Burnout vs. No Burnout BNB) are displayed in **Table 2**. The *Baseline* row corresponds to a model that predicts the label *no burnout* for all samples. Such a model achieves 97% accuracy due to the class imbalance in Dataset 1 (BNB). Indeed, accuracy is a misleading measure in such a situation: all classifiers in this experiment demonstrated an accuracy of approximately 97% despite large differences in performance. For this reason, balanced accuracy provides a more meaningful metric for model performance.

Only logistic regression and SVM linear demonstrated significant improvement over the baseline, although roughly 50% of burnout samples were incorrectly classified as *no burnout*.

3.1.2. Burnout vs. No Burnout (Balanced)

Here, the results of classifiers trained using Dataset 2 (BNB-balanced) are presented. Aside from the SVM poly degree 3 classifier, the models in **Table 3** appear to demonstrate good performance.[7] It was noted that these results are dependent on the random sample of *no burnout* data points that are used to construct Dataset 2 (BNB-balanced). While random forest classifiers demonstrated the best performance in this instance, there were also cases in which logistic regression performed best. For the best models, accuracies and F1 scores approximately distributed between 0.90 and 0.97 were observed.

3.1.3. Burnout vs. No Burnout (No Keywords) (BNB-No-Keywords)

The data collection process applied in this work explicitly searches for burnout-related keywords. It is, therefore, possible

[5] Here, p is a threshold that may vary. Values of $0.4 \leq p \leq 0.99$ were used.

[6] Here, we do not need to take the balanced accuracy because the submodels are trained on balanced datasets.

[7] The ensemble experiments revealed that logistic regression classifiers trained on Datasets 2 and 3 do not perform as well on a highly unbalanced data sampled from Dataset 1. This will be further discussed in Section 3.2.

TABLE 2 | Results for Burnout vs. No Burnout – Dataset 1 (BNB) (no. test samples = 4071).

Model	Mean CV Bal. Acc.	Mean CV F1	Test Bal. Acc.	Test F1	Test Recall
Logistic Regression	0.72	0.48	0.75	0.49	0.50
SVM Linear	0.72	0.40	0.75	0.45	0.51
SVM RBF	0.51	0.04	0.51	0.03	0.01
SVM Poly Degree 3	0.55	0.16	0.56	0.18	0.12
SVM Sigmoid	0.57	0.23	0.56	0.21	0.12
Random Forest	0.50	0.02	0.51	0.04	0.02
Baseline	0.50	0.0	0.50	0.0	0.0

Baseline refers to a model predicting no burnout for all samples. The mean CV statistics are computed by taking an average of overall 10 folds cross-validation (CV) during training.

TABLE 3 | Results for Burnout vs. No Burnout (Balanced)—Dataset 2 (BNB-balanced) (no. test samples = 234).

Model	Mean CV Accuracy	Mean CV F1	Test Accuracy	Test F1
Logistic regression	0.91	0.91	0.87	0.88
SVM Linear	0.89	0.89	0.84	0.85
SVM RBF	0.88	0.88	0.89	0.89
SVM Poly degree 3	0.60	0.35	0.60	0.41
SVM Sigmoid	0.85	0.85	0.82	0.82
Random Forest	0.92	0.92	0.88	0.89

The mean CV statistics are computed by taking an average of overall 10 folds CV during training.

TABLE 4 | Results for Burnout vs. No Burnout (no keywords)—Dataset 3 (BNB-no-key-words) (no. test samples = 234).

Model	Mean CV Accuracy	Mean CV F1	Test Accuracy	Test F1
Logistic regression	0.88	0.88	0.86	0.87
SVM Linear	0.85	0.85	0.82	0.83
SVM RBF	0.85	0.83	0.85	0.86
SVM Poly degree 3	0.59	0.34	0.59	0.40
SVM Sigmoid	0.79	0.80	0.81	0.82
Random Forest	0.88	0.88	0.87	0.88

The mean CV statistics are computed by taking an average of overall 10 folds CV during training.

that trained models identify the presence of such keywords as a key defining feature for posts belonging to the *burnout* class. The effect of the presence of such keywords was measured, and it was tested whether they provided a significant basis for the models' predictions. Therefore, all keywords related to burnout were removed from Dataset 2 (BNB-balanced) to obtain Dataset 3 (BNB-no-keywords) and the experiment was repeated. The corresponding results are displayed in **Table 4**. As one might expect, the removal of keywords resulted in decreased model performance. However, the decrease was not very important, providing evidence that the presence of keywords is not an overly important factor in any of our other experiments.

3.1.4. Burnout vs. Depression (BD)

In this experiment, as shown in **Table 5**, the Burnout vs. Depression dataset (Dataset 4, BD) was classified by using the models described previously. Again, it was found that logistic regression and SVM linear performed best, with random forest following closely. The datasets are balanced, and the random baseline for both accuracy and F1 score is set at 50%.

TABLE 5 | Results for Burnout vs. Depression—Dataset 4 (BD).

Model	Mean CV Accuracy	Mean CV F1	Test Accuracy	Test F1
Logistic regression	0.87	0.87	0.84	0.82
SVM Linear	0.84	0.85	0.82	0.78
SVM RBF	0.84	0.85	0.78	0.77
SVM Poly degree 3	0.59	0.42	0.65	0.43
SVM Sigmoid	0.82	0.82	0.78	0.76
Random Forest	0.85	0.86	0.81	0.80

The mean CV statistics are computed by taking an average of overall 10 folds CV during training.

Although these models perform well, an across-the-board decrease of roughly 0.04 points is observed compared to the results listed in **Table 3**.

3.2. Ensemble Models

Tables 6, **7** record metrics and statistics that pertain exclusively to the submodels and not to the overall ensembles. They are meant as a means to compare the performance of the individual submodels to that of the ensembles (**Table 8**).

Table 6 records the average CV performance metrics over the submodels within each of the ensemble classifiers. Recall that each of these submodels is a logistic regression classifier trained on a balanced dataset. The *Mean CV Accuracy* and *Mean CV F1* columns in **Table 6** are thus comparable to the corresponding columns in **Table 3**.

Table 7 records the average test statistics for the submodels. The *Mean test Bal. Acc.* and *Mean test F1* columns refer to the average performance of the submodels on the unbalanced test set consisting of 4,071 samples, of which 106 belong to the *burnout* class. It also provides the corresponding SDs. The random batching submodels were much more consistent than the subreddit batching submodels, the latter of which demonstrated greater variance and lower average balanced accuracy and F1 score while achieving higher recall scores.

TABLE 6 | Submodel CV averages.

Model	Mean CV Accuracy	Mean CV F1
Random Batching	0.91	0.90
Subreddit Batching	0.96	0.96

The test results reveal the limitations of the previously presented non-ensemble models trained on balanced data. Those models appeared to demonstrate very good performance on unseen test data (**Table 3**), but were tested on small balanced test sets consisting of only 234 posts. In comparison, the mean test F1 scores in **Table 7** are relatively low, which shows that the high test performance observed in **Table 3** does not imply similar performance on the unbalanced dataset of 4,071 samples. The high test recall in **Table 7** indicates that the test F1 scores are primarily reduced due to low precision, i.e., a relatively large number of false positives.

The mean test metrics in **Table 7** corresponding to random batching give an indication of how the logistic regression model trained on Dataset 2 (BNB-balanced) (**Table 3**) would perform on the large unbalanced test dataset used in our ensemble experiments. Note that the performance of the balanced data model depends on the random sample of 352 *no burnout* posts used to construct Dataset 2 (BNB-balanced), and significant fluctuations in performance were observed depending on the sample, encapsulated in the SDs recorded in **Table 7**. Indeed, the pursuit of ensemble approaches presented in this work was driven partially by the desire for a model with more stable performance. Effectively, **Table 7** portrays the average performance of logistic regression models trained on balanced No Burnout vs. Burnout data over $n = 20$ disjoint random samples of *no burnout* data. It was observed that the submodels trained *via* subreddit batching demonstrated lower performance on the test data than those trained *via* random batching.

Table 8 shows the test results of the two ensemble models. The logistic regression (LR) model trained on Dataset 1 (BNB) was used as a first baseline, which demonstrated the best overall performance on the unbalanced test data among the single-model classifiers. As a second baseline, a single logistic regression classifier trained on a balanced dataset obtained by randomly undersampling from the *no burnout* class was considered, as was done to construct Dataset 2 (BNB-balanced). The model demonstrated performance similar to the averages recorded in **Table 7**.

The ensemble models demonstrated substantially improved balanced accuracy and recall relative to the baseline unbalanced LR model. However, the unbalanced LR model achieved the second-highest F1 score. Both ensembles and the baseline random undersampling LR demonstrated similar performance, with the random batching ensemble exhibiting a trade-off between F1 and recall. In comparing **Tables 7**, **8**, one can see that the random batching ensemble demonstrates a relatively modest performance improvement over the submodels composing it. On the other hand, the subreddit batching ensemble performs markedly better than its component submodels.

The confusion matrices in **Figures 4**, **5** describe the distribution of the ensemble models' test predictions. Both the submodels and ensembles had test recall scores near 1. However, the high recall of the submodels came at the cost of a large number of false positives. It was observed that each submodel identified approximately 400–500 (random batching) to 1,000–3,000 (subreddit batching) test samples as belonging to the *burnout* class, whereas the correct number was 106. In contrast, with a majority vote threshold of $p = 80\%$, the ensemble models placed 216 (random batching). There were 486 (subreddit batching) test samples in the burnout class while maintaining a recall score close to 1. The majority vote ensemble rule is thus an effective method for eliminating false positives while preserving true positives.

The effect of modifying the voting threshold on performance was also tested. The results are depicted in **Figure 6**.

With random batching, a trade-off between recall/balanced accuracy and F1 score was experienced; while subreddit batching demonstrated a trade-off between recall and F1 score, balanced accuracy, and F1 score could be simultaneously improved, with increased voting threshold.

In practice, such approaches are interested in capturing as many burnout samples as possible while maintaining a manageable number of false positives. The threshold can be modified accordingly, for example, aiming to maximize the F1 score under the condition that recall is greater than 0.9. The subreddit batching ensemble with $p = 0.85$ and the random batching ensemble with $p = 0.5$ both demonstrated performance close to such an optimum (as shown in **Table 9**). Both of these ensembles achieve better results than either of the baseline models.

Finally, a qualitative analysis of test samples incorrectly classified as burnout by the ensemble models revealed posts from the no-burnout dataset that contained topics similar to burnout posts, e.g., work-related, stress, depression, and anxiety. This indicates that the classifiers presented in this work are indeed identifying features related to burnout. It even appears that, in some cases, it may be the labels rather than the predictions that are incorrect, i.e., posts from scraped sub-breddits where users write about experience with burnout.

4. DISCUSSION

The work presented in this article makes the following contributions:

- It demonstrates that NLP methods applied to free text are an effective means to detect indicators for burnout, measured against both a control group of general text and a group composed of text samples related to depression.
- A machine learning ensemble classifier trained on data from Reddit posts to detect burnout indicators with a promising accuracy is presented.
- A range of machine learning classifiers trained to detect burnout indicators are compared, showing in particular that the presented ensemble classifier outperforms two single classifier baselines: logistic regression classifiers trained on

TABLE 7 | Submodel test statistics (no. test samples = 4,071).

Model	Mean Test Bal. Accuracy	Std. Dev. Test Bal. Acc.	Mean Test F1	Std. Dev. Test F1	Mean Test Recall	Std. Dev. Test Recall
Random Batching	0.91	0.01	0.35	0.02	0.91	0.02
Subreddit Batching	0.78	0.08	0.13	0.04	0.96	0.02

either a large unbalanced dataset or an undersampled balanced dataset. The best-performing model attained a balanced accuracy of 93%, F1 score of 0.43, and recall of 93% on unbalanced test data.

These findings have a large potential to be further developed with an interdisciplinary approach toward a new generation of smart tools for clinical psychology, eventually supporting a wider array of conditions and mental health diagnoses in the future.

4.1. Burnout Detection for a Clinical Setting

Extracting data from social media is one of the most commonly used methods in research in this area (e.g., Shen and Rudzicz, 2017; Thorstad and Wolff, 2019). The research presented in this article also relies primarily on data extracted from the social media website Reddit, particularly because it was easy to obtain a large quantity of data to train our model. Nonetheless, clinical data are a more reliable source for detecting burnout due to the certainty of labeling. Clinical data also have the advantage of more closely resembling the data such models are expected to be applied to in the future. A first attempt of working with clinical data to detect burnout has shown promising results. By presenting a dataset from real-world burnout patient data, Nath and Kurpicz-Briki (2021) managed to go beyond typical burnout detection approaches, which usually includes the use of inventories with scaling questions and worked on applying NLP to mental health. The dataset consisted of data extracted from German-language interviews with burnout patients, a control group, and experts. The authors proceeded to train an SVM classifier on the dataset and ended up achieving accuracy greater than their original baseline.

4.2. Burnout vs. Depression

A poorer classifier performance on Dataset 4 than on Datasets 2 or 3 was observed. This is likely due to the fact that depression- and burnout-related texts share many similar characteristics. Indeed, depression and burnout are not disjoint categories, and some degree of classification ambiguity is inevitable. This overlap is a significant object of scientific investigation, e.g., by Schonfeld and Bianchi (2016). The work in this article provides evidence of the non-trivial nature of differentiating burnout and depression. Ongoing work of the authors aims to more closely analyze the markers that indicate and differentiate depression and burnout in free text first-person accounts.

4.3. Methods for Dealing With Unbalanced Data

Class imbalance is a natural phenomenon in many real-world applications (e.g., fraud detection, tumor detection, software defect prediction). It is well-documented in machine learning literature that unbalanced training data impairs the classification performance of many machine learning models (e.g., Chawla et al., 2004; García et al., 2010). For example, in cases of extreme class imbalance, models can tend toward placing all samples in the majority class. For a detailed survey on the unbalanced data problem, refer to He and Garcia (2009). Class imbalance is considered to be intrinsic to the task of burnout detection from real-world (clinical) data, rather than being an artifact of the data collection methods used in this article, and, therefore, it was aimed to address the problem in this work.

TABLE 8 | Ensemble vs. baseline performance (Threshold $p = 80\%$, no. test samples = 4,071).

Model	Test Bal. Acc.	Test F1	Test Recall
Random Batching Ensemble	0.91	0.56	0.84
Subreddit Batching Ensemble	0.93	0.34	0.95
Baseline 1: Unbalanced LR	0.75	0.49	0.50
Baseline 2: Random Undersampling LR	0.90	0.33	0.91

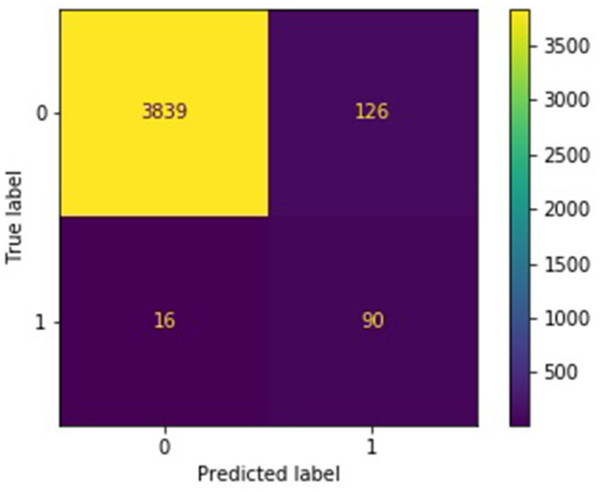


FIGURE 4 | Confusion matrix for random batching ensemble, $p = 0.8$.

Common solutions involve oversampling the minority class or undersampling the majority class to achieve class balance or using cost-sensitive methods that apply a higher penalty to the incorrect classification of samples from the minority class. A number of ensemble methods use oversampling and/or undersampling to train separate models and aggregate their predictions. Successful ensemble methods for unbalanced learning include EasyEnsemble (Liu et al., 2008), SMOTE-Boost

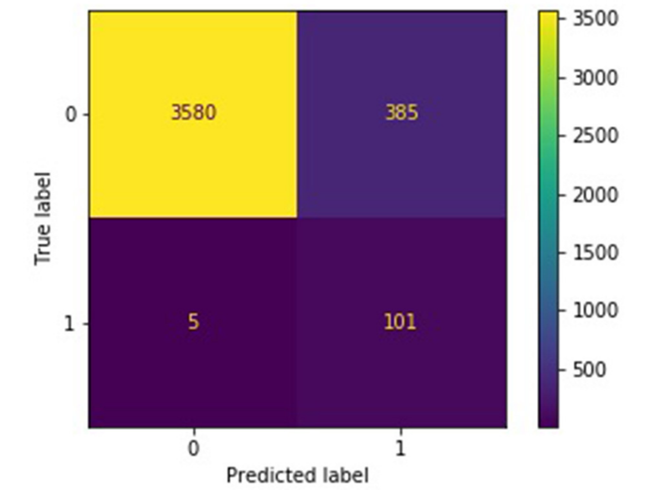


FIGURE 5 | Confusion matrix for subreddit batching ensemble, $p = 0.8$.

(Chawla et al., 2003), UnderBagging (Barandela et al., 2003), and Cluster/SplitBal (Sun et al., 2015).

Sun et al. (2015) argue that most existing methods might suffer from the loss of potentially useful information and/or overfitting by altering the original data distribution. Of the ensemble methods explored in this work, only UnderBagging, ClusterBal, and SplitBal do not discard data or change the data distribution. These three methods differ mainly in how balanced data batches are constructed and how the predictions of the submodels are aggregated. The method presented in this article is most similar to that of UnderBagging, which was chosen for the ease of implementation in the given setting and the fact that (Sun et al., 2015) found that it performs well across several classifier types. The method presented in this article differs only in that different voting thresholds are considered, not all of the majority class samples are exhausted, and balanced batches based on subgroupings inherent in the presented dataset (subreddits) are constructed.

The single model experiments reflect some of the problems of class imbalance. The best classifiers trained on Dataset 1 reached lower benchmark metrics but demonstrated more consistent performance between training and test data. This is consistent with the expectation that larger training datasets generalize better. Many of the classifiers trained on the unbalanced Dataset 1 performed very poorly, essentially predicting only the majority class. On the other hand, classifiers trained on the balanced Dataset 2 attained a high benchmark performance on relatively small balanced data batches, but that performance dropped considerably (as measured by F1 scores) when applied to highly unbalanced test data. The balanced data models use undersampling and demonstrate the drawbacks of throwing out data points: much of the variance in the *no burnout* dataset is not accounted for, and the undersampling-based models incorrectly classified a relatively large number of more general *no burnout* data. As one would expect, this effect is most pronounced in the models trained using a single subreddit, where a very specialized sample of *no burnout* data were used for training.

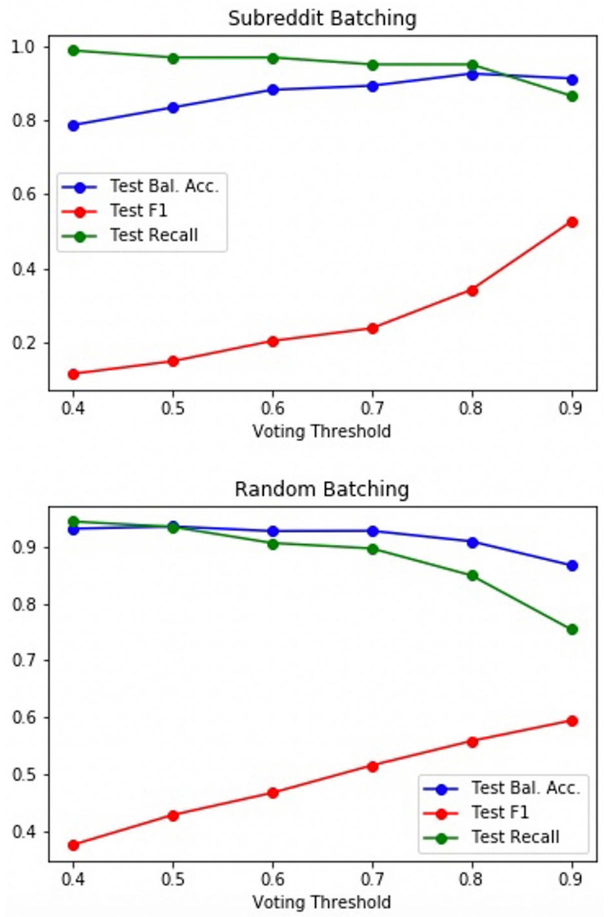


FIGURE 6 | Ensemble performance vs. voting threshold.

Overall, the presented results provide evidence that both undersampling—as long as attention is paid to maintaining the variance in the majority class data—and ensemble methods are viable approaches to handling the unbalanced data problem in this context. The single logistic regression classifiers trained on undersampled, balanced data performed at a level similar to the ensembles, although the subreddit batching ensemble with $p = 0.85$ and random batching ensemble with $p = 0.5$ both outperformed the single random batching classifier in all three metrics. Undersampling does have the advantage of requiring many fewer training data with both faster training and inference, although this speed difference can be erased by running ensemble submodels in parallel. However, better performance was achieved with ensembles. The ensemble methods provide additional advantages: the voting threshold hyperparameter allows to easily fine-tune the ensemble model according to the relative importance placed on recall and F1 score; in addition, the performance of the ensemble model is more stable, i.e., immune to fluctuations according to the subsample of *no burnout* data used for training.

TABLE 9 | Optimal ensembles (no. test samples = 4,071).

Model	Test Bal. Acc.	Test F1	Test Recall
Subreddit Batching ($p = 0.85$)	0.93	0.43	0.93
Random Batching ($p = 0.5$)	0.93	0.42	0.93

4.4. Random vs. Subreddit Batching

As similar performance with both methods for creating balanced data batches was achieved, the experiments do not indicate which, if either, of the two procedures is preferable. However, it was noted that several differences between the two methods exist. Perhaps the most important difference is that the subreddit batching ensembles required fewer training data to achieve the same performance. In addition, as **Table 6** shows that the submodels in the subreddit batching ensemble achieved higher accuracy and F1 score during CV, which might result from the relative ease of distinguishing between burnout-related posts and a single specialized topic with little relation to the condition. This results in overfitting, as reflected in the gap between CV and test results recorded in **Tables 6**, **7**. **Table 7** also shows that the performance of the individual submodels in the subreddit batching ensemble varied much more than for random batching; a comparison with **Table 8** also shows that the relative gain achieved by using ensemble methods over single classifiers was much greater in the case of subreddit batching. This is consistent with expectations and findings in the literature, which suggest that ensembles are an effective method for combining weak learners with considerable variance in their predictions into a strong learner (Schapire, 1990). It is also possible that the subreddits that were excluded from the ensemble due to an insufficient number of posts are over-represented among the misclassified samples and that performance could be improved by including more subreddits. Experiments in this direction are suggested for future work.

4.5. Recall as an Evaluation Metric

The use of recall as an evaluation metric was chosen because it is assumed that recall is of great importance in real-world applications. In the case of burnout detection, it is better to capture most or all of the true positives at the cost of a manageable number of false positives than to miss positive cases. In practice, marking individuals who are potentially experiencing burnout should help mental health professionals decide which cases should be subjected to further analysis. For this reason, even though the Baseline 1: Unbalanced LR model attained an F1 score better than or on par with the other models (**Table 8**), the significantly lower recall score makes this model unequivocally the least desirable. A tool that misses half of the patients demonstrating potential burnout is not useful.

4.6. Limitations

In this work, data procured from Reddit posts were used largely because of the ease in obtaining large quantities of data for use in model training. It is expected in the future to apply these methods to the verbal responses of patients in clinical interventions in order to train models to their destined target application. Therefore, the data origin is a limitation of this study. Obtaining a sufficient quantity (and in different local languages) of data for machine learning-based methods poses a significant challenge and will be addressed in future work by other data collection methods, involving also clinical institutions. The authors intend to collaborate with researchers and practitioners in psychology for data collection and to aid in developing a beneficial, easy-to-use clinical tool as well as expanding their work toward other areas of mental health. Another limitation of this work is the diversity in the available data. Being completely anonymous data from online forums, no information about gender, origin, socio-economic background, or similar is available. Therefore, the classifiers presented in this work may not work with the same efficiency for different groups of society. In future work, and before implementing such methods into a product, further validations and potentially additional training data will be required.

4.7. Future Work

In future work, the authors would like to experiment with more sophisticated ensemble methods, such as those outlined in Sun et al. (2015), where the general superiority of ensembles over other methods for addressing the class imbalance in several experiments was demonstrated. Since undersampling also showed promising results, more sophisticated methods for undersampling should be explored, such as clustering-based methods (Lin et al., 2017). However, the low variance observed among the random batching submodels may delineate the limits of undersampling-based methods. Furthermore, the use of classifier types beyond logistic regression could be explored, perhaps by incorporating neural network-based models and using other methods for creating balanced data batches for submodel training. Mixing different types of classifiers within an ensemble could be a means to capture *burnout* samples that are otherwise overlooked by logistic regression. It should also be considered to experiment with other vectorization methods in the future, particularly the use of word embeddings learned from deep learning-based language models, such as Word2Vec (Mikolov et al., 2017), GLoVe (Pennington et al., 2014), BERT (Devlin et al., 2019), and fastText (Joulin et al., 2016).

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because the collected online data can be subject to change (e.g., deletions) over time. A similar dataset can be created following the instructions in the article.

AUTHOR CONTRIBUTIONS

MK-B is the principal investigator of the project and brought the original idea. SN was the main contributor for data collection and the first experiments, in particular, the single classifier models. AP and GM contributed by extending the single classifier models and by developing the ensemble models in collaboration with MK-B. All authors contributed to idea development, experimental setup, and paper writing/editing. All authors contributed to the article and approved the submitted version.

FUNDING

The authors gratefully acknowledge the funding for this work by the Swiss National Science Foundation (SNSF) through an SNSF Spark grant.

REFERENCES

ADP, A. D. P. (2018). *The Workforce View in Europe 2018.* (accessed December 29, 2021).

American Psychological Association (2019). *Stress in America: Stress and Current Events.* Stress in America^{TM} Survey. Available online at: https://www.apa.org/news/press/releases/stress/2019/stress-america-2019.pdf

Azoulay, E., De Waele, J., Ferrer, R., Staudinger, T., Borkowska, M., Povoa, P., et al. (2020). Symptoms of burnout in intensive care unit specialists facing the covid-19 outbreak. *Ann. Intensive Care* 10, 1–8. doi: 10.1186/s13613-020-00722-3

Barandela, R., Valdovinos, R., and Sánchez, J. (2003). New applications of ensembles of classifiers. *Pattern Anal. Appl.* 6, 245–256. doi: 10.1007/s10044-003-0192-z

Boe, B. (2011). *PRAW the Python Reddit Api Wrapper.* (accessed January 14, 2022).

Brulé, G., and Veenhoven, R. (2017). The '10 excess' phenomenon in responses to survey questions on happiness. *Soc. Indicators Res.* 131, 853–870. doi: 10.1007/s11205-016-1265-x

Burisch, M. (2014). *Das Burnout-Syndrom*. Berlin; Heidelberg: Springer.

Chawla, N. V., Japkowicz, N., and Kotcz, A. (2004). Special issue on learning from imbalanced data sets. *ACM SIGKDD Explor. Newslett.* 6, 1–6. doi: 10.1145/1007730.1007733

Chawla, N. V., Lazarevic, A., Hall, L. O., and Bowyer, K. W. (2003). "Smoteboost: improving prediction of the minority class in boosting," in *European Conference on Principles of Data Mining and Knowledge Discovery* (Cavtat: Springer), 107–119.

Coppersmith, G., Dredze, M., and Harman, C. (2014). "Quantifying mental health signals in twitter," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (Baltimore, MD), 51–60.

De Choudhury, M., Counts, S., and Horvitz, E. (2013). "Predicting postpartum changes in emotion and behavior via social media," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY), 3267–3276.

De Choudhury, M., Counts, S., Horvitz, E. J., and Hoff, A. (2014). "Characterizing and predicting postpartum depression from shared facebook data," in *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing* (New York, NY), 626–638.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). "Bert: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT (1)* (Minneapolis, MN), 4171–4186.

Diener, E., Sandvik, E., Pavot, W., and Gallagher, D. (1991). Response artifacts in the measurement of subjective well-being. *Soc. Indicators Res.* 24, 35–56.

García, V., Mollineda, R. A., and Sánchez, J. S. (2010). "Theoretical analysis of a performance measure for imbalanced data," in *2010 20th International Conference on Pattern Recognition* (Istanbul: IEEE), 617–620.

Greenleaf, E. A. (1992). Measuring extreme response style. *Publ. Opin. Q.* 56, 328–351.

Havigerová, J. M., Haviger, J., Kučera, D., and Hoffmannová, P. (2019). Text-based detection of the risk of depression. *Front. Psychol.* 10, 513. doi: 10.3389/fpsyg.2019.00513

He, H., and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* 21, 1263–1284. doi: 10.1109/TKDE.2008.239

Holden, R. R. (2007). Socially desirable responding does moderate personality scale validity both in experimental and in nonexperimental contexts. *Can. J. Behav. Sci./Revue canadienne des sciences du comportement* 39, 184. doi: 10.1037/cjbs2007015

Jaggi, F. (2019). *Burnout Praxisnah*. Lehmanns Media.

Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., and Mikolov, T. (2016). Fasttext. zip: Compressing text classification models. *arXiv preprint* arXiv:1612.03651.

Just, M. A., Pan, L., Cherkassky, V. L., McMakin, D. L., Cha, C., Nock, M. K., et al. (2017). Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth. *Nat. Hum. Behav.* 1, 911–919. doi: 10.1038/s41562-017-0234-y

Lambert, C. E. (2013). *Identifying Faking on Self-Report Personality Inventories: Relative Merits of Traditional Lie Scales, New Lie Scales, Response Patterns, and Response Times* (Kingston, ON: Queen's University).

Lin, W.-C., Tsai, C.-F., Hu, Y.-H., and Jhang, J.-S. (2017). Clustering-based undersampling in class-imbalanced data. *Inf. Sci.* 409, 17–26. doi: 10.1016/j.ins.2017.05.008

Liu, X.-Y., Wu, J., and Zhou, Z.-H. (2008). Exploratory undersampling for class-imbalance learning. *IEEE Trans. Syst. Man Cybern. B (Cybern.)* 39, 539–550. doi: 10.1109/TSMCB.2008.2007853

Mäntylä, M., Adams, B., Destefanis, G., Graziotin, D., and Ortu, M. (2016). Mining valence, arousal, and dominance - possibilities for detecting burnout and productivity? *CoRR*, abs/1603.04287.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. (2017). Advances in pre-training distributed word representations. *arXiv preprint* arXiv:1712.09405.

Morales, M., Dey, P., Theisen, T., Belitz, D., and Chernova, N. (2019). "An investigation of deep learning systems for suicide risk assessment," in *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology* (Minneapolis, MN), 177–181.

Moreno, M. A., Jelenchick, L. A., Egan, K. G., Cox, E., Young, H., Gannon, K. E., et al. (2011). Feeling bad on facebook: depression disclosures by college students on a social networking site. *Depress. Anxiety* 28, 447–455. doi: 10.1002/da.20805

Nath, S., and Kurpicz-Briki, M. (2021). "Burnoutwords - detecting burnout for a clinical setting," in *Proceedings of the 10th International Conference on Soft Computing, Artificial Intelligence and Applications (SCAI 2021), CS & IT Conference Proceedings* (Zurich), vol. 11, 177–191.

Pennington, J., Socher, R., and Manning, C. D. (2014). "Glove: global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Doha), 1532–1543.

Rui, Y. (2017). From artificial intelligence to augmented intelligence. *IEEE MultiMedia* 24, 4–5. doi: 10.1109/MMUL.2017.8

Schapire, R. E. (1990). The strength of weak learnability. *Mach. Learn.* 5, 197–227.

Schonfeld, I. S., and Bianchi, R. (2016). Burnout and depression: two entities or one? *J. Clin. Psychol.* 72, 22–37. doi: 10.1002/jclp.22229

Schwartz, H. A., Eichstaedt, J., Kern, M., Park, G., Sap, M., Stillwell, D., et al. (2014). "Towards assessing changes in degree of depression through facebook," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From linguistic Signal to Clinical Reality* (Baltimore, MD), 118–125.

SECO (2015). *The Sixth European Working Conditions Survey (EWCS).* (accessed December 29, 2021).

Shen, J. H., and Rudzicz, F. (2017). "Detecting anxiety through reddit," in *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology-From Linguistic Signal to Clinical Reality* (Vancouver, BC), 58–65.

Sun, Z., Song, Q., Zhu, X., Sun, H., Xu, B., and Zhou, Y. (2015). A novel ensemble method for classifying imbalanced data. *Pattern Recogn.* 48, 1623–1637. doi: 10.1155/2017/1827016

Tadesse, M. M., Lin, H., Xu, B., and Yang, L. (2019). Detection of depression-related posts in reddit social media forum. *IEEE Access* 7, 44883–44893. doi: 10.1109/ACCESS.2019.2909180

Thorstad, R., and Wolff, P. (2019). Predicting future mental illness from social media: a big-data approach. *Behav. Res. Meth.* 51, 1586–1600. doi: 10.3758/s13428-019-01235-z

Warriner, A. B., Kuperman, V., and Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behav. Res. Meth.* 45, 1191–1207. doi: 10.3758/s13428-012-0314-x

Williams, M. M., Rogers, R., Sharf, A. J., and Ross, C. A. (2019). Faking good: an investigation of social desirability and defensiveness in an inpatient sample with personality disorder traits. *J. Pers. Assess.* 101, 253–263. doi: 10.1080/00223891.2018.1455691

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.

SYSTEMATIC REVIEW
published: 29 June 2022
doi: 10.3389/fdata.2022.927520

The Oral Microbiome and Its Role in Systemic Autoimmune Diseases: A Systematic Review of Big Data Analysis

Lu Gao[1,2,3†], *Zijian Cheng*[1,2,3†], *Fudong Zhu*[1,2,3], *Chunsheng Bi*[1,2,3], *Qiongling Shi*[1,2,3] *and Xiaoyan Chen*[1,2,3*]

[1] Stomatology Hospital, School of Stomatology, Zhejiang University School of Medicine, Hangzhou, China, [2] Zhejiang Provincial Clinical Research Center for Oral Diseases, Hangzhou, China, [3] Key Laboratory of Oral Biomedical Research of Zhejiang Province, Cancer Center of Zhejiang University, Hangzhou, China

Introduction: Despite decades of research, systemic autoimmune diseases (SADs) continue to be a major global health concern and the etiology of these diseases is still not clear. To date, with the development of high-throughput techniques, increasing evidence indicated a key role of oral microbiome in the pathogenesis of SADs, and the alterations of oral microbiome may contribute to the disease emergence or evolution. This review is to present the latest knowledge on the relationship between the oral microbiome and SADs, focusing on the multiomics data generated from a large set of samples.

Methodology: By searching the PubMed and Embase databases, studies that investigated the oral microbiome of SADs, including systemic lupus erythematosus (SLE), rheumatoid arthritis (RA), and Sjögren's syndrome (SS), were systematically reviewed according to the PRISMA guidelines.

Results: One thousand and thirty-eight studies were found, and 25 studies were included: three referred to SLE, 12 referred to RA, nine referred to SS, and one to both SLE and SS. The 16S rRNA sequencing was the most frequent technique used. HOMD was the most common database aligned to and QIIME was the most popular pipeline for downstream analysis. Alterations in bacterial composition and population have been found in the oral samples of patients with SAD compared with the healthy controls. Results regarding candidate pathogens were not always in accordance, but *Selenomonas* and *Veillonella* were found significantly increased in three SADs, and *Streptococcus* was significantly decreased in the SADs compared with controls.

Conclusion: A large amount of sequencing data was collected from patients with SAD and controls in this systematic review. Oral microbial dysbiosis had been identified in these SADs, although the dysbiosis features were different among studies. There was a lack of standardized study methodology for each study from the inclusion criteria, sample type, sequencing platform, and referred database to downstream analysis pipeline and cutoff. Besides the genomics, transcriptomics, proteomics, and metabolomics

Edited by:
Jianhua Wu,
University of Leeds, United Kingdom

Reviewed by:
Jonathan Vernon,
University of Leeds, United Kingdom
Ying An,
Fourth Military Medical University, China

***Correspondence:**
Xiaoyan Chen
ortho_chenxy@zju.edu.cn

[†] These authors have contributed equally to this work

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

Received: *24 April 2022*
Accepted: *23 May 2022*
Published: *29 June 2022*

Citation:
Gao L, Cheng Z, Zhu F, Bi C, Shi Q and Chen X (2022) The Oral Microbiome and Its Role in Systemic Autoimmune Diseases: A Systematic Review of Big Data Analysis. Front. Big Data 5:927520. doi: 10.3389/fdata.2022.927520

technology should be used to investigate the oral microbiome of patients with SADs and also the at-risk individuals of disease development, which may provide us with a better understanding of the etiology of SADs and promote the development of the novel therapies.

Keywords: oral microbiome, systemic autoimmune disease, systemic lupus erythematosus, rheumatoid arthritis, Sjögren's syndrome, high-throughput analysis

INTRODUCTION

Autoimmune diseases are a heterogeneous group of multifactorial disorders characterized by abnormal immune responses to the body's own cells or tissues (Bolon, 2012). Generally, the immune system can distinguish foreign pathogens from the body's own cells and tissues and thus does not respond to the biomolecules expressed in endogenous tissues, which is so called "self-tolerance" (Ahsan, 2017). When the self-tolerance is damaged, the immune system will produce autoantibodies binding to the target tissues and cause destruction (Xiao et al., 2021). Autoimmune diseases can be classified into organ-specific and systemic autoimmune diseases based on the range of tissues targeted by autoantibodies (Inanç, 2020). The common systemic autoimmune diseases (SADs) include rheumatoid arthritis (RA), systemic lupus erythematosus (SLE), and Sjögren's Syndrome (SS), affecting more than 5% of people worldwide (Van Loveren et al., 2001), women predominantly (Credendino et al., 2020; Willame et al., 2021). SADs can cause chronic, systemic, excessive immune response and inflammation, resulting in a series of mild to life-threatening symptoms, such as fatigue, dizziness, malaise, fever, neurological problems, anemia, and thrombocytopenia (Wang et al., 2015). Although the symptoms can be managed by the treatment, there are no cures for SADs currently. Treatment depends on the type of disease but often includes immune suppression, which can lead to compromised immunity and vulnerability to other diseases after long-term use (Ostrov, 2015). Although a complex interplay of variable genetic risks, environmental factors, and hormonal factors is thought to contribute to breaking the immunological tolerance, the etiology of SADs remain undefined, and more effective therapies are needed (Wahren-Herlenius and Dörner, 2013).

Autoimmunity develops in the context of the human microbiome, which is defined as the full complement of microorganisms and its collective genetic materials at a particular location (Ursell et al., 2012). Inside the human body, the oral microbiome is considered to be the second largest and diverse microbiome following the gut microbiome (Verma et al., 2018). The oral microbiome comprises billions of microorganisms composed of more than 700 species of bacteria, as well as fungi, viruses, and protozoa (Deo and Deshmukh, 2019). The oral microbiome can have an impact on the general health of an individual (Lamont et al., 2018). Periodontitis, a microbially-induced inflammatory condition that causes damage to the supporting tissues of the teeth, alongside its related pathogens, may be a risk factor for cardiovascular diseases (Tonetti and Van Dyke, 2013), preterm or low birth weight babies (Teshome and Yitayeh, 2016), rheumatoid arthritis (de Molon et al., 2019), or diabetes (Sanz et al., 2018). Oral bacteria can act as opportunistic pathogens at distant sites in the body, e.g., following entry to the bloodstream (bacteraemia) or aspiration into the lungs (Potgieter et al., 2015).

To date, with the development of high-throughput techniques and the availability of multi-omics data generated from a large set of samples, increasing studies have tried to investigate the link between microbiome and SADs, suggesting that perturbations of the oral microbiome may influence the emergence or evolution of autoimmunity (Chu et al., 2021; Doaré et al., 2021). However, it is undefined whether the oral microbial dysbiosis is a consequence of bad oral hygiene or periodontitis. There are many different high-throughput techniques, analysis pipelines, and bioinformatics tools available to use but no agreement has been reached to set a standard methodology. Big data analysis after sequencing is also a significant challenge for researchers because it is highly computationally demanding.

The aim of this review is to present the latest knowledge on the relationship between the oral microbiome and SADs, focusing on the multi-omics data generated from a large set of samples.

METHODS

Information Sources and Search Process

By searching the PubMed and Embase databases, systematic research was performed according to the PRISMA guidelines (Page et al., 2021). All articles published from 1 January 2000 to 1 January 2022 were taken into account. The search queries follow: ["oral" AND "microbiota" OR "microbiome" OR "dysbiosis" OR "flora"] AND ["systemic lupus erythematosus" OR "Lupus Erythematosus, Systemic" OR "Libman Sacks Disease" OR "rheumatoid arthritis" OR "Sjogren's Syndrome" OR "Sicca Syndrome" OR "SS"].

Eligibility Criteria

To be eligible for inclusion, studies should provide the evaluation of oral microbiome (e.g., the composition and/or diversity of the oral microbial community) from oral samples (rinsing samples, subgingival dental plaque, buccal swab, saliva, etc.) in patients with SADs by multi-omics approaches.

All patients with SLE within the studies should satisfy one of the classification criteria of the American College of Rheumatology (ACR) 1982/1997 criteria (Hochberg, 1997) or the Systemic Lupus International Collaborating Clinics (SLICC) 2012 criteria (Petri et al., 2012). All patients with RA within the studies should satisfy the classification criteria of the American Rheumatism Association (ARA) 1987 criteria (Arnett et al., 1988) or the American College of Rheumatology/European League

Against Rheumatism (ACR/EULAR) 2010 criteria (Aletaha et al., 2010). All patients with SS should satisfy the classification criteria of the ACR/EULAR 2016 criteria (Shiboski et al., 2017) or the American-European Consensus Group (AECG) 2002 criteria (Vitali et al., 2002) or the ACR 2012 criteria (Shiboski et al., 2012).

Studies were excluded if they (1) did not clarify the diagnosis criteria; (2) included patients secondary to other diseases; (3) only evaluated the oral microbiome by bacterial culture or DNA hybridization technology; (4) only evaluated the gut microbiome; (5) were reviews; (6) were not written in English; (7) were *in vitro* studies.

Study Selection

The studies were selected by two authors (L.G. and ZJ.C.) based on the inclusion/exclusion criteria and by considering titles and abstracts, with any disputes resolved by a third author (CS.B.). Then the authors analyzed the full-text selected studies again and determined the eligible articles.

Data Collection

Standardized extraction was used to extract the features of the included studies. The following data were extracted: (1) oral sample type, (2) region, (3) sample size, (4) confounding variables, (5) dental status, (6) use of antibiotics, (7) sequencing platform, (8) pipeline for data analysis, (9) referred database, and (10) specific changes in the oral microbiome associated with SADs.

RESULTS

Study Search

One thousand and thirty-eight studies were identified from the Embase and PubMed databases. Duplicate references ($n = 282$) were removed and 624 were excluded by title and abstract. Of the remaining 132 studies, 107 were excluded through full-text selection. A total of 25 studies were finally included and their data were extracted. Among these, three studies were referred to the SLE, 12 referred to RA, nine referred to SS, and one study referred to both SLE and SS (**Figure 1**).

General Population Characteristics

In total, 137 SLE, 760 RA, and 189 primary SS (pSS) patients were included with information. The control group consisted of healthy volunteers free of any autoimmunity diseases for most studies (22/25) (**Table 1**). In addition, patients with osteoarthritis (OA) (Chen et al., 2018; Mikuls et al., 2018) and at-risk individuals of RA development who have no clinical symptoms of RA (Tong et al., 2019; Cheng et al., 2021; Kroese et al., 2021) were included for comparison with RA patients. Non-SS sicca patients were compared with SS patients (van der Meulen et al., 2018a; Rusthen et al., 2019; Alam et al., 2020).

Most studies considered gender (22/25), age (20/25), smoking status (11/25), use of antibiotics (14/25), and dental status (17/25) as confounding variables. The exclusion criteria about the use of antibiotics varied from 2 to 12 weeks before the sample collection.

Although 68% studies (17/25) (**Table 1**) took the dental status into consideration, the method of dental assessment was different across studies. Two studies used the self-reported symptoms for assessment (Tong et al., 2019; de Jesus et al., 2021). Nine studies provided a full periodontal examination to assess the parameters including probing depth (PD), clinical attachment level (CAL), and bleeding on probing (BOP). Three studies performed a detailed caries-related registration on decayed, missing, and filled teeth/-surfaces (DMFT/DMFS; Zhou et al., 2018; Rusthen et al., 2019; Sembler-Møller et al., 2019). However, the information about dental treatment was not always considered. Only two studies claimed that the volunteers were free of treatment for periodontal disease within the last 6 months (Corrêa et al., 2017, 2019).

The oral sample type differed between studies (**Table 1**). Saliva was collected in 10 studies and subgingival dental plaque was collected in nine studies using sterile paper points. Oral washings, sterile cotton swabs on buccal mucosa, and dorsum of the tongue were also employed.

All the individuals included in each study were local residents (**Table 1**). Among them, nine studies analyzed the oral microbiome of Asians, of which 77.8% (7/9) studies referred to Chinese people. Eight studies investigated the oral microbiome of Europeans and the other eight studies focused on Americans.

General Analysis Characteristics

The most common analysis method was 16S rRNA gene sequencing, which was used in 92% of studies (23/25) (**Table 2**). Only two studies (Zhang et al., 2015; Cheng et al., 2021) used a shotgun metagenomics approach to investigate the oral microbiome of patients with RA. The Human Oral Microbiome Database (HOMD) was the most popular database used for taxonomic assignment, although the similarity threshold was different between studies ranging from 95 to 100% identity. Most 16S rRNA gene sequencing analyses (13/23) were performed with at least 97% similarity when clustering the sequences for operational taxonomic unit (OTU), while the shotgun metagenomics used a less stringent cutoff (95%) instead (**Table 2**). QIIME was the most widely used pipeline (16/25) for the downstream analysis and sometimes was used along with other software such as Mothur, PhyloToAST, and LoTuS.

Oral Microbial Dysbiosis Features

Oral microbial dysbiosis has been identified in the three SADs included in our review, although inconsistent results exist (**Tables 3–6**). Sembler-Møller et al. (2019) reported that there was no significant difference in the oral bacterial diversity or relative abundance on the genus and species level between SS and non-SS controls, indicating that changes in the salivary microbiome was not related to the SS itself.

Among the 25 articles included, *Selenomonas* and *Veillonella* were found significantly increased in the three SADs covered by this review, and *Streptococcus* was significantly decreased in the SADs compared with controls (**Figure 2**). At the species level (**Figure 3**), *Rothia aeria* was significantly decreased in all three diseases. *Prevotella nigrescens*, *Prevotella oulorum*, *Prevotella pleuritidis,* and *Selenomonas noxia* were identified enriched in both RA and SLE compared with healthy controls. *Prevotella salivae*, *Prevotella histicola*, *Lactobacillus salivarius*,

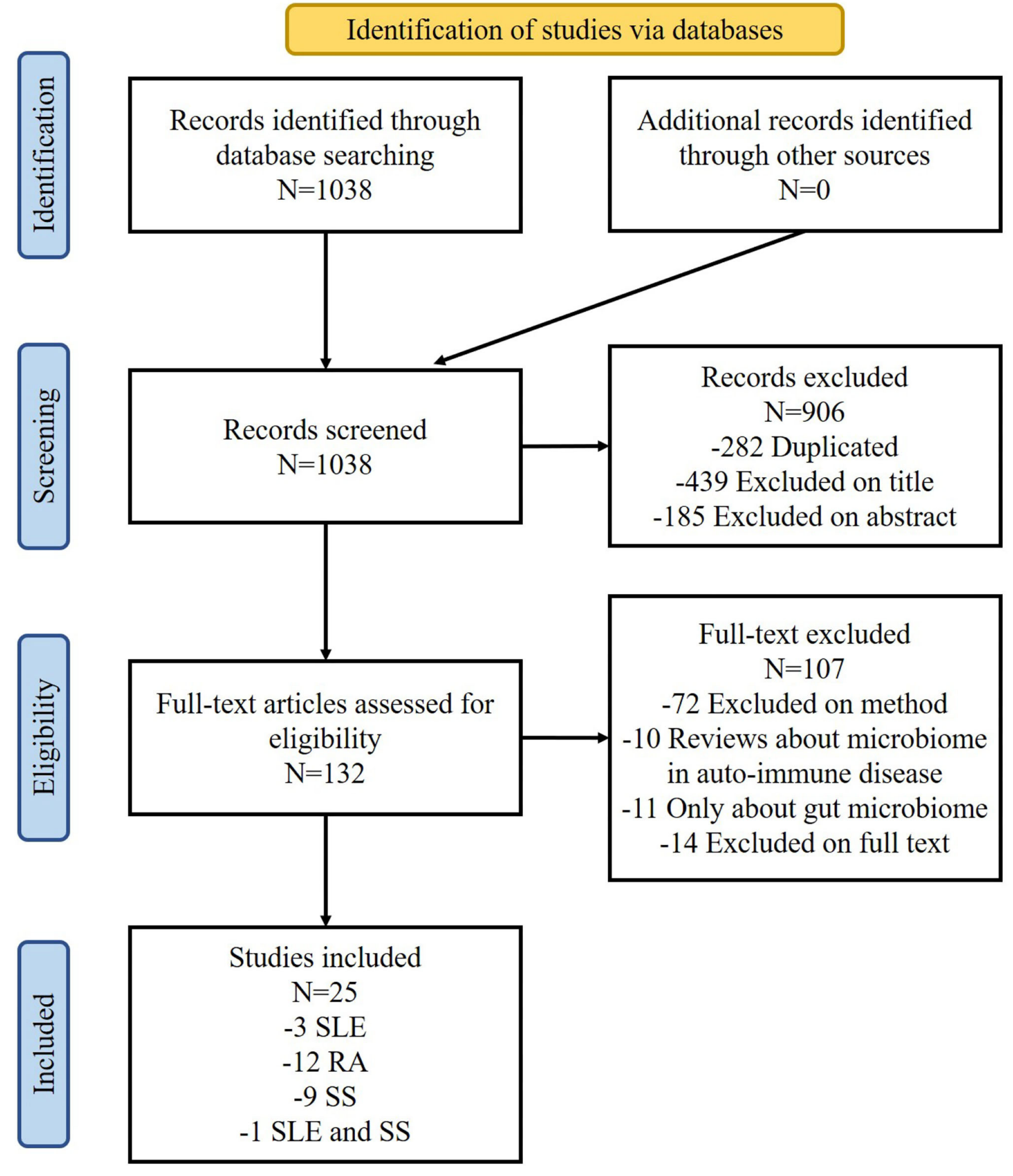

Identification of studies via databases
Identification
Screening
Eligibility
Included
Records identified through database searching N=1038
Additional records identified through other sources N=0
Records screened N=1038
Records excluded N=906 -282 Duplicated -439 Excluded on title -185 Excluded on abstract
Full-text articles assessed for eligibility N=132
Full-text excluded N=107 -72 Excluded on method -10 Reviews about microbiome in auto-immune disease -11 Only about gut microbiome -14 Excluded on full text
Studies included N=25 -3 SLE -12 RA -9 SS -1 SLE and SS

FIGURE 1 | Flow-chart diagram of the selection process.

Prevotella melaninogenica, *Streptococcus parasanguinis*, and *Porphyromonas endodontalis* were more abundant in patients with RA and SS.

Systemic Lupus Erythematosus

The oral microbial dysbiosis features in the SLE patients are summarized in **Tables 3**, **4**.

TABLE 1 | List of the general population characteristics.

Disease	References	Oral sample type	Region	Sample size	Confounding variables	Dental status
SLE	Liu et al., 2021	Saliva	Asia	35 SLE 35 HCs	No antibiotics Sex- and age-matched SLE group currently receives low-dose prednisone and hydroxychloroquine	Without oral disease
SLE	Li et al., 2020	Buccal swab	Asia	20 SLE 19 HCs	Similar age, BMI and diet	No data
SLE	van der Meulen et al., 2019	Oral washings Buccal swab	Europe	30 SLE 39 SS 965 HCs	Similar age, sex, ethnic background, BMI and smoking status SS vs. SLE HCs were not matched to SS or SLE patients	No data
SLE	Corrêa et al., 2017	Subgingival dental plaque	America	52 SLE (17 NCP and 35 CP) 52 non-SLE (24 NCP and 28 CP)	Similar age, sex and oral hygiene habits no difference in smoking status	PD, CAL, BOP, PI, TL
RA	Esberg et al., 2021	Saliva	America	61 eRA 59 HCs	No antibiotics Similar gender and age	PD, TL
RA	Kroese et al., 2021	Tongue Saliva Subgingival dental plaque	Europe	50 eRA 50 at-risk of RA 50 HCs	Gender- and age- matched Similar smoking status, alcohol consumption, use of drugs, use of antibiotics within 3 months and oral hygiene status	PD, BOP, PISA
RA	Cheng et al., 2021	Subgingival dental plaque	Europe	26 eRA 48 at-risk of RA 32 HCs	No antibiotics Balanced for age, gender, and smoking status	PD, CAL, BOP, PI, TL
RA	Lehenaff et al., 2021	Subgingival dental plaque from shallow and deep sites	America	8 RA 10 household members of the RA patients	No antibiotics Similar age, gender, race, number of caries, and periodontal health status	CAL, PD, BOP
RA	de Jesus et al., 2021	Buccal swab	America	35 RA 64 non-RA	No antibiotics Similar oral health status, smoking status	Self-reported oral health symptoms denture, gum bleeding
RA	Tong et al., 2019	Saliva	Asia	27 RA 29 at-risk of RA 23 HCs	No antibiotics Similar age, gender, and smoking status	Self-reported questionnaire
RA	Corrêa et al., 2019	Subgingival dental plaque	America	42 RA (21 CP and 21 NCP) 47 HCs (20 CP and 27 NCP)	No antibiotics Gender- and age- matched Similar smoking status	PD, CAL, BOP, PI
RA	Mikuls et al., 2018	Subgingival dental plaque	America	260 RA 296 OA	No antibiotics Similar age, gender and race	Full mouth periodontal evaluation
RA	Lopez-Oliva et al., 2018	Subgingival dental plaque	Europe	22 RA 19 HCs (both periodontally healthy)	No antibiotics Similar gender, race, smoking history and alcohol consumption	CAL, PD, BOP
RA	Chen et al., 2018	Saliva	Asia	110 RA 67 OA 155 HCs	Gender and age not matched	No data
RA	Zhang et al., 2015	Dental plaque Saliva	Asia	54 RA 51 HCs 51 RA 47 HCs	No antibiotics Age-, gender-, and ethnicity-matched	No data
RA	Scher et al., 2012	Subgingival dental plaque	America	31 NORA 34 CRA 18 HCs	No antibiotics Age-, gender-, and ethnicity-matched	CAL, PD, BOP
SS	Sharma et al., 2020	Saliva	Asia	37 SS 35 HCs	No antibiotics No smoking Similar gender	No data

(Continued)

TABLE 1 | Continued

Disease	References	Oral sample type	Region	Sample size	Confounding variables	Dental status
SS	Alam et al., 2020	Oral washings	Asia	8 SS without oral dryness 17 SS with dryness 11 sicca 14 HCs	No smoking, no antibiotics and steroids Similar gender, age	No data
SS	Rusthen et al., 2019	Saliva	Europe	15 SS 15 sicca 15 HCs	Similar gender, age, smoking and dental status	Missing and decayed teeth, number of mobile teeth and gingivitis, dental caries experience
SS	Sembler-Møller et al., 2019	Saliva	Europe	24 SS 34 sicca	No smoking, no antibiotics Similar age, gender, general health, oral health status	DMFT and DMFS, dental plaque, gingival inflammation and periodontal pocket depth
SS	Zhou et al., 2018	Oral washings	Asia	22 SS 23 HCs	Similar gender and age	DMFT and DMFS
SS	van der Meulen et al., 2018a	Buccal swab	Europe	37 SS 86 sicca 24 HCs	Gender matched Age not matched	Own teeth, oral dryness
SS	de Paiva et al., 2016	Tongue	America	10 SS 11 HCs	Similar gender and age	No data
SS	Siddiqui et al., 2016	Saliva	Europe	9 SS 9 HCs	Similar gender and age No hyposalivation	No data
SS	Li et al., 2016	Buccal swab	Asia	10 SS 10 HCs	No smoking, no antibiotics Similar gender and age, number of teeth, periodontal and mucosal status	Oral mucosa, number of teeth and stimulated/ unstimulated secretion rat

SLE, systemic lupus erythematosus; RA, rheumatoid arthritis; SS, Sjögren's syndrome; HCs, healthy controls; NCP, non-chronic periodontitis; CP, chronic periodontitis; OA, osteoarthritis; NORA, new-onset rheumatoid arthritis, disease duration of >6 weeks and absence of any treatment with disease-modifying anti-rheumatic drug (DMARD) or steroids (ever); eRA, early onset of RA, symptom duration ≤12 months; CRA, chronic RA with minimum disease duration of 6 months; PD, probing depth; CAL, clinical attachment level; BOP, bleeding on probing; PI, plaque index; PISA, periodontal inflamed surface area; TL, tooth loss; DMFT, decayed, missing and filled teeth; DMFS, decayed, missing and filled surfaces.

With regard to the studies about the oral microbiome in patients with SLE, all the four studies assessed alpha- and beta- diversity and found that there were significant differences between the SLE patients and controls (**Table 3**). But the results were not consistent among the studies, which may be due to the different sample types relied on. One study analyzing the subgingival dental plaque found higher alpha diversity in patients with SLE compared with healthy controls (Corrêa et al., 2017), while another study focusing on the buccal swabs found decreased bacterial diversity in patients with SLE compared with healthy controls (Li et al., 2020).

As shown in **Table 4**, *Veillonella*, *Prevotella*, *Selenomonas*, *Blautia*, *Barnesiella*, *Pyramidobacter*, *Alistipes,* and *Lactobacillus* were more abundant in patients with SLE compared with healthy controls when analyzing the oral microbiome at the genus level. There was only one study that analyzed the subgingival dental plaque of patients with SLE and presented changes in subgingival microbiome at the species level (Corrêa et al., 2017). By periodontal assessment of the participants, species associated with SLE had been identified in the non-periodontitis group. *Prevotella nigrescens*, *Prevotella oulorum*, *Prevotella oris,* and *Selenomonas noxia* were more abundant in the patients with SLE compared with healthy controls. The results of this study indicated that oral microbial dysbiosis was associated with SLE, independent of periodontal status.

Rheumatoid Arthritis

The oral microbial dysbiosis features in the patients with RA are summarized in **Tables 3**, **5**.

Among the 11 studies that compared patients with RA with healthy controls, nine studies analyzed oral microbial diversity and richness of patients with RA, seven of which found a significant difference between patients with RA and healthy controls (**Table 3**), while the other two studies found no significant changes in oral microbial diversity in patients with RA (Scher et al., 2012; Lehenaff et al., 2021).

Eight studies investigated the microbiome at the genus level, and half of them found *Prevotella* significantly increased in patients with RA (**Table 5**). Some genera were identified with evidently different abundance in different studies. For example,

TABLE 2 | Analysis of the methodology of the included studies.

Disease	References	Sequencing platform	Database	Pipeline	Identity
SLE	Liu et al., 2021	16S rRNA sequencing/Illlumina MiSeq platform	Greengenes V.13-8	QIIME 2	99%
SLE	Li et al., 2020	16S rRNA sequencing	SILVA 128 database	Mothur	97%
SLE	van der Meulen et al., 2019	16S rRNA sequencing	SILVA 128 database	QIIME	No data
SLE	Corrêa et al., 2017	16S rRNA sequencing/Illumina MiSeq platform	CORE	QIIME	No data
RA	Esberg et al., 2021	16S rRNA sequencing/Illlumina MiSeq platform	eHOMD	QIIME 2	>98.5%
RA	Kroese et al., 2021	16S rDNA sequencing/Illumina MiSeq platform	HOMD	QIIME v1.8.0	No data
RA	Cheng et al., 2021	Shotgun metagenomics sequencing/Illumina HiSeq 3000 platform	MG-RAST	Refseq	95%
RA	Lehenaff et al., 2021	16S rRNA sequencing/Illumina Miseq platform	HOMD	QIIME2	97%
RA	de Jesus et al., 2021	16S rRNA sequencing/Illumina Miseq PE250 platform	HOMD	QIIME2	No data
RA	Tong et al., 2019	16S rRNA sequencing/Illumina Miseq platform	SILVA 128 database	/	97%
RA	Corrêa et al., 2019	16S rRNA sequencing/Illumina MiSeq platform	CORE	QIIME	97%
RA	Mikuls et al., 2018	16S rRNA sequencing/Illumina MiSeq platform	HOMD	/	97%
RA	Lopez-Oliva et al., 2018	16S rRNA sequencing/Illumina MiSeq platform	HOMD	QIIME PhyloToAST	97%
RA	Chen et al., 2018	16S rRNA sequencing/HiSeq 2500 platform	Greengenes ribosomal database	QIIME 1.9.1	97%
RA	Zhang et al., 2015	Metagenomic shotgun sequencing and a metagenome-wide association study (MGWAS)/Illumina platform	Microbial Genomes (IMG, v400) database	in-house pipeline	95%
RA	Scher et al., 2012	16S rRNA sequencing/454 GS FLX Titanium platform	SILVA 128 database	Mothur	97%
SS	Sharma et al., 2020	16S rRNA sequencing/HiSeq 2500 platform	Greengene database SILVA 128 database	QIIME LoTuS	97%
SS	Alam et al., 2020	16S rRNA sequencing/454 GS FLX titanium pyrosequencer	The EzTaxon-e database	No data	No data
SS	Rusthen et al., 2019	16S rRNA sequencing/Roche 454 GS Junior platform	SILVA 128 database HOMD	QIIME 1.8.0	99–100%
SS	Sembler-Møller et al., 2019	16S rRNA sequencing/Illumina Miseq platform	eHOMD	DADA2 R	No data
SS	Zhou et al., 2018	16S rRNA sequencing/Illumina Miseq PE300 platform	HOMD	Mothur QIIME 1.9.1	97%
SS	van der Meulen et al., 2018a	16S rRNA sequencing/Illumina MiSeq platform	HOMD	QIIME V.1.9.1	97%
SS	de Paiva et al., 2016	16S rRNA sequencing/MiSeq platform	UPARSE and the SILVA 128 database	No data	97%
SS	Siddiqui et al., 2016	16S rRNA sequencing/454 GS Junior system	HOMDEXTGG set the NCBI 16S rRNA reference sequence set	QIIME 1.9.1	98%
SS	Li et al., 2016	16S rRNA sequencing/NGS illumine Miseq 2 × 300 bp platform	SILVA dataset	Mothur	97%

SLE, systemic lupus erythematosus; RA, rheumatoid arthritis; SS, Sjögren's syndrome; HOMD, Human Oral Microbiome Database; eHOMD, expanded Human Oral Microbiome Database.

TABLE 3 | Major changes in microbial community associated with SADs.

Disease	References	Oral sample type	Alpha diversity	Beta diversity
SLE	Liu et al., 2021	Saliva	No significant change	Increased bacterial diversity in SLE patients compared with HCs
SLE	Li et al., 2020	Buccal swab	Lower alpha- diversity in SLE patients compared with HCs	Higher beta- diversity in SLE patients compared with HCs
SLE	van der Meulen et al., 2019	Oral washings Buccal swab	Higher alpha- and beta- diversity in SLE patients compared with SS patients	
SLE	Corrêa et al., 2017	Subgingival dental plaque	Higher alpha-diversity in SLE patients compared with HCs	Lower beta-diversity in SLE patients compared with HCs
RA	Esberg et al., 2021	Saliva	Higher alpha- and beta- diversity in SLE patients compared with HCs	
RA	Kroese et al., 2021	Tongue Saliva Subgingival dental plaque	/	/
RA	Cheng et al., 2021	Subgingival dental plaque	Lower richness and diversity in CCP+ at-risk group and the eRA group compared with HCs	
RA	Lehenaff et al., 2021	Subgingival dental plaque	No significant difference between RA patients and HCs	
RA	de Jesus et al., 2021	Buccal swab	Similar Shannon diversity index of bacterial species among RA compared with non-RA controls	Significant difference between RA and controls
RA	Tong et al., 2019	Saliva	Lower alpha- diversity in high-risk group compared with HCs	A tendency of gradual lower change from HCs, high-risk group to RA patients
RA	Corrêa et al., 2019	Subgingival dental plaque	Higher bacterial richness than controls without periodontitis	Increased microbial diversity compared with controls
RA	Mikuls et al., 2018	Subgingival dental plaque	No difference between RA and OA patients	/
RA	Lopez-Oliva et al., 2018	Subgingival dental plaque	/	/
RA	Chen et al., 2018	Saliva	Higher diversity in RA and OA compared with HCs, but no difference between RA and OA	Higher diversity in RA and OA compared with HCs, lower diversity in RA compared with OA
RA	Zhang et al., 2015	Dental plaque Saliva	Increased richness and diversity in RA patients compared with HCs	
RA	Scher et al., 2012	Subgingival dental plaque	The oral microbiota is equally rich and diverse in NORA, CRA and control groups	
SS	Sharma et al., 2020	Saliva	No difference between SS patients and HCs	
SS	Alam et al., 2020	Oral washings	Higher diversity in SS patients compared with HCs	/
SS	Rusthen et al., 2019	Saliva	No difference in SS, sicca and HCs	
SS	Sembler-Møller et al., 2019	Saliva	No difference between SS and sicca	
SS	Zhou et al., 2018	Oral washings	Lower oral bacterial community evenness and diversity in SS patients compared with HCs	No difference between SS and HCs
SS	van der Meulen et al., 2018a	Buccal swab	No difference among SS, sicca and HCs, but showed a trend towards lower richness and diversity compared with HCs	
SS	de Paiva et al., 2016	Tongue	Lower Shannon diversity in SS compared with HCs	/
SS	Siddiqui et al., 2016	Saliva	Lower species richness, alpha- diversity in SS compared with HCs	/
SS	Li et al., 2016	Buccal swab	No difference between SS patients and HCs	

SLE, systemic lupus erythematosus; RA, rheumatoid arthritis; SS, Sjögren's syndrome; HCs, healthy controls; OA, osteoarthritis; eRA, early onset of RA, symptom duration ≤12 months; NORA, new-onset rheumatoid arthritis, disease duration of >6 weeks and absence of any treatment with disease-modifying anti-rheumatic drug (DMARD) or steroids (ever); CRA, chronic RA with minimum disease duration of 6 months.

TABLE 4 | Specific changes in the oral microbiome of SLE patients.

References	Enriched genus	Decreased genus	Enriched species	Decreased species
Liu et al. (2021)	*Prevotella*, *Selenomonas*, and *Veillonella*	*Bacteroides* and *Streptococcus*	/	/
Li et al. (2020)	*Barnesiella*, *Blautia*, *Lactobacillus*, *Pyramidobacter* and *Veillonella*	/	/	/
van der Meulen et al. (2019)	**SLE vs. HCs:** *Alistipes*	/	/	/
Corrêa et al. (2017)	/	**In NCP group:** *Sphingomonas* **In CP group:** *Clostridiales*	**In NCP group:** *Prevotella* (*P. nigrescens*, *P. oulorum*, *P. oris*), and *Selenomonas noxia* **In CP group:** *Prevotella* (*P. oulorum*, *P. pleuritidis*), *Pseudomonas spp.*, *Treponema maltophilum* and *Actinomyces* IP073	**In CP group:** *Rothia aeria*, *Capnocytophaga gingivalis*, *Rasltonia* oral taxon 027, *Leptotrichia* oral taxon A71, *Streptococcus sanguinis* and *Haemophilus parainfluenzae*

SLE, systemic lupus erythematosus; NCP, non-chronic periodontitis; CP, chronic periodontitis.

by analyzing the subgingival dental plaque of patients with RA, *Streptococcus* was found with significantly higher relative abundance compared with healthy controls by Cheng et al. (2021), but was identified at a lower level in the other two studies (Corrêa et al., 2019; Tong et al., 2019). Ten studies presented the results at the species level and demonstrated different specific dysbiosis features associated with RA, of which two studies found a higher level of *Rothia mucilaginosa* in the patients with RA (Zhang et al., 2015; de Jesus et al., 2021). Two studies had performed periodontal examination on the participants, and thus were able to identify the alterations of oral microbiome in patients with RA without periodontitis (Lopez-Oliva et al., 2018; Cheng et al., 2021).

In addition, potential functions of oral microbiome were also analyzed by shotgun sequencing studies (Zhang et al., 2015; Cheng et al., 2021). Functional units were found altered in the oral microbiome of patients with RA including ATP-dependent 26S proteasome regulatory subunit, component of SCF ubiquitin ligase and anaphase-promoting complex, cysteine synthase, DNA helicase TIP49, TBP-interacting protein, serine/threonine protein phosphatase 2A, regulatory subunit, the redox environment, transport and metabolism of iron, sulfur, zinc, and arginine.

Oral microbial dysbiosis had also been discovered in the at-risk individuals of RA development, indicating that these species may be related with the RA initiation (Tong et al., 2019; Cheng et al., 2021; Kroese et al., 2021).

Sjögren's Syndrome

SS and Healthy Controls

Eight studies analyzed the alpha-diversity between patients with SS and healthy controls (**Table 3**). Three studies found a significantly decreased bacterial richness and alpha-diversity in patients with SS compared with healthy controls by analyzing saliva, oral washings, and tongue samples (de Paiva et al., 2016; Siddiqui et al., 2016; Zhou et al., 2018), while Alam et al. (2020) reported a significantly higher diversity in the saliva microbiome of patients with SS compared with healthy controls. Other studies found no significant differences when investigating saliva and buccal mucosa samples between the groups (Li et al., 2016; van der Meulen et al., 2018a; Rusthen et al., 2019; Sharma et al., 2020).

At genus level (**Table 6**), *Bifidobacterium*, *Lactobacillus,* and *Dialister* were found significantly increased in the saliva and buccal mucosa of patients with SS (van der Meulen et al., 2018a; Sharma et al., 2020). *Haemophilus* and *Neisseria* were found significantly decreased in four studies (Li et al., 2016; van der Meulen et al., 2018a; Zhou et al., 2018; Rusthen et al., 2019).

Only three studies reported results at the species level (Siddiqui et al., 2016; Rusthen et al., 2019; Alam et al., 2020). Thirty-five species, including *Streptococcus mutans*, *Prevotella melaninogenica*, and *Veillonella rogosae* were significantly more abundant in patients with SS compared with healthy controls (Siddiqui et al., 2016; Rusthen et al., 2019; Alam et al., 2020) and nine species were less abundant (Rusthen et al., 2019; Alam et al., 2020).

SS and Sicca Patients

Four studies analyzed the alpha-diversity between SS and non-SS sicca patients (**Table 3**). In accordance with the results of comparing SS with healthy controls, Alam et al. reported significantly a higher diversity in patients with SS compared with sicca patients (Alam et al., 2020). But others found no significant differences between patients with SS and sicca (van der Meulen et al., 2018a; Rusthen et al., 2019; Sembler-Møller et al., 2019).

At the genus level (**Table 6**), *Bergeyella* and *Granulicatella* were found significantly decreased in patients with SS compared with sicca patients, which were also decreased when compared with healthy controls (van der Meulen et al., 2018a). At the species level, six species were identified as significantly more abundant in patients with SS than sicca patients. Among those species, *Veillonella parvula*, *Lactobacillus salivarius*, *Lactobacillus fermentum*, *Prevotella nanceiensis,* and *Veillonella rodentium* were also found to be increased when comparing SS patients with healthy controls (Rusthen et al., 2019; Alam et al., 2020).

TABLE 5 | Specific changes in the oral microbiome of RA patients.

References	Enriched genus	Decreased genus	Enriched species	Decreased species
Esberg et al. (2021)	/	/	*Prevotella pleuritidis*, *Porphyromonas endodontalis*, *Filifactor alocis* and *Treponema denticola*	*Oribacterium sinus*, *Catonella morbi*, *Veillonella rogosae* and *Campylobacter concisus*
Kroese et al. (2021)	*Veillonella*, *Prevotella*	/	*Prevotella salivae*	*Neisseria flavescens*, *Streptococcus dentisani*, *Porphyromonas pasteri* and *Veillonella parvula*
Cheng et al. (2021)	**Periodontally healthy site** *Cardiobacterium*, *Bifidobacterium*, *Porphyromonas*, *Capnocytophaga*, *Neisseria* and *Streptococcus* **Diseased site** *Cardiobacterium*, *Capnocytophaga*, *Neisseria* and *Streptococcus*	/	**Periodontally healthy site:** *Acinetobacter baumannii*, *Acinetobacter johnsonii*, *Acinetobacter lwoffii*, *Alistipes putredinis*, *Cardiobacterium hominis*, *Caulobacter segnis*, *Clostridium phytofermentans*, *Enhydrobacter aerosaccus*, *Enterococcus casseliflavus*, *Methylobacterium extorquens*, *Methylobacterium nodulans*, *Methylobacterium populi*, *Methylobacterium radiotolerans*, *Pseudomonas stutzeri*, *Shewanella sp. ANA-3*, *Sphingopyxis alaskensis*, *Thiomonas intermedia*, *Xanthobacter autotrophicus* and *Xanthomonas campestris* **Diseased site:** *Capnocytophaga gingivalis*, *Cardiobacterium hominis*, *Eikenella corrodens*, *Neisseria gonorrhoeae*, *Neisseria mucosa*, *Neisseria sicca*, *Neisseria subflava*, *Streptococcus mitis*, *Streptococcus oralis*, *Streptococcus pneumoniae*, *Streptococcus sanguinis* and *Streptococcus sp. M143*	/
Lehenaff et al. (2021)	/	/	*Actinomyces meyeri* and *Streptococcus parasanguinis*	*Gemella morbillorum*, *Kingella denitrificans*, *Prevotella melaninogenica and Leptotrichia spp.*
de Jesus et al. (2021)	*Streptococcus*, *Rothia* and *Leptotrichia*	*Fusobacterium*, *Porphyromonas*, *Aggregatibacter* and *Capnocytophaga*	*Streptococcus salivarius*, *Rothia mucilaginosa*, *Prevotella spp.*, *Leptotrichia spp.* and *Selenomonas fueggei*	*Prevotella melaninogenica*, *Fusobacterium periodonticum*, *Granulicatella elegans* and *Porphyromonas endodontalis*
Tong et al. (2019)	**RA vs. HCs:** *Prevotella*_6 and *Selenomonas*_3 **RA vs. at-risk:** *Rothia*	**RA vs. HCs:** *Neisseria*, *Haemophilus*, and *Parvimonas* **RA vs. at-risk:** *Filifactor*	/	**RA and at-risk vs. HCs:** *Defluviitaleaceae UCG-011* and *Neisseria oralis*
Corrêa et al. (2019)	*Prevotella*	*Streptococcus*, *Haemophilus* and *Actinomyces*	*Prevotella* (*P. melaninogenica*, *P. denticola*, *P. histicola*, *P. nigrescens*, *P. oulorum*, *and P. maculosa*), *Selenomonas noxia*, S. *sputigena*, *Anaeroglobus geminatus*, *Aggregaticbacter actinomycetemcomitans* and *Parvimonas micra*	*Rothia aeria* and *Kingella oralis*
Mikuls et al. (2018)	**RA vs. OA:** *Prevotella*	/	/	/
Lopez-Oliva et al. (2018)	/	/	*Actinomyces spp.*, *Cryptobacterium spp.*, *Dialister spp.*, *Desulfovibrio spp.*, *Fretibacterium spp.*, *Leptotrichia spp.*, *Prevotella spp.*, *Selenomonas spp.*, *Treponema spp.* (119), *Cryptobacterium curtum* and *Veillonellaceae* [G1]	*Aggregatibacter spp.*, *Gemella spp.*, *Granulicatella spp.*, *Hemophilus spp.*, *Neisseria spp.* and *Streptoccoci spp.* (110)

(Continued)

TABLE 5 | Continued

References	Enriched genus	Decreased genus	Enriched species	Decreased species
Chen et al. (2018)	**RA vs. OA:** *Neisseria*, *Haemophilus*, *Prevotella*, *Veillonella*, *Fusobacterium*, *Aggregatibacter* and *Actinobacillus* **RA and OA vs. HCs:** *Prevotella*, *Neisseria*, *Porphyromonas*, *Veillonella*, *Haemophilus*, *Rothia*, *Streptococcus*, *Actinomyces*, *Granulicatella*, *Leptotrichia*, *Lautropia and Fusobacterium*	**RA vs. OA:** *Streptococcus*, *Actinomyces*, *Lautropia*, *Rothia*, *Granulicatella*, *Ruminococcus*, *Oribacterium* and *Abiotrophia*	**RA vs. OA:** *Neisseria subflava*, *Haemophilus parainfluenzae*, *Veillonella dispar*, *Prevotella tannerae*, and *Actinobacillus parahaemolyticus* **RA and OA vs. HCs:** *Prevotella melaninogenica*, and *Veillonella dispar*	**RA vs. OA:** *Rothia dentocariosa*, and *Ruminococcus gnavus*
Zhang et al. (2015)	/	*Haemophilus*, *Aggregatibacter*, *Cardiobacterium*, *Eikenella* and *Kingella*	*Rothia mucilaginosa*, *Rothia dentocariosa*, *Lactobacillus salivarius*, *Atopobium spp.* and *Cryptobacterium curtum*	*Rothia aeria*, *Porphyromonas gingivalis*, *Lactococcus spp.*, and *Cardiobacterium hominis*
Scher et al. (2012)	**NORA and CRA vs. HCs:** *Anaeroglobus*, *Uncl. Prevotellaceae* and *Phocaeiola*	**NORA and CRA vs. HCs:** *Corynebacterium*, *Mitsuokella* and *Streptococcus*	**NORA and CRA vs. HCs:** *Anaeroglobus* OTU99, *Leptotrichia* OTU87, *Prevotella* OTU60, *Selenomonas* OTU168, *Phocaeiola* OTU92, *Prevotella* OTU31, *Prevotella* OTU134, *Neisseria* OTU16, and *Porphyromonas* OTU1	**NORA and CRA vs. HCs:** *Leptotrichia* OTU12, *Leptotrichia* OTU86, *Leptotrichia* OTU9, *Capnocytophaga* OTU74, *Corynebacterium* OTU4 and *Uncl.*TM7 OTU58

RA, rheumatoid arthritis; OA, osteoarthritis; NORA, new-onset rheumatoid arthritis, disease duration of >6 weeks and absence of any treatment with disease-modifying anti-rheumatic drug (DMARD) or steroids (ever); CRA, chronic RA with minimum disease duration of 6 months.

DISCUSSION

Currently, only symptomatic treatments are available for patients with SAD because of the unknown etiology (Zampeli et al., 2015; Fava and Petri, 2019; Ramos-Casals et al., 2020). In a healthy state, a balance is sustained between the oral microbiome and the host immune response, as well as inside the oral microbial community (Lamont et al., 2018). Therefore, the oral microbiome plays an important role in maintaining the health of the host, as well as the immune system and metabolic stability. Under the pathological conditions, the homeostasis is broken and the oral dysbiosis occurs, which usually manifests as the changes in composition and/or function of the oral microbiome (Lamont et al., 2018). Elucidating the role of the oral microbiome in the initiation and development of SADs may present new possibilities for the treatment and prevention of these diseases.

In this systematic review, we reviewed 25 studies covering 137 patients with SLE, 760 patients with RA, and 189 patients with SS with information on their oral microbiome. Oral microbial dysbiosis has been identified in the SADs in this review by comparing bacterial diversity and richness, as well as abundance of genus or species between patients and healthy controls. Significantly altered microbial diversity has been reported in patients with SLE, RA, and SS, although the inconsistent results exist, which could be due to the different sample sites of the oral cavity. Bacterial diversity of saliva microbiome, which consists mostly of gram-positive aerobes, was found elevated in patients with RA compared with controls (Chen et al., 2018; Esberg et al., 2021), while in the subgingival dental plaque that colonized predominantly by the gram-negative anaerobes or facultative anaerobes, decreased or similar diversity was reported in patients with RA compared with controls (Scher et al., 2012; Mikuls et al., 2018; Cheng et al., 2021). These findings suggested that regular periodontal maintenance or oral hygiene behavior may play an important role in the prevention and treatment of SADs. Understanding the exact association between oral microbial dysbiosis and SADs may help to develop novel combined therapies for both physicians and dentists.

Selenomonas and *Veillonella* were found significantly increased in the SADs covered in this review. In addition to the SADs, increased *Selenomonas* has also been identified to be associated with other systemic diseases, for example, human diabetes (Tsuzukibashi et al., 2017). Reduction of *Streptococcus*, a health-associated genus, was observed in the SADs, indicating that these SADs may disturb the oral microbiome, the mechanisms of which still need further investigation.

At the species level, significant alterations in the abundance of *Rothia aeria*, a gram-positive aerobe from the family *Micrococcaceae*, had been discovered in the three SADs, which could be explained by the abnormal immune status of those patients and also by the effect of treatment of SADs. *R. aeria* is a part of the normal human oral microbiome occasionally related with periodontal and dental infections, but has also been reported in osteomyelitis, endocarditis, and joint infections (Graves et al., 2019).

In the studies included in this review, population characteristics were not always considered, especially the smoking status, periodontal status, and oral hygiene conditions, which can explain the inconsistent results to some extent. In fact, it is well-established that smoking (Al Bataineh et al., 2020), oral hygiene (Radaic and Kapila, 2021), and periodontal

TABLE 6 | Specific changes in the oral microbiome of SS patients.

References	Enriched genus	Decreased genus	Enriched species	Decreased species
Sharma et al. (2020)	*Bifidobacterium*, *Lactobacillus* and *Dialister*	*Leptotrichia*	/	/
Alam et al. (2020)	/	/	**SS with dryness vs. sicca:** *Veillonella parvula*, *Lactobacillus salivarius*, *Veillonella tobetsuensis*, *Lactobacillus fermentum* and *Veillonella rodentium* **SS vs. HCs:** *Prevotella melaninogenica*, *Veillonella rogosae*, *Streptococcus* HQ748137, *Streptococcus* HQ762034, *Prevotella histicola*, *Streptococcus parasanguinis*, *Streptococcus* 4P003152, *Streptococcus* uc, *Streptococcus mutans*, *Haemophilus* HQ807753, *Veillonella parvula*, *Prevotella* FM995711, *Streptococcus sobrinus*, *Prevotella salivae*, *Lactobacillus salivarius*, *Veillonella rodentium*, *Haemophilus haemolyticus*, *Lactobacillus fermentum*, *Prevotella* 4P003758 and *Streptococcus* HQ757980	**SS with dryness vs. sicca:** *Haemophilus sputorum*, *Neisseria* AY005028, *Neisseria* uc, *Capnocytophaga gingivalis*, *Leptotrichia wadei*, *Porphyromonas gingivalis*, *Porphyromonas* AM420091, *Lachnoanaerobaculum orale*, *Lautropia mirabilis*, *Neisseria elongata*, *Rothia aeria*, *Neisseria sicca group*, *Neisseria mucosa*, *Neisseria subflava*, *Streptococcus* CP006776 and *Neisseria* perflava **SS vs. HCs:** *Eikenella corrodens*
Rusthen et al. (2019)	/	**SS and sicca vs. HCs:** *Haemophilus* and *Neisseria*	**SS and sicca vs. HCs:** *Porphyromonas endodontalis*, *Prevotella nancensis*, *Tannerella spp*. and *Treponema spp*. (12) **SS vs. sicca (with hyposalivation):** *Prevotella nanceiensis*	**SS and sicca vs. HCs:** *Actinomyces lingnae*, *Fusobacterium nucleatum subspvincentii*, *Lachnoanaerobaculum orale* and *Megasphaera micronuciformis*, *Oribacterium asaccharolyticum*, *Prevotella nanceiensis*, *Stomatobaculum longum* and *Streptococcus intermedius* **SS vs. sicca (with hyposalivation):** *Capnocytophaga leadbetteri*, *Granulicatella adiacens*, *Neisseria flavescens*, and *RuminococcaceaeG1spt*
Sembler-Møller et al. (2019)	No significant difference	No significant difference	No significant difference	No significant difference
Zhou et al. (2018)	*Veillonella*	*Actinomyces*, *Haemophilus*, *Neisseria*, *Rothia*, *Porphyromonas* and *Peptostreptococcus*	/	/
van der Meulen et al. (2018a)	**SS vs. HCs:** *Alloscardovia*, *Bifidobacterium*, *Scardovia*, *Atopobium*, *Lactobacillus*, *Parvimonas*, *Peptostreptococcaceae*, *Anaeroglobus*, and *Dialister*	**SS vs. HCs:** *Alloprevotella*, *Bergeyella*, *Abiotrophia*, *Granulicatella*, *Enterococcus*, *Ruminococcaceae*, *Lautropia*, *Neisseria* and *Haemophilus* **SS vs. sicca:** *Bergeyella* and *Granulicatella*	/	/
de Paiva et al. (2016)	*Streptococcus*	*Leptotrichia* and *Fusobacterium*	/	/
Siddiqui et al. (2016)	*Streptococcus* and *Veillonella*	/	*Veillonella* sp. Oral Taxon 917	/
Li et al. (2016)	*Leucobacter*, *Delftia*, *Pseudochrobactrum*, *Ralstonia* and *Mitsuaria*	*Haemophilus*, *Neisseria*, *Comamona*, *Granulicatella* and *Limnohabitans*	/	/

SS, Sjögren's syndrome; HCs, healthy controls.

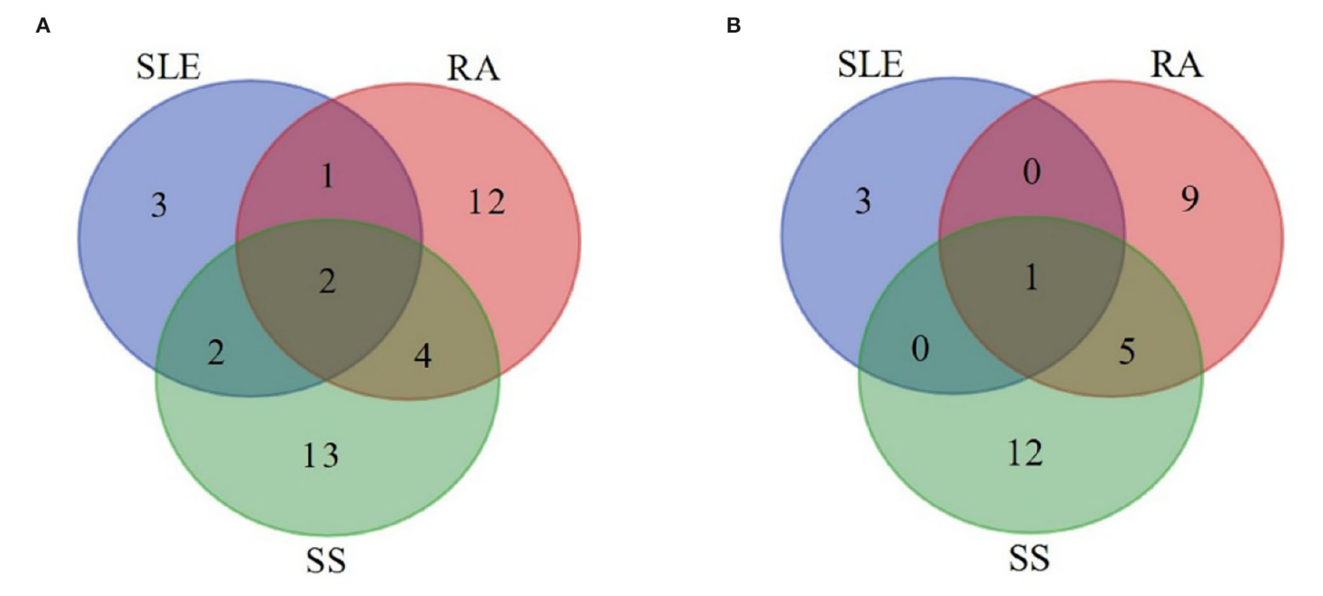


FIGURE 2 | Overlap analysis of the significantly increased **(A)** and decreased **(B)** genera in systemic autoimmune diseases. Numbers of the increased **(A)** and decreased **(B)** genera were visualized for SLE, RA, and SS patients. SLE, systemic lupus erythematosus; RA, rheumatoid arthritis; SS, Sjögren's syndrome.

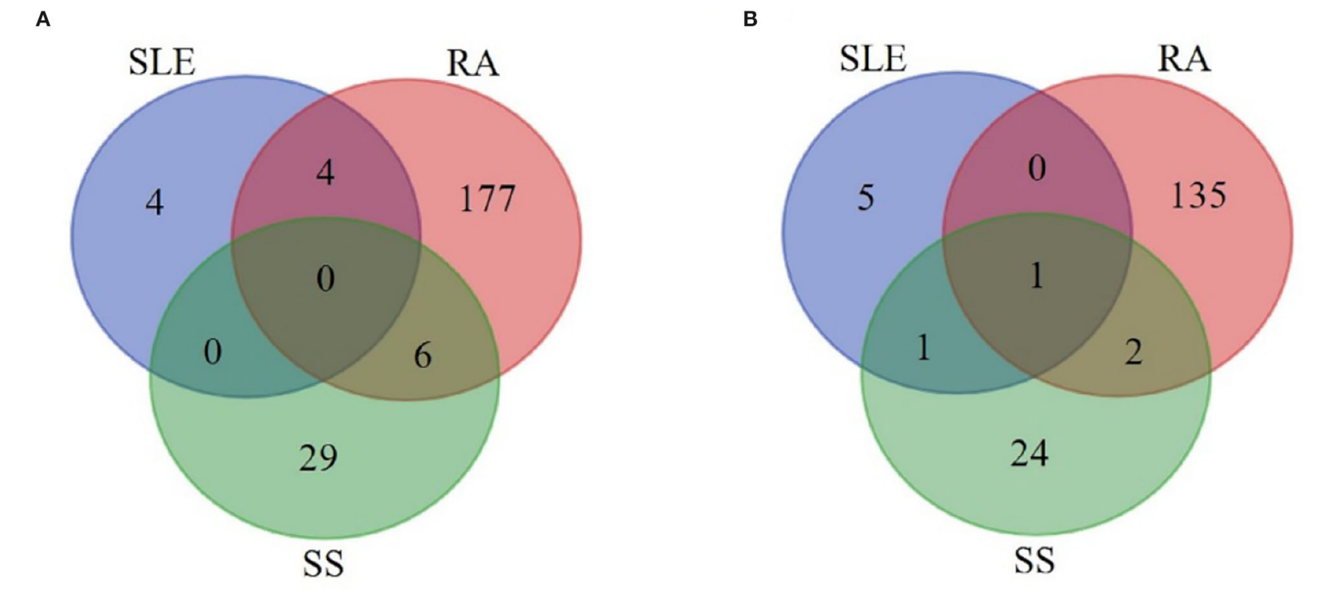


FIGURE 3 | Overlap analysis of the significantly increased **(A)** and decreased **(B)** species in systemic autoimmune diseases. Numbers of the increased **(A)** and decreased **(B)** species were visualized for SLE, RA, and SS patients. SLE, systemic lupus erythematosus; RA, rheumatoid arthritis; SS, Sjögren's syndrome.

disease (Kumar et al., 2006) can influence the oral microbiome. It would not be sensible to evaluate the oral microbiome without considering the above factors. When the periodontal status of the participants was unknown, the results would be somewhat ambiguous as observations might have been due to the influence of periodontal disease (Corrêa et al., 2017). There were two studies performed in the periodontal examination of the participants; thus they were able to analyze the samples from periodontal healthy sites or individuals and confirm that the observed alterations of the oral microbiome were related with the SAD itself (Lopez-Oliva et al., 2018; Cheng et al., 2021).

Also, the effect of medications, especially the antibiotics, should be taken into consideration. Individuals with a history of antibiotics treatment in the last 2 weeks to 3 months were excluded in most studies (14/25), while in some studies the participants were undergoing treatment. Li et al. (2016)

investigated the effect of prednisone on the oral microbiota in SS and found that *Lactobacillus* and *Streptococcus* were more affected by corticosteroids than the disease itself. RA therapy with potential antibacterial properties, such as methotrexate or hydroxychloroquine (Greenstein et al., 2007; Rolain et al., 2007), may also influence the oral microbiome. Therefore, future studies with treatment-naive individuals will be needed to clearly determine the role of oral microbiome in SADs.

There are many other confounding variables that should be considered. Decreased salivary secretion has a negative impact on the quantity of oral microorganisms, which can be seen in patients with SS. Thus, it is not clear whether the changed oral microbiome was caused by SS disease itself or the decreased salivary secretion. Interestingly, Siddiqui et al. (2016) have evaluated the microbiome of saliva in patients with SS with normal salivation and suggested that SS can lead to oral microbial dysbiosis independently of oral dryness. van der Meulen et al. (2018a) found that SS disease status and salivary secretion rate contributed almost equally to the variation of bacterial composition (3.8 vs. 4.3%). While another study observed that the reduction of salivary secretion contributed more to the changes in oral microbiome in patients with SS than the disease itself (van der Meulen et al., 2018b).

From this review, we found that it was difficult to prove a causal link between the oral microbial dysbiosis and disease by investigating the established SADs patients. At-risk individuals with RA development were included in some studies and dysbiosis were identified in their oral microbiome, indicating these perturbations may be related to the RA initiation (Tong et al., 2019; Cheng et al., 2021; Kroese et al., 2021). Cheng et al. found that a higher relative abundance of *Porphyromonas gingivalis* preceded the onset of clinical arthritis, supporting the hypothesis that oral microbial dysbiosis may be a cause of RA initiation (Cheng et al., 2021). However, for SLE and SS the current data was not sufficient to determine whether oral microbial dysbiosis is the consequence or the cause of diseases. Thus, the prospective cohorts of at-risk individuals should be included in the future study to elucidate the mechanisms underlying the potential link between the oral microbial dysbiosis and SADs.

In this systematic review, 92% studies (23/25) relied on the 16S rRNA sequencing technology, which is cost-effective and efficient to detect alterations in bacterial populations. However, a major limitation of this method is that only a single region of the bacterial genome can be analyzed and it is difficult to distinguish the species when their 16S rRNA gene sequences have high similarities (Větrovský and Baldrian, 2013). The shotgun metagenomics approach can provide information on the taxonomic composition of the ecosystem but also on functional genes in the sample, displaying several advantages over the 16S amplicon method, such as more confident identification of bacterial species, increased detection of diversity, and prediction of genes (Ranjan et al., 2016; Durazzi et al., 2021). However, it has been employed only in the two studies to investigate the oral microbiome of patients with RA (Zhang et al., 2015; Cheng et al., 2021). The changes in functional capability in the oral microbiome of patients with RA have been identified, although the actual function gene expression could not be determined by such a method. Besides the genomics, to the best of our knowledge, there was one study conducted by Konig et al. (2016) who analyzed the subgingival microbiome of patients with periodontitis using proteomic techniques and found that the citrullinome in periodontitis mirrored patterns of hypercitrullination observed in the rheumatoid joint. Periodontal pathogen *Aggregatibacter actinomycetemcomitans* has been identified as a candidate bacterial trigger of autoimmunity in RA. More proteomics, transcriptomics, and metabolomics technologies should be used for future studies and may provide a better understanding of the mechanisms underlying the association between oral microbiome and SADs.

In addition to RA, SLE, and SS, there are also other SADs not covered by this review, and few studies have investigated their oral microbiome. To the best of our knowledge, there was one study carried out by Zorba et al., who analyzed the smear samples from oral lesions of patients with pemphigus vulgaris (PV) using 16S rRNA sequencing and found that *Fusobacterium nucleatum* was the most dominant species (Zorba et al., 2021). In the future, high-throughput analysis could be used more widely to study the oral microbiome of other SADs.

CONCLUSION

In this article, we presented a systematic review of literature that is focused on the big data analysis of oral microbiome of SADs patients. Oral microbial dysbiosis has been identified in all the SADs included in our review, by detecting the alterations in microbial composition and populations, as well as the function capabilities. Most dysbiosis features were different between studies, which could be due to a lack of standardized study methodology for each study, from the inclusion criteria, sample type, sequencing platform, referred database, to downstream analysis pipeline and cutoff. Besides the genomics, transcriptomics, proteomics and metabolomics technology should be used to investigate the oral microbiome of SADs patients and also the at-risk individuals of disease development, which may provide us with a better understanding of the etiology of SADs and promote the development of the novel therapies.

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

LG: conceptualization, methodology, validation, formal analysis, data curation, and writing. ZC: conceptualization, methodology, validation, formal analysis, data curation, writing, and funding. CB and QS: conceptualization, methodology, and writing. FZ: conceptualization, methodology, validation, writing, and funding. XC: conceptualization, writing, supervision, and administration. All authors contributed to the article and approved the submitted version.

FUNDING

This study was supported by the Fundamental Research Funds for the Zhejiang Provincial Universities (2021XZZX033). ZC was supported by Young Elite Scientist Support Program by CSA (2020PYRC001). Support was also provided by the NSFC (82001048).

REFERENCES

Ahsan, H. (2017). Selfie: autoimmunity, boon or bane. *J. Immunoassay Immunochem.* 38, 235–246. doi: 10.1080/15321819.2017.1319861

Al Bataineh, M. T., Dash, N. R., Elkhazendar, M., Alnusairat, D., a,.M. H., Darwish, I. M. I., et al. (2020). Revealing oral microbiota composition and functionality associated with heavy cigarette smoking. *J. Transl. Med.* 18, 421–421. doi: 10.1186/s12967-020-02579-3

Alam, J., Lee, A., Lee, J., Kwon, D. I., Park, H. K., Park, J. H., et al. (2020). Dysbiotic oral microbiota and infected salivary glands in Sjögren's syndrome. *PLoS ONE* 15, e0230667. doi: 10.1371/journal.pone.0230667

Aletaha, D., Neogi, T., Silman, A. J., Funovits, J., Felson, D. T., Bingham, C. O., et al. (2010). 2010 Rheumatoid arthritis classification criteria: an American College of Rheumatology/European League Against Rheumatism collaborative initiative. *Arthritis Rheum.* 62, 2569–2581. doi: 10.1002/art.27584

Arnett, F. C., Edworthy, S. M., Bloch, D. A., McShane, D. J., Fries, J. F., Cooper, N. S., et al. (1988). The American Rheumatism Association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis Rheum.* 31, 315–324. doi: 10.1002/art.1780310302

Bolon, B. (2012). Cellular and molecular mechanisms of autoimmune disease. *Toxicol. Pathol.* 40, 216–229. doi: 10.1177/0192623311428481

Chen, B., Zhao, Y., Li, S., Yang, L., Wang, H., Wang, T., et al. (2018). Variations in oral microbiome profiles in rheumatoid arthritis and osteoarthritis with potential biomarkers for arthritis screening. *Sci. Rep.* 8, 17126. doi: 10.1038/s41598-018-35473-6

Cheng, Z., Do, T., Mankia, K., Meade, J., Hunt, L., Clerehugh, V., et al. (2021). Dysbiosis in the oral microbiomes of anti-CCP positive individuals at risk of developing rheumatoid arthritis. *Ann. Rheum. Dis.* 80, 162–168. doi: 10.1136/annrheumdis-2020-216972

Chu, X. J., Cao, N. W., Zhou, H. Y., Meng, X., Guo, B., Zhang, H. Y., et al. (2021). The oral and gut microbiome in rheumatoid arthritis patients: a systematic review. *Rheumatology* 60, 1054–1066. doi: 10.1093/rheumatology/keaa835

Corrêa, J. D., Calderaro, D. C., Ferreira, G. A., Mendonça, S. M., Fernandes, G. R., Xiao, E., et al. (2017). Subgingival microbiota dysbiosis in systemic lupus erythematosus: association with periodontal status. *Microbiome* 5, 34. doi: 10.1186/s40168-017-0252-z

Corrêa, J. D., Fernandes, G. R., Calderaro, D. C., Mendonça, S. M. S., Silva, J. M., Albiero, M. L., et al. (2019). Oral microbial dysbiosis linked to worsened periodontal condition in rheumatoid arthritis patients. *Sci. Rep.* 9, 8379. doi: 10.1038/s41598-019-44674-6

Credendino, S. C., Neumayer, C., and Cantone, I. (2020). Genetics and epigenetics of sex bias: insights from human cancer and autoimmunity. *Trends Genet.* 36, 650–663. doi: 10.1016/j.tig.2020.06.016

de Jesus, V. C., Singh, M., Schroth, R. J., Chelikani, P., and Hitchon, C. A. (2021). Association of bitter taste receptor T2R38 polymorphisms, oral microbiota, and rheumatoid arthritis. *Curr. Issues Mol. Biol.* 43, 1460–1472. doi: 10.3390/cimb43030103

de Molon, R. S., Rossa, C., Thurlings, R. M., Cirelli, J. A., and Koenders, M. I. (2019). Linkage of periodontitis and rheumatoid arthritis: current evidence and potential biological interactions. *Int. J. Mol. Sci.* 20, 4541. doi: 10.3390/ijms20184541

de Paiva, C. S., Jones, D. B., Stern, M. E., Bian, F., Moore, Q. L., Corbiere, S., et al. (2016). Altered mucosal microbiome diversity and disease severity in Sjögren syndrome. *Sci. Rep.* 6, 23561. doi: 10.1038/srep23561

Deo, P. N., and Deshmukh, R. (2019). Oral microbiome: unveiling the fundamentals. *J. Oral Maxillofac. Pathol.* 23, 122–128. doi: 10.4103/jomfp.JOMFP_304_18

Doaré, E., Héry-Arnaud, G., Devauchelle-Pensec, V., and Alegria, G. C. (2021). Healthy patients are not the best controls for microbiome-based clinical studies: example of Sjögren's syndrome in a systematic review. *Front. Immunol.* 12, 699011. doi: 10.3389/fimmu.2021.699011

Durazzi, F., Sala, C., Castellani, G., Manfreda, G., Remondini, D., and De Cesare, A. (2021). Comparison between 16S rRNA and shotgun sequencing data for the taxonomic characterization of the gut microbiota. *Sci. Rep.* 11, 3030. doi: 10.1038/s41598-021-82726-y

Esberg, A., Johansson, L., Johansson, I., and Dahlqvist, S. R. (2021). Oral microbiota identifies patients in early onset rheumatoid arthritis. *Microorganisms* 9, 1657. doi: 10.3390/microorganisms9081657

Fava, A., and Petri, M. (2019). Systemic lupus erythematosus: diagnosis and clinical management. *J. Autoimmun.* 96, 1–13. doi: 10.1016/j.jaut.2018.11.001

Graves, D. T., Corrêa, J. D., and Silva, T. A. (2019). The oral microbiota is modified by systemic diseases. *J. Dent. Res.* 98, 148–156. doi: 10.1177/0022034518805739

Greenstein, R. J., Su, L., Haroutunian, V., Shahidi, A., and Brown, S. T. (2007). On the action of methotrexate and 6-mercaptopurine on *M. avium* subspecies paratuberculosis. *PLoS ONE* 2, e161. doi: 10.1371/journal.pone.0000161

Hochberg, M. C. (1997). Updating the American College of Rheumatology revised criteria for the classification of systemic lupus erythematosus. *Arthritis Rheum.* 40, 1725. doi: 10.1002/art.1780400928

Inanç, B.B. (2020). Different point of view to the autoimmune diseases and treatment with acupuncture. *J. Pharmacopunct.* 23, 187–193. doi: 10.3831/KPI.2020.23.4.187

Konig, M. F., Abusleme, L., Reinholdt, J., Palmer, R. J., Teles, R. P., Sampson, K., et al. (2016). Aggregatibacter actinomycetemcomitans-induced hypercitrullination links periodontal infection to autoimmunity in rheumatoid arthritis. *Sci. Transl. Med.* 8, 369ra176. doi: 10.1126/scitranslmed.aaj1921

Kroese, J. M., Brandt, B. W., Buijs, M. J., Crielaard, W., Lobbezoo, F., Loos, B. G., et al. (2021). Differences in the oral microbiome in patients with early rheumatoid arthritis and individuals at risk of rheumatoid arthritis compared to healthy individuals. *Arthritis Rheumatol.* 73, 1986–1993. doi: 10.1002/art.41780

Kumar, P. S., Leys, E. J., Bryk, J. M., Martinez, F. J., Moeschberger, M. L., and Griffen, A. L. (2006). Changes in periodontal health status are associated with bacterial community shifts as assessed by quantitative 16S cloning and sequencing. *J. Clin. Microbiol.* 44, 3665–3673. doi: 10.1128/JCM.00317-06

Lamont, R. J., Koo, H., and Hajishengallis, G. (2018). The oral microbiota: dynamic communities and host interactions. *Nat. Rev. Microbiol.* 16, 745–759. doi: 10.1038/s41579-018-0089-x

Lehenaff, R., Tamashiro, R., Nascimento, M. M., Lee, K., Jenkins, R., Whitlock, J., et al. (2021). Subgingival microbiome of deep and shallow periodontal sites in patients with rheumatoid arthritis: a pilot study. *BMC Oral Health* 21, 248. doi: 10.1186/s12903-021-01597-x

Li, B. Z., Zhou, H. Y., Guo, B., Chen, W. J., Tao, J. H., Cao, N. W., et al. (2020). Dysbiosis of oral microbiota is associated with systemic lupus erythematosus. *Arch. Oral Biol.* 113, 104708. doi: 10.1016/j.archoralbio.2020.104708

Li, M., Zou, Y., Jiang, Q., Jiang, L., Yu, Q., Ding, X., et al. (2016). A preliminary study of the oral microbiota in Chinese patients with Sjögren's syndrome. *Arch. Oral Biol.* 70, 143–148. doi: 10.1016/j.archoralbio.2016.06.016

Liu, F., Ren, T., Li, X., Zhai, Q., Xu, X., Zhang, N., et al. (2021). Distinct microbiomes of gut and saliva in patients with systemic lupus erythematous and clinical associations. *Front. Immunol.* 12, 626217. doi: 10.3389/fimmu.2021.626217

Lopez-Oliva, I., Paropkari, A. D., Saraswat, S., Serban, S., Yonel, Z., Sharma, P., et al. (2018). Dysbiotic subgingival microbial communities in periodontally healthy patients with rheumatoid arthritis. *Arthritis Rheumatol.* 70, 1008–1013. doi: 10.1002/art.40485

Mikuls, T. R., Walker, C., Qiu, F., Yu, F., Thiele, G. M., Alfant, B., et al. (2018). The subgingival microbiome in patients with established rheumatoid arthritis. *Rheumatology* 57, 1162–1172. doi: 10.1093/rheumatology/key052

Ostrov, B. E. (2015). Immunotherapeutic biologic agents in autoimmune and autoinflammatory diseases. *Immunol. Invest.* 44, 777–802. doi: 10.3109/08820139.2015.1093912

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *J. Clin. Epidemiol.* 134, 178–189. doi: 10.1016/j.jclinepi.2021.03.001

Petri, M., Orbai, A. M., Alarcón, G. S., Gordon, C., Merrill, J. T., Fortin, P. R., et al. (2012). Derivation and validation of the Systemic Lupus International Collaborating Clinics classification criteria for systemic lupus erythematosus. *Arthritis Rheum.* 64, 2677–2686. doi: 10.1002/art.34473

Potgieter, M., Bester, J., Kell, D. B., and Pretorius, E. (2015). The dormant blood microbiome in chronic, inflammatory diseases. *FEMS Microbiol. Rev.* 39, 567–591. doi: 10.1093/femsre/fuv013

Radaic, A., and Kapila, Y. L. (2021). The oralome and its dysbiosis: new insights into oral microbiome-host interactions. *Comput. Struct. Biotechnol. J.* 19, 1335–1360. doi: 10.1016/j.csbj.2021.02.010

Ramos-Casals, M., Brito-Zerón, P., Bombardieri, S., Bootsma, H., De Vita, S., Dörner, T., et al. (2020). EULAR recommendations for the management of Sjögren's syndrome with topical and systemic therapies. *Ann. Rheum. Dis.* 79, 3–18. doi: 10.1136/annrheumdis-2019-216114

Ranjan, R., Rani, A., Metwally, A., McGee, H. S., and Perkins, D. L. (2016). Analysis of the microbiome: advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochem. Biophys. Res. Commun.* 469, 967–977. doi: 10.1016/j.bbrc.2015.12.083

Rolain, J. M., Colson, P., and Raoult, D. (2007). Recycling of chloroquine and its hydroxyl analogue to face bacterial, fungal and viral infections in the 21st century. *Int. J. Antimicrob. Agents* 30, 297–308. doi: 10.1016/j.ijantimicag.2007.05.015

Rusthen, S., Kristoffersen, A. K., Young, A., Galtung, H. K., Petrovski, B., Palm, Ø., et al. (2019). Dysbiotic salivary microbiota in dry mouth and primary Sjögren's syndrome patients. *PLoS ONE* 14, e0218319. doi: 10.1371/journal.pone.0218319

Sanz, M., Ceriello, A., Buysschaert, M., Chapple, I., Demmer, R. T., Graziani, F., et al. (2018). Scientific evidence on the links between periodontal diseases and diabetes: consensus report and guidelines of the joint workshop on periodontal diseases and diabetes by the International Diabetes Federation and the European Federation of Periodontology. *J. Clin. Periodontol.* 45, 138–149. doi: 10.1111/jcpe.12808

Scher, J. U., Ubeda, C., Equinda, M., Khanin, R., Buischi, Y., Viale, A., et al. (2012). Periodontal disease and the oral microbiota in new-onset rheumatoid arthritis. *Arthritis Rheum.* 64, 3083–3094. doi: 10.1002/art.34539

Sembler-Møller, M. L., Belstrøm, D., Locht, H., Enevold, C., and Pedersen, A. M. L. (2019). Next-generation sequencing of whole saliva from patients with primary Sjögren's syndrome and non-Sjögren's sicca reveals comparable salivary microbiota. *J. Oral Microbiol.* 11, 1660566. doi: 10.1080/20002297.2019.1660566

Sharma, D., Sandhya, P., Vellarikkal, S. K., Surin, A. K., Jayarajan, R., Verma, A., et al. (2020). Saliva microbiome in primary Sjögren's syndrome reveals distinct set of disease-associated microbes. *Oral Dis.* 26, 295–301. doi: 10.1111/odi.13191

Shiboski, C. H., Shiboski, S. C., Seror, R., Criswell, L. A., Labetoulle, M., Lietman, T. M., et al. (2017). 2016 American College of Rheumatology/European League Against Rheumatism classification criteria for primary Sjögren's syndrome. A consensus and data-driven methodology involving three international patient cohorts. *Ann. Rheum. Dis.* 76, 9–16. doi: 10.1136/annrheumdis-2016-210571

Shiboski, S. C., Shiboski, C. H., Criswell, L., Baer, A., Challacombe, S., Lanfranchi, H., et al. (2012). American College of Rheumatology classification criteria for Sjögren's syndrome: a data-driven, expert consensus approach in the Sjögren's International Collaborative Clinical Alliance cohort. *Arthritis Care Res.* 64, 475–487. doi: 10.1002/acr.21591

Siddiqui, H., Chen, T., Aliko, A., Mydel, P. M., Jonsson, R., and Olsen, I. (2016). Microbiological and bioinformatics analysis of primary Sjogren's syndrome patients with normal salivation. *J. Oral Microbiol.* 8, 31119. doi: 10.3402/jom.v8.31119

Teshome, A., and Yitayeh, A. (2016). Relationship between periodontal disease and preterm low birth weight: systematic review. *Pan Afr. Med. J.* 24, 215. doi: 10.11604/pamj.2016.24.215.8727

Tonetti, M. S., and Van Dyke, T. E. (2013). Periodontitis and atherosclerotic cardiovascular disease: consensus report of the Joint EFP/AAP Workshop on Periodontitis and Systemic Diseases. *J. Periodontol.* 84, S24–S29. doi: 10.1111/jcpe.12089

Tong, Y., Zheng, L., Qing, P., Zhao, H., Li, Y., Su, L., et al. (2019). Oral microbiota perturbations are linked to high risk for rheumatoid arthritis. *Front. Cell. Infect. Microbiol.* 9, 475. doi: 10.3389/fcimb.2019.00475

Tsuzukibashi, O., Uchibori, S., Kobayashi, T., Umezawa, K., Mashimo, C., Nambu, T., et al. (2017). Isolation and identification methods of *Rothia* species in oral cavities. *J. Microbiol. Methods* 134, 21–26. doi: 10.1016/j.mimet.2017.01.005

Ursell, L. K., Metcalf, J. L., Parfrey, L. W., and Knight, R. (2012). Defining the human microbiome. *Nutr. Rev.* 70(Suppl. 1), S38–S44. doi: 10.1111/j.1753-4887.2012.00493.x

van der Meulen, T. A., Harmsen, H. J. M., Bootsma, H., Liefers, S. C., Vich Vila, A., Zhernakova, A., et al. (2018a). Dysbiosis of the buccal mucosa microbiome in primary Sjögren's syndrome patients. *Rheumatology* 57, 2225–2234. doi: 10.1093/rheumatology/key215

van der Meulen, T. A., Harmsen, H. J. M., Bootsma, H., Liefers, S. C., Vich Vila, A., Zhernakova, A., et al. (2018b). Reduced salivary secretion contributes more to changes in the oral microbiome of patients with primary Sjögren's syndrome than underlying disease. *Ann. Rheum. Dis.* 77, 1542–1544. doi: 10.1136/annrheumdis-2018-213026

van der Meulen, T. A., Harmsen, H. J. M., Vila, A. V., Kurilshikov, A., Liefers, S. C., Zhernakova, A., et al. (2019). Shared gut, but distinct oral microbiota composition in primary Sjögren's syndrome and systemic lupus erythematosus. *J. Autoimmun.* 97, 77–87. doi: 10.1016/j.jaut.2018.10.009

Van Loveren, H., Vos, J. G., Germolec, D., Simeonova, P. P., Eijkemanns, G., and McMichael, A. J. (2001). Epidemiologic associations between occupational and environmental exposures and autoimmune disease: report of a meeting to explore current evidence and identify research needs. *Int. J. Hyg. Environ. Health* 203, 483–495. doi: 10.1078/1438-4639-00057

Verma, D., Garg, P. K., and Dubey, A. K. (2018). Insights into the human oral microbiome. *Arch. Microbiol.* 200, 525–540. doi: 10.1007/s00203-018-1505-3

Větrovský, T., and Baldrian, P. (2013). The variability of the 16S rRNA gene in bacterial genomes and its consequences for bacterial community analyses. *PLoS ONE* 8, e57923. doi: 10.1371/journal.pone.0057923

Vitali, C., Bombardieri, S., Jonsson, R., Moutsopoulos, H. M., Alexander, E. L., Carsons, S. E., et al. (2002). Classification criteria for Sjögren's syndrome: a revised version of the European criteria proposed by the American-European Consensus Group. *Ann. Rheum. Dis.* 61, 554–558. doi: 10.1136/ard.61.6.554

Wahren-Herlenius, M., and Dörner, T. (2013). Immunopathogenic mechanisms of systemic autoimmune disease. *Lancet* 382, 819–831. doi: 10.1016/S0140-6736(13)60954-X

Wang, L., Wang, F. S., and Gershwin, M. E. (2015). Human autoimmune diseases: a comprehensive update. *J. Intern. Med.* 278, 369–395. doi: 10.1111/joim.12395

Willame, C., Dodd, C., van der Aa, L., Picelli, G., Emborg, H. D., Kahlert, J., et al. (2021). Incidence rates of autoimmune diseases in European healthcare databases: a contribution of the ADVANCE project. *Drug Saf.* 44, 383–395. doi: 10.1007/s40264-020-01031-1

Xiao, Z. X., Miller, J. S., and Zheng, S. G. (2021). An updated advance of autoantibodies in autoimmune diseases. *Autoimmun. Rev.* 20, 102743. doi: 10.1016/j.autrev.2020.102743

Zampeli, E., Vlachoyiannopoulos, P. G., and Tzioufas, A. G. (2015). Treatment of rheumatoid arthritis: unraveling the conundrum. *J. Autoimmun.* 65, 1–18. doi: 10.1016/j.jaut.2015.10.003

Zhang, X., Zhang, D., Jia, H., Feng, Q., Wang, D., Liang, D., et al. (2015). The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. *Nat. Med.* 21, 895–905. doi: 10.1038/nm.3914

Zhou, Z., Ling, G., Ding, N., Xun, Z., Zhu, C., Hua, H., et al. (2018). Molecular analysis of oral microflora in patients with primary Sjögren's syndrome by using high-throughput sequencing. *PeerJ* 6, e5649. doi: 10.7717/peerj.5649

Zorba, M., Melidou, A., Patsatsi, A., Poulopoulos, A., Gioula, G., Kolokotronis, A., et al. (2021). The role of oral microbiome in pemphigus vulgaris. *Arch. Microbiol.* 203, 2237–2247. doi: 10.1007/s00203-021-02199-5

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

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

frontiers | Frontiers in Big Data

REVIEW
published: 08 July 2022
doi: 10.3389/fdata.2022.833196

Blockchain for Electronic Vaccine Certificates: More Cons Than Pros?

Raphaëlle Toubiana[1], *Millie Macdonald*[2], *Sivananda Rajananda*[3], *Tale Lokvenec*[3], *Thomas C. Kingsley*[4,5] *and Santiago Romero-Brufau*[1,6]*

[1] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Harvard University, Boston, MA, United States, [2] University of Queenland, Saint Lucia, QLD, Australia, [3] Institute for Applied Computational Science, Graduate School of Arts and Sciences, Harvard University, Cambridge, MA, United States, [4] Department of Medicine, Mayo Clinic, Rochester, MN, United States, [5] Department of Biomedical Informatics, Mayo Clinic, Rochester, MN, United States, [6] Department of Otolaryngology - Head and Neck Surgery, Mayo Clinic, Rochester, MN, United States

Electronic vaccine certificates (EVC) for COVID-19 vaccination are likely to become widespread. Blockchain (BC) is an electronic immutable distributed ledger and is one of the more common proposed EVC platform options. However, the principles of blockchain are not widely understood by public health and medical professionals. We attempt to describe, in an accessible style, how BC works and the potential benefits and drawbacks in its use for EVCs. Our assessment is BC technology is not well suited to be used for EVCs. Overall, blockchain technology is based on two key principles: the use of cryptography, and a distributed immutable ledger in the format of blockchains. While the use of cryptography can provide ease of sharing vaccination records while maintaining privacy, EVCs require some amount of contribution from a centralized authority to confirm vaccine status; this is partly because these authorities are responsible for the distribution and often the administration of the vaccine. Having the data distributed makes the role of a centralized authority less effective. We concluded there are alternative ways to use cryptography outside of a BC that allow a centralized authority to better participate, which seems necessary for an EVC platform to be of practical use.

Keywords: blockchain (BC), electronic vaccination record, electronic vaccine certificate, cryptography, COVID-19, clinical informatics

OPEN ACCESS

Edited by:
Juan Zhao,
Vanderbilt University Medical Center,
United States

Reviewed by:
Tsung-Ting Kuo,
University of California, San Diego,
United States
Jiyong Park,
University of North Carolina at
Greensboro, United States

Correspondence:
Santiago Romero-Brufau
romerobrufau.santiago@mayo.edu

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

Received: *10 December 2021*
Accepted: *31 May 2022*
Published: *08 July 2022*

Citation:
Toubiana R, Macdonald M,
Rajananda S, Lokvenec T, Kingsley TC
and Romero-Brufau S (2022)
Blockchain for Electronic Vaccine
Certificates: More Cons Than Pros?
Front. Big Data 5:833196.
doi: 10.3389/fdata.2022.833196

INTRODUCTION

The Rise of COVID-19 Electronic Vaccine Certificates

The requirement of proof-of-vaccination to COVID-19 is gaining traction in government agencies and the private sector, despite vocal opposition. The European Commission on December 21st, 2021 created regulations around the use of European Union Digital COVID Certificates (EUDCC) (EU Digital COVID Certificate, 2022). These regulations apply to all nations (non-EU included) that choose to adopt the EUDCC. Its primary use is to open travel between EU countries, but some nations are using it domestically to control entry to public places such as restaurants or sporting events. As of February 1st 2022, 42 countries are already connected to the EUDCC, and many more are considering joining (EU Digital COVID Certificate, 2022). The EUDCC uses a technology called distributed identity. The United States (US) federal government has taken a more limited role in regulating and mandating proof-of-vaccination through EVC platforms. This has left the responsibility to the private sector and state governments. Employers such as airlines, hospitals, and restaurants are increasingly requiring proof-of-vaccination for their patrons and employees (Eldred, 2021). Other non-EU countries are also evaluating EVC technology platforms to use domestically.

Blockchain Technology as a Solution

Blockchain has been a commonly proposed technology solution for COVID EVC platforms (Mithani et al., 2021). Although awareness of blockchain has increased because of the rise of digital currency such as Bitcoin and Ethereum, the majority of the public and decision makers have little understanding of the technology, especially in non-currency-based uses. Moreover, despite vocal opposition to proof-of-vaccination measures, it seems likely some versions of them will stay and become more widespread as COVID becomes more endemic, especially if COVID remains a deadly disease in those unvaccinated.

Blockchain use in EVCs is commonly proposed but there is a paucity of literature or real-world examples of its use for this purpose. As pressure increases for decision makers to choose amongst the various technology options, the authors of this paper thought it was important to review this topic.

Ten Important Characteristics of an EVC Technology Platform

As governments and the private sector are evaluating EVC platforms for deployment there are multiple considerations. Through discussion, our team identified 10 key considerations: (**1) data privacy and security** (patient health information, demographic data, location, etc), (**2) data verifiability and fidelity** (data remains auditable and accurate over time), (**3) data retrievability** (data can be queried and retrieved with accuracy and within a timeframe that is useful for its application), (**4) technology accessibility** (how easy it is for the public to access it as users), **(5) equitable (**regardless of socioeconomic, racial, or cultural differences), (**6) interoperability** with other public health and healthcare system information technology, (**7**) **scalability** (to be broadly available to the public within a short time period) **(8) cost effective** to maintain and operate (**9) potential for public adoption** (important factors include understandability, trust, and public perception of the technology), and (**10) feasibility** of development and operationalization (e.g., prior examples of the technology platform being successfully deployed in similar contexts).

BACKGROUND

Databases

A data storage application like an EVC system would traditionally use a database (generally what is called a relational database) to store patient and vaccination data. A relational database can be compared to a Microsoft Excel or Google Sheets document - data is stored in tables with rows of entries similar to a spreadsheet, and may contain multiple, possibly interlinked tables similar to the tabs in an Excel or Sheets document. Data can be retrieved from the database by writing queries in the appropriate query language, similar to the functions that can be used with Excel and Sheets. There are other types of databases that do not use blockchain technology, and the main benefit of databases is that they can also be optimized for specific use-cases, such as minimizing the size of the data and increasing the speed of updating or querying the database.

Theoretically, any kind of data can be represented in a database in any way, with any kind of relationships between different pieces of data. For example, for an EVC, there might be one table where each row contains the full private data of a patient and a vaccination they received. Alternatively, for a vaccine that requires multiple shots, data that is duplicated between each entry, such as a patient's details, could be entered into its own table which can then be linked to a second table that contains only the data for each shot. This way, the amount of data stored for each patient is reduced, and therefore so is the overall size of the database. This can lead to various improvements to the overall system, including the hard drive space required to store the database.

Generally, the security of the data in a database depends on the security of the systems it is connected to, unless the data itself is encrypted (see glossary). For example, most applications that use a database would have a user interface (UI) to make it easier for users to view and update the data in the database. Permission systems (such as usernames and passwords) can be used to control who can do what with the database data - for example, perhaps anyone with a login can read the data entries that pertain to themselves, but only some people can add or change data. The security of such an application then depends on factors like who has permission to do what operations, and how easy it is for a malicious entity to gain access to the database (e.g., by hacking the system or stealing login information from a user and using it to access the data *via* the user interface). Cryptographic techniques are commonly used at various points in an application in order to add layers of security.

WHAT IS BLOCKCHAIN TECHNOLOGY

Blockchain is a distributed ledger technology for storing and transmitting information. Its main characteristics are transparency, security, and decentralization (operating without a centralized control body) of both data and authority (Cawrey, 2021). A common application is money transfers that can be performed without the need for trusted third parties or banks. This is how Bitcoin or Ethereum work: thanks to blockchain, there is peer-to-peer (P2P) review that permits direct transfers between individuals.

The blockchain can therefore be compared to a public, anonymous and unforgeable accounting ledger. We can also think of this technology as a way to securely store private information such as vaccination records. In this section we describe what's known as a public blockchain, which is the original design by Nakamoto (2008). There are other variations of blockchain called "permissioned blockchains" that we will describe in the next section.

The first step is to initiate the transfer.

Let's say Mike wants to do a transaction with Santiago. If we consider Bitcoin for example, Mike would like to transfer money to Santiago; in that case we would have a record that says: "Mike pays Santiago 2 Bitcoins (transaction signed by Mike)." If we consider vaccination records, we could record the vaccination similarly: "Mike vaccinates Santiago (transaction

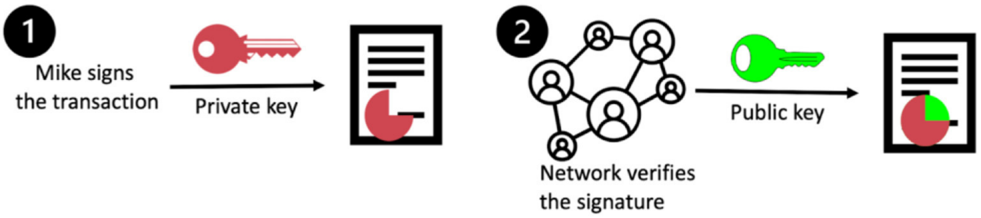


FIGURE 1 | Verification of Mike's identity.

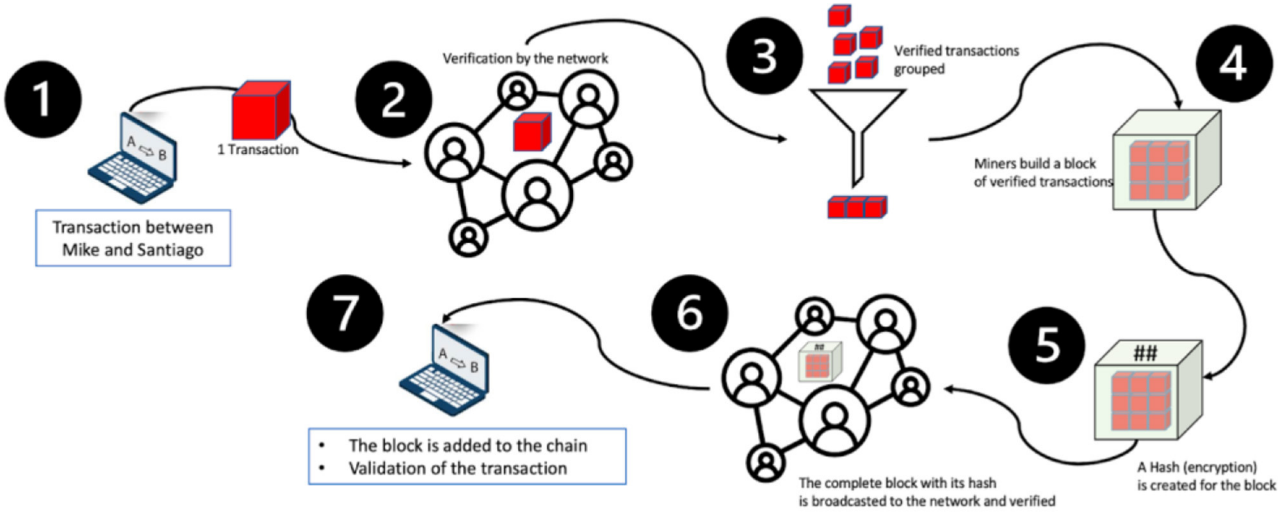


FIGURE 2 | Process of adding a transaction to the blockchain in 7 steps. The ## indicates the hash that was created for the block between steps 4 and 5.

signed by Mike)," with Mike being a vaccinator. A vaccinator is anyone approved to administer the vaccine, often a licensed healthcare provider or a public health official.

In step two, the transaction is sent to the network, composed of all the people using the blockchain, for verification. The first verification concerns the identity of the individuals involved in the transaction: is it really Mike that wants to do the transaction with Santiago?

How does this validation step work? Mike has to sign the transaction with an electronic signature called a private key. Only Mike has access to this key. The rest of the network has a public key that can only be used to decode Mike's private key. When the transaction is sent by Mike, several people in the network will verify that their public key decodes Mike's private key (**Figure 1**). If the public key doesn't decode Mike's private key, it means that it is not really Mike that sent the transaction. The transaction is thus canceled.

In the case of money transfers, the verification consists of verifying the identity of Mike with his electronic signature, as explained above, and verifying if Mike has enough money on his account to send to Santiago. In the case of vaccination records, one could envision a similar verification process using two keys to verify the identity of the parties.

The transaction is approved only if more than half of the people on the network accept it. This way, since there is a vast number of users, it is very unlikely that a compromised transaction will be approved.

Once the transaction is verified by the network, it is grouped together with other transactions to form a block (**Figure 2**, step 3).

On step four (**Figure 2**), a block is built for the group of transactions.

In Bitcoin and other proof-of-work systems, the "validators" of the chain, also called "miners," must spend computational work to find the solution to a mathematical problem, and that solution links the block to the chain. In systems using proof-of-stake or proof-of-authority, the miners only need to produce a digital signature that authenticates it to the network.

Once the block is validated, a timestamp is added to the block, i.e., the approximate date and time when the block was found.

Step five (**Figure 2**) is called hashing. Each block has an identifier, which is a unique cryptographic fingerprint, resulting from the hash of the data that this block contains: the transactions, the timestamp and the hash of the previous block. If someone attempts to modify the information stored in a block,

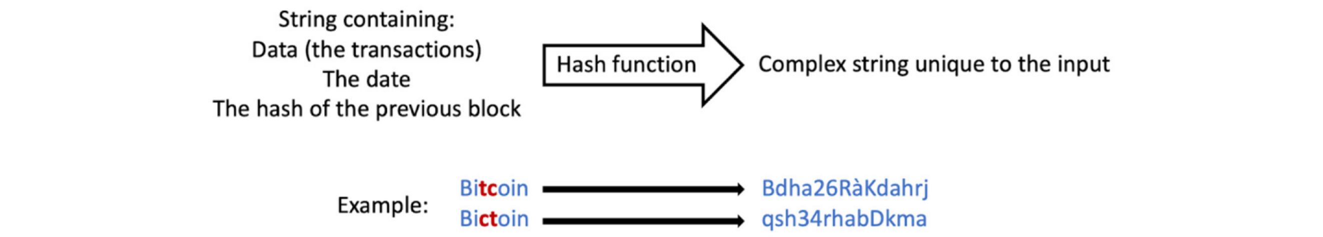


FIGURE 3 | Hashing.

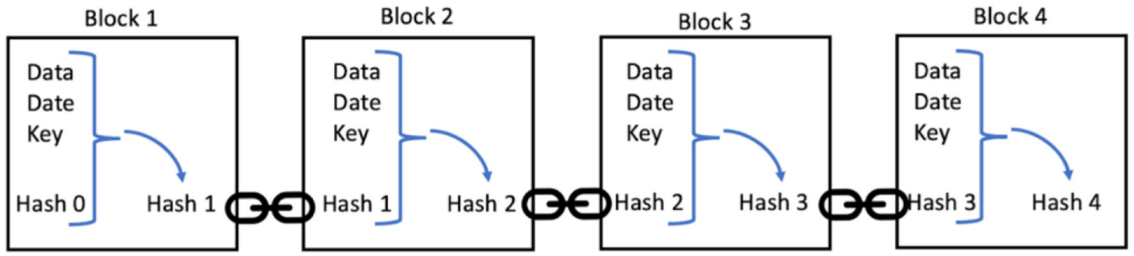


FIGURE 4 | Schematic of a Blockchain.

the hash will change drastically, and the fraud will be detected (see **Figure 3**).

The block is then broadcast to the network and is verified one last time before being added to the chain. We call this technology blockchain, because each block of transactions is linked to the previous one through the hash, as shown in **Figure 4**.

PERMISSIONED BLOCKCHAINS

In the previous section we have described the general functioning of blockchain technology. However, there are multiple variations, which can change critical aspects of the technology.

In general, there are three types of blockchains: public, consortium, and private (Zhang and Lin, 2018). Public blockchains such as Bitcoin allow anyone to participate: there are no restrictions on who can read or write to the blockchain. Consortium blockchains are permissioned blockchains where a consortium of entities are able to validate blocks. Access to the blockchain may vary between public or restricted (e.g., *via* APIs). Private blockchains are permissioned blockchains where a single entity has complete authority over the network and that entity fully controls both read and write permissions.

In the context of vaccination records, public blockchains will likely not suffice since vaccination records in the chain must be trustworthy (i.e., they should be added to the chain by a trusted medical entity). This then naturally leads to a private or consortium blockchain, where the ability to add to the chain and validate blocks can be restricted to only trusted entities, such as vaccinators (doctors and professionals in the medical community). In this scenario, we can imagine a certain trusted entity, such as the Health Ministry of one or several countries, having control over who is allowed to add vaccination records to the blockchain. A system like the European Union Digital Covid Certificate allows any of several countries to add vaccination records.

Proof-of-Work Validation

We have described how and when a block is validated. After this occurs, it is then added to the blockchain (Step 4 on **Figure 2**). However, there are many different consensus algorithms for validating blocks. The most popular, due to its use in Bitcoin and the way it incentivizes participation, is the proof-of-work algorithm.

In proof-of-work blockchains, a block is validated by performing a task that is computationally expensive, but easy to confirm. For example, in Bitcoin this task is finding a sequence when added to the block that will result in a hash that ends in a certain number of zeroes. This requires miners to use trial-and-error to find a sequence that will result in a certain hash. However, once that sequence is found, it is very easy to confirm that its hash has the required number of zeroes. Proof-of-work systems often need to provide an incentive to the agent who solved the problem. In currency-focused blockchains, this is easily solved by rewarding that agent with a certain amount of currency.

However, in an EVC system there isn't a clear reward that could be provided to the agent that validated the block. For these reasons, a proof-of-work validation algorithm would not be appropriate for this application, and other validation systems would need to be used. An algorithm which relies on a majority consensus between parties may be best, and especially, if used in a permissioned blockchain system, where the various entities are trusted.

TABLE 1 | Differences between public and permissioned blockchains.

Property	Public	Permissioned
Access restrictions	No restrictions inherent to the blockchain	Ability to read and write data to the blockchain is controlled
Trust	Doesn't require trust between agents in the network	Requires trust, due to agents having different read, write and validation permissions
Risk of takeover by majority of authoritative nodes	Anyone can join the network and validate transactions	Only some nodes are authoritative (can validate transactions)
Security	Malicious entities can easily gain access, and data is public	Permissions control who can do what, including viewing the data
Validation	Anyone can validate blocks, but validation is computationally expensive, so an incentive is generally needed	Trusted entities can be assigned the duty of validating blocks which removes the need for an incentive
Consensus algorithm	Can operate in an environment with low trust between entities, and may need to handle faults and malicious entities	Trust allows the consensus algorithm to be simplified

In **Table 1** we summarize the differences between a public blockchain and a permissioned blockchain. An EVC system would likely use a permissioned blockchain.

In some ways a permissioned blockchain is more similar to a traditional database compared to a public blockchain. For example, fewer authoritative entities means that an entity or group of entities could theoretically gain authority more easily, allowing them to block new transactions and rewrite their past transactions. However, in a permissioned blockchain, like in a traditional database, those entities need to be externally permissioned, which increases security.

However, a blockchain-based application will generally have more components than just the blockchain, such as user management and other data storage. A permissioned blockchain may allow for security trade-offs to be made elsewhere, such as choosing a less secure but faster consensus algorithm.

Considerations for Cooperative Applications

Decentralized authority may be an appealing solution when multiple entities are collectively using a system and each one is unwilling to let others have more authority over the system (such as countries sharing a common vaccination record system). This then could incentivize additional entities to join the blockchain.

However, a major hurdle for using blockchain technology on such a large scale is agreeing on a common protocol for the chain. These include the consensus mechanism, privacy standards, incentives for maintaining the chain, and managing write access to the chain. In addition, there has to be some level of trust that the other entities are managing their write access to the chain properly and those records can be trusted.

Some technical designs using consortium blockchains for EVC have been described (Haque et al., 2021).

In the case where multiple countries share the same blockchain, a consortium blockchain could theoretically be employed. This would allow each country to control the permissions to their respective medical institutions to write to the chain. Since no one country would have complete authority over the blockchain, the core benefit of decentralized authority would be preserved.

With regards to suitable blockchain platforms, Bitcoin and Ethereum are public, not consortium. Other platforms such as Multichain, Hyperledger Fabric and Hyperledger Sawtooth are likely more appropriate (Chowdhury et al., 2018; Chowdhury et al., 2019).

DIFFERENCES BETWEEN DATA STORAGE IN BLOCKCHAIN AND DATABASES

The biggest difference between blockchain and other types of distributed ledger technologies is the use of cryptographic techniques to add a layer of security to the data. While cryptography is often used for secrecy, in the context of blockchain the technology is used to make it significantly harder to change the transaction history, as described above. This is how cryptocurrency got its name. It is currency that is traded on the blockchain, many of the advantages of which come from the cryptographic techniques it utilizes.

As mentioned above, databases are based around storing data in tables with various methods for optimizing a database. This flexibility, especially combined with the various innovations in database technology and other fields over the last few decades, means there is very little to which databases are not suitable, with the right configuration.

Blockchain technology, in comparison, is designed to store individual data entries in a chronological manner. Innovations such as Ethereum have greatly improved what kinds of data can be stored on a blockchain, but the chronological nature of the technology and the fact that each data entry is independent of any other entry are core to blockchain.

With cryptocurrencies such as Bitcoin, people who use the currency do not directly access the blockchain to make transactions. Each user has a "wallet" which contains a list of their private keys, usually combined with a software interface with which users can manage keys and make transactions (Frankenfield, 2022). The data within a wallet is not stored on a blockchain. Instead, there are various data storage methods that are used, and one common option is a traditional database.

TABLE 2 | Properties of blockchain and how they relate to the EVC use case.

Property	Advantage	Disadvantage	Mitigation	Counterfactual
Decentralized authority (public blockchain)	**Safe operations** of applications **Incentivize co-operation** of shared authority	**Agreement** on protocols, etc. Can't control who has access	Use **private or consortium permissioned** blockchain	Standard databases can be **permissioned**
Decentralized data storage	**Less risk of data loss** with redundancy of data	The dataset for each authority can become **extremely large** blockchain	**Limit which entities** require the full blockchain **Limit on-chain** data storage	Minimization of data loss risk in traditional databases through **backups or other redundancy methods**
Immutability, data handling	**Improved data security** thanks to limited data operations (create, read)	**No updates or deletion** of data **Overhead** introduced to create and read operations	Data **can not be erroneous**, or **policies** must be created for changing chain history	All operations **allowed** in databases and can be **controlled** through permissions. Possible **performance optimization**
Timeline verification	**Reliable verification** of timeline	N/A	N/A	**Similar timeline verification** functionality with database encryption methods
Resource usage (energy and computation)	Usage **controlled by blockchain implementation** choices, e.g., consensus algorithm	**Significant energy consumption**, particularly of popular blockchain properties	**Architect** blockchain to reduce resource usage, e.g., choice of less energy-intensive **consensus** algorithm	Databases can be **optimized to minimize resource usage**
Pseudonymous identities	**Tracking** of transactions by entities	IDs (usernames) **can not be linked** to real-world identities without integration with **external systems**	Integrate with **external identity systems**	Standard databases can use **any identity verification system** and completely control the creation of identities
Performance	**Validity** of data and **ordering** thereof ensured	**Block validation speed** affects performance	**Carefully select properties** such as block size limit	Standard databases are **faster and more optimized**

Bold is for emphasis.

ANALYSIS OF BLOCKCHAIN TECHNOLOGY FOR EVC USE

Pros and Cons of BC Compared to Traditional Databases

Many blockchain platforms now exist, but most are designed for specific use cases or are too early in development or adoption for a use case as important as EVCs. The following therefore generalizes blockchain systems, based mainly on popular platforms Bitcoin and Ethereum. On **Table 2** we provide a summary of the characteristics of blockchain and how they relate to the EVC use case.

Decentralized Data Storage

Decentralized data storage means that, theoretically, every node would have a complete copy of the blockchain. However, blockchain data can grow quickly to gigabytes or even terabytes of data. For example, as of January 20th 2022, the blockchain size of Bitcoin was 386 GB for its 704 million transactions (Blockchain Charts). The full Ethereum chain was 1178.68 GB (Ethereum Chain Full Sync Data Size, 2022).

The full blockchain is required for authorities who validate blocks, but usually not required just to create transactions. It is also unrealistic that every entity would be willing to store the full chain. Therefore, these blockchains can create light nodes, which only store the data necessary to create transactions and rely on full nodes for other data as well as validation (Wackerow, 2022).

The blockchain size scales with the number of transactions and the data size of each transaction. Databases scale in a similar way, but as a more mature technology are optimized to reduce the impact. Data redundancy is another benefit of decentralized data but can also be achieved with databases using backups.

For context for the EVC use case, the population of the USA is 329.5 million with 551 million doses given. The population of the European Union is 447 million with roughly 848 million doses given (Daily COVID-19 vaccine doses administered, 2021; Ritchie et al., 2022). These vaccinations have been done in the last year, compared to Bitcoin's transaction history which goes back to 2009. This means that not only would vaccination records quickly exceed the size of Bitcoin transaction history, it would also present problems with record entry speeds.

Blockchain systems tend to limit how fast entries can be added by controlling how long or how big blocks can get. For example, Bitcoin is designed so a new block is mined every 6–10 min. This restriction on the system may be a significant problem with EVCs, whether they are set up at the beginning of vaccinations or, like now, potentially having to catch up with a significant number of past vaccinations.

Immutability, Data Handling and Performance

Databases support operations to create, read, update and delete (CRUD), and who performs each of these operations can be managed with permissions. Blockchain only supports create and read operations. As past transactions cannot be easily changed, this theoretically creates an immutable record. Rewriting the chain is technically possible, but extremely difficult. It would require changing past transactions, propagating the changes through the chain, then getting majority acceptance from the authoritative nodes. This would require recomputing blocks, which may be costly and slow. The majority agreement may also be difficult. Other options for changing the chain may be viable but depend on the specific blockchain implementation.

Databases can be optimized for the most used operations. Blockchain's "create" and "read" operations are slower due to the overhead of the validation and consensus mechanisms. Bottlenecks can also happen, such as block validation delays slowing transaction processing.

Databases are also designed to allow for any data to be queried based on any relationship between the data points. For example, an EVC database could likely be easily queried for "one patient's records," or "everyone vaccinated with a specific vaccine lot." Blockchain data is not designed to be queried in this way, as it is structured based on individual transactions and metadata about the entities doing transactions. It is possible to replicate such queries with blockchain technology, but due to it not being designed for such purposes this requires additional effort to implement and compute.

Whether immutability is beneficial for an application can depend on the risk of human error. For instance, is the data generated by a trusted program, or is it entered by humans who may make mistakes? If reading data very soon after it is created is important, databases may be preferable to blockchain.

Some existing blockchain applications try to get around some of the limitations of blockchain by using a combination of blockchain and databases. This requires careful implementation. A recent incident with OpenSea, a blockchain application that allows users to trade in images and other media, which used a hybrid blockchain and database approach to avoid Ethereum's high transaction fees. A bug was found where the blockchain and database got out of sync. This allowed an attacker to buy several items at an older, lower price, then sell them at the more recent price for a substantial profit (Cimpanu).

Timeline Verification

A major advantage of blockchain is that transaction validity and order can be easily verified. This is due to it being an immutable and chronological ledger. Databases can store timestamps for entries, security techniques can be applied to achieve immutability, and there are methods of encrypting database information to provide similar functionality.

In the case of EVCs, the specific order of the records is not critical. For example, it does not really matter whether Sue was vaccinated before or after Mary.

Pseudonymous Identities

An EVC system will require integration with real-world identification systems. A common example is using Social Security Numbers in the US to link the blockchain records with real-world people. This would apply to vaccinators, patients, and anyone else involved. There must also be checks to ensure individuals are not duplicated in the system.

Implementing these required checks in the blockchain system may be difficult for the same reasons querying data is difficult. Additionally, the existing identity systems are traditional databases, and integration with a blockchain-based system would add complexity and challenges.

Resource Usage

Blockchains can require a significant amount of computation and energy. Different blockchain implementations require different amounts due to factors like choice of consensus algorithm. In proof-of-work verification, nodes race to complete the computation of each block for a reward, but as a winner-takes-all contest, energy used by the losing nodes is wasted. Other consensus algorithms tend to use less energy (Chowdhury et al., 2018), thereby lowering the energy cost of the entire system.

Another consideration is the resource usage of everyone using the blockchain application. Because of its distributed nature, all full nodes who are capable of validating transactions. This requires each entity to have a computer storing the full blockchain and capable of validating nodes, which most likely must run continuously. This requirement may affect adoption in the case of EVCs, as it is an added cost and burden on those entities who would have authority to validate blocks. Light nodes at least must only store part of the blockchain, and do not need the computation ability to validate nodes. So careful organization of who requires a full node and who can use a light node can minimize this distributed cost.

Databases, in comparison, due to their centralized nature, only use the energy required to run their servers (including those used for backups) and external systems such as air conditioning (Sedlmeir et al., 2020). Users of the application would connect to it *via* the Internet, so no special machines or systems are needed. This also allows for low-cost backups that can be performed routinely but do not require to be constantly connected and computing.

Hype and Public Opinion

Blockchain, with regards to its use in cryptocurrencies, NFTs, and games, has been appearing in the news more often in recent months and years. It is a technology that is drawing a lot of attention and is often described as being "hyped" (Litan, 2021), meaning that the amount of attention and public expectations may surpass its actual delivery of progress. There have been reports of publicly-traded companies adding the term "blockchain" to their name and having their shares surge (What is in a name UK stock surgers 394% on blockchain rebrand, 2017). This points toward significant expectations associated with the term, regardless of its actual feasibility.

However, as with any novel term, its valence in the public opinion can quickly turn. For example, several companies in the software and gaming industries announced blockchain-related projects near the end of 2021, usually receiving mixed feedback from the general public. For example, when the CEO of Discord, a popular chat program, hinted at blockchain integration, there were supporters but also many users who were publicly against the move on Twitter, Reddit and Discord's own forum, and an unknown number canceled their paid subscriptions in protest (Orland, 2021). Molly White's timeline of problems with "web3" (a catch-all term for blockchain-based innovations), while focused on negative news, is a good indicator of what

TABLE 3 | Comparison of blockchain and alternative technologies regarding EVC requirements.

EVC platform technology feature	Optimal blockchain configuration compared to alternative technology solutions	Comments
Data privacy and security	**Equivalent or uncertain** based on current information	Both blockchain and standard databases can use similar cryptographic techniques [Transparent data encryption (TDE), 2022].
Data verifiability and fidelity	**Superior**	Harder to forge records without leaving a trace of it in blockchains
Data retrievability	**Inferior**	Blockchain's data structure is not designed for flexible data queries, databases are
Technology accessibility	**Equivalent or uncertain** based on current information	Depends on the front-end design and not much affected by the underlying data storage technology
Equitable	**Equivalent or uncertain** based on current information	Same as above. Mainly depends on accessibility.
Interoperability	**Inferior**	Blockchain is a less mature technology, and by design harder to modify? combining data registries or changing data standards is much harder
Scalability	**Inferior**	Traditional databases can be more easily scaled in transaction rate and storage
Cost effectiveness	**Inferior**	Blockchain's distributed nature makes it more costly to maintain. Traditional databases have been optimized for efficiency.
Potential for public adoption	**Equivalent or uncertain** based on current information	As a novel technology, public perception of blockchain can change quickly
Feasibility	**Inferior**	Blockchain is a less mature technology compared to time-tested database solutions.

Bold is for emphasis.

is happening in the space, especially in terms of its effects on the general public (White, 2022). It highlights that scams and hack are abundant in the web3 sphere, and many people are suffering losses, usually monetary, because of blockchain-based applications.

A question then, regarding adopting blockchain for EVCs, is "Will the public trust their data is safe on a blockchain-based solution?" Blockchain is known for being difficult to understand, not helped by the complexities around all the variations and different use cases it can be used for. If public opinion of the technology - informed or otherwise - becomes negative, will people be willing to have their private medical data stored using such a technology?

Assessment of Blockchain for EVC

In **Table 3** we summarize our assessment of the comparison between blockchain technology and traditional database solutions regarding the 10 key considerations presented in the introduction. As can be seen, blockchain only seems superior in the Data verifiability and fidelity domain, with all other aspects being either clearly inferior, equivalent, or uncertain.

CURRENT BLOCKCHAIN-BASED EVC SOLUTIONS

Some existing EVC solutions do claim to be using blockchain as part of their technology. A recent review by Mithani et al. (2021) listed eight such applications, including IBM's Digital Health Pass. However, most of these solutions have not made public the technical details of how blockchain is used. In fact, the solutions proposed in this article published in March 2021 are not operational today. Some of the webpages are not even functional. Raising the question whether would the projects are still active?

For these solutions, the question remains of whether blockchain is really a key part of the technology, or if the name is being used for the "hype factor." Given the lack of transparency it is hard to estimate the number of truly functional blockchain platforms in use for EVC, but from our teams estimate it appears to be none.

DISCUSSION

In this paper we have described the conceptual framework of blockchain technology as it could apply to storing electronic vaccine certificates (EVC). We have also discussed some of the advantages and drawbacks. Overall, blockchain technology seems to have more cons than pros for this use case. In line with our assessment, some widely-respected cyber-security companies have also assessed that blockchain is not necessary for EVCs, taking the example of the European COVID certificates system (Schubert, 2021).

A recent review of blockchain applications for COVID-19 (Ng et al., 2021) found that "vaccine passport monitoring" was one of the most common applications described in blockchain papers. However, most papers were limited to the technical description or reports of technical performance. Several blockchain system designs for vaccine supply management have also been described (Peng et al., 2020; Yong et al., 2020; Antal et al., 2021).

There have been other attempts to use blockchain technology for the storage and access to vaccination records using what is known as "smart contracts" (Zhao and Ma, 2022). In these approaches, the common idea is that the vaccination data (including vaccine certificates) is stored publicly but in encrypted form. The blockchain "smart contract" is then used to manage access to the key that would allow to decrypt the public data or a portion of it (Abubakar et al., 2021). This has been shown to significantly increase speed and convenience of data retrievability compared to scanning the blocks in the blockchain to find the vaccination information (Abuhashim et al., 2021).

As mentioned earlier, some of the main principles that inspired the creation of blockchain technology run counter to the EVC use case. For example, one of the key principles is decentralized authority. However, with vaccination records it makes sense to have one, or a few, central authorities who certify that an approved vaccine was administered. In a blockchain that stores information about money, the agreement in the network that a certain person has X amount can be enough to make that judgment meaningful. However, vaccines must correlate with an external event in the real world (the person's immunity status against a virus). That requires a central authority to determine, at least, that what was administered was a vaccine. This centralized assessment could be delegated to each "physician" agent in the network.

The aspect of blockchain technology that makes the most sense for the vaccination record use case, is the use of cryptography, which is closely linked to privacy. However, as we have discussed, a centralized or federated system to record and store vaccinations using cryptography can be designed without the use of blockchain, possibly using another distributed ledger technology. For example, a very simple system could store hashed records and make them publicly accessible. In the simplest form, there would be one hash per vaccination record. In this case the patient would go get their vaccine at a point of care and would have privileged access to the public record. After confirming the patient's identity, they would put information about the patient (e.g., patient full name and date of birth), the vaccine administered (e.g., vaccine name, provider, and lot number), and the date of administration, and create a hash with that information. Because this cryptographic hash is a one-way function that can't be tracked back, the hash can be posted publicly without loss of patient privacy. The provider would then upload this information into a public repository maintained by the authorized central agency (either the CDC or a similar organization). Then, to verify the patient's vaccination status, the patient would only need to present the information that was used to create the hash (which includes their identification), and the verifier could run it by the hashing function and compare to the public list of hashes posted in the trusted public repository. This hashing and comparison step could be easily automated into a phone app that would either read the patient's information from a printed vaccination card, or from a QR code that the patient would carry. There is a similar idea to that described in recent papers (Haque et al., 2021).

There are other questions that would need to be resolved almost independently of the technology used to store the vaccination records. There are several COVID vaccines available, with varying degrees of effectiveness. Ideally, the technology would store the information that is the most primary. In the case of an EVC, that's probably the record of which vaccine was administered, and when. This way, the rules of what constitutes a "fully vaccinated" patient can be flexible for different uses and can even be adjusted as more information becomes available. For example, if evidence becomes clear that vaccine efficacy wanes significantly with time, some countries may choose to include the time from the last dose in the definition of "fully vaccinated." However, even in this scenario, a central body still needs to decide whether some vaccines are not considered effective enough to even include in the record.

CONCLUSION

While blockchain has some useful applications, it does not seem to have clear advantages for electronic vaccine certificates (EVC) compared to more traditional database technologies. There is significant hype associated with blockchain that could be motivating its utilization for use cases in which it is not necessary. The existing EVC solutions that claim to use blockchain do not provide enough detail to assess whether blockchain is a core component of the system.

AUTHOR CONTRIBUTIONS

SR-B, RT, SR, and TL conceptualized the paper. SR, RT, TL, TK, and SR-B drafted the initial manuscript. MM significantly revised the manuscript critically. SR, RT, TL, and MM performed the review. RT composed the figures. SR-B and TK provided general guidance. All authors contributed to the article and approved the submitted version.

REFERENCES

Abubakar, M., McCarron, P., Jaroucheh, Z., and Buchanan, A. A. D. W. J. (2021). Blockchain-based Platform for Secure Sharing and Validation of Vaccination Certificates. *arXiv preprint arXiv*:211 2.10124.

Abuhashim, A. A., Shafei, H. A., and Tan, C. C. (2021). "Block-VC: a blockchain-based global vaccination certification," in *2021 IEEE International Conference on Blockchain* (Melbourne, VIC: IEEE), 347–352. Available online at: https://ieeexplore.ieee.org/abstract/document/9680556

Antal, C., Cioara, T., Antal, M., and Anghel, I. (2021). Blockchain platform For COVID-19 vaccine supply management. *IEEE Open J. Comput. Soc.* 2, 164–178. doi: 10.1109/OJCS.2021.3067450

Blockchain Charts. (2022). Available online at: https://www.blockchain.com/charts (accessed April, 2022).

Cawrey, L. L. D. (2021). *Mastering Blockchain, Vol. 1*. Sebastopol, CA: O'Reilly Media, Inc.

Chowdhury, M. J. M., Colman, A., Kabir, M. A., Han, J. and Sarda, P. (2018). "Blockchain versus database: a critical analysis," in *2018 17th IEEE International Conference on Trust, Security and Privacy in Computing and*

Communications/12th IEEE International Conference on Big Data Science and Engineering (IEEE), 1348–1353.

Chowdhury, M. J. M., Ferdous, M. S., Biswas, K., Chowdhury, N., Kayes, A. S. M., Alazab, M., et al. (2019). A comparative analysis of distributed ledger technology platforms. *IEEE Access* 7, 167930–167943. doi: 10.1109/ACCESS.2019.2953729

Cimpanu, C. (2022). *Hacker abuses OpenSea to buy NFT at older, cheaper prices.* The Record. Available online at: https://therecord.media/hacker-abuses-opensea-to-buy-nfts-at-older-cheaper-prices/ (accessed April, 2022).

Daily COVID-19 vaccine doses administered (2021). *Daily COVID-19 vaccine doses administered*. Available online at: https://ourworldindata.org/grapher/daily-covid-19-vaccination-doses (accessed April, 2022).

Eldred, S. M. (2021). "Coronavirus FAQ: is there an app that'll prove i'm vaccinated, or is paper the best?," in *NPR*. Online

Ethereum Chain Full Sync Data Size (2022). YCHARTS. Available online at: https://ycharts.com/indicators/ethereum_chain_full_sync_data_size (accessed April, 2022).

EU Digital COVID Certificate (2022). Available online at: https://ec.europa.eu/info/live-work-travel-eu/coronavirus-response/safe-covid-19-vaccines-europeans/eu-digital-covid-certificate_en (accessed April, 2022).

Frankenfield, J. (2022). *Bitcoin Wallet*. Available online at: https://www.investopedia.com/terms/b/bitcoin-wallet.asp (accessed April, 2022).

Haque, A. B., Naqvi, B., Islam, A. K. M., and Hyrynsalmi, S. (2021). Towards a GDPR-compliant blockchain-based COVID vaccination passport. *Appl. Sci.* 11, 6132. doi: 10.3390/app11136132

Litan, A. (2021). *Hype Cycle for Blockchain 2021; More Action than Hype*. Available online at: https://blogs.gartner.com/avivah-litan/2021/07/14/hype-cycle-for-blockchain-2021-more-action-than-hype/ (accessed April, 2022).

Mithani, S. S., Bota, A. B., Zhu, D. T., and Wilson, K. (2021). A scoping review of global vaccine certificate solutions for COVID-19. *Hum. Vaccin. Immunother.* 18, 1–12. doi: 10.1080/21645515.2021.1969849

Nakamoto, S. (2008). *Bitcoin: A Peer-to-Peer Electronic Cash System*. Available online at: https://www.debr.io/article/21260-bitcoin-a-peer-to-peer-electronic-cash-system (accessed April, 2022).

Ng, W. Y., Tan, T. E., Movva, P. V., Fang, A. H. S., Yeo, K. K., Ho, D., et al. (2021). Blockchain applications in health care for COVID-19 and beyond: a systematic review. *Lancet Digit. Health* 3, e819–e829. doi: 10.1016/S2589-7500(21)00210-7

Orland, K. (2021). *Discord CEO backs away from hinted NFT integration after backlash*. Available online at: https://arstechnica.com/gaming/2021/11/discord-ceo-backs-away-from-hinted-nft-integration-after-backlash/ (accessed April, 2022).

Peng, S., Hu, X., Zhang, J., Xie, X., Long, C., Tian, Z., et al. (2020). An efficient double-layer blockchain method for vaccine production supervision. *IEEE Trans. NanoBiosci.* 19, 579–587. doi: 10.1109/TNB.2020.2999637

Ritchie, H., Mathieu, E., Rodés-Guirao, L., Appel, C., Giattino, C., Ortiz-Ospina, E., et al. (2022). *Coronavirus (COVID-19) Vaccinations*. Available online at: https://ourworldindata.org/coronavirus (accessed April, 2022).

Schubert, I. (2021). *The New Technology Powering Europe's COVID Certificates*. Available online at: https://www.securid.com/en-us/blog/the-new-technology-powering-european-covid-certificates/ (accessed April, 2022).

Sedlmeir, J., Buhl, H. U., Fridgen, G., and Keller, R. (2020). The energy consumption of blockchain technology: beyond myth. *Bus. Inf. Syst. Eng.* 62, 599–608. doi: 10.1007/s12599-020-00656-x

Transparent data encryption (TDE) (2022). *SQL Docs*. Available online at: https://docs.microsoft.com/en-us/sql/relational-databases/security/encryption/transparent-data-encryption?view=sql-server-ver15 (accessed April, 2022).

Wackerow, P. (2022). *NODES AND CLIENTS*. Available online at: https://ethereum.org/en/developers/docs/nodes-and-clients/#node-types (accessed April, 2022).

What is in a name UK stock surgers 394% on blockchain rebrand (2017). Available online at: https://www.bloomberg.com/news/articles/2017-10-27/what-s-in-a-name-u-k-stock-surges-394-on-blockchain-rebrand (accessed April, 2022).

White, M. (2022). *Web3 is going great*. Available online at: https://web3isgoinggreat.com/ (accessed April, 2022).

Yong, B., Shen, J., Liu, X., Li, F., Chen, H., and Zhou, Q. (2020). A blockchain based system for safe vaccine supply and supervision. *Faculty Eng. Inf. Sci*. 52, 102024. doi: 10.1016/j.ijinfomgt.2019.10.009

Zhang, A., and Lin, X. (2018). Towards secure and privacy-preserving data sharing in e-health systems via consortium blockchain. *J. Med. Syst.* 42, 140. doi: 10.1007/s10916-018-0995-5

Zhao, Z., and Ma, J. (2022). Application of blockchain in trusted digital vaccination certificates. *China CDC Weekly* 4, 106–110. doi: 10.46234/ccdcw2022.021

Conflict of Interest: TK has a role in Pathcheck Foundation, a non-profit involved in the development of public health related technology that partially uses cryptography.

The remaining 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.

GLOSSARY

Cryptography is the study of techniques with which communications can be secured such that only the sender and intended recipient can understand the message. Encryption is a technique that is part of cryptography, where data is scrambled so that it is unintelligible, then sent to the recipient who knows how to unscramble it.

Encryption is the process of codifying the data so that it cannot be immediately read without an "decryption key". The data is scrambled (as with a hash) and can only be unscrambled into an understandable form by using the decryption key. Data that is encrypted is more secure because, even if a malicious agent manages to access the data storage, they won't be able to read the data itself unless they also have access to the decryption key.

Hashing is a method of scrambling data that is often used in encryption as it creates a fixed-length series of characters which are usually shorter than the original data. It is possible for different input data to produce the same hash, however choosing the correct hashing algorithm will mean that chances of that happening are considered too unlikely to be a risk. In this way, it can be compared to a fingerprint. Hashing is also a one-way function - given a hash, it is computationally infeasible (i.e., near to impossible given current computing technology) to calculate the original data, which gives us a secure way to represent a piece of data without using the data directly.

Public and private keys also come from cryptography, where the public-private key pairs are used as described in the previous section to scramble and unscramble data.

frontiers | Frontiers in Big Data

TYPE Opinion
PUBLISHED 27 September 2022
DOI 10.3389/fdata.2022.979465

The ABC recommendations for validation of supervised machine learning results in biomedical sciences

Davide Chicco [1]* and Giuseppe Jurman [2]

[1]Institute of Health Policy Management and Evaluation, University of Toronto, Toronto, ON, Canada,
[2]Data Science for Health Unit, Fondazione Bruno Kessler, Trento, Italy

KEYWORDS

supervised machine learning, computational validation, recommendations, data mining, best practices in machine learning

Check for updates

OPEN ACCESS

EDITED BY
Sujata Dash,
North Orissa University, India

REVIEWED BY
Ashish Luhach,
Papua New Guinea University of Technology, Papua New Guinea

*CORRESPONDENCE
Davide Chicco
davidechicco@davidechicco.it

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

RECEIVED 27 June 2022
ACCEPTED 12 September 2022
PUBLISHED 27 September 2022

CITATION
Chicco D and Jurman G (2022) The ABC recommendations for validation of supervised machine learning results in biomedical sciences. *Front. Big Data* 5:979465.
doi: 10.3389/fdata.2022.979465

1. Introduction

Supervised machine learning has become pervasive in the biomedical sciences nowadays (Larrañaga et al., 2006; Tarca et al., 2007), and its validation has obtained a key role in all these scientific fields. We therefore read with great interest the article by Walsh et al. (2021), which reported a list of DOME recommendations to properly validate results achieved with supervised machine learning, according to the authors. In the past, several studies already listed common best practices and recommendations for the proper usage of machine learning (Bhaskar et al., 2006; Domingos, 2012; Chicco, 2017; Cearns et al., 2019; Stevens et al., 2020; Artrith et al., 2021; Cabitza and Campagner, 2021; Larson et al., 2021; Whalen et al., 2021; Lee et al., 2022) and computational statistics (Benjamin et al., 2018; Makin and de Xivry, 2019), but the comment by Walsh et al. (2021) has the merit to highlight the importance of computational validation, which is a key step perhaps even more important than the machine learning algorithm design itself.

Although interesting and complete, that article describes numerous of steps and aspects in a way that we find complicated, especially for beginners. We believe that the 21 questions of the Box 1 of the DOME article (Walsh et al., 2021) can be adequate for a data mining expert, but they might scare and discourage an inexperienced practitioner. For example, the recommendations about the *meta-predictions* and about the hyper-parameters' optimization might not be understandable by a machine learning beginner or by a wet lab biologist. And it should not be a problem: a robust machine learning analysis can be performed, in fact, without using meta-predictions or hyper-parameters, too. A beginner, in front of so many guidelines of that article, some of which being so complex, might even decide to abandon the computational intelligence analysis, to avoid making any mistake in their scientific project. Moreover, the DOME (Walsh et al., 2021) authors present the 21 questions of the article Box 1 with the same level of importance. In contrast, we think that three key aspects to keep in mind for computational validation are pivotal and can be sufficient, if verified correctly. So we believe that a practitioner would better focus all their attention and energy on accurately respecting these three recommendations.

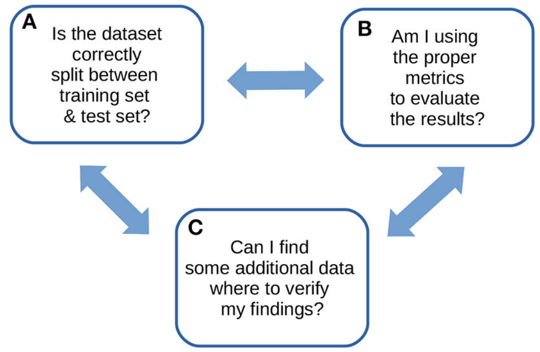

A
Is the dataset correctly split between training set & test set?
B
Am I using the proper metrics to evaluate the results?
C
Can I find some additional data where to verify my findings?

FIGURE 1
ABC recommendations checklist. An overview of our ABC recommendations, to keep in mind for any machine learning study.

We therefore wrote this note to propose our own recommendations for the computational validation of supervised machine learning results in the biomedical sciences: just three, explained easily and clearly, that alone can pave the way for a successful machine learning validation phase. We designed these simple quick tips from our experience gained on tens of biomedical projects involving machine learning phases. We call these recommendations ABC to highlight their essential role in any computational validation (Figure 1).

2. The ABC recommendations

(A) Always divide the dataset carefully into separate training set and test set

This rule must become your obsession: verify and double-check that no data element is shared by both the training set and the test set. They must be completely independent.

You then can do anything you want on the training set, including the hyper-parameter optimization, but make sure you do not touch the test set. Leave the test set alone until your supervised machine learning model training has finished (and its hyper-parameters are optimized, if any). If you have enough data, consider also allocating a subset of it (such as 10% of data elements, randomly selected) as a holdout set (Skocik et al., 2016), to use as an alternative test set to confirm your findings and to avoid over-validation (Wainberg et al., 2016).

This important separation will allow you to avoid *data snooping* (White, 2000; Smith, 2021), that is a common mistake in multiple studies involving computational intelligence (Jensen, 2000; Sewell, 2021). Data snooping, also known as *data dredging* and called "the dark side of data mining" (Jensen, 2000), happens in fact when some data elements of the training set are present in the test set, too, and therefore over-optimistically improve the results obtained by the trained machine learning model on the test set. Sometimes, this problem can happen even when different data elements of the same patients (for example, radiography images in digital pathology) are shared between training set and test set, and is usually called *data leakage* (Bussola et al., 2021). This mistake is dangerous for every machine learning study, because it can give the illusion of success to an unaware researcher. In this situation, you need to keep in mind the famous quote by Richard Feynman: "The first principle is that you must not fool yourself, and you are the easiest person to fool" (Chicco, 2017).

Data snooping does exactly that: it makes you fool yourself and makes you believe you obtained excellent results, while actually machine learning performance was flawed. Once you make sure the training set and the test set are independent from each other, you can use traditional cross-validation methods such as *k*-fold cross-validation, leave-one-out cross-validation, and nested cross-validation (Yadav and Shukla, 2016), or bootstrap validation (Efron, 1992; Efron and Tibshirani, 1994), to mitigate over-fitting (Dietterich, 1995; Chicco, 2017). Moreover, over-fitting can be tackled through calibration methods such as calibration curves (Austin et al., 2022) or calibration-in-the-large (Crowson et al., 2016), which can also help measuring the robustness of model performance.

Moreover, it is important to notice that sometimes splitting the dataset into two subsets (training set and test set) might not be enough (Picard and Berk, 1990). Even for shallow machine learning models, a correct splitting methodology should be enforced: for instance, see the Data Analysis Protocol strategy introduced by the MAQC/SEQC initiatives led by the US Food and Drug Administration (FDA) (MAQC Consortium, 2010; Zhang et al., 2015). And when there are hyper-parameters to optimize (Feurer and Hutter, 2019), such as the number of hidden layers and the number of hidden units in artificial neural networks, it is advisable to split the dataset into three subsets: training set, validation set, and test set (Chicco, 2017). In these cases, sometimes in scientific literature the names *validation set* and *test set* are used interchangeably; in this report, we call *validation set* the part of the dataset employed to evaluate the algorithm configuration with a particular hyper-parameter value, and we call *test set* the portion of the dataset to keep untouched and eventually use to verify the algorithm with the optimal hyper-parameters' configuration.

(B) Broadly use multiple rates to evaluate your results

Evaluate your results with various rates, and definitely include the Matthew's correlation coefficient (MCC) (Matthews, 1975) for binary classifications (Chicco and Jurman, 2020; Chicco et al., 2021a) and the coefficient of

TABLE 1 Recap of the suggested metrics for evaluating results of binary classifications and regression analyses.

Analysis type	Always include	We suggest to include
Binary classification	MCC	TPR, TNR, PPV, NPV, accuracy, F_1 score, Cohen's Kappa, ROC AUC, and PR AUC
Regression analysis	R^2	SMAPE, MAPE, MAE, MSE, and RMSE

The formulas of the binary classification rates can be found in Chicco and Jurman (2020) and Chicco et al. (2021a,c) and the formulas of the regression analysis rates can be found in Chicco et al. (2021b).

determination (R^2) (Wright, 1921) for regression analyses (Chicco et al., 2021b). Moreover, make sure you include at least accuracy, F_1 score, sensitivity, specificity, precision, negative predictive value, Cohen's Kappa, and the area under the curve (AUC) of the receiving operating characteristic curve (ROC) and of the prediction-recall curve (PR) for binary classifications. For regression analyses, make sure you incorporate at least mean absolute error (MAE), mean absolute percentage error (MAPE), mean square error (MSE), root mean square error (RMSE), and symmetric mean absolute percentage error (SMAPE), in addition to the already-mentioned R^2. We recap our suggestions in Table 1.

It is necessary to include all these scores because each of them provides a singular, useful piece of information about your supervised machine learning results. The more statistics you include, the more chances you have to spot any possible flaw in your predictions. All these rates work like dashboard indicator lamps in a car: if something somewhere in your machine (learning) did not work out the way it was supposed to, a lamp (rate) will inform you about it.

The Matthew's correlation coefficient, in particular, has a fundamental role in binary classification evaluation: it has a high score only if the classifier correctly predicted most of the positive elements and of the negative elements, and only if the classifier made mostly correct positive predictions and mostly correct negative predictions (Chicco and Jurman, 2020, 2022; Chicco et al., 2021; Chicco et al., 2021a). That means, a high MCC corresponds to a high score for all the four basic rates of a 2 × 2 confusion matrix: sensitivity, specificity, precision, and negative predictive value (Chicco et al., 2021a). Because of its efficacy, the MCC has been employed as the standard metric in several scientific projects. For example, the USFDA agency used the MCC as the main evaluation rate in the MicroArray II/Sequencing Quality Control (MAQC/SEQC) projects (MAQC Consortium, 2010; SEQC/MAQC-III Consortium, 2014).

Regarding regression analysis assessment, the coefficient of determination R-squared (R^2) is the only rate that generates a high score only if the predictive algorithm was able to correctly predict most of the elements of each data class, considering their distribution (Chicco et al., 2021b). Additionally, R^2 allows the comparison of models applied to datasets having different scales (Chicco et al., 2021b). Because of its effectiveness, the coefficient of determination has been employed as the standard evaluation metric for several international scientific projects, such as the Overhead Geopose DrivenData Challenge (DrivenData.org, 2022) and the Breast Cancer Prognosis DREAM Education Challenge (Bionetworks, 2021).

(C) Confirm your findings with external data, if possible

If you can, use data coming from a different data source and made of a different data type from the main dataset to verify your discoveries. Obtaining the same results you achieved on the main original dataset on an external dataset coming from another scientific research centre would be a strong confirmation of your scientific findings. Moreover, if this external data were in a data type different from the original data, it would even increase the level of independence between the two datasets, and even more strongly confirm your scientific outcomes.

In a bioinformatics study, for example, Kustra and Zagdanski (2008) employed a data fusion approach to cluster microarray gene expression data and associate the derived clusters to Gene Ontology annotations (Gene Ontology Consortium, 2019). For validating their results, instead of using a different microarray dataset, the authors decided to take advantage of an external database made of a different data type: a protein–protein database called General Repository for Interaction Data Sets (GRID) (Breitkreutz et al., 2003). This way, the authors were able to find in external data a strong confirmation of the results they obtained on the original data, and therefore were able to claim their study outcomes as robust and reliable in their manuscript's conclusions.

Moving from bioinformatics to health informatics, a call for external data validation has recently been raised in machine learning and computational statistics applied to heart failure prediction as well (Shin et al., 2021).

That being said, we are aware that obtaining compatible additional data and integrating them might be difficult for some biomedical studies, but we still invite all the machine learning practitioners to make an attempt and to try to collect confirmatory data for their analyses anyway. In some cases, there are plenty of public datasets available for free use that can be downloaded and integrated easily.

Bioinformaticians working on gene expression analysis, for example, can take advantage of the thousands of different datasets available on the Gene Expression Omnibus (GEO) (Edgar et al., 2002). Tens of compatible datasets of a particular cancer type can be found by specifying the microarray platform, for example, through the recently released `geoCancerPrognosticDatasetsRetriever` (Alameer and Chicco, 2022) bioinformatics tool. Researchers can take advantage of these compatible datasets (for example, built on the GPL570 Affymetrix platform) to verify their findings, after applying some quality-control and preprocessing steps such as batch correction (Chen et al., 2011) and data normalization, if needed.

Moreover, public data repositories for biomedical domains, such as ophthalmology images (Khan et al., 2021), cancer images (Clark et al., 2013), or neuroblastoma electronic health records (Chicco et al., in press), can provide additional datasets that can be used as validation cohorts. Additional public datasets can be found on the University of California Irvine Machine Learning Repository (University of California Irvine, 1987), on the DREAM Challenges platform (Kueffner et al., 2019; Sage Bionetworks, 2022), or on Kaggle (Kaggle, 2022), for example.

When using external data, an aspect to keep in mind is checking and correcting issues like dataset shift (Finlayson et al., 2021) and model underspecification (D'Amour et al., 2020), which might jeopardize the coherence of the learning pipeline when moving from training and testing and validation.

3. Discussion

Computational intelligence makes computers able to identify trends in data that otherwise would be difficult or impossible to notice by humans. With the spread of new technologies and electronic devices able to save and store large amounts of data, data mining has become a ubiquitous tool in numerous scientific studies, especially in biomedical informatics. In these studies, the validation of the results obtained through supervised machine learning has become a crucial phase, especially because of the high risk of achieving over-optimistic, inflated results, that can even lead to false discoveries (Ioannidis, 2005).

In the past, several studies proposed rules and guidelines to develop more effective and efficient predictive models in medical informatics and computational epidemiology (Steyerberg and Vergouwe, 2014, Riley et al., 2016, 2021; Bonnett et al., 2019; Wolff et al., 2019; Navarro et al., 2021; Van Calster et al., 2021). Most of them however, provided complicated lists of steps and tips which might be hard to follow by machine learning practitioners, especially by beginners.

In this context, the article of Walsh et al. (2021) plays its part by describing thoroughly several DOME recommendations and steps for validating supervised machine learning results, but in our opinion it suffers from excessive complexity and might be difficult to follow by beginners. In this note, we propose our own simple, easy, essential ABC tips to keep in mind when validating results obtained with data mining methods.

We believe our ABC recommendations can be an effective tool to follow for all the machine learning practitioners, both by beginners and experienced ones, and can pave the way to stronger, more robust, more reliable scientific results in all the biomedical sciences.

Author contributions

DC conceived the study and wrote most of the article. GJ reviewed and contributed to the 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

Alameer, A., and Chicco, D. (2022). geoCancerPrognosticDatasetsRetriever, a bioinformatics tool to easily identify cancer prognostic datasets on Gene Expression Omnibus (GEO). *Bioinformatics* 2021:btab852. doi: 10.1093/bioinformatics/btab852

Artrith, N., Butler, K. T., Coudert, F. -X., Han, S., Isayev, O., Jain, A., et al. (2021). Best practices in machine learning for chemistry. *Nat. Chem.* 13, 505–508. doi: 10.1038/s41557-021-00716-z

Austin, P. C., Putter, H., Giardiello, D., and van Klaveren, D. (2022). Graphical calibration curves and the integrated calibration index (ICI) for competing risk models. *Diagn. Progn. Res.* 6, 1–22. doi: 10.1186/s41512-021-00114-6

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. -J., Berk, R., et al. (2018). Redefine statistical significance. *Nat. Hum. Behav.* 2, 6–10. doi: 10.1038/s41562-017-0189-z

Bhaskar, H., Hoyle, D. C., and Singh, S. (2006). Machine learning in bioinformatics: a brief survey and recommendations for practitioners. *Comput. Biol. Med.* 36, 1104–1125. doi: 10.1016/j.compbiomed.2005.09.002

Bionetworks, S. (2021). *Breast Cancer Prognosis DREAM Education Challenge*. Available online at: https://www.synapse.org/#!Synapse:syn8650663/wiki/436447 (accessed August 12, 2021).

Bonnett, L. J., Snell, K. I. E., Collins, G. S., and Riley, R. D. (2019). Guide to presenting clinical prediction models for use in clinical settings. *BMJ* 365:l737. doi: 10.1136/bmj.l737

Breitkreutz, B. -J., Stark, C., and Tyers, M. (2003). The GRID: the general repository for interaction datasets. *Genome Biol.* 4:R23. doi: 10.1186/gb-2003-4-2-p1

Bussola, N., Marcolini, A., Maggio, V., Jurman, G., and Furlanello, C. (2021). "AI slipping on tiles: data leakage in digital pathology," in *Proceedings of ICPR 2021 – The 25th International Conference on Pattern Recognition. ICPR International Workshops and Challenges* (Berlin: Springer International Publishing), 167–182.

Cabitza, F., and Campagner, A. (2021). The need to separate the wheat from the chaff in medical informatics: introducing a comprehensive checklist for the (self)-assessment of medical AI studies. *Int. J. Med. Inform.* 153:104510. doi: 10.1016/j.ijmedinf.2021.104510

Cearns, M., Hahn, T., and Baune, B. T. (2019). Recommendations and future directions for supervised machine learning in psychiatry. *Transl. Psychiatry* 9:271. doi: 10.1038/s41398-019-0607-2

Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L., et al. (2011). Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLoS ONE* 6:e17238. doi: 10.1371/journal.pone.0017238

Chicco, D. (2017). Ten quick tips for machine learning in computational biology. *BioData Min.* 10:35. doi: 10.1186/s13040-017-0155-3

Chicco, D., Cerono, G., Cangelosi, D. (in press). A survey on publicly available open datasets of electronic health records (EHRs) of patients with neuroblastoma. *Data Sci. J.* 1–15.

Chicco, D., and Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 21:6. doi: 10.1186/s12864-019-6413-7

Chicco, D., and Jurman, G. (2022). An invitation to greater use of Matthews correlation coefficient in robotics and artificial intelligence. *Front. Robot. AI* 9:876814. doi: 10.3389/frobt.2022.876814

Chicco, D., Starovoitov, V., and Jurman, G. (2021). The benefits of the Matthews correlation coefficient (MCC) over the diagnostic odds ratio (DOR) in binary classification assessment. *IEEE Access.* 9, 47112–47124. doi: 10.1109/ACCESS.2021.3068614

Chicco, D., Tötsch, N., and Jurman, G. (2021a). The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Min.* 14:13. doi: 10.1186/s13040-021-00244-z

Chicco, D., Warrens, M. J., and Jurman, G. (2021b). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Comput. Sci.* 7:e623. doi: 10.7717/peerj-cs.623

Chicco, D., Warrens, M. J., and Jurman, G. (2021c). The Matthews correlation coefficient (MCC) is more informative than Cohens Kappa and Brier score in binary classification assessment. *IEEE Access.* 9, 78368–78381. doi: 10.1109/ACCESS.2021.3084050

Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., et al. (2013). The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. *J. Digit. Imaging* 26, 1045–1057. doi: 10.1007/s10278-013-9622-7

Crowson, C. S., Atkinson, E. J., and Therneau, T. M. (2016). Assessing calibration of prognostic risk scores. *Stat. Methods Med. Res.* 25, 1692–1706. doi: 10.1177/0962280213497434

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., et al. (2020). Underspecification presents challenges for credibility in modern machine learning. *arXiv Preprint arXiv:2011.03395*. doi: 10.48550/arXiv.2011.03395

Dietterich, T. (1995). Overfitting and undercomputing in machine learning. *ACM Comput. Surveys* 27, 326–327. doi: 10.1145/212094.212114

Domingos, P. (2012). A few useful things to know about machine learning. *Commun. ACM* 55, 78–87. doi: 10.1145/2347736.2347755

DrivenData.org (2022). *Overhead Geopose Challenge.* Available online at: https://www.drivendata.org/competitions/78/competition-overhead-geopose/page/372/ (accessed August 12, 2021).

Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucl. Acids Res.* 30, 207–210. doi: 10.1093/nar/30.1.207

Efron, B. (1992). "Bootstrap methods: another look at the jackknife," in *Breakthroughs in Statistics*, eds S. Kotz and N. L. Johnson (New York, NY: Springer), 569–593. doi: 10.1007/978-1-4612-4380-9_41

Efron, B., and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. New York, NY: CRC Press. doi: 10.1201/9780429246593

Feurer, M., and Hutter, F. (2019). "Hyperparameter optimization," in *Automated Machine Learning*, eds F. Hutter, L. Kotthoff, and J. Vanschoren (Berlin: Springer), 3–33. doi: 10.1007/978-3-030-05318-5_1

Finlayson, S. G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., et al. (2021). The clinician and dataset shift in artificial intelligence. *N. Engl. J. Med.* 385, 283–286. doi: 10.1056/NEJMc2104626

Gene Ontology Consortium (2019). The Gene Ontology resource: 20 years and still GOing strong. *Nucl. Acids Res.* 47, D330–D338. doi: 10.1093/nar/gky1055

Ioannidis, J. P. (2005). Why most published research findings are false. *PLOS Med.* 2:e124. doi: 10.1371/journal.pmed.0020124

Jensen, D. (2000). Data snooping, dredging and fishing: the dark side of data mining a SIGKDD99 panel report. *ACM SIGKDD Explor. Newsl.* 1, 52–54. doi: 10.1145/846183.846195

Kaggle (2022). *Kaggle.com – Find Open Datasets*. Available online at: https://www.kaggle.com/datasets (accessed March 27, 2022).

Khan, S. M., Liu, X., Nath, S., Korot, E., Faes, L., Wagner, S. K., et al. (2021). A global review of publicly available datasets for ophthalmological imaging: barriers to access, usability, and generalisability. *Lancet Digit. Health* 3, e51–e66. doi: 10.1016/S2589-7500(20)30240-5

Kueffner, R., Zach, N., Bronfeld, M., Norel, R., Atassi, N., Balagurusamy, V., et al. (2019). Stratification of amyotrophic lateral sclerosis patients: a crowdsourcing approach. *Sci. Reports* 9:690. doi: 10.1038/s41598-018-36873-4

Kustra, R., and Zagdanski, A. (2008). Data-fusion in clustering microarray data: balancing discovery and interpretability. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 7, 50–63. doi: 10.1109/TCBB.2007.70267

Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., et al. (2006). Machine learning in bioinformatics. *Brief. Bioinform.* 7, 86–112. doi: 10.1093/bib/bbk007

Larson, D. B., Harvey, H., Rubin, D. L., Irani, N., Justin, R. T., and Langlotz, C. P. (2021). Regulatory frameworks for development and evaluation of artificial intelligence–based diagnostic imaging algorithms: summary and recommendations. *J. Amer. Coll. Radiol.* 18, 413–424. doi: 10.1016/j.jacr.2020.09.060

Lee, B. D., Gitter, A., Greene, C. S., Raschka, S., Maguire, F., Titus, A. J., et al. (2022). Ten quick tips for deep learning in biology. *PLoS Comput. Biol.* 18:e1009803. doi: 10.1371/journal.pcbi.1009803

Makin, T. R., and de Xivry, J.-J. O. (2019). Science forum: ten common statistical mistakes to watch out for when writing or reviewing a manuscript. *eLife* 8:e48175. doi: 10.7554/eLife,.48175.005

MAQC Consortium (2010). The MicroArray quality control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nat. Biotechnol.* 28, 827–838. doi: 10.1038/nbt.1665

Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta Prot. Struct.* 405, 442–451. doi: 10.1016/0005-2795(75)90109-9

Navarro, C. L. A., Damen, J. A., Takada, T., Nijman, S. W., Dhiman, P., Ma, J., et al. (2021). Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review. *BMJ* 375:n2281. doi: 10.1136/bmj.n2281

Picard, R. R., and Berk, K. N. (1990). Data splitting. *Amer. Stat.* 44, 140–147. doi: 10.1080/00031305.1990.10475704

Riley, R. D., Debray, T. P. A., Collins, G. S., Archer, L., Ensor, J., Smeden, M., et al. (2021). Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat. Med.* 40, 4230–4251. doi: 10.1002/sim.9025

Riley, R. D., Ensor, J., Snell, K. I. E., Debray, T. P. A., Altman, D. G., Moons, K. G. M., et al. (2016). External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. *BMJ* 353:i3140. doi: 10.1136/bmj.i3140

Sage Bionetworks (2022). *DREAM Challenges Publications*. Available online at: https://dreamchallenges.org/publications/ (accessed January 17, 2022).

SEQC/MAQC-III Consortium (2014). A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the sequencing quality control consortium. *Nat. Biotechnol.* 32, 903–914. doi: 10.1038/nbt.2957

Sewell, M. (2021). *Data Snooping.* Available online at: http://data-snooping.martinsewell.com (accessed August 6, 2021).

Shin, S., Austin, P. C., Ross, H. J., Abdel-Qadir, H., Freitas, C., Tomlinson, G., et al. (2021). Machine learning vs. conventional statistical models for predicting heart failure readmission and mortality. *ESC Heart Fail.* 8, 106–115. doi: 10.1002/ehf2.13073

Skocik, M., Collins, J., Callahan-Flintoft, C., Bowman, H., and Wyble, B. (2016). I tried a bunch of things: the dangers of unexpected overfitting in classification. *bioRxiv* 2016:078816. doi: 10.1101/078816

Smith, M. K. (2021). *Data snooping.* Available online at: https://web.ma.utexas.edu/users/mks/statmistakes/datasnooping.html (accessed August 5, 2021).

Stevens, L. M., Mortazavi, B. J., Deo, R. C., Curtis, L., and Kao, D. P. (2020). Recommendations for reporting machine learning analyses in clinical research. *Circ. Cardiovasc. Qual. Outcomes* 13:e006556. doi: 10.1161/CIRCOUTCOMES.120.006556

Steyerberg, E. W., and Vergouwe, Y. (2014). Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur. Heart J.* 35, 1925–1931. doi: 10.1093/eurheartj/ehu207

Tarca, A. L., Carey, V. J., Chen, X.-W., Romero, R., and Drăghici, S. (2007). Machine learning and its applications to biology. *PLoS Comput. Biol.* 3:e116. doi: 10.1371/journal.pcbi.0030116

University of California Irvine (1987). *Machine Learning Repository*. Available online at: https://archive.ics.uci.edu/ml (accessed January 12, 2021).

Van Calster, B., Wynants, L., Riley, R. D., van Smeden, M., and Collins, G. S. (2021). Methodology over metrics: current scientific standards are a disservice to patients and society. *J. Clin. Epidemiol.* 138, 219–226. doi: 10.1016/j.jclinepi.2021.05.018

Wainberg, M., Alipanahi, B., and Frey, B. J. (2016). Are random forests truly the best classifiers? *J. Mach. Learn. Res.* 17, 3837–3841. doi: 10.5555/2946645.3007063

Walsh, I., Fishman, D., Garcia-Gasulla, D., Titma, T., Pollastri, G., Capriotti, E., et al. (2021). DOME: recommendations for supervised machine learning validation in biology. *Nat. Methods* 5, 1122–1127. doi: 10.1038/s41592-021-01205-4

Whalen, S., Schreiber, J., Noble, W. S., and Pollard, K. S. (2021). Navigating the pitfalls of applying machine learning in genomics. *Nat. Rev. Genet.* 23, 169–181. doi: 10.1038/s41576-021-00434-9

White, H. (2000). A reality check for data snooping. *Econometrica* 68, 1097–1126. doi: 10.1111/1468-0262.00152

Wolff, R. F., Moons, K. G., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., et al. (2019). PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann. Intern. Med.* 170, 51–58. doi: 10.7326/M18-1376

Wright, S. (1921). Correlation and causation. *J. Agric. Res.* 557–585.

Yadav, S., and Shukla, S. (2016). "Analysis of *k*-fold cross-validation over hold-out validation on colossal datasets for quality classification," in *Proceedings of IACC 2016—the 6th International Conference on Advanced Computing* (Bhimavaram), 78–83.

Zhang, W., Yu, Y., Hertwig, F., Thierry-Mieg, J., Zhang, W., Thierry-Mieg, D., et al. (2015). Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. *Genome Biol.* 16:133. doi: 10.1186/s13059-015-0694-1

Frontiers in Big Data

TYPE Original Research
PUBLISHED 10 October 2022
DOI 10.3389/fdata.2022.1021518

Imbalanced ECG signal-based heart disease classification using ensemble machine learning technique

Adyasha Rath[1], Debahuti Mishra[1] and Ganapati Panda[2]*

[1]Department of Computer Science and Engineering, Siksha O Anusandhan (Deemed to be) University, Bhubaneswar, Odisha, India, [2]Department of Electronics and Tele Communication, C. V. Raman Global University, Bhubaneswar, Odisha, India

OPEN ACCESS

EDITED BY
Sujata Dash,
North Orissa University, India

REVIEWED BY
Sitanshu Sekhar Sahu,
Birla Institute of Technology,
Mesra, India
Soumya Ranjan Nayak,
Amity University, India

*CORRESPONDENCE
Ganapati Panda
ganapati.panda@gmail.com

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

RECEIVED 17 August 2022
ACCEPTED 07 September 2022
PUBLISHED 10 October 2022

CITATION
Rath A, Mishra D and Panda G (2022) Imbalanced ECG signal-based heart disease classification using ensemble machine learning technique. *Front. Big Data* 5:1021518. doi: 10.3389/fdata.2022.1021518

The machine learning (ML)-based classification models are widely utilized for the automated detection of heart diseases (HDs) using various physiological signals such as electrocardiogram (ECG), magnetocardiography (MCG), heart sound (HS), and impedance cardiography (ICG) signals. However, ECG-based HD identification is the most common one used by clinicians. In the current investigation, the ECG records or subjects have been sampled and are used as inputs to the classification model to distinguish between normal and abnormal patients. The study has employed an imbalanced number of ECG samples for training the various classification models. Few ML methods such as support vector machine (SVM), logistic regression (LR), and adaptive boosting (AdaBoost) which have been rarely used for HD detection have been selected. The performance of the developed model has been evaluated in terms of accuracy, F1-score, and area under curve (AUC) values using ECG signals of subjects given in publicly available (PTB-ECG, MIT-BIH) datasets. Ranking of the models has been assigned based on these performance metrics and it is found that the AdaBoost and LR classifiers stand in first and second positions. These two models have been ensembled based on the majority voting principle and the performance measure of this ensemble model has also been determined. It is, in general, observed that the proposed ensemble model demonstrates the best HD detection performance of 0.946, 0.949, and 0.951 for the PTB-ECG dataset and 0.921, 0.926, and 0.950 for the MIT-BIH dataset in terms of accuracy, F1-score, and AUC, respectively. The proposed methodology can also be employed for the classification of HD using ICG, MCG, and HS signals as inputs. Further, the proposed methodology can also be applied to the detection of other diseases.

KEYWORDS

ensemble model-based HD detection, classification of HD using imbalanced ECG records, SVM, AdaBoost, LR

Introduction

Cardiovascular disease (CVD) is a generalized term that includes diseases relating to the heart as well as blood vessels (Anooj, 2012). The various types of CVDs are: coronary artery disease (CAD), VHD, heart failure (HF), coronary heart disease (CHD), peripheral artery disease, and angina (Anooj, 2012; Dwivedi, 2018). These variants of CHDs are diagnosed either by clinical test data, ECG, HS, echocardiography, ICG, or MCG signal (Dwivedi, 2018; Kumar and Gandhi, 2018). It is observed that the ECG is mostly used by physicians for detecting the HD of a subject. The classification / detection of CVD mostly employs soft computing, evolutionary computing, ML as well as deep learning (DL)-based approaches. In this section, a detailed review of the existing literature on CVD detection is presented.

In Anooj (2012), the authors have developed a clinical decision support system for HD risk prediction from the clinical test data using the fuzzy logic technique. The experimental results using the University of California, Irvine (UCI) repository show that the proposed method outperforms neural network-based classifiers in terms of accuracy, sensitivity, and specificity. In an interesting article (Dwivedi, 2018), the author investigated HD prediction using different ML techniques. It is reported that the LR classifier provides highest accuracy, sensitivity, and specificity of 85, 89, and 81%, respectively. A non-invasive internet of things (IoT) platform-based HD detection scheme has been proposed in Kumar and Gandhi (2018) by employing clinical data. The proposed scheme involves a three-tier IoT architecture. The author has also made a receiver operating characteristic (ROC) analysis to find the significant clinical parameters responsible for detection. The random forest (RF), as well as hidden Markov model (HMM)-based HD classification models, have been suggested in Meng et al. (2019) by employing activity tracker data. It is found that the HMM model provides higher AUC of 0.79 compared to that (0.76) provided by the RF model. For the prediction of HD, a hybrid scheme (Mohan et al., 2019) using the linear model (LM) and RF has been developed under the IoT platform. It is shown that the proposed hybrid scheme yields an accuracy of 88.7%. For the diagnosis of CAD, the binary-real particle swarm optimization (PSO)-based hybrid scheme using two different feature selection methods has been employed in Zomorodi-moghadam et al. (2021). It is observed that the selected 11 features outperform the classification results compared to the 13 feature-based models. A novel approach to HD prediction has been reported (Magesh and Swarnalatha, 2021) by using the cluster-based decision tree (DT) and RF classifier from UCI repository data. The suggested approach provides higher classification accuracy of 89.3% compared to 76.70% accuracy yielded by the same classifier without cluster-based DT learning. In an interesting article (Li et al., 2020), five different ML-based HD identification models have been reported. The classifiers used in these methods are k-nearest neighbor (k-NN), DT, LR, and artificial neural network (ANN). The authors have introduced a fast conditional mutual information-based feature selection approach (FCMIM). In addition, other feature selection algorithms such as relief, least absolute shrinkage selection operator (LASSO), minimal redundancy maximal relevance (MRMR), and local learning-based methods have been employed for comparing the performance measures. It is reported that the proposed FCMIM-based support vector machine (SVM) classifier produces highest accuracy of classification. Using the clinical test data, a non-invasive CHD detection method is proposed in Wang J. et al. (2020a). This method employs base and meta-level stacking. It is reported that the suggested scheme provides specificity, sensitivity, and accuracy of 94.44, 95.84, and 95.43%, respectively. The ML techniques such as k-NN, NB, and binary logistics have been used to develop the individual as well as ensemble models using the principle of bagging, boosting, and stacking for the detection of CHD from clinical data (Shorewala, 2021). The boosted models provide highest AUC score of 0.73. But the stacked model is found to be the best with an accuracy of 75.1%. In another work (Oresko et al., 2010), the authors have proposed a real-time CVD detection method from the ECG sample. It can be implemented in a smartphone-based platform. A long short-term memory (LSTM) network has been trained (Ganguly et al., 2020) using ECG signal for the automatic classification of arrhythmia. It is shown that a bi-directional LSTM (b-LSTM) network outperforms another LSTM model. The CHD risk detection using ECG samples has been achieved under a mobile cloud computing environment (Venkatesan et al., 2018). The proposed method has employed wavelet transform (WT) for the detection of R-peaks. The adaptive neuro-fuzzy inference system (ANFIS) approach has been followed to develop as a classifier. A hybrid approach using WT and b-LSTM has been employed for the classification of ECG signal (Yildirim, 2018). It is shown that the proposed model provides a recognition performance of 99.39%. A CVD classifier employing ECG signal has been developed (Deng et al., 2018) following the dynamical neural learning mechanism. The effectiveness of the proposed scheme has been proved using (PTB-ECG datasets, (2004)). A modified RF along with an improved LM for detecting HD on the internet of medical things platform (IoMT) has been developed in Guo et al. (2020). The proposed scheme provides 96.6, 96.8, and 96.7% of accuracy, stability ratio, and F1-score, respectively. An automated convolutional neural network (CNN)-based heartbeat classifier has been developed (Wang H. et al., 2020c) using ECG records and its various performance measures have been evaluated. It is reported that the suggested model detects arrhythmia with an accuracy of 99.06%. In another article (Hussain et al., 2020), the authors have developed a model to detect HF. To achieve this, they have employed SVM, DT, k-NN, and ensemble classification models and multi-dimensional features. It is observed that the SVM classifier provides a sensitivity of 96%, specificity of 89%, and accuracy of 93.1%. The HD has been diagnosed using deep learning neural network (DLNN) and CNN-based models (Rath et al., 2021a). It is found that the accuracy of classification, sensitivity, and specificity varies between 89–99, 91–97.5, and 92.83–99.2%, respectively. Most of the ML and DL models provide satisfactory CVD detection from balanced ECG samples. However, in Rath et al. (2021c), the authors have suggested generative adversarial network (GAN) and LSTM models to detect CHD from two types of imbalanced datasets. It is shown that the GAN model outperforms all other models but the GAN-LSTM ensemble model provides the best CHD detection performance from the imbalanced datasets. In another interesting article (Sengur and Turkoglu, 2008), an artificial immune system-based fuzzy k-NN classifier has been suggested to detect heart valve disorders using Doppler HSs. It is reported that the proposed method yields 95.9 and 96% sensitivity and specificity rates, respectively. The incremental self-organizing map (ISOM) as well as Kohonen's SOM have been used as classifiers of HS (Dokur and Ölmez, 2008). The WT has been employed for segmentation as well as for the extraction of features. It is found that the ISOM model satisfactorily classifies the HS in the noisy environment. A radial wavelet neural network (RWNN) with an extended Kalman filter (EKF)-based training scheme has been used (Guillermo et al., 2015) as a classifier for detecting the heart murmur. The results of this model have been compared with an ANN model using Levenberg–Marquardt training. The authors in Liu et al. (2019) have developed an extreme learning machine (ELM) classifier for the identification of HF from the characteristics of HS. They have used 11 features extracted from the HS. The proposed method provides 96.32, 95.48, and 97.10% accuracy, sensitivity, and specificity, respectively. The SVM classifier has been used (Abduh et al., 2020) for classifying HS using mel-frequency spectral coefficients. It is shown that the proposed scheme offers a sensitivity of 0.8735 and specificity of 0.9666. The detection of HD from the HS signal of children has been obtained by employing an ANN classifier. The HS has been segmented using discrete wavelet transform (DWT) as well as the Hadamard product (Wang J. et al., 2020b). It is observed that the detection accuracy, specificity, and sensitivity of 93, 91.7, and 93.5%, respectively, have been achieved by the proposed model. Very few works have been carried out on HD detection employing MCG signal. In Tao et al. (2018), the authors have employed the SVM-extreme gradient boost (XGBoost) hybrid model providing the best performance metrics compared to other methods. Three different classifiers (DT, RF, and SVM) have been chosen to diagnose (Salah et al., 2020) the VHD from the ICG signal. The authors have extracted the statistical, morphological, and spectral features from the ICG samples. Subsequently, principal component analysis (PCA) has been used to reduce the number of features. It is observed that the combination of these three features-based RF classifiers provides highest accuracy of 96.34%. Many DL-based classifiers have been employed for the detection of CVD from mammograms (Wang J. et al., 2017). A 12-layer CNN has been trained to identify breast arterial calcification (BAC). It is observed that the proposed approach achieves a detection efficiency similar to human experts. A critical review article (Rath et al., 2021b) has been reported on the diagnosis of HD using various clinical data, ECG, and HS samples. It also presents various types of datasets, different feature extraction and reduction techniques, and various ML and DL classifiers for HD detection.

The analysis of the literature review reveals that many standard ML methods have already been used for CVD detection from ECG signal of subjects. However, it is observed that many ML methods such as AdaBoost (Wang J. et al., 2020a) and LR (Dwivedi, 2018) have been employed as a classifier in a few cases. Further, the validation task of the detection model has been carried out using only one source of standard ECG samples (Oresko et al., 2010; Ganguly et al., 2020). Third, in few articles, the ensemble model has been suggested (Hussain et al., 2020; Shorewala, 2021) using the ML models for achieving enhanced detection performance. In most of the articles, the training and validation operations of the ML and DL models have been carried out using a balanced number of ECG signals of subjects. These observations have encouraged developing of ML-based detection models using LR and AdaBoost classifiers. Further, to assess the consistent performance of the proposed models, the standard MIT-BIH and PTB-ECG-based ECG datasets have been chosen both during the training and validation phases.

The imbalanced data mean the number of normal and abnormal patients is not equal. When the number of normal and abnormal cases is not equal, the model is trained with a bias toward higher number of the two classes. The model which is developed under such conditions provides a lower accuracy of detection. So, the challenge is to achieve improved training and testing results under the imbalanced condition of the input data.

Many ML methods exhibit poor detection performance when the training and testing datasets are imbalanced. Therefore, in this article, imbalanced ECG samples have been employed to examine the performance potentiality of the classifier. With an objective to further improve the detection accuracy, an ensemble model has been developed by choosing the best of the three ML classifiers.

Based on the motivation and objectives of the proposed work, the article has been organized in the following way. Section "Materials and mehods" deals with the materials and methodology required for CHD detection from imbalanced ECG samples. It provides the details of the standard data sources used as well as the training and testing schemes of SVM, LR and AdaBoost and ensemble version of LR and AdaBoost classifiers. Section "Simulation based experiments" outlines the simulation-based experiments obtained using the trained models of Section "Materials and methods." The analysis and discussions on various results have also been made in Section "Discussions." It also presents the contribution of the article. Finally, Section "Conclusion" provides the concluding remarks of the investigation and suggests the scopes of future research work.

Materials and methods

This section presents the details of materials in terms of ECG recordings of normal and abnormal subjects available from two standard ECG datasets. The block diagram/flowchart of three classifiers and the corresponding training and testing steps are provided in this section.

Materials

The two datasets which are used for obtaining ECG samples are MIT-BIH and PTB-ECG (Bousseljot et al., 1995; Goldberger et al., 2000; George Moody and Mark Roger, 2001). The details of these two datasets are available in MIT-BIH ECG datasets, (2005) and PTB-ECG datasets, (2004). The details of the numbers of normal and abnormal cases and numbers of training (70%) and testing sets (30%) are shown in Table 1. As evident from this table, the number of normal and abnormal cases chosen is imbalanced.

Methodology

From each subject, 12 ECG recordings have been taken and averaged to achieve a smooth ECG waveform. The average ECG waveform of each subject has been sampled to produce 1,024 discrete samples. At a time, all the samples of a subject have been fed to each classification model for training and validation purposes. In this study, the 1,024 samples of each ECG signal are considered. The 1,024-dimensional sample vectors for the ECG signals are used for the training and testing of the classifiers for HD detection. Four classification models such as logistic regression (LR), SVM, AdaBoost, and LR-AdaBoost are used in this work.

Logistic regression

It is a predictive classification algorithm that assigns a class to the set of measurements or observations (Scott, 2002). It employs a sigmoid function to limit the output between 0 and 1. The output of the LR equation is computed as

$$z = \alpha_0 + \alpha_1 (x) \tag{1}$$

$$sig(z) = \frac{1}{1 + e^{-z}} \tag{2}$$

$$f_\theta (x) = sig(z) \tag{3}$$

The cost function is given by (Scott, 2002).

$$J(\theta) = \frac{1}{2} \sum_{k=1}^{K} \left(f_\theta \left(x^{(k)} \right) - y^{(k)} \right)^2 \tag{4}$$

For a two-class problem, *y is equal to either* 1 *or* 0. The cost function is minimized with respect to θ, to obtain the update equation (Scott, 2002).

$$\theta_j = \theta_j - \alpha \left(f_\theta (x) - y \right) x \tag{5}$$

The symbol α denotes the learning rate which lies between 0 and 1 and needs to be suitably adjusted during the training phase. After the completion of the training, the performance metrics of the model are evaluated.

Support vector machine

The principle of the SVM classifier is explained in steps. Let X represents the samples of ECG recording of subjects and Y represents the corresponding class vector (Cortes and Vladimir, 1995). The key steps of SVM classifier during the training phase are:

Step 1: Compute $Y^T Y$ and XX^T
Step 2: Compute the matrix $H = Y^T Y.X^T X$
Step 3: Compute the Lagrangian Multipliers, α.
Step 4: Compute decision hyperplane normal vector, $W = (\alpha.Y)^T.X$
Step 5: Compute bias, $b = 1 - w^T x_1$

During the testing phase, the class of unknown ECG samples, z is evaluated by computing $sign(w^T z + b)$. If it is positive, then the test dataset belongs to class 1 (Cortes and Vladimir, 1995).

AdaBoost

The AdaBoost algorithm is an ensemble method of ML (Schapire, 2013). In this case, higher weights are assigned to wrongly classify instances. The boosting is used to minimize the bias and the variance for supervised learning. Excluding the first one, each subsequent learner is developed from the previous ones. The AdaBoost is based on the principle that weak learners are transformed into strong ones. The block diagram of the AdaBoost-based classifier is shown in Figure 3. The ECG samples of subjects are fed to the first model (in this case DT). In this case, the first model is built and the errors from this model are noted. The ECG record which is incorrectly classified is fed to the next model (Schapire, 2013). This process is continued until a pre-specified condition is made. In this case, the algorithm only

TABLE 1 Materials used for training and testing the ML classification models.

Datasets	No. of ECG datasets	No. of abnormal cases	No. of normal cases	Training sets (70%)	Testing sets (30%)
MIT-BIH	268	104	164	188 normal-115 abnormal-73	80 normal-49 abnormal-31
PTB-ECG	200	54	146	140 normal-104 abnormal-36	60 normal-42 abnormal-18

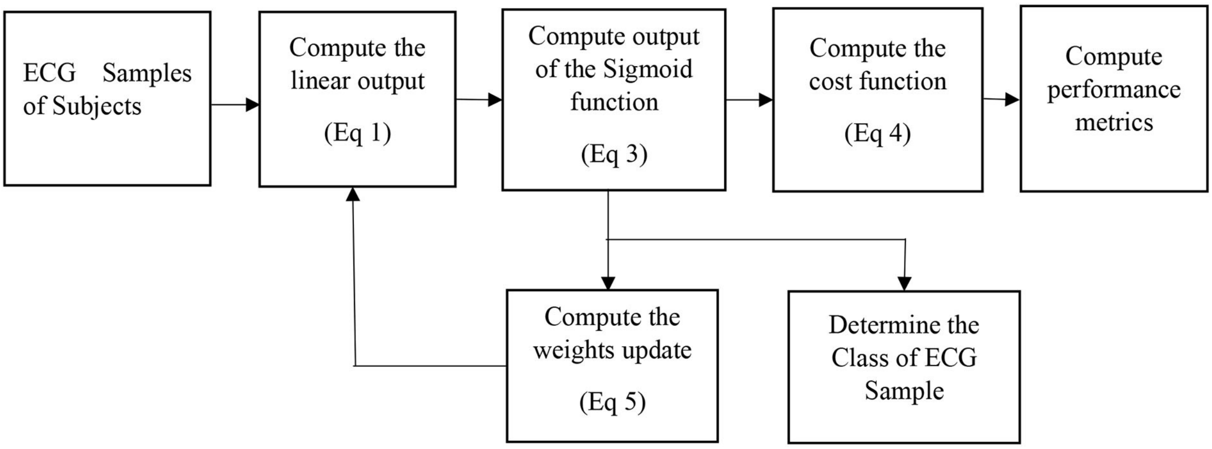

ECG Samples of Subjects
Compute the linear output (Eq 1)
Compute output of the Sigmoid function (Eq 3)
Compute the cost function (Eq 4)
Compute performance metrics
Compute the weights update (Eq 5)
Determine the Class of ECG Sample

FIGURE 1
Block diagram of the logistic regression-based classification model.

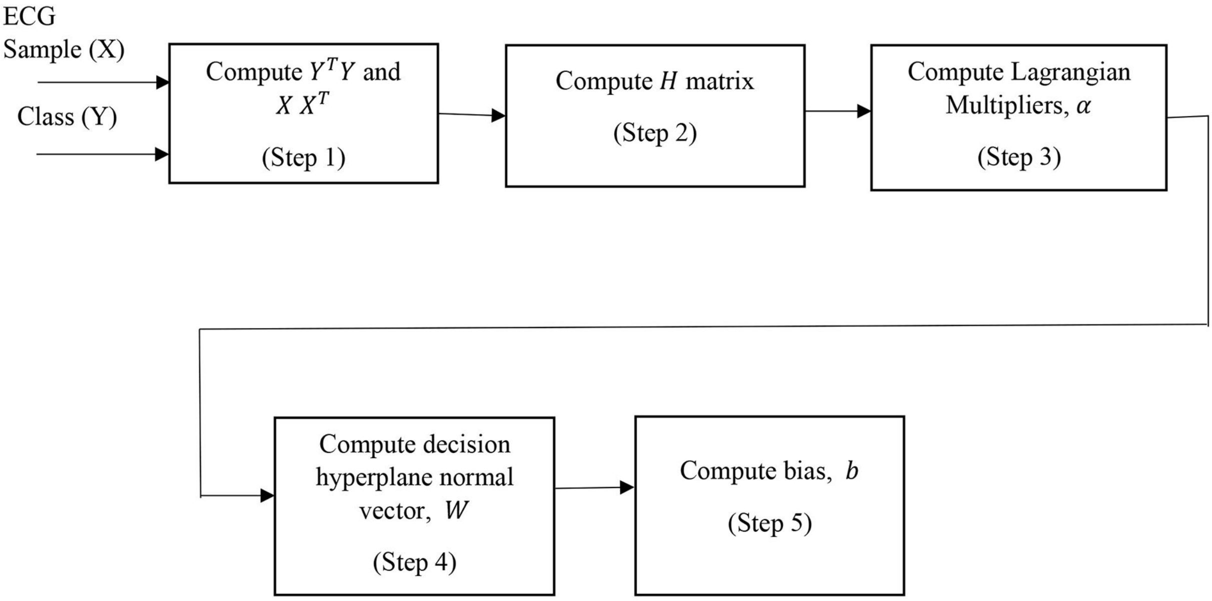

ECG Sample (X)
Class (Y)
Compute Y^TY and $X X^T$ (Step 1)
Compute H matrix (Step 2)
Compute Lagrangian Multipliers, α (Step 3)
Compute decision hyperplane normal vector, W (Step 4)
Compute bias, b (Step 5)

FIGURE 2
Block diagram of SVM classification model.

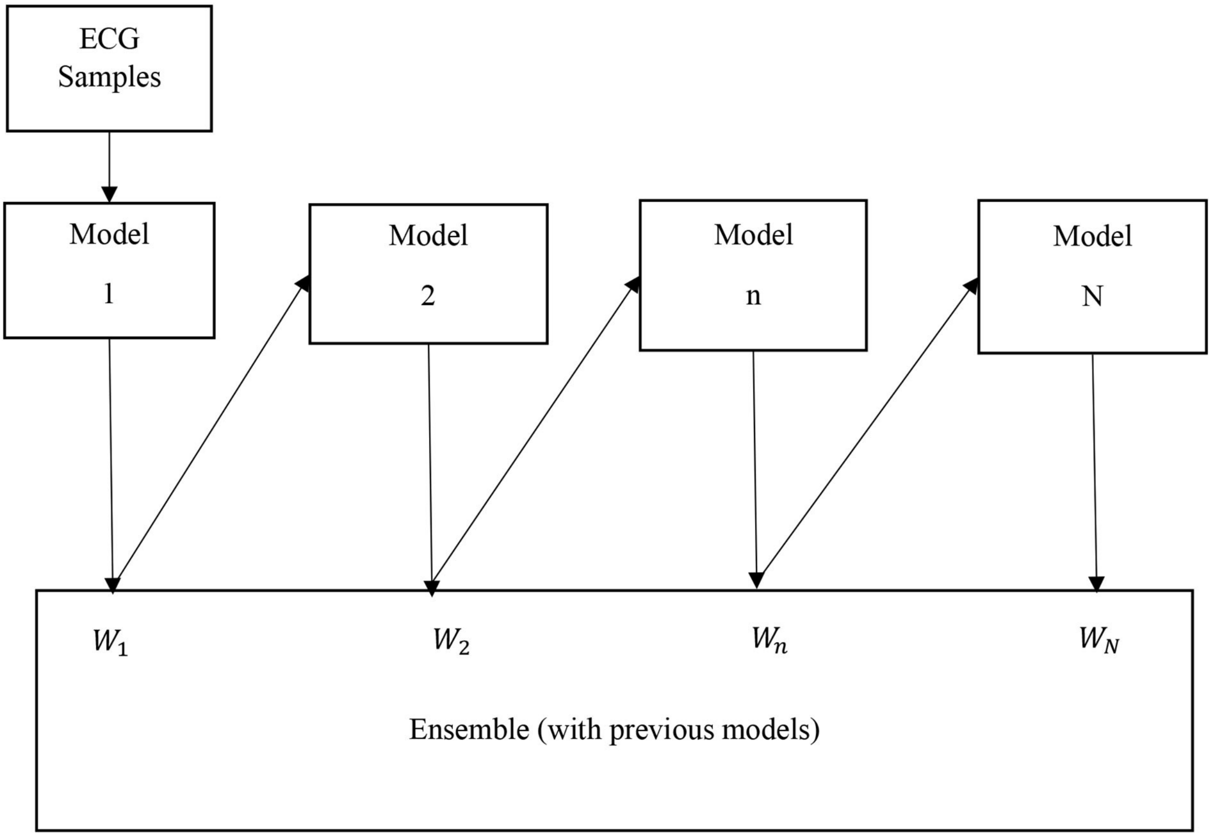


FIGURE 3
Block diagram of AdaBoost classification model.

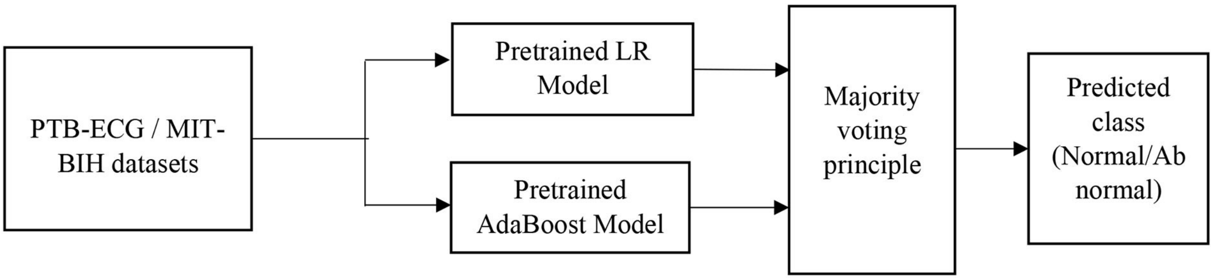

PTB-ECG / MIT-BIH datasets
Pretrained LR Model
Pretrained AdaBoost Model
Majority voting principle
Predicted class (Normal/Ab normal)

FIGURE 4
Block diagram of LR-AdaBoost ensemble model.

makes a node with two leaves which is called a stump. The major steps of building the classifier are

Step 1: To create the first base learner by taking the first feature and the process is continued for all features. So, the number of base learners or stumps is equal to the number of features.

Step 2: To calculate the total error is $E = 1/N$, where N is equal to the number of records.

Step 3: To compute the performance (P) of the stump according to

$$P = \frac{1}{2} \ln \frac{(1 - E)}{E}$$

where, ln denotes the natural log.

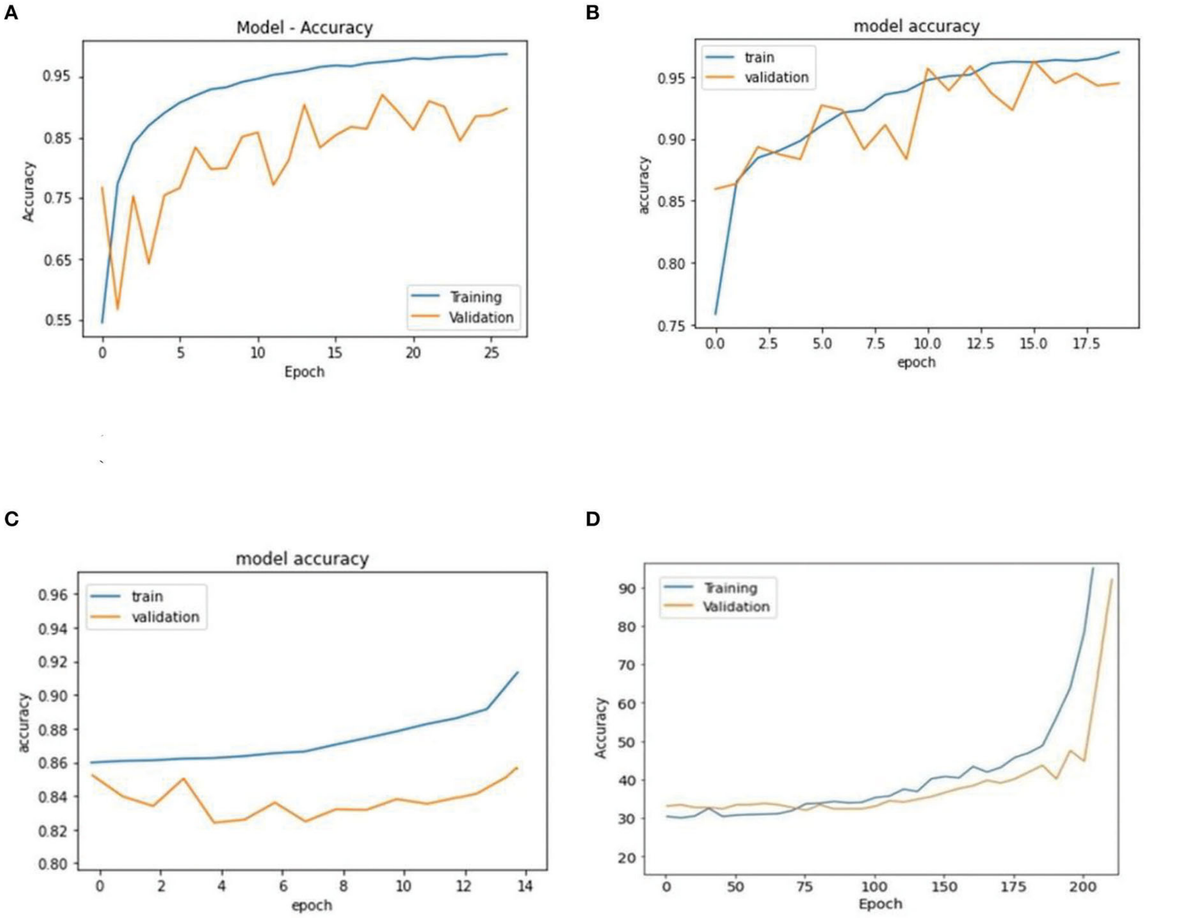


FIGURE 5
Comparison of accuracy achieved during training and validation phases of LR, AdaBoost, SVM, and ensemble of AdaBoost and LR models (PTB-ECG dataset). **(A)** Logistic regression. **(B)** AdaBoost. **(C)** Support vector machine (SVM). **(D)** AdaBoost – logistic regression based ensemble model.

Step 4: To update the sample weights according to

$$New\ Weight = Old\ weight\ \times e^{-(Performance)}$$

where, the initial weight $= \frac{1}{N}$

Step 5: To create a new dataset by choosing incorrectly classified records as well as a few correct ones.

Step 6: To create a set of new DTs (stump) and continue the process until the last error is produced.

LR-AdaBoost ensemble model

So, in the present case, the ML techniques are primarily used to develop prediction or classification tasks. Each of the developed model provides the accuracy of classification based on their potentiality. To further improve the accuracy of performance, the ensemble model is developed using each of the basic model. In this process, the outcome of the overall ensemble model becomes better than the individual model which is part of the combination model. The challenging case of the ensemble model is to determine the connecting weights of each individual model. Mostly this is achieved by majority voting or bio-inspired-based optimization techniques.

To improve the classification performance, the ensemble model is developed by choosing the two best models (Polikar, 2006). In the present case, the LR and AdaBoost models are first trained and these pre-trained models are connected in parallel as shown in Figure 4. The input to the ensemble scheme is the samples of each record of the standard ECG dataset. The output of each of these models is fed to the majority voting scheme. The final predicted class (normal/abnormal) refers to the output of the majority voting scheme (Polikar, 2006). This principle

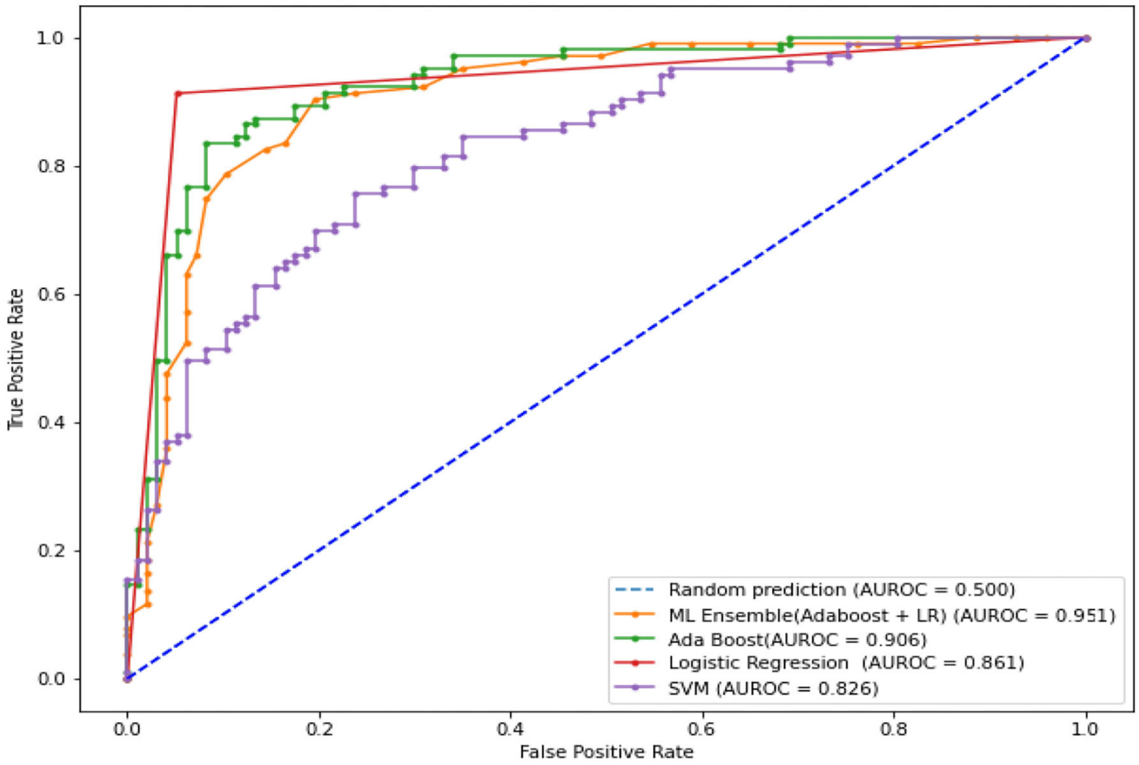


FIGURE 6
Comparison of ROC plots and AUC values of LR, AdaBoost, SVM, and LR-AdaBoost ensemble models (PTB-ECG dataset).

classifies the input records in a superior way compared to each individual model.

The various results obtained from the simulation study of the three ML and one ensemble models have been obtained and have been tabulated and plotted in the next section.

Simulation-based experiments

The LR, SVM, and AdaBoost classification models shown in Figures 1, 2, 3 have been simulated following the training principle of each of the model. Separate models have been simulated for each imbalanced PTB-ECG and MIT-BIH datasets as inputs. Similarly, the ensemble model shown in Figure 4 has been simulated using the same inputs. Each ECG record provides 1,024 samples which are simultaneously fed to the model both during the training and testing phases. In case of LR, the sigmoid function is used to keep the output between 0 and 1. In the simulation study, the learning rate alpha has been chosen to be 0.1. In case of AdaBoost, the decision tree has been used as the base estimator. In the present case, 30 decision trees have been used in the simulation study. In this case, the learning rate has been taken as 0.05. In case of the SVM classifier, the linear kernel has been used. For the PTB-ECG dataset, the plots of variation in accuracy with change in epochs during training and validation phases for LR, AdaBoost, SVM, and ensemble model are shown in Figures 5A–D, respectively. Further, the comparison of ROC plots obtained by LR, AdaBoost, SVM, and LR-AdaBoost ensemble model for the PTB-ECG dataset is shown in Figure 6. The same figure also provides the AUC values of these models. During the validation phase, the accuracy, F1-score, and AUC values of LR, SVM, AdaBoost, and ensemble (LR-AdaBoost) model have been determined and listed in Tables 2, 3 for PTB-ECG, and MIT-BIH datasets, respectively. The various results shown in the Tables and plotted in the graphs have been analyzed in the next section.

Discussions

This section presents the interpretation of the various results presented in the previous sections. It is observed from the plots of Figure 5 that as the number of epochs increases, the accuracy value also increases and remains constant at the end of the training phase. Further, it is found that for any given epoch the training accuracy is higher than the corresponding validation accuracy. The observation of ROC plots of Figure 6 (PTB-ECG) dataset reveals that the ensemble model provides the highest AUC of 0.951. It is then followed by AdaBoost, LR, and SVM classification models. It is interesting to observe

TABLE 2 Comparison of three performance metrics of different models using the PTB-ECG dataset.

Performance measures	Logistic regression	SVM	AdaBoost	Ensemble model LR - AdaBoost
Accuracy	0.898	0.864	0.927	0.946
	(III)	(IV)	(II)	(I)
F1 Score	0.902	0.852	0.936	0.949
	(III)	(IV)	(II)	(I)
AUC	0.861	0.826	0.906	0.951
	(III)	(IV)	(II)	(I)

TABLE 3 Comparison of three performance metrics of different models using the MIT-BIH dataset.

Performance measures	Logistic regression	SVM	AdaBoost	Ensemble model LR-AdaBoost
Accuracy	0.869	0.821	0.894	0.921
	(III)	(IV)	(II)	(I)
F1 Score	0.884	0.782	0.845	0.926
	(III)	(IV)	(II)	(I)
AUC	0.858	0.819	0.910	0.950
	(III)	(IV)	(II)	(I)

that the order in terms of magnitude of AUC values of different models is the same for both datasets as evident from Tables 2, 3. The observation of three important performance metrics (Accuracy, F1-score, AUC values) obtained from the simulation study of LR, SVM, AdaBoost, and ensemble model is shown in Table 2 (PTB-ECG dataset) and in Table 3 (MIT-BIH dataset). The observation shows that the ensemble model outperforms the individual three ML models. The bracketed terms such as (I), (II), etc. in Tables 2, 3 indicate the rank of respective classification models which are assigned based on the performance metrics. This is also evidenced by the individual and overall ranking assigned to these models in Tables 2, 3. In general, it is found that based on the three-performance metrics of all the four classification models and by employing imbalanced ECG data samples from two standard datasets, the rankings assigned are I, II, III, and IV for ensemble, AdaBoost, LR, and SVM models, respectively.

Based on the above analysis, the major contributions of investigation on HD detection are the following:

i. All the proposed classification models for HD detection using two imbalanced ECG recordings as subjects exhibit consistent performance following the imbalanced number of inputs both during the training and testing phases.
ii. As expected, the ensemble model developed using LR-AdaBoost has demonstrated the best performance among all the four models yielding accuracy, F1-score, and AUC values of 0.946, 0.949, and 0951 for the PTB-ECG dataset and 0.921, 0.926, and 0.950 for MIT-BIH dataset.
iii. These four models show similar performance for both the datasets as well as the following imbalanced number of ECG records as inputs to training and validation phases.

Conclusion

This article has investigated the classification potentiality of HD using three ML algorithms and one ensemble model. The development of these models is based on imbalanced training ECG records. The accuracy plots and three performance measures reveal that the AdaBoost performs better than the SVM, LR-based classification models. This observation is true for both datasets. The LR-AdaBoost ensemble model based on majority voting principle demonstrates the best performance in terms of accuracy, F1-score, and AUC values compared to individual models. The numerical performance results also show that the order of the performance is consistent for both datasets. The present methodology can also be applied for HD detection using ICG, MCG, and HS signals. The HD detection results obtained from these three types of signals as inputs can be analyzed and compared with the results obtained from the present study. There are different kinds of ensemble techniques that can be employed for developing the ensemble model. The results of these ensemble models can be compared based on the performance and the best model can be chosen. The proposed approaches can also be applied to the detection of other diseases.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: MIT-BIH ECG datasets - https://physionet.org/content/mitdb/1.0.0/ and PTB-ECG datasets - https://www.physionet.org/content/ptbdb/1.0.0/.

Author contributions

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

Conflict of interest

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

Publisher's note

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

Abduh, Z., Nehary, E. A., Wahed, M. A., and Kadah, Y. M. (2020). Classification of heart sounds using fractional fourier transform based mel-frequency spectral coefficients and traditional classifiers. *Biomed. Signal Process. Control.* 57, 101788. doi: 10.1016/j.bspc.2019.101788

Anooj, P. K. (2012). Clinical decision support system: risk level prediction of heart disease using weighted fuzzy rules. *J. King Saud Univ. - Comput. Inf. Sci.*. 24, 27–40. doi: 10.1016/j.jksuci.2011.09.002

Bousseljot, R., Kreiseler, D., and Schnabel, A. (1995). Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das *Internet*. 317–318. doi: 10.1515/bmte.1995.40.s1.317

Cortes, C., and Vladimir, V. (1995). Support-vector networks. *Mach. Learn.* 20, 273–297. doi: 10.1007/BF00994018

Deng, M., Wang, C., Tang, M., and Zheng, T. (2018). Extracting cardiac dynamics within ECG signal for human identification and cardiovascular diseases classification. *Neural Netw.* 100, 70–83. doi: 10.1016/j.neunet.2018.01.009

Dokur, Z., and Ölmez, T. (2008). Heart sound classification using wavelet transform and incremental self-organizing map. *Digit. Signal Process.* 18, 951–959. doi: 10.1016/j.dsp.2008.06.001

Dwivedi, A. K. (2018). Performance evaluation of different machine learning techniques for prediction of heart disease. *Neural. Comput. Appl.* 29, 685–693. doi: 10.1007/s00521-016-2604-1

Ganguly, B., Ghosal, A., Das, A., Das, D., Chatterjee, D., Rakshit, D., et al. (2020). Automated detection and classification of arrhythmia from ECG signals using feature-induced long short-term memory network. *IEEE Sens. Lett.* 4, 1–4. doi: 10.1109/LSENS.2020.3006756

George Moody, B., and Mark Roger, G. (2001). The impact of the MIT-BIH arrhythmia database. *IEEE Eng. Med. Biol. Mag.* 20, 45–50. doi: 10.1109/51.932724

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., et al. (2000). PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* 101, 215–220. doi: 10.1161/01.CIR.101.23.e215

Guillermo, J. E., Castellanos, L. J. R., Sanchez, E. N., and Alanis, A. Y. (2015). Detection of heart murmurs based on radial wavelet neural network with Kalman learning. *Neurocomputing*. 164, 307–317. doi: 10.1016/j.neucom.2014.12.059

Guo, C., Zhang, J., Liu, Y., Xie, Y., Han, Z., Yu, J., et al. (2020). Recursion enhanced random forest with an improved linear model (RERF-ILM) for heart disease detection on the internet of medical things platform. *IEEE Access* 8, 59247–59256. doi: 10.1109/ACCESS.2020.2981159

Hussain, L., Awan, I. A., Aziz, W., Saeed, S., Ali, A., Zeeshan, F., et al. (2020). Detecting congestive heart failure by extracting multimodal features and employing machine learning techniques. *BioMed Res. Int.* 2020, 1–9. doi: 10.1155/2020/4281243

Kumar, P. M., and Gandhi, U. D. (2018). A novel three-tier Internet of Things architecture with machine learning algorithm for early detection of heart diseases. *Comput. Electr. Eng.* 65, 222–235. doi: 10.1016/j.compeleceng.2017.09.001

Li, J. P., Haq, A. U., Din, S. U., Khan, J., Khan, A., Saboor, A., et al. (2020). Heart disease identification method using machine learning classification in e-healthcare. *IEEE Access* 8, 107562–107582. doi: 10.1109/ACCESS.2020.3001149

Liu, Y., Guo, X., and Zheng, Y. (2019). An automatic approach using ELM classifier for HFpEF identification based on heart sound characteristics. *J. Med. Syst.* 43, 1–8. doi: 10.1007/s10916-019-1415-1

Magesh, G., and Swarnalatha, P. (2021). Optimal feature selection through a cluster-based DT learning (CDTL) in heart disease prediction. *Evol. Intell.* 14, 583–593. doi: 10.1007/s12065-019-00336-0

Meng, Y., Speier, W., Shufelt, C., Joung, S., Van Eyk, J. E., et al. (2019). A machine learning approach to classifying self-reported health status in a cohort of patients with heart disease using activity tracker data. *IEEE J. Biomed. Health Inform.* 24, 3, 878–884. doi: 10.1109/JBHI.2019.2922178

MIT-BIH ECG datasets. (2005). Available online at: https://physionet.org/content/mitdb/1.0.0/ (accessed September 10, 2020).

Mohan, S., Thirumalai, C., and Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access* 7, 81542–81554. doi: 10.1109/ACCESS.2019.2923707

Oresko, J. J., Jin, Z., Cheng, J., Huang, S., Sun, Y., Duschl, H., et al. (2010). A wearable smartphone-based platform for real-time cardiovascular disease detection via electrocardiogram processing. *IEEE Trans. Inf.Technol. Biomed.* 14, 734–740. doi: 10.1109/TITB.2010.2047865

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits Syst. Mag.* 6, 21–45. doi: 10.1109/MCAS.2006.1688199

PTB-ECG datasets. (2004). Available online at: https://www.physionet.org/content/ptbdb/1.0.0/ (accessed September 10, 2020).

Rath, A., Mishra, D., and Panda, G. (2021a). Deep learning neural network and CNN-based diagnosis of heart diseases. *Tech. Adv. Mach. Learn. Healthc.* 936, 169. doi: 10.1007/978-981-33-4698-7_9

Rath, A., Mishra, D., Panda, G., and Satapathy, S. C. (2021b). An exhaustive review of machine and deep learning-based diagnosis of heart diseases. *Multimed. Tools Appl.* 1, 1–59. doi: 10.1007/s11042-021-11259-3

Rath, A., Mishra, D., Panda, G., and Satapathy, S. C. (2021c). Heart disease detection using deep learning methods from imbalanced ECG samples. *Biomed. Signal Process. Control* 68, 102820. doi: 10.1016/j.bspc.2021.102820

Salah, I. B., De la Rosa, R., Ouni, K., and Salah, R. B. (2020). Automatic diagnosis of valvular heart diseases by impedance cardiography signal processing. *Biomed. Signal Process. Control* 57, 101758. doi: 10.1016/j.bspc.2019.101758

Schapire, E. R. (2013). "Explaining adaboost," in *Empirical Inference*, ed V. N. Vapnik (Berlin; Heidelberg: Springer), 37–52. doi: 10.1007/978-3-642-41136-6_5

Scott, M. (2002). *Applied Logistic Regression Analysis*, Vol. 106. London: Sage.

Sengur, A., and Turkoglu, I. (2008). A hybrid method based on artificial immune system and fuzzy k-NN algorithm for diagnosis of heart valve diseases. *Expert Syst. Appl*. 35, 1011–1020. doi: 10.1016/j.eswa.2007.08.003

Shorewala, V. (2021). Early detection of coronary heart disease using ensemble techniques. *Inform. Med. Unlocked*. 26, 100655. doi: 10.1016/j.imu.2021.100655

Tao, R., Zhang, S., Huang, X., Tao, M., Ma, J., et al. (2018). Magnetocardiography-based ischemic heart disease detection and localization using machine learning methods. *IEEE Trans. Biomed. Eng*. 66, 1658–1667. doi: 10.1109/TBME.2018.2877649

Venkatesan, C., Karthigaikumar, P., and Satheeskumaran, S. (2018). Mobile cloud computing for ECG telemonitoring and real-time coronary heart disease risk detection. *Biomed. Signal Process. Control* 44, 138–145. doi: 10.1016/j.bspc.2018.04.013

Wang, H., Shi, H., Chen, X., Zhao, L., Huang, Y., Liu, C., et al. (2020c). An improved convolutional neural network-based approach for automated heartbeat classification. *J. Med. Syst*. 44, 1–9. doi: 10.1007/s10916-019-1511-2

Wang, J., Ding, H., Bidgoli, F. A., Zhou, B., Iribarren, C., Molloi, S., et al. (2017). Detecting cardiovascular disease from mammograms with deep learning. *IEEE Trans. Med. Imaging*. 36, 1172–1181. doi: 10.1109/TMI.2017.2655486

Wang, J., Liu, C., Li, L., Li, W., Yao, L., Li, H., et al. (2020a). A stacking-based model for non-invasive detection of coronary heart disease. *IEEE Access* 8, 37124–37133. doi: 10.1109/ACCESS.2020.2975377

Wang, J., You, T., Yi, K., Gong, Y., Xie, Q., et al. (2020b). Intelligent diagnosis of heart murmurs in children with congenital heart disease. *J. Healthc. Eng*. 2020, 1–9. doi: 10.1155/2020/9640821

Yildirim, Ö. (2018). A novel wavelet sequence based on deep bidirectional LSTM network model for ECG signal classification. *Comput. Biol. Med*. 96, 189–202. doi: 10.1016/j.compbiomed.2018.03.016

Zomorodi-moghadam, M., Abdar, M., Davarzani, Z., Zhou, X., Pławiak, P., and Acharya, U. R. (2021). Hybrid particle swarm optimization for rule discovery in the diagnosis of coronary artery disease. *Expert Systems*. 38, 12485. doi: 10.1111/exsy.12485

frontiers | Frontiers in Big Data

TYPE Original Research
PUBLISHED 31 October 2022
DOI 10.3389/fdata.2022.603429

Check for updates

OPEN ACCESS

EDITED BY
Issam El Naqa,
University of Michigan, United States

REVIEWED BY
Gregory R. Hart,
Yale University, United States
Sunyoung Jang,
Princeton Radiation Oncology Center,
United States

*CORRESPONDENCE
Konstantin Sharafutdinov
ksharafutdin@ukaachen.de

†These authors share first authorship

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

RECEIVED 06 September 2020
ACCEPTED 03 October 2022
PUBLISHED 31 October 2022

CITATION
Sharafutdinov K, Bhat JS, Fritsch SJ, Nikulina K, E. Samadi M, Polzin R, Mayer H, Marx G, Bickenbach J and Schuppert A (2022) Application of convex hull analysis for the evaluation of data heterogeneity between patient populations of different origin and implications of hospital bias in downstream machine-learning-based data processing: A comparison of 4 critical-care patient datasets. *Front. Big Data* 5:603429.
doi: 10.3389/fdata.2022.603429

Application of convex hull analysis for the evaluation of data heterogeneity between patient populations of different origin and implications of hospital bias in downstream machine-learning-based data processing: A comparison of 4 critical-care patient datasets

Konstantin Sharafutdinov[1,2,3]*†, Jayesh S. Bhat[1,2†], Sebastian Johannes Fritsch[3,4,5], Kateryna Nikulina[1,2,3], Moein E. Samadi[1,2], Richard Polzin[1,2,3], Hannah Mayer[3,6], Gernot Marx[3,4], Johannes Bickenbach[3,4] and Andreas Schuppert[1,2,3]

[1]Institute for Computational Biomedicine, RWTH Aachen University, Aachen, Germany, [2]Joint Research Center for Computational Biomedicine, RWTH Aachen University, Aachen, Germany, [3]SMITH Consortium of the German Medical Informatics Initiative, Leipzig, Germany, [4]Department of Intensive Care Medicine, University Hospital RWTH Aachen, Aachen, Germany, [5]Juelich Supercomputing Centre, Forschungszentrum Juelich, Juelich, Germany, [6]Systems Pharmacology and Medicine, Bayer AG, Leverkusen, Germany

Machine learning (ML) models are developed on a learning dataset covering only a small part of the data of interest. If model predictions are accurate for the learning dataset but fail for unseen data then generalization error is considered high. This problem manifests itself within all major sub-fields of ML but is especially relevant in medical applications. Clinical data structures, patient cohorts, and clinical protocols may be highly biased among hospitals such that sampling of representative learning datasets to learn ML models remains a challenge. As ML models exhibit poor predictive performance over data ranges sparsely or not covered by the learning dataset, in this study, we propose a novel method to assess their generalization capability among different hospitals based on the convex hull (CH) overlap between multivariate datasets. To reduce dimensionality effects, we used a two-step approach. First, CH analysis was applied to find mean CH coverage between each of the two datasets, resulting in an upper bound of the prediction range. Second, 4 types of ML models were trained to classify the origin of a dataset (i.e., from which hospital) and to estimate differences in datasets with respect to underlying distributions. To demonstrate the applicability of our method, we used 4 critical-care patient datasets from different hospitals in Germany and USA. We estimated the similarity of these populations and investigated whether ML models developed on one dataset can be reliably applied to another one. We show that the strongest drop in performance was associated with the poor intersection of convex hulls in the corresponding hospitals' datasets and with a high performance of ML methods for dataset discrimination. Hence, we suggest the application of our pipeline as a first tool to assess the transferability of trained models. We emphasize that datasets from different hospitals represent heterogeneous data sources, and the transfer from one database to another should be performed with utmost care to avoid implications during real-world applications of the developed models. Further research is needed to develop methods for the adaptation of ML models to new hospitals. In addition, more work should be aimed at the creation of gold-standard datasets that are large and diverse with data from varied application sites.

KEYWORDS

dataset-bias, data pooling, ARDS, convex hull (CH), generalization error

Introduction

Driven by giant leaps in compute performance, the availability of huge datasets, and new algorithms for the training of deep neural networks (DNN), Machine Learning (ML) has seen a renaissance during the last 10 years. Today, ML approaches help us discover patterns in large swaths of data, predominantly on an automated or semi-automated basis. They have revolutionized how we process images, video, and text. The primary advantage of ML when compared to traditional modeling approaches for the input-output behavior of complex systems is the unbiased learning from data without a priori knowledge about the system to be learned (black-box modeling approach). Mathematically, ML algorithms are designed as universal machines mapping a high dimensional input space onto a low dimensional output space up to an order of error without restrictions. The algorithms enable unrestricted learning by a modeling strategy with a priori unrestricted complexity of the model, e.g., expressed by the unrestricted number of parameters to be adapted to the data. For large classes of functions, ML algorithms, e.g., neural networks, provide superior approximation performance compared to all linear series expansions (Barron and Klusowski, 2018). Recently, the equivalence of DNN learning with wavelet-based approximations indicated the superior performance of DNN for applications with close association with image recognition and time-series analysis (Mallat, 2016).

Data-driven models, such as ML methods, aim to represent systems solely from available measurement data. Hence, a critical conceptual issue of such models is their limited performance in the case of extrapolation into data regions sparsely covered by the data samples used for learning the model. These models handle test data better if they come from the same dataset used for training and generalize worse on the data obtained from other sources (Torralba and Efros, 2011; AlBadawy et al., 2018; Pooch et al., 2019). Model performance drops if data used to train and test a model come from different distributions. This difference is referred to as a domain shift (Pooch et al., 2019). Unless strong assumptions are posed on the learned function, data-driven models, not depending on the output to be predicted, can only be valid in regions where they have sufficiently dense coverage of training data points, which is referred to as the validity domain (Courrieu, 1994). This can be approximated by the convex hull spanned by the data, which represents an upper bound of the validity domain for any ML application. The convex hull (CH) of a set of data points is defined as the smallest polytope with dimensionality equal to the number of attributes containing the points in such a way that every straight line connecting a pair of points lies inside the polytope (Graham, 1972; Shesu et al., 2021). One approach to estimate the ability of a model to generalize is to consider the CH of the points used in a training set. Generalization tends to fail with the increase in the distance of a new point to the CH of the training set (Zhou and Shi, 2009). Therefore, the coverage of the CH of a test set by the CH of a training set represents an upper bound for the generalization ability of

Abbreviations: ARDS, Acute respiratory distress syndrome; CH, Convex hull; ICU, Intensive care unit; FiO_2, Fraction of inspired oxygen; MV, Mechanical ventilation; PaO_2, Arterial partial pressure of oxygen; PEEP, Positive end-expiratory pressure; ROC AUC, Area under receiver operating characteristic curve.

any ML-based model. In the case of learning from different populations, the mutual coverage of the convex hulls can serve as a measure for the sufficient similarity of heterogeneous populations enabling the first estimate for the reliability of the generalization of ML models. Hence, one possible approach to examine different populations for homogeneity concerning the predictive performance of ML models is to perform a convex hull analysis of the available data to be used for training and prediction, respectively (Ostrouchov and Samatova, 2005; Zhou and Shi, 2009).

However, even if the convex hulls of training and test sets intersect to a large extent, there might be differences in the underlying distributions of some parameters. For instance, when data of one dataset lay in a region which shows a low density of samples in the other dataset. An extreme example is a dataset consisting of two clusters of data apart from each other; the convex hull envelopes all dataset values, including the space between them. If the majority of samples of the second dataset fall inside the gap area between the two clusters, the generalization capacity of a model will be impaired, as there is not enough training data in that region. Although the intersection values are high, in this case, it does not allow us to judge the generalization ability of the trained model. Therefore, the CH analysis provides necessary, but not sufficient conditions for a proper generalization of ML models.

Consequently, a second step in the analysis is needed to investigate datasets for diverging underlying distributions. If there are no such differences, two datasets form a homogeneous population and are indistinguishable, otherwise, it would be possible to differentiate the datasets. Therefore, if ML classifiers can identify the origin of a drawn sample with high accuracy, we postulate that there are diverging underlying distributions of parameters forming different areas with a high density of samples in two datasets. Thus, training an ML model in one dataset and applying it in the other one would mean interpolation into areas sparsely covered by training data and could impair the generalization of respective models. However, ML methods do not provide the direction of impaired generalization (i.e., model trained on one dataset and applied in the other one and vice versa).

In contrast, CH analysis provides a model-agnostic a priori data assessment and more importantly direction of impaired generalization. The CH of one dataset may completely cover the CH of the other dataset, meaning no restrictions for generalization from the CH perspective. However, in the opposite case (the second dataset covering the first one) the CH coverage may be modest suggesting generalization issues once models developed in the second dataset will be applied to the first dataset. Furthermore, the CH analysis proposed in this paper is computationally inexpensive and is an order of magnitude faster than ML methods. Therefore, we suggest an application of the CH method for universal generalization assessment supported by the application of ML methods to reveal the scope of differences in underlying distributions. Combining the results of these 2 methods, one receives a complete vision of potential generalization issues.

In medicine, the application of ML promises to provide solutions for unmet needs in clinical practice which have partly been hampered by a missing mechanistic understanding of the underlying processes. Medical applications, like an early diagnosis of rare or complex diseases, optimization of therapeutic strategies or the surveillance of patients, and resource planning are expected to benefit from the advantages of ML significantly (Komorowski et al., 2018; Miotto et al., 2018; Shillan et al., 2019; Ghassemi et al., 2020). However, despite promising results for image-analysis-based medical applications or time-series monitoring (Arcadu et al., 2019; Tomasev et al., 2019), the superiority of DNN when compared to traditional approaches has not been proven yet (Chen et al., 2019). Moreover, it has been demonstrated that the design and integration of complex data analytics workflows play a key role in the performance of ML algorithms in biomedical applications (Schatzle et al., 2020). The realization of the promises of ML in medicine requires further innovations in a huge variety of challenges, ranging from data availability and learning strategies up to the integration of a priori knowledge into ML setup (Frohlich et al., 2018).

A highly crucial issue of ML application in medicine arises, when a model developed and trained on high-quality data of one hospital and showing good predictive performance, does not deliver adequate performance when applied to data of other hospitals. Hidden biases between hospitals could be caused by different admission strategies, guidelines for treatment, patients' baseline values, protocols on settings of medical support devices, or definitions of cut-off values (Kelliny et al., 2008). As an example, in 2019, Yan et al. built a simple data-driven model from electronic health records of 485 patients infected with SARS-CoV2 in the region of Wuhan, China (Yan et al., 2020). The authors claimed that their model could predict the outcome for patients with >90% accuracy using the values of three laboratory parameters only. However, the model failed to deliver the same high accuracy on patient datasets from hospitals in France, the USA, and the Netherlands (Barish et al., 2021; Dupuis et al., 2021; Quanjel et al., 2021).

In this work, we developed a pipeline for the comparison of populations and assessment of an ML model's generalization ability. First, we applied our CH analysis to find CH coverage values between datasets. Second, 4 types of ML models were trained to classify from which hospital a patient's sample originated. The performance of these models was assessed to judge, which datasets differ the most in terms of underlying data distributions. We applied our pipeline to 4 critical-care patient datasets of different origins: three datasets from German hospitals generated within the SMITH project (Marx et al., 2021) and the American "Medical Information Mart for Intensive Care" III (later referred to as MIMIC) dataset (Johnson et al.,

TABLE 1 Clinical characteristics of the analyzed patient cohorts in four hospitals under consideration.

	Hosp A	Hosp B	Hosp C	MIMIC
Total number of patients, n (%)	13,067 (100)	2,976 (100)	1,368 (100)	7,683 (100)
Age, years (mean ± SD)	67.3 ± 14.5	67.3 ± 13.8	68.7 ± 13.0	64.1 ± 15.5
Male gender, n (%)	8,529 (65.3)	1,957 (65.8)	961 (70.2)	4,416 (57.5)
Length of stay ICU, days (mean ± SD)	17.3 ± 19.4	21.2 ± 20.1	18.7 ± 18.1	13.5 ± 12.4
Mortality, n (%)	3,742 (28.6)	828 (27.8)	608 (44.4)	1,277 (16.6)

2016). First, the pipeline was applied to every pair of hospitals to find mean CH coverages and performances of ML models for classification for a data source. Second, we investigated the applicability of the developed pipeline using the example of acute respiratory distress syndrome (ARDS)—a potentially life-threatening condition leading to respiratory insufficiency with possible multi-organ failure and fatal outcomes (Cochi et al., 2016; Raymondos et al., 2017). We showed that drops in the performance of models developed for the classification of ARDS on the first day in the Intensive Care Unit (ICU) were attributed to the poor intersection of convex hulls and to the large differences in underlying data distributions of corresponding hospitals.

Methods

Data

Three German hospitals (later referred to as Hosp A, Hosp B, and Hosp C) provided retrospective, fully anonymized data of ICU patients within the context of the use case "Algorithmic surveillance of ICU patients with acute respiratory distress syndrome" (ASIC) (Marx et al., 2021) of the SMITH consortium which is part of the German Medical Informatics Initiative. The ASIC project was approved by the independent Ethics Committee (EC) at the RWTH Aachen Faculty of Medicine (local EC reference number: EK 102/19). Patient inclusion criteria were age above 18 years and a cumulative duration of mechanical ventilation for at least 24 h. In addition, MIMIC was used as an independent dataset with different geographical origins. To identify the duration of invasive mechanical ventilation (MV) of patients from this dataset, a special MIMIC view was used[1] Each patient's data included routinely charted ICU parameters collected over the whole ICU stay. The full list of parameters is given in Supplementary List S1. Data from all 4 sites were brought to the same units of measurement and were checked for consistency. The final number of patients in corresponding hospitals is given in Table 1.

Data for further analysis were prepared in the following way: first, the median values of routinely charted ICU parameters collected over the first day of ICU stay were extracted as features for the analysis. Features with values missing in more than 30% of patients were omitted. We considered features, that were present in all 4 hospitals after the data feature omission step. The final list of features (21 features overall) used in the analysis can be found in Supplementary List S2. Missing values of features were filled with the hospital-wide median value for that feature.

Use case example: Classification for ARDS on the first day of treatment in ICU

To demonstrate the applicability of the developed pipeline, we considered the following typical use case of the application of ML models in healthcare: classification for a critical condition based on the first-day data. We used the presence of ARDS on the first day in the ICU as an endpoint for classification. The criteria for the diagnosis of an ARDS episode are defined in the Berlin criteria (ARDS Definition Task Force et al., 2012). However, in our use case scenario, only the criteria for oxygenation were taken into account. To be able to assess these criteria, only patients having parameters of MV [positive end-expiratory pressure (PEEP), a fraction of inspiratory oxygen (FiO_2)] and blood gas analysis measurements [partial pressure of oxygen (PaO_2)] during the first 24 h were selected.

ARDS patients were chosen based on ICD-10 codes (J80), where available. In the MIMIC database, ICD-9 coding system was used, which does not contain a specific code for ARDS. Therefore, the ARDS label was assigned to patients having ICD-9 codes for pulmonary insufficiency or respiratory failure (Reynolds et al., 1998): 5,185, 51,851, 51,852, 51,853, and 51,882. ARDS onset time was defined as a time point when the Horowitz index drops below 300 for the first time and stays below this threshold for at least 24 h. To ensure that information on the ARDS/non-ARDS status of patients is present in the data, only first-day ARDS patients were chosen as a case group. The Control group comprised all non-ARDS patients and patients with ARDS onset later than on the first day. A total number of day1-ARDS/non-ARDS patients in corresponding hospitals is given in Table 2.

[1] https://github.com/MIT-LCP/mimic-code/blob/62102b08040ac5db96af7922db8d7832ce30a813/etc/ventilation-durations.sql

TABLE 2 Number of day 1 non-ARDS/ARDS patients in hospitals.

Hospital	Non-ARDS	ARDS (%)
Hosp A	9,471	639 (6.3)
Hosp B	1,123	86 (7.1)
Hosp C	924	88 (8.7)
MIMIC	4,555	237 (4.9)

In this use case, we evaluated how a ML model trained in one hospital behaves in terms of performance if it is applied in another hospital. A Random Forest Classifier was trained in each of the four hospitals separately to classify ARDS and non-ARDS patients and tested in the other unseen hospitals. Performance in all datasets was assessed with ROC AUC.

Convex hull analysis

CH coverage for a new dataset was defined as the ratio of data points of a new dataset that lay inside of the CH of the initial dataset in the pair. An example of CH intersections for hospitals (Hosp B, Hosp C) and for the pair of features, arterial oxygen saturation (SaO_2) and arterial bicarbonate, is shown in Figure 1. It should be noted that CH coverage is not a symmetric measure, i.e., CH coverage of Hosp A by Hosp B can differ from CH coverage of Hosp B by Hosp A. CH coverage for each feature combination was assessed in 2 dimensions, i.e., for each combination of pair of features the coverage of CH of one hospital was calculated for all other hospitals. For instance, if hospitals Hosp A and Hosp B were considered, for each pair of features, CH coverage of Hosp A by Hosp B and CH coverage of Hosp B by Hosp A were calculated. CH coverages were assessed using bootstrapping of underlying data (100 times). The equation for a CH coverage for a Hosp A by Hosp for a pair of features *(feature^i , feature^j)*, where *i* and *j* denote feature indeces, is given by:

$$CH_{cov}\left(feature^{i}, feature^{j}\right) = \frac{\sum_{k\in Hosp\,A1}\left[\left(feature^{i}_{k}, feature^{j}_{k}\right) \in CH_{ij}\left(Hosp\,B\right)\right]}{\sum_{k\in Hosp\,A1}} \quad (1)$$

where $CH_{ij}(Hosp\ B)$ corresponds to the CH of the dataset of Hosp B in 2 dimensions *(feature^i, feature^j)*.

In higher dimensions intersections of CHs identified from datasets of sizes, which are usually available in single hospitals, tend to shrink even for datasets drawn from the same distribution due to the curse of dimensionality. Hence, we tested overlapping data by means of the overlaps of projections onto subspaces spanned by all combinations of 2 features. In case of overlapping CHs, the CHs of all projections will overlap as well. The opposite holds only in the case of homogeneous data distributions within the box in full data space spanned by the intersection of all projections. We assume that this is the case for real-world data available in healthcare and our approach delivers an acceptable approximation for the estimation of translational predictivity for practical use.

CH coverage for a feature was calculated as the median CH coverage value of all feature pairs that contain this feature:

$$CH_{cov}\left(feature^{j}\right) = med\left(CH_{cov}\left(feature^{i}, feature^{j}\right), \ldots, CH_{cov}\left(feature^{n}, feature^{j}\right)\right). \quad (2)$$

Next, the distribution of CH coverages for all features was computed. Finally, mean CH coverages for each pair of hospitals were calculated as the mean CH coverage among all features:

$$CH_{cov}\left(Hosp\,A\ by\ Hosp\,B\right) = \frac{\sum_{i\in n} CH_{cov}\left(feature^{i}\right)}{n} \quad (3)$$

where *n* is the number of features. Additionally, we specified the value of the first quartile minus 1.5*interquartile range of the distribution as a threshold for low-coverage features. A low-coverage feature was defined as a feature with a CH coverage value that lies below the threshold. Such features were identified for each pair of datasets.

To eliminate the influence of noisy data on the CH analysis, a density-based data clustering algorithm DBSCAN[2] was applied to the data. Before each run of the CH algorithm, outliers were removed using the DBSCAN method.

Machine learning method for classification of a dataset, including an algorithm to derive important features to differentiate two datasets

The prepared dataset was split into the train (80%) and test (20%) sets. The classification task was to distinguish patients between two hospitals. Four classifiers, namely Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and AdaBoost (ADA) were used. Since the target label (hospital source identifier) was imbalanced, the "class weight" hyperparameter for LR, RF, and SVM was set to the "balanced" option. An optimal set of model hyperparameters were found using grid search with stratified 5-fold cross-validation on the train set. A ROC AUC score was used to evaluate the performance of the chosen model. Predictions on the test set were evaluated with ROC AUC, precision, recall, and F1 score metrics.

[2] https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html

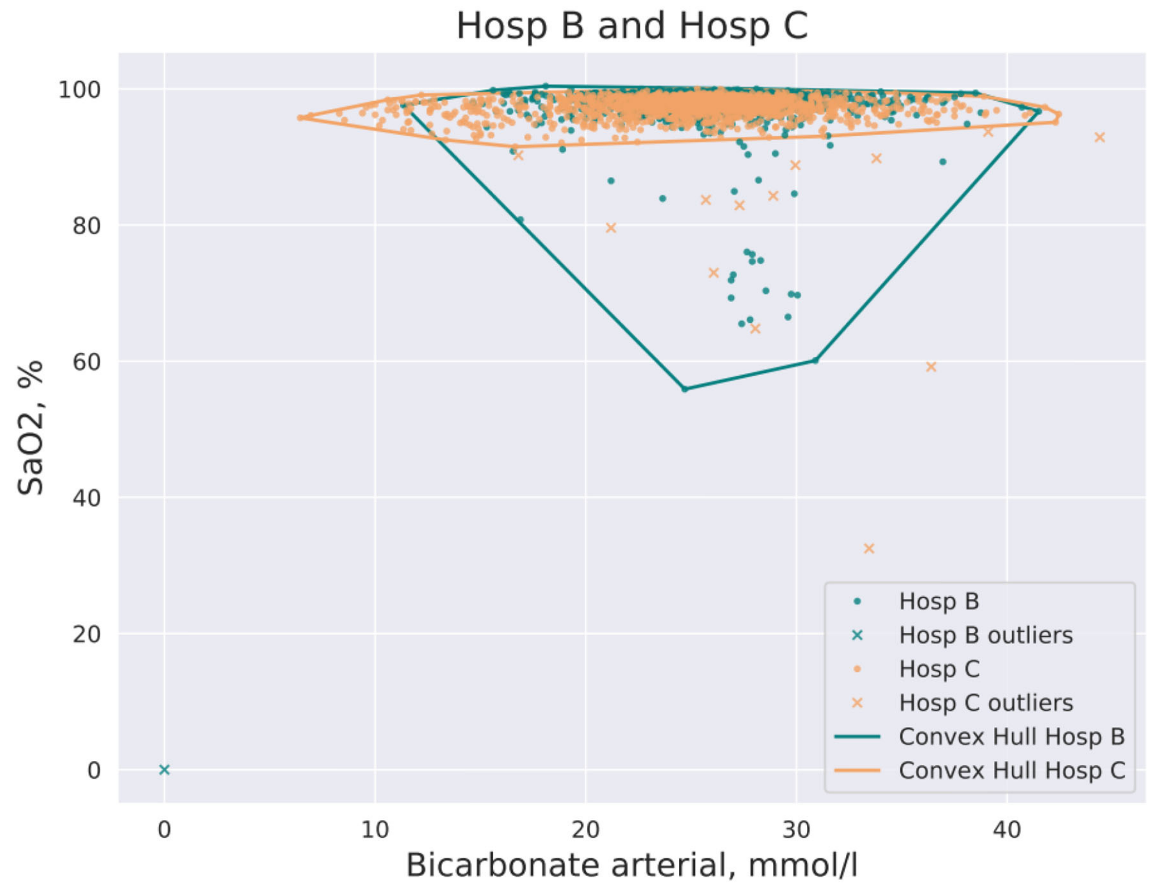


FIGURE 1
Example of CH intersection for the pair of hospitals (Hosp A, Hosp B) and the pair of features: SaO_2 and bicarbonate. Some data points are filtered out by the DBSCAN method prior to the construction of the CH.

ML methods were trained twice. First, all features were used to train ML models. Second, features with low CH coverage were omitted from the analysis and ML models were retrained. This allowed judging, whether the discriminating ability of ML models was predominantly caused by different CHs of underlying data or by differences in underlying data distributions of corresponding hospitals.

Python 3 modules used in this study and system requirements

In this study, the SciPy Python 3 spatial library with the Quickhull algorithm and the Delaunay class (Virtanen et al., 2020) was used for CH analysis. And the Scikit-learn implementations of ML classification methods (Pedregosa et al., 2011) were svm.SVC, linear_model. LogisticRegression, ensemble.RandomForestClassifier and ensemble.AdaBoostClassifier. CH and ML analysis was performed on the computational cluster of the RWTH Aachen University using 1 node with 40 cores, 2.66 GHz, 4 GB RAM. The longest runtime for the CH analysis was 16 min. The runtime for the ML script comprised 24 h. Analysis was tested as well on the 2018 quadcore laptop i7-8565U CPU @ 1.80 GHz × 8. It could be run as it is on most modern CPUs with minimal RAM usage. No GPU is required.

CH and ML methods used in this study are available as a python package "chgen". Example scripts on how to use this package are available in the repository: https://git.rwth-aachen.de/jrc-combine/chgen.

Results

Application of CH analysis to each pair of hospitals

Figure 2 shows the mean CH coverage for each pair of hospitals. For each German hospital, minimum coverage was found when data of the corresponding hospital were covering the MIMIC dataset (last column in Figure 2). However, that was not the case for the opposite situation. Maximum mean CH coverage was found for cases when MIMIC data covers data from German hospitals (last row in Figure 2).

Features with low CH coverage values were identified for each pair of hospitals. These features are shown in Table 3. The table is not symmetric since features with low coverage values when the first hospital's data cloud is covering the second one may be different from features in the opposite coverage situation. Results of the mean CH coverage are accompanied by the number of features with low CH coverage in each case of the datasets' comparison. For each German hospital, a maximum number of such features was found when data of German hospitals were covering the MIMIC dataset (3 or 2 features correspondingly, last column in Table 3). CH coverages for all features in the case of MIMIC coverage are given in Supplementary Table S1.

Application of ML routines for classification of the hospital

Results of the application of ML routines to classify the hospital for every pair of hospitals are shown in Supplementary Figure S1A. Results of the ADA method are shown, as it gained the highest performance in terms of ROC AUC in all cases. In each pair of hospitals, the hospital where the patient samples were derived from could be almost perfectly classified (ROC AUC $\geq$ 0.94). The best separation was obtained between the MIMIC cohort and German hospitals. German hospitals looked more alike to classifiers. The worst separation was observed between Hosp B and Hosp C.

After the exclusion of the features with low CH coverage values, and retraining with the best-performing ML classifiers, the largest ROC AUCs were still observed between the MIMIC cohort and German hospitals (see Supplementary Figure S1B).

Use case example: Classification for ARDS on the first day of treatment in ICU

The results of the classification task are shown in Figure 3. Diagonal cells represent the performance of a specialized model which was trained and tested in the same hospital. The performance of specialized models strongly differed among hospitals under consideration, with the lowest ROC AUC of 0.79 in MIMIC and the highest of 0.94 in Hosp B. To test the generalization ability of developed models, they were tested on other unseen datasets, i.e., other hospitals (non-diagonal cells).

If the population of the new hospital is similar to or more homogeneous than the one of the original hospitals concerning the condition under consideration, the performance of the model will stay on a similar level or can be even higher than in the original hospital. However, if the population differs from the original one, performance will be impaired. For each specialized model trained in German hospitals the largest drop in performance was observed when the respective model was applied in the MIMIC dataset with the strongest drop of 0.26 for a model trained in Hosp B. Overall, models developed in Germany, showed impaired performance compared to the specialized MIMIC model. The opposite was not the case, as the MIMIC model showed similar performance in German hospitals to the performance in the original cohort.

Discussion

"Internal" model performance on structurally similar, previously unseen data, gathered from the same source used for model training, can be contrasted with "external" model performance on new, previously unseen data from other sources. ML models perform worse in external cohorts due to several reasons such as different protocols, confounding variables, or heterogeneous populations (Cabitza et al., 2017; Zech et al., 2018; Martensson et al., 2020; Goncalves et al., 2021). Moreover, medical data can be biased by a variety of factors such as admission policies, hospital treatment protocols, country-specific guidelines, clinician discretion, healthcare economy, etc. Furthermore, labeling or coding criteria of a certain disease or syndrome and treatment guidelines evolve with time (Kunze et al., 2020). Since ML models for healthcare are predominantly developed on retrospective data, it remains unclear how the performance of such models is affected by the temporal separation of the target group even within one hospital.

Similarly, model reproducibility and model transportability have distinct objectives (Justice et al., 1999). While reproducibility focuses on the performance of the model in the same target population, transportability refers to performance in different but related source populations. Nevertheless, the closeness of this relationship between populations must be ascertained to achieve valid results of external validation. The performance will be poor in a sample that is too different from the data used for development. Conversely, a test sample that is too similar will overestimate the predictive performance showing reproducibility rather than transportability. To address these different aspects, an elaborate validation approach as described by Debray et al. seems necessary. They recommend the examination of the validation datasets in the first step to ensure adequate relatedness using a case-mix of a dataset and subsequent evaluation of the model with respect to the perceived relatedness (Debray et al., 2015).

In this study, we have introduced another method for population comparison and assessment of a model's generalizability. First, it estimates the similarity of the underlying populations in terms of mean CH coverage. Second,

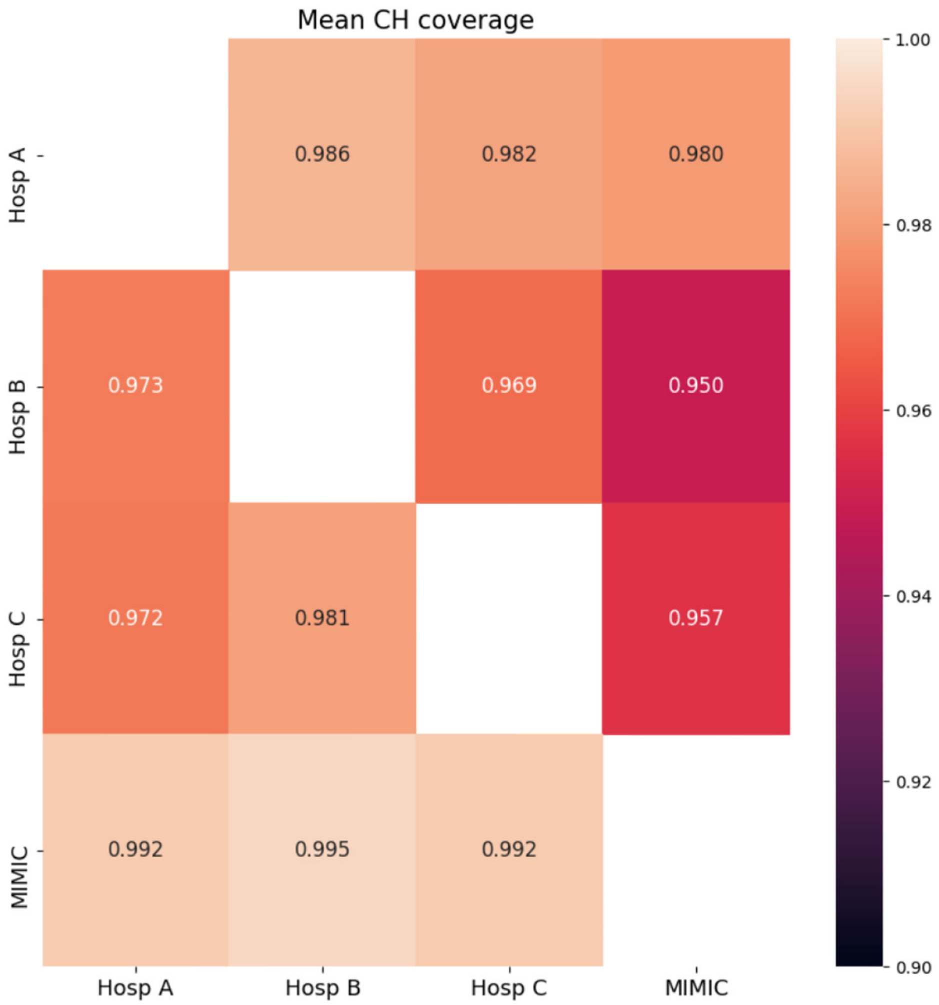


FIGURE 2
CH analysis results for data from four hospitals. Mean CH coverage over all features is shown. Rows-initial population, columns-population, whose CH is covered by the CH of the initial population.

it estimates the differences in datasets in terms of underlying data distributions. These two tasks are accomplished by the application of 2 methods–first the CH analysis and followed by the ML classifiers.

During the application of the pipeline on the datasets obtained from 4 hospitals, we found that there were significant differences in CH coverage among pairs of hospitals. The lowest CH coverages for each of the German hospitals were observed when the MIMIC dataset was covered by data obtained from the corresponding hospital. However, in the opposite case i.e., Hosp A/Hosp B/Hosp C covered by MIMIC, the coverages were large. This shows that Hosp B/Hosp C and to a lesser extent Hosp A represented a part of the data space, spanned by data of MIMIC. In other words, data from German hospitals comprised, in greater or lesser proportions, parts of the MIMIC data cloud.

All four datasets exhibited differences in underlying data distributions. Once trained, ML classifiers were able to distinguish data coming from different sources with ROC AUC larger than 0.94, suggesting nearly perfect identification of the hospital from where the patient data originated from. After the omission of features with low CH coverages, the performance of retrained models dropped. However, the performance of models distinguishing MIMIC from German hospitals was still largely supporting the finding, that the MIMIC dataset significantly differed from German hospitals.

To demonstrate that our pipeline can be used to assess the generalization ability of ML models, we considered a use case of classification for the first day of ARDS data. A specialized model was trained for each of the four hospitals' data. Then it was applied to unseen hospital data and the performance of

TABLE 3 Lists of parameters with low CH intersections for all pairs of hospitals.

	Hosp A	Hosp B	Hosp C	MIMIC
Hosp A	-	Tidal volume, PEEP	Tidal volume	PaO_2, Tidal volume, PTT
Hosp B	PaO_2, Respiratory rate	-	Bicarbonate arterial, Respiratory rate, PTT	PaO_2, Bicarbonate arterial, PTT
Hosp C	FiO_2, PEEP	-	-	PaO_2, PEEP
MIMIC	FiO_2, Lactate arterial	Lactate arterial	Lactate arterial	-

Rows-initial population, columns-population, CH of which is covered by the CH of the initial population.

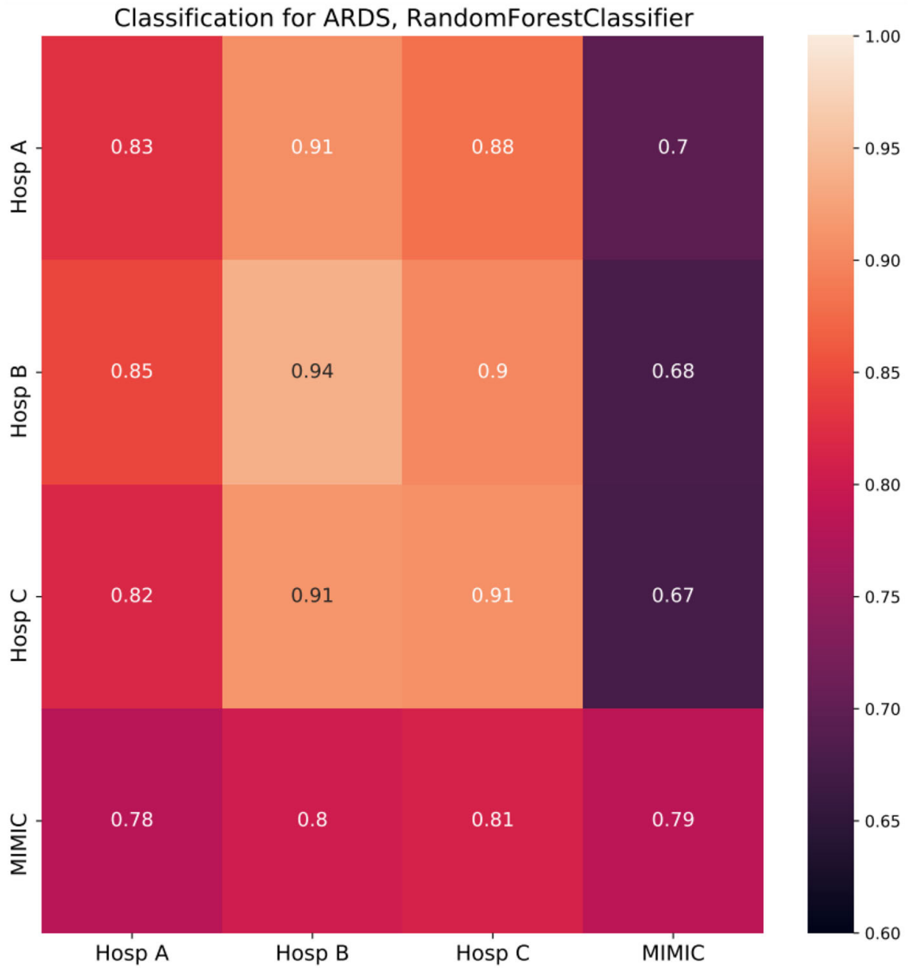


FIGURE 3
Random forest classifier classification results (cross-prediction matrix) for ARDS on the first day in ICU. RF trained in each of the four hospitals (row name) and applied in each of the four hospitals (column name). Diagonal cells represent the performance of specialized models which were trained and tested in the same hospital. Non-diagonal cells represent the performance of such models once they are applied in other hospitals and reflect ability of a model to generalize to the unseen population of another hospital. Twenty-one features common for all four hospitals were used to build corresponding RF models. Performance is depicted in terms of ROC AUC.

the model on the original data was compared to those of the new data. We observed 2 clusters of datasets, namely German hospitals and MIMIC. Models developed for German hospitals' data exhibited the largest drop in performance once applied to MIMIC. That was not the case in the opposite situation, i.e., application of the MIMIC model to German hospitals data, where almost no drops were observed. CH analysis fully supported these findings. First, for each of the German hospitals, the lowest CH coverages were observed when the MIMIC dataset was covered by data from corresponding hospitals suggesting the impaired performance of models developed in German hospitals and applied in MIMIC. Second, mean CH coverages of German datasets by MIMIC data were found close to 1, suggesting full CH coverage and thus, the absence of limitations for generalization.

Moreover, smaller drops in performance were observed when models developed on data from Hosp B or Hosp C were applied to data from Hosp A. This is in line with corresponding CH coverages (Hosp A by Hosp B/Hosp C), which are in the medium range. Interestingly, when models, developed in Hosp A or MIMIC were applied in Hosp B or Hosp C we did not observe any drop in performance, but even a slight increase. It could be the case if the population of the new hospital is similar to or more homogeneous than the one of the original hospitals concerning the condition under consideration. In our case, it would mean, that fewer non-ARDS patients with low Horowitz index are present in Hosp B/C compared to Hosp A/MIMIC. On the other hand, the necessary condition for the proper generalization, in this case, is satisfied by the fact, that CH coverages of Hosp B/C by Hosp A/MIMIC are among the largest in our study. Overall, the results of cross-prediction for ARDS were found to be in accordance with the results of the CH analysis of corresponding datasets.

Application of ML routines for classification for a hospital also supported the finding, which suggests that the MIMIC data significantly differed from German datasets, as the best separation with ROC AUCs > 0.99 was obtained between the MIMIC cohort and German hospitals. Nearly perfect separation was still possible after the exclusion of features with low CH coverage. This result indicated that the MIMIC cohort is not only less covered by German data, but exhibits diverging underlying data distributions once compared to German hospitals. However, while ML methods indicated, that there were significant differences in underlying distributions and performance of a model could be impaired, they did not point in the direction of proper or poor generalization, i.e., models trained in dataset A and applied in dataset B and vice versa. This constitutes an advantage of the CH method, as it is originally asymmetric and allows to assessment direction of impaired generalization. Moreover, the CH assessment is universal and does not depend on the particular ML classification method.

However, there could be multiple other reasons for such strong discrepancies in models' performance. First, some of the features with low CH coverage (PEEP, FiO_2) belong to parameters, which are set by physicians in the ICU, thus suggesting different treatment strategies in underlying hospitals. Second, diverging ARDS labeling criteria (ICD-10 in Germany vs. ICD-9 in MIMIC) might contribute to label uncertainty in ARDS classification. Finally, the timespans of data collection overlap only partially. MIMIC data were collected between 2001 and 2012, Hosp A data between 2009 and 2019. Data from Hosp B and Hosp C were collected after 2012. This is relevant since in 2012 the American European Consensus Conference (AECC) definition of ARDS changed to the currently accepted Berlin definition (Kunze et al., 2020).

Nevertheless, the main observation is valid regardless of particular ARDS labeling: MIMIC data do significantly differ from all three other hospitals in this study. Given that this database is considered nearly a gold standard of open ICU databases, an external validation for models developed on this database is absolutely necessary. In the best case, a special pipeline for the assessment of the transferability of trained models should be included in the data preparation step before a model development, so that generated models might exhibit significantly better performance.

Our study has other limitations that have to be considered. It is known that CH analysis is very sensitive to noise in the data (Worton, 1995). To eliminate the influence of noisy data on the convex hull analysis, a density-based data clustering algorithm DBSCAN (Schubert et al., 2017) was applied to the data so that during each run of the CH algorithm, outliers were removed using the DBSCAN method. Additionally, to increase the robustness of the CH analysis results, each CH analysis execution was averaged over 100 runs with bootstrapped data. Another potential weakness of our study design is that imputation was done without taking into account multidimensional parameter distribution. However, this could not significantly influence the main conclusions on differentiating parameters in this study, as we specifically have chosen patients with charted data of the main parameters important in the ARDS state: PaO_2, FiO_2, and PEEP. Another important question is how to define cutoff values between good and bad performance for both CH and ML analysis. We estimated CH coverages between train and test sets for the same hospital (see Supplementary Table S2). These can be used as benchmarks for CH intersections for reasonable generalization. However, these also differed among hospitals, but here a clear correlation with the sample size of the cohort was observed. For instance, in Hosp C, a test set of 202 patients was covered by a train set of 810 patients. Therefore, the estimates for proper CH coverage should also depend on the sample size under consideration. For large datasets (Hosp A/MIMIC), where test set sizes were comparable to the sizes of smaller datasets in the study (Hosp C) they comprise 0.987/0.972. For ML routines, there is no rule of thumb to define minimum ROC AUC to judge whether hospitals cannot be distinguished. Usually, values of ROC AUC < 0.7 are considered poor discrimination performance.

Additionally, sample size can potentially be a factor, while considering convex hull intersections and machine learning results. However, there are some pieces of evidence, that this is at least not a dominant factor for generalization differences. First, models developed in small cohorts of Hosp B/Hosp C for ARDS classification deliver similar performance in Hosp A, as a specialized model of that hospital. Second, the model developed in Hosp A has a high generalization error in MIMIC (0.13), but a model developed in MIMIC shows the opposite behavior in Hosp A having a small generalization error (0.01). Third, a model developed in the smallest Hosp C does not exhibit any generalization error in a dataset of completely different Hosp B. Therefore, we are of the strong opinion, that different sample sizes in underlying hospitals cannot explain such strong discrepancies in models' performance in different hospitals.

Another important remark is that as the dimension of a dataset grows, then a trained ML model will almost always lay in the extrapolation range once applied to unseen data (Balestriero et al., 2021). This is a consequence of the curse of dimensionality and has to be considered in all ML applications and especially in deep learning where models are dealing with hundreds or thousands of features. However, ML models that utilize continuous time series data and are applied in real healthcare settings usually require a degree of interpretability and therefore contain a limited number of features (Chen et al., 2019). This was also the case in our study, where the number of features did not exceed 30.

Conclusions

Currently, with the ever-growing number of AI and ML models in healthcare, there is a huge challenge in the translation of such models into clinical practice. In healthcare, new data are often different from those used in the training of the model. To achieve a clinical implementation, a model must be able to perform with sufficient accuracy on previously unseen data. Hospitals may have different policies, guidelines, or protocols, but even within one hospital, guidelines could change over time causing altering patients' responses.

Therefore, the validation of developed models before a potential application at the bedside plays a key role in translation research. With the pipeline introduced in this study, we contribute to the solution of this issue. Given the training data and a retrospective dataset from a hospital, where the model is intended to be used, we can judge the generalization ability in another hospital. On the use case of classification for the first day ARDS, we showed that the strongest drop in performance is associated with the poor intersection of convex hulls of corresponding hospitals and with differences in underlying data distributions. Therefore, we suggest the application of our pipeline as a first tool to assess the transferability of trained models.

Based on our analysis of four different hospital datasets, we conclude that datasets from different hospitals represent heterogeneous data sources and the transfer from one database to another should be performed with care to avoid implications during real-world applications of the developed models. Further research is needed to develop methods for the adaptation of ML models to new hospitals. In addition, more work should be aimed at the creation of gold-standard datasets that are large and diverse with data from varied application sites.

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

HM, SF, KS, RP, and KN worked on data acquisition and harmonization. KS, MS, and KN developed and implemented CH analysis scripts. KS, JSB, and KN developed and implemented ML routines. KS, JSB, and AS designed the research, performed analysis, analyzed the patient data, and developed the ARDS prediction model. SF gave medical advice during the development of the pipeline. SF, GM, and JB interpreted the results from a medical perspective. KS, JSB, SF, and AS wrote the manuscript. All authors read and approved the final manuscript.

Funding

This publication of the SMITH consortium was supported by the German Federal Ministry of Education and Research, Grant Nos. 01ZZ1803B, 01ZZ1803I, and 01ZZ1803M.

Acknowledgments

We would like to thank for the funding support that Moein Einollahzadeh Samadi received as a member of the Helmholtz School for Data Science in Life Earth and Energy (HDS-LEE).

Conflict of interest

HM is an employee of Bayer AG, Germany. HM has stock ownership with Bayer AG, Germany.

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

Publisher's note

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

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdata.2022.603429/full#supplementary-material

SUPPLEMENTARY FIGURE S1
ROC AUC for classification for a hospital. **(A)**: Performance of an ML learning algorithm for classification for a hospital. **(B)**: Performance of an ML learning algorithm for classification for a hospital after removal of features with low CH coverage values. Numbers in cells reflect the ROC AUC of the classifier trained to separate between hospital 1 (row name) and hospital 2 (column name).

SUPPLEMENTARY LIST S1
Diagnostic parameters used in this study. Overall, 54 diagnostic parameters routinely assessed in the ICU were used in this study. Additionally, 6 biometric parameters were used.

SUPPLEMENTARY LIST S2
List of parameters used for classification ARDS on the first day in ICU.

SUPPLEMENTARY TABLE S1
CH coverages for all features. MIMIC data covered by other hospitals.

SUPPLEMENTARY TABLE S2
CH coverage of the test set by the train set in the same hospital, where ML models were developed.

References

AlBadawy, E. A., Saha, A., and Mazurowski, M. A. (2018). Deep learning for segmentation of brain tumors: Impact of cross-institutional training and testing. *Med. Phys.* 45, 1150–1158. doi: 10.1002/mp.12752

Arcadu, F., Benmansour, F., Maunz, A., Willis, J., Haskova, Z., and Prunotto, M. (2019). Deep learning algorithm predicts diabetic retinopathy progression in individual patients. *NPJ Digit. Med.* 2, 92. doi: 10.1038/s41746-019-0172-3

ARDS Definition Task Force, Ranieri, V. M., Rubenfeld, G. D., Thompson, B. T., Ferguson, N. D., Caldwell, E., et al. (2012). Acute respiratory distress syndrome: the Berlin Definition. *JAMA* 307, 2526–2533. doi: 10.1001/jama.2012.5669

Balestriero, R., Pesenti, J., and LeCun, Y. (2021). Learning in high dimension always amounts to extrapolation. *arXiv preprint arXiv:2110.09485.* doi: 10.48550/arXiv.2110.09485

Barish, M., Bolourani, S., Lau, L. F., Shah, S., and Zanos, T. P. (2021). External validation demonstrates limited clinical utility of the interpretable mortality prediction model for patients with COVID-19. *Nat. Mach. Intell.* 3, 25–27. doi: 10.1038/s42256-020-00254-2

Barron, A. R., and Klusowski, J. M. (2018). Approximation and estimation for high-dimensional deep learning networks. *arXiv preprint arXiv:1809.03090.* doi: 10.48550/arXiv.1809.03090

Cabitza, F., Rasoini, R., and Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. *JAMA* 318, 517–518. doi: 10.1001/jama.2017.7797

Chen, D., Liu, S., Kingsbury, P., Sohn, S., Storlie, C. B., Habermann, E. B., et al. (2019). Deep learning and alternative learning strategies for retrospective real-world clinical data. *NPJ Digit. Med.* 2, 43. doi: 10.1038/s41746-019-0122-0

Cochi, S. E., Kempker, J. A., Annangi, S., Kramer, M. R., and Martin, G. S. (2016). mortality trends of acute respiratory distress syndrome in the United States from 1999 to 2013. *Ann. Am. Thorac. Soc.* 13, 1742–1751. doi: 10.1513/AnnalsATS.201512-841OC

Courrieu, P. (1994). Three algorithms for estimating the domain of validity of feedforward neural networks. *Neural Netw.* 7, 169–174. doi: 10.1016/0893-6080(94)90065-5

Debray, T. P., Vergouwe, Y., Koffijberg, H., Nieboer, D., Steyerberg, E. W., and Moons, K. G. (2015). A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J. Clin. Epidemiol.* 68, 279–289. doi: 10.1016/j.jclinepi.2014.06.018

Dupuis, C., De Montmollin, E., Neuville, M., Mourvillier, B., Ruckly, S., and Timsit, J. F. (2021). Limited applicability of a COVID-19 specific mortality prediction rule to the intensive care setting. *Nat. Mach. Intell.* 3, 20–22. doi: 10.1038/s42256-020-00252-4

Frohlich, H., Balling, R., Beerenwinkel, N., Kohlbacher, O., Kumar, S., Lengauer, T., et al. (2018). From hype to reality: data science enabling personalized medicine. *BMC Med* 0.16, 150. doi: 10.1186/s12916-018-1122-7

Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., Chen, I. Y., and Ranganath, R. (2020). "A review of challenges and opportunities in machine learning for health," in *AMIA Joint Summits on Translational Science Proceedings. AMIA Joint Summits on Translational Science 2020* (Bethesda, MD: American Medical Informatics Association), 191–200.

Goncalves, J., Yan, L., Zhang, H.-T., Xiao, Y., Wang, M., Guo, Y., et al. (2021). Li Yan et al. reply. *Nat. Mach. Intell.* 3, 28–32. doi: 10.1038/s42256-020-00251-5

Graham, R. L. (1972). An efficient algorith for determining the convex hull of a finite planar set. *Inf. Process. Lett.* 1, 132–133. doi: 10.1016/0020-0190(72)90045-2

Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L. W., Feng, M., Ghassemi, M., et al. (2016). MIMIC-III, a freely accessible critical care database. *Sci. Data* 3, 160035. doi: 10.1038/sdata.2016.35

Justice, A. C., Covinsky, K. E., and Berlin, J. A. (1999). Assessing the generalizability of prognostic information. *Ann. Internal Med.* 130, 515–524. doi: 10.7326/0003-4819-130-6-199903160-00016

Kelliny, C., William, J., Riesen, W., Paccaud, F., and Bovet, P. (2008). Metabolic syndrome according to different definitions in a rapidly developing country of the African region. *Cardiovasc. Diabetol.* 7, 27. doi: 10.1186/1475-2840-7-27

Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C., and Faisal, A. A. (2018). The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat. Med* 0.24, 1716–1720. doi: 10.1038/s41591-018-0213-5

Kunze, J., Fritsch, S., Peine, A., Maaßen, O., Marx, G., and Bickenbach, J. (2020). Management of ARDS: from ventilation strategies to intelligent technical support–connecting the dots. *Trends Anaesth. Crit. Care* 34, 50–58. doi: 10.1016/j.tacc.2020.05.005

Mallat, S. (2016). Understanding deep convolutional networks. *Philos. Trans. A Math. Phys. Eng. Sci.* 374, 20150203. doi: 10.1098/rsta.2015.0203

Martensson, G., Ferreira, D., Granberg, T., Cavallin, L., Oppedal, K., Padovani, A., et al. (2020). The reliability of a deep learning model in clinical out-of-distribution MRI data: a multicohort study. *Med. Image Anal.* 66, 101714. doi: 10.1016/j.media.2020.101714

Marx, G., Bickenbach, J., Fritsch, S. J., Kunze, J. B., Maassen, O., Deffge, S., et al. (2021). Algorithmic surveillance of ICU patients with acute respiratory distress syndrome (ASIC): protocol for a multicentre stepped-wedge cluster randomised quality improvement strategy. *BMJ Open* 11, e045589. doi: 10.1136/bmjopen-2020-045589

Miotto, R., Wang, F., Wang, S., Jiang, X., and Dudley, J. T. (2018). Deep learning for healthcare: review, opportunities and challenges. *Brief. Bioinform.* 19, 1236–1246. doi: 10.1093/bib/bbx044

Ostrouchov, G., and Samatova, N. F. (2005). On FastMap and the convex hull of multivariate data: toward fast and robust dimension reduction. *IEEE Trans. Pattern Anal. Mach. Intell.* 27, 1340–1343. doi: 10.1109/TPAMI.2005.164

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830. doi: 10.48550/arXiv.1201.0490

Pooch, E. H., Ballester, P. L., and Barros, R. C. (2019). Can we trust deep learning models diagnosis? The impact of domain shift in chest radiograph classification. *arXiv preprint arXiv:1909.01940.* doi: 10.1007/978-3-030-62469-9_7

Quanjel, M. J. R., van Holten, T. C., Gunst-van der Vliet, P. C., Wielaard, J., Karakaya, B., Söhne, M., et al. (2021). Replication of a mortality prediction model in Dutch patients with COVID-19. *Nat. Mach. Intell.* 3, 23–24. doi: 10.1038/s42256-020-00253-3

Raymondos, K., Dirks, T., Quintel, M., Molitoris, U., Ahrens, J., Dieck, T., et al. (2017). Outcome of acute respiratory distress syndrome in university and non-university hospitals in Germany. *Crit. Care* 21, 122. doi: 10.1186/s13054-017-1687-0

Reynolds, H. N., McCunn, M., Borg, U., Habashi, N., Cottingham, C., and Bar-Lavi, Y. (1998). Acute respiratory distress syndrome: estimated incidence and mortality rate in a 5 million-person population base. *Crit. Care* 2, 29–34. doi: 10.1186/cc121

Schatzle, L. K., Hadizadeh Esfahani, A., and Schuppert, A. (2020). Methodological challenges in translational drug response modeling in cancer: a systematic analysis with FORESEE. *PLoS Comput. Biol.* 16, e1007803. doi: 10.1371/journal.pcbi.1007803

Schubert, E., Sander, J., Ester, M., Kriegel, H. P., and Xu, X. (2017). DBSCAN revisited, revisited: why and how you should (Still) use DBSCAN. *ACM Trans. Database Syst.* 42, 19. doi: 10.1145/3068335

Shesu, R. V., Bhaskar, T. V. S. U., Rao, E. P. R., Ravichandran, M., and Rao, B. V. (2021). An improved method for quality control of in situ data from Argo floats using α convex hulls. *MethodsX* 8, 101337. doi: 10.1016/j.mex.2021.101337

Shillan, D., Sterne, J. A. C., Champneys, A., and Gibbison, B. (2019). Use of machine learning to analyse routinely collected intensive care unit data: a systematic review. *Crit. Care* 23, 284. doi: 10.1186/s13054-019-2564-9

Tomasev, N., Glorot, X., Rae, J. W., Zielinski, M., Askham, H., Saraiva, A., et al. (2019). A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* 572, 116–119. doi: 10.1038/s41586-019-1390-1

Torralba, A., and Efros, A. A. (2011). Unbiased look at dataset bias. *CVPR* 2011, 1521–1528. doi: 10.1109/CVPR.2011.5995347

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* 17, 261–272. doi: 10.1038/s41592-019-0686-2

Worton, B. J. (1995). A convex hull-based estimator of home-range size. *Biometrics* 51, 1206–1215. doi: 10.2307/2533254

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

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., and Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS Med.* 15, e1002683. doi: 10.1371/journal.pmed.1002683

Zhou, X., and Shi, Y. (2009). Nearest neighbor convex hull classification method for face recognition. *Comput. Sci.* 2009, 570–577. doi: 10.1007/978-3-642-01973-9_64

TYPE Original Research
PUBLISHED 04 January 2023
DOI 10.3389/frai.2022.1035805

OPEN ACCESS

EDITED BY
Subhendu Kumar Pani,
Krupajal Engineering College, India

REVIEWED BY
Gokhan Altan,
Iskenderun Technical University,
Turkey
Evaldas Vaiciukynas,
Kaunas University of Technology,
Lithuania

*CORRESPONDENCE
Soumya Mishra
✉ soumya@cgu-odisha.ac.in
Tusar Kanti Dash
✉ tusarkantidash@gmail.com
Ganapati Panda
✉ ganapati.panda@gmail.com

SPECIALTY SECTION
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

RECEIVED 03 September 2022
ACCEPTED 18 November 2022
PUBLISHED 04 January 2023

CITATION
Mishra S, Dash TK and Panda G (2023) Speech phoneme and spectral smearing based non-invasive COVID-19 detection. *Front. Artif. Intell.* 5:1035805. doi: 10.3389/frai.2022.1035805

Speech phoneme and spectral smearing based non-invasive COVID-19 detection

Soumya Mishra*, Tusar Kanti Dash* and Ganapati Panda*

Department of Electronics and Communication Engineering, C. V. Raman Global University, Bhubaneswar, India

COVID-19 is a deadly viral infection that mainly affects the nasopharyngeal and oropharyngeal cavities before the lung in the human body. Early detection followed by immediate treatment can potentially reduce lung invasion and decrease fatality. Recently, several COVID-19 detections methods have been proposed using cough and breath sounds. However, very little study has been done on the use of phoneme analysis and the smearing of the audio signal in COVID-19 detection. In this paper, this problem has been addressed and the classification of speech samples has been carried out in COVID-19-positive and healthy audio samples. Additionally, the grouping of the phonemes based on reference classification accuracies have been proposed for effectiveness and faster detection of the disease at a primary stage. The Mel and Gammatone Cepstral coefficients and their derivatives are used as the features for five standard machine learning-based classifiers. It is observed that the generalized additive model provides the highest accuracy of 97.22% for the phoneme grouping “/t//r//n//g//l/.” This smearing-based phoneme classification technique can also be used in the future to classify other speech-related disease detections.

KEYWORDS

COVID-19 detection, machine learning, spectral smearing, phoneme analysis, COVID-19

1. Introduction

COVID-19 was publicly avowed as an epidemic demanding leading nations with medical prowess to develop faster and more accurate testing mechanisms. Flu, cough, exhaustion, asthma, and pneumonia with fatality have been primarily the clinical symptoms of the affected patients (Peng, 2020). To alleviate the dearth of RT-PCR testing sets, medicos and testing centers had to discover alternate options such as Computed Tomography scans (CT scans) for COVID-19 diagnosis of suspected patients. Some improved COVID-19 detection schemes are used such as contrast limited adaptive histogram equalization and local histogram equalization for extracting significant information from raw chest X-ray images (Narlı, 2021; Narli and Altan, 2022). The velcro-like lung sounds and lung ultrasound readings are also used for the successful detection of COVID-19 (Kiamanesh et al., 2020; Pancaldi et al., 2022). Radiologists have been found to be heavily engaged during the epidemic of COVID-19. They somehow lacked the capacity to decipher a variety of CT scans in due time (Afshar et al., 2021).

In addition, clinicians could not as well distinguish COVID-19 from CT scans in remote villages, such as rural regions, because this disease is relatively recent. The importance of reducing the dose of radiation in radiological studies, particularly concerning CT, had become a point of apprehension based on its numerous and dependable medical applications across the globe.

Corona Virus has been primarily hosted on the intra-nasal, bronchial, and lung systems of the human body (Gallo, 2021), and therefore, audio analysis of speech segments from infected samples could potentially indicate respiratory, articulatory, and breathing aberrations as compared with healthy speech samples. Speech-based audio detection of COVID-19 would not only be non-invasive and cost-friendly but can be performed with huge flexibility and portability from any location, adhering to social distancing norms. Speech-based disease recognition has gained immense admiration in recent times predominantly in diagnosing neurodegenerative diseases affecting regular speech patterns. Audio features are explicitly extracted from the concerned database samples, assigned markers for classification, and fed into the system model for training followed by validation and performing an accuracy check (Sharma G. et al., 2020). Phoneme-based disease classification has showcased progressive accuracy with minimum latency in diagnosing several diseases such as stroke, amyotrophic lateral sclerosis (ALS), Parkinson's disease (PD), cleft lip and palate (CLP), primary progressive aphasia, spasmodic dysphonia, Alzheimer's disease, and dementia.

The conventional speech features considered are high-frequency local field potential, zero crossing rate, mean and standard deviation, spikes in the audio signal, Mel-frequency Cepstral coefficients (MFCC), Jitter, shimmer, and voice breaks (Zhang and Wu, 2020). Perceptual linear prediction (PLP), relative spectra (RASTA), and linear prediction coefficients (LPC) have also been reported as instrumental in classification (Moro-Velazquez et al., 2019). Prospective artificial intelligence/machine learning and deep-learning phoneme classification methodologies have been the topic of interest in research advancements for decades (Lamba et al., 2021). Phonemes in the process of articulation can be distinctively segregated into six categories such as stop, affricate, fricative, nasal, and lateral. Subsequently, they can be sub-categorized to the next level of distinction based on modes of sound articulation originating in the vocal tract forming a tubal resonance effect while producing speech (Katamba, 1989). Phonemes, irrespective of dialects, spoken language, or vocabulary adhered across diversities, can alone suffice to be a powerful speech segment for processing speech-based recognition applications. Researchers have actively formulated words made up of relevant phonemes to trigger the appropriate vocal parametric articulations for detecting speech disorders, indicating anomalies (Wielgat, 2008).

1.1. Motivation

In previous research outcomes, it has been apparent that variations in phoneme lengths and frequency, as well as changes in phoneme-dependent tone and formant gradients, represent the phonemic segment reliance on phonation and articulation shifts with Parkinson's severity. Yet, there has been a preliminary study on speech-based COVID-19 detection focusing mainly on cough, breath, and vowels (Han et al., 2021; Kumar and Alphonse, 2021) and a generalized comparison of the COVID-19 assessment of phoneme-vowel categories (Boothroyd et al., 1996). Not every affected patient might show cough and shortness of breath as potential symptoms. In this case, phonemes may emerge as worthy indicators for early detection of the disease. The best bet to utilize phonemes as an efficient classification strategy is based on the fact that a speaker need not necessarily generate his samples to train all words in the vocabulary list but only the phonetic segments need to be processed.

1.2. Research objective

An effort is initiated in this article to classify COVID-19-affected positive and healthy candidates by disintegrating the audio speech sentence spoken by the concerned specimen into relevantly available English phonemes. The various phonemes are then labeled as positive and healthy classes as demarcated in the referred corpus. In an attempt to enhance classification accuracy, the individual phoneme audio wave has been smeared using low-pass filter noise. Most importantly, the phonemes acquiring the highest classification performance have been concatenated to propose a phoneme group called "buzzword." The so-called buzzword may be used in the future to detect the disease, evading the dependency on cough or breath samples. In this article, 16 distinct English phonemes with three vowels have been utilized on the available datasets, using 78 feature-sets comprising MFCC, GTCC, and its variant features with five machine-learning classification techniques. The findings of the investigation are as follows:

- Selection of appropriate smearing bandwidth for improving the classification accuracy for different feature sets.
- Use of smearing signal for enhancing the classification accuracy.
- Application of Phoneme-based Buzzwords to assist clinicians and patients with more precise and focused detection mechanism.

TABLE 1 Phoneme database prepared for this study.

Sl.	Phoneme	Phoneme category	No of speech samples (C-19 p +n)
1	/b/	Stop	112+104
2	/d/	Stop	108+110
3	/v/	Fricative	108+110
4	/m/	Nasal	108+108
5	/l/	Alveolar Lateral approximant	104+104
6	/f/	Fricative	112+110
7	/Oy/	Diphthong vowel	105+108
8	/r/	Post-alveolar fricative/voiced approximant liquid	108+110
9	/w/	Labio-velar approximant	110+110
10	/p/	Stop	112+112
11	/n/	Nasal	105+104
12	/s/	Fricative	110+110
13	/t/	Stop	112+112
14	/k/	Stop	108+110
15	/h/	Voiceless glottal fricative/Approximants	110+108
16	/g/	Stop	108+110
17	/a/	Vowel	100+100
18	/e/	Vowel	100+100
19	/o/	Vowel	100+100

*(C-19 p +n) denotes (COVID-19 Positive + Healthy).

2. Materials and methods

2.1. Dataset

The proposed non-invasive COVID-19 detection scheme is trained and tested in a combined speech dataset, which is prepared from speech samples collected from the Telephone band speech dataset (Ritwik et al., 2020) and Coswara dataset (Sharma N. et al., 2020). A total of 19 speakers' voice has been used in the Telephone band speech dataset, out of which 10 are COVID-19 positive and 9 are healthy. The original speech samples are recorded with 44.1 kHz sampling frequency. But it has been observed that most of the relevant speech components are present within the frequency range of 300 Hz to 3.4 kHz (Jax and Vary, 2004). In the next step, the filtered speech samples are segmented into different phoneme categories using the Audacity Toolkit[1]. There are a total of 432 speech samples in 16 phoneme categories and the details are mentioned in Table 1. From the Coswara dataset, three vowel sounds are taken and the samples are down sampled to 8 kHz sampling frequency. The speech samples are combined and labeled into 19 phoneme categories belonging to vowels, diphthongs, stops, fricatives, glides, liquids, approximants, and nasals. To deal with the insufficient speech samples, the existing speech phoneme samples are processed by an audio data augmentation scheme (Salamon and Bello, 2017). The details of the prepared dataset are listed in Table 1.

2.2. Proposed methodology

The proposed method is implemented in the following steps: dataset preparation, spectral smearing, extraction of cepstral features, and training and testing of the classification model. The proposed COVID-19 detection scheme is shown in Figure 1.

2.3. Smearing of phonemes

It has been observed that various speech components respond differently to spectral and temporal cues which can be helpful in speech recognition (Xu et al., 2005). The process of spectral smearing is obtained by multiplying the signal with a low-pass filter noise. The approach is known to replace the individual tone factor of the audio-spectrum with a noise band whose center-frequency collides with the particular tone. By this, the bandwidth of the modulated tone is increased twice the tone factor. It has been reported that the effect of smearing has enhanced phoneme detection accuracy (Boothroyd et al., 1996). In Golestani et al. (2009), the authors have conducted experiments on native-language detection to emphasize that certain words can be more

[1] http://www.audacityteam.org/

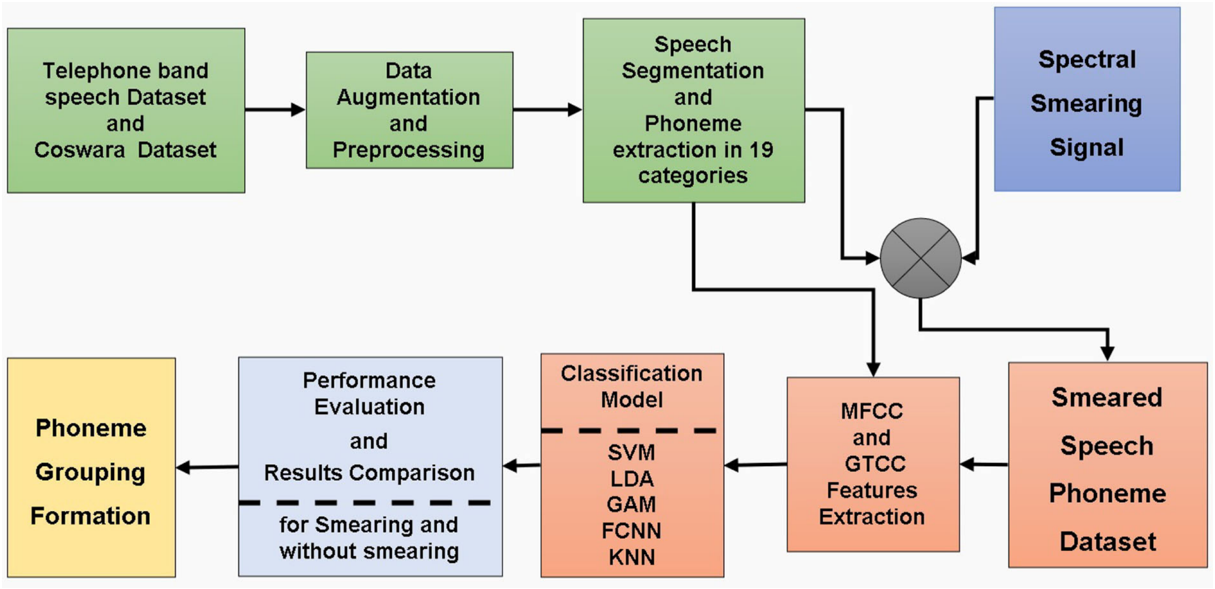

Telephone band speech Dataset and Coswara Dataset
Data Augmentation and Preprocessing
Speech Segmentation and Phoneme extraction in 19 categories
Spectral Smearing Signal
Smeared Speech Phoneme Dataset
MFCC and GTCC Features Extraction
Classification Model
SVM
LDA
GAM
FCNN
KNN
Performance Evaluation and Results Comparison
for Smearing and without smearing
Phoneme Grouping Formation

FIGURE 1
Block diagram of the proposed model.

conveniently detected at a particular noise configuration than others. This has been accounted for the differential-phoneme-recognition outcomes in a noisy environment. In this case, the speech signals have been smeared using varying SNR levels and it has been observed that this technique provides superior performance as compared to the phonemes without smearing. In yet another study (Shannon, 2005), speech detection has been shown to be possible with highly distorted and degraded audio signals. The spectral information can be modified by smearing to a considerable level till it starts degrading the classification outcome. A study by Goldsworthy (Goldsworthy et al., 2013) has demonstrated evaluating psycho-acoustic phoneme-based identification methods in normal hearing vs. cochlear-implant subjects. The presence of fluctuating noise-makers has shown better interpretation for normal hearing participants. By varying the range of low pass cut-off frequencies, vowel, and consonant recognition scores have shown marked differences illustrating the relativity of spectral resolution (Xu et al., 2005).

In the present study, an attempt has been made to apply spectral smearing to increase phoneme recognition without affecting signal perception by the addition of noise. In the first step, the smearing signal is generated by combining a sinusoidal signal with varied center frequencies and additive white Gaussian noise. This signal is passed through low-pass filters having cut-off frequencies ranging from 10 Hz to 10 kHz. The smearing signal is then multiplied by the phoneme signal to generate the smeared phoneme. The best values of these center frequencies and cut-off frequencies of low-pass filters are calculated based on the classification accuracies from the support vector machine-based classifier. The corresponding values are listed in Tables 2, 3.

2.4. Feature extraction

The objective of signifying an audio signal through its features is primarily to represent a huge data set through a compact form without compromising its vital information. The cepstral features are one of the effective features that are widely used in speech signal processing and mechanical engineering. These features are specially designed by considering the perceptual quality of the human hearing system (Dash et al., 2021a). The following steps are usually performed in cepstral feature extraction:

- Short-time Fourier transforms of windowed speech frames of the input signals.
- Calculation of the short-time energy of speech frame.
- Application of auditory filter bank on the power spectrum.
- Calculation of logarithm and Discrete cosine transform.
- Extraction of specific cepstral features based on the auditory filter bank used.

The third step is the crucial step that works on the conversion between the linear frequency scale and to perceptual frequency scale. Depending on the conversion, two cepstral features such as Mel and Gammatone cepstral Features are used in the proposed implementation scheme. The conversion scale

TABLE 2 Values of center frequency of sinusoidal signal and cut-off frequencies of low-pass filter for before and after tuning SVM.

Phonemes	Cf/LPBW pre-tuning	Accuracy pre-tuning	Cf/LPBW post tuning	Kernel function/ gamma/C	Accuracy post tuning
STOPS /b/	6.3/6.2	64.25	1.4/4.9	Quadratic /1/ 8.73	89.2
NASALS /m/	4.2/7.9	76.85	1.1/4.3	Gaussian /3.63/238.6	84.2
DIPHTONGS /Oy/	3.1/6	82.6	1.9/2	Gaussian /0.007 /635	84.3
GLIDES + /r/	8.6/8.5	67.7	2.1/4.8	Gaussian/0.001/6523.4	89.9
FRICATIVES /s/	5.6/4.4	67.7	9.9/2.9	Quadratic/1/0.1	85.3
Vowel a	9/7.1	63.6	8/6.6	linear/1/12.6013	79.4

*Cf and LPBW denote Cosine frequency Low Pass Bandwidth in kHz, GLIDES+ denotes the GLIDES, APPROXIMATES, and LIQUIDS.

of the Mel scale is mentioned in Equation (1)

$$f_{mel} = 2595 \times \log_{10}\left(1 + \frac{f_{lin}}{700}\right)$$
$$f_{lin} = 700 \times \left(10^{\left(\frac{f_{mel}}{2595}\right)} - 1\right) \qquad (1)$$

Where, $f_l$and $f_m$are the linear scale and mel scale frequencies, respectively.

2.4.1. Mel-scale cepstral features

Studies have shown that short-time speech-based Mel-Cepstral features have been noise evasive, and have significantly detected the pathologies on the vocal tract and vocal folds in past years. The MFCC feature considers human hearing by warping the frequency onto the Mel scale (Milner, 2002). It computes the cepstrum to separate the glottal source and vocal tract filtering information (Quatieri, 2002). The MFCCs have been chosen for this study because, in the presence of voice issues, these have the inherent ability to reflect either irregular movements of the vocal folds or a lack of closures produced by an increase in size or a variation in the attributes of the tissue covering the vocal folds. In this study, 13 feature-based MFCC coefficients, 13 MFCC Delta coefficients, and 13 MFCC Delta-Delta coefficients have been extracted. The delta values represent the first and second derivatives that depict the dynamics of variation in MFCC feature values.

2.4.2. Gammatone cepstral features

Gammatone Cepstral coefficients (GTCCs) are physiologically inspired adaptations that use Gammatone filters and have comparable rectangular bandwidth bands. Several papers (Cheng et al., 2005; Lee et al., 2014) have examined the benefits and use of the Gammatone function in the modeling of the human auditory filter response. The Gammatone filter impulse response is calculated by multiplying a Gamma distribution function by a pure sine wave tone. The delta and double delta GTCC variants (Cheng et al., 2005) are also taken into consideration. In essence, 13 feature-based GTCC coefficients, 13 GTCC Delta coefficients, and 13 GTCC Delta-Delta Coefficients.

TABLE 3 Best values of the center frequency of the sinusoidal signal and cut-off frequencies of low-pass filter for the smearing of different phonemes.

Phonemes	Center frequency (kHz)	Low-pass filter cut-off frequency (kHz)
/b/	1.4	4.9
/d/	3.2	1
/v/	8.2	1.6
/m/	1.1	4.3
/l/	4.3	9.6
/f/	6.3	2
/Oy/	1.9	2
/r/	2.1	4.8
/w/	3.1	2.9
/p/	3.7	9.9
/n/	4.4	2
/s/	9.9	2.9
/t/	8.1	4.7
/k/	8.4	9.6
/h/	8.2	4.2
/g/	0.3	6.8
Vowel /a/	8	6.6
Vowel /e/	9.1	4.5

2.5. Classification

Machine learning-based (ML) classifiers working along with time and frequency extracted features have made substantial progress in this field. Even in noisy conditions, this combination exhibited outstanding accuracies for discrete sound categorization (Dash et al., 2021b). To initiate

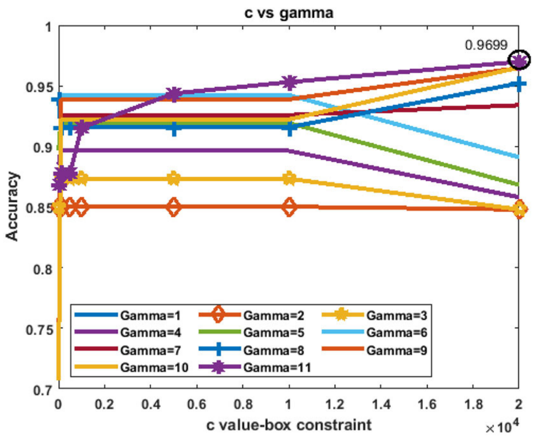


FIGURE 2
Selection of the values of C and gamma in the SVM classifier.

classification, all the above-mentioned 78 features were extracted from the speech signal and were provided as inputs to the following classifiers. The smeared phonemes were Short-Time Fourier transformed (STFT) using the hamming window of a length of 1,024, having a 30 ms analysis window with a 20 ms overlap. As the noise level varies during the time of recording of different speech samples, the speech enhancement algorithms are widely used to reduce the interfering noise. In the proposed implementation, one of the popular speech enhancement algorithms called the multi-band spectral subtraction method is used in the preprocessing stage before feature extraction (Kamath and Loizou, 2002).

2.5.1. Support vector machines

The primary objective of a support vector machine (SVM) classifier is to obtain the most feasible hyperplanes to assess a proposed model for classification (Soumaya et al., 2021). SVMs have been widely used in speech classification tasks and have shown superior performance (Dash and Solanki, 2019). In this study, bayesian optimization has been applied to select the best SVM parameters. The best values of c and gamma are taken from the comparative analyses between the values of c and gamma vs. classification accuracy as plotted in Figure 2 for the "rbf" kernel.

2.5.2. Linear discriminant analysis

Linear discriminant analysis (LDA) has been employed in multiple speech disease detection or health anomalies through audio analysis (Fredouille et al., 2009; Akbari and Arjmandi, 2014). Fisher's approach is commonly used in linear discriminant analysis. This approach is based on the sample averages and covariance matrices generated from the several groupings that comprise the training sample. Based on the training sample, a discriminant rule is developed and used to classify fresh occurrences into one of the categories. Fisher's linear discriminant analysis is a basic and widely used discriminating approach (Croux et al., 2008).

2.5.3. Generalized additive model

For analyzing the data set and picturing the affiliation of a dependent variable with an independent variable, the generalized additive model (GAM) is used, which evolves from a class of generalized linear models (GLM) (Liu, 2008). Previous studies have shown that the GLM classifier has given appreciable results in temporal feature integration based on music genre classification (Meng et al., 2007). In this case, the boosted tree is used as a shape function for each predictor to capture a nonlinear relation between a predictor and the response variable.

2.5.4. Feed-forward fully connected neural network

Neural network-based classifier models are widely used in speech processing for improved performance (Lopez-Moreno et al., 2016; Dash et al., 2020). In this case, feed-forward fully connected neural network (FCNN) is used with the input layer connected to a fully connected layer of 10 neurons, a ReLU function, followed by a second fully connected layer, a softmax function. A memory-limited device based loss function minimization approach used here is the Broyden-Flecter-Goldfarb-Shanno quasi-Newton algorithm (LBFGS) (Nocedal and Wright, 2006; Hui et al., 2019), where the cross-entropy loss is reduced during the training phase.

2.5.5. K-nearest neighbor

K-nearest neighbor (KNN) is one of the effective and popular classifiers that are used for speech-based applications (Alsmadi and Kahya, 2008). The categorization process is divided into two stages: the first is determining the closest neighbors, and the second is determining the class based on those neighbors. The K-nearest neighbors are selected using the Grid search method that provides the best value of k as 5.

2.6. Validation

K-fold cross-validation is a commonly applied validation approach (He et al., 2018). The entire set of voice samples is randomly divided into k equal-sized subgroups. Each fold has an equal proportion of two different types of class labels (glottal and normal stop speech). One of the subsamples is engaged for testing, while the remaining k-1 subsamples can be utilized for training (Altan, 2021, 2022). The process is replayed k times (the folds), for each of the k subsamples serving as testing data.

TABLE 4 Performance comparison of classifiers on different phoneme categories.

Smeared phoneme category	Model	Accuracy	AUC	Precision	Recall	F-2 Score
STOPS	SVM	0.9 ± 0.0045	0.87	0.92 ± 0.0012	0.9 ± 0.004	0.75 ± 0.002
/b/,/d/,	LDA	0.81±0.025	0.83	0.81 ± 0.007	0.86 ± 0.0063	0.70 ±0.0069
/g/,/k/,	GAM	0.9± 0.02	0.96	0.9 ± 0.0033	0.9 ± 0.0047	0.75 ± 0.0033
/t/,/p/	FCNN	0.85 ±0.016	0.89	0.87 ± 0.0022	0.86 ± 0.0033	0.72 ± 0.0022
	KNN	0.80± 0.023	0.79	0.86 ± 0.0067	0.77 ± 0.0031	0.65 ± 0.0071
FRICAT	SVM	0.92± 0.01	0.8	0.97 ± 0.0032	0.92 ± 0.0058	0.69 ± 0.0041
IVES	LDA	0.72 ±0.02	0.74	0.67 ± 0.0015	0.70 ± 0.0033	0.57 ± 0.0011
/f/,/s/,/v/	GAM	0.89± 0.2	0.94	0.89 ± 0.0073	0.92 ± 0.0064	0.76 ± 0.0022
	FCNN	0.64 ±0.04	0.59	0.70 ± 0.0017	0.68 ± 0.001	0.57 ± 0.0046
	KNN	0.82± 0.015	0.77	0.84 ± 0.0069	0.82 ± 0.0022	0.69 ± 0.004
NASALS	SVM	0.87 ±0.02	0.88	0.87 ± 0.0033	0.87 ± 0.0071	0.73 ± 0.006
/m/,/n/	LDA	0.67 ± 0.06	0.68	0.70 ± 0.004	0.63 ± 0.0022	0.53 ± 0.0011
	GAM	0.94 ± 0.01	0.98	0.95 ± 0.001	0.93 ± 0.0023	0.77 ± 0.0066
	FCNN	0.87± 0.01	0.91	0.77 ± 0.0014	0.89 ± 0.008	0.72 ± 0.004
	KNN	0.77 ±0.02	0.76	0.78 ± 0.0012	0.76 ± 0.0011	0.63 ± 0.0032
VOWELS	SVM	0.78 ± 0.0012	0.77	0.78 ± 0.0046	0.78 ± 0.0010	0.70 ± 0.0067
/a/,/e/, /o/	LDA	0.63 ± 0.0071	0.68	0.59 ± 0.0012	0.73 ± 0.0012	0.58 ± 0.0033
	GAM	0.84 ± 0.0045	0.91	0.79 ± 0.0033	0.85 ± 0.004	0.69 ± 0.0012
	FCNN	0.85 ± 0.0023	0.90	0.89 ± 0.0047	0.83 ± 0.0033	0.69 ± 0.001
	KNN	0.64 ± 0.0017	0.64	0.55 ± 0.0014	0.68 ± 0.004	0.54 ± 0.0064
GLIDES+,	SVM	0.81 ± 0.006	0.81	0.81 ± 0.0035	0.81 ± 0.0010	0.73 ± 0.0022
/l/ /w/	LDA	0.80 ± 0.0011	0.74	0.75 ± 0.0044	0.86 ± 0.006	0.7 ± 0.004
/r/ /h/	GAM	0.96 ± 0.0014	0.98	0.95 ± 0.008	0.95 ± 0.0010	0.8 ± 0.0041
	FCNN	0.57 ± 0.0079	0.66	0.55 ± 0.0011	0.57 ± 0.001	0.5 ± 0.007
	KNN	0.82 ± 0.0015	0.83	0.85 ± 0.0066	0.79 ± 0.002	0.67 ± 0.006
DIPTHO	SVM	0.79 ± 0.0028	0.76	0.78 ± 0.0044	0.78 ± 0.0035	0.75 ± 0.008
NGS	LDA	0.63 ± 0.0067	0.54	0.68 ± 0.0033	0.54 ± 0.0022	0.47 ± 0.006
/Oy/	GAM	0.87 ± 0.0011	0.93	0.85 ± 0.0014	0.85 ± 0.001	0.71 ± 0.0022
	FCNN	0.88 ± 0.0036	0.80	0.88 ± 0.0026	0.86 ± 0.006	0.72 ± 0.001
	KNN	0.67 ± 0.0044	0.67	0.73 ± 0.007	0.57 ± 0.007	0.5 ± 0.0041

The classification accuracy is calculated for each operation. The mean classification accuracies are calculated using 10 times in 10-fold cross-validation (Muthusamy et al., 2015) for this study. The validation accuracy is computed from confusion metrics as shown below

$$Classification\ Accuracy\ = \left(\frac{T_P + T_N}{T_P + T_N + F_P + F_N}\right) \quad (2)$$

where T_P stands for True-Positives, T_N stands for True-Negatives, F_P for False-Positives, and F_N for False-Negatives. The Precision and Recall are calculated as mentioned below.

$$Precision = \frac{T_P}{T_P + F_P}$$
$$Recall = \frac{T_P}{T_P + F_N} \quad (3)$$

The F-2 score is calculated as

$$F2 - Measure = \frac{(5\ \times Precision\ \times\ Recall)}{(4\ \times Precision\ +\ Recall)} \quad (4)$$
$$= \frac{T_P}{T_P + 0.2F_P + 0.8F_N}$$

The F-2 score is one of the important parameters in medical diagnosis since it indicates the cases who are False Negative (who have COVID-19 infection but have been incorrectly classified as healthy by the model).

3. Results and discussions

After completing the experimental setup, the simulations study has been performed on the MATLAB platform using a Core i5, 12GB RAM processor. The results are analyzed in three

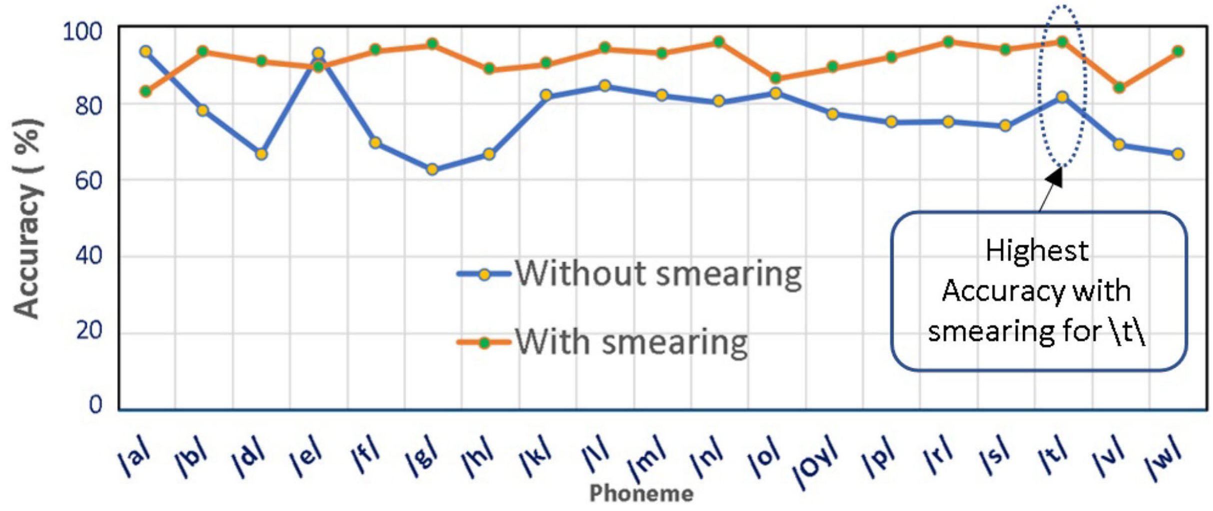


FIGURE 3
Performance comparison of phonemes in non-smearing and smearing conditions.

TABLE 5 Comparison of GAM Classification performance for best phoneme categories and groupings.

Phoneme grouping	Classification accuracy	Precision	Recall	F-2 Score
/t/	0.95 ± 0.01	0.95 ± 0.0021	0.93 ± 0.030	0.94 ± 0.014
/r/	0.94 ± 0.01	0.94 ± 0.001	0.94 ± 0.021	0.94 ± 0.017
/n/	0.94 ± 0.012	0.94 ± 0.0024	0.94 ± 0.003	0.94 ± 0.01
/g/	0.93 ± 0.012	0.94 ± 0.013	0.94 ± 0.001	0.93 ± 0.008
/l/	0.92 ± 0.016	0.93 ± 0.015	0.93 ± 0.0012	0.93 ± 0.006
/t//r//n/	0.96 ± 0.0011	0.97 ± 0.001	0.96 ± 0.001	0.96 ± 0.004
/t//r//n//g//l/	0.97 ± 0.0005	0.97 ± 0.001	0.97 ± 0.001	0.97 ± 0.0013

broad categories: the selection of the best classification model, the effect of smearing, and the formation of the grouping of phonemes.

3.1. Performance comparison of different classifiers on smeared phoneme detection

For the selection of the best performing classifier for COVID-19 detection using phoneme and smearing, the performance of the five different classifiers (SVM, LDA, GAM, FCNN, and k-NN) are compared. For this, the classification accuracy, area under the curve (AUC), precision, recall, and F-2 score are used and the results are plotted in Table 4. The average classification performances are listed for six broad categories of phonemes including stops, fricatives, nasals, vowels, voiced, and dipthongs.

In terms of classification accuracies, /t/, /a/, /f/, /k/, /l/, /m/, /n/, /o/, and /r/ have obtained the best results under GAM Classifier. Similarly, /b/, /e/, /g/, and /oy/ have achieved their highest classification accuracies under FCNN Classifier. LDA Classifier outperformed the rest for /p/, and /v/. SVM offered the highest classification accuracies for both /w/, and /s/. Finally, KNN achieved the best performance in the case of /h/ phoneme. Conclusively, GAM delivers an overall best performance for all phonemes as compared to other classifiers.

3.2. Comparison of classification accuracy between non-smeared and smeared phonemes

To detect the effect of smearing on the classification performance, a comparative analysis is carried out between the phonemes with and without smearing. For the classification of the best performing model from the classification analysis, GAM is used. The same 78-dimensional feature vector sets have been extracted from corresponding phoneme samples. The simulation results are shown in Figure 3.

It is evident from the above figure that the smearing of phonemes yields appreciably better classification accuracies in the majority of the cases. The phoneme /t/ exhibits the highest classification accuracy of 95.92%, and phoneme/a/ exhibits the lowest accuracy of 83.08% under the smeared conditions.

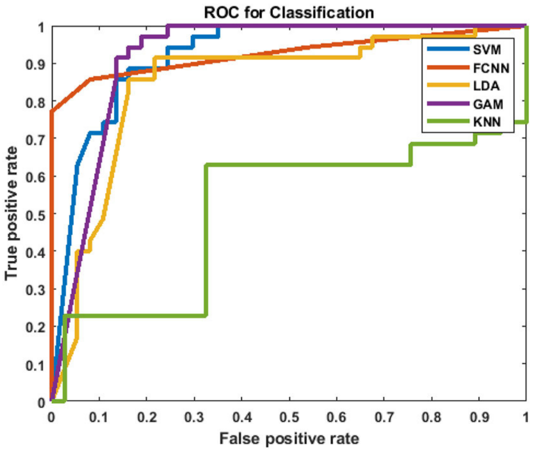


FIGURE 4
ROC-AUC plot for the phoneme group /t//r//n//g//l/.

3.3. Phoneme groupings

After analyzing the classification performance of smearing and individual phonemes, a phoneme grouping based approach is adopted. Based on the individual classification accuracy of phonemes, the 3-tuple and 5-tuple phoneme buzzwords are created by combining the high-performing individual phonemes (Moro-Velazquez et al., 2019). By taking the first reference level of 95.67% classification accuracy, the first phoneme group of "/t/-/r/-/n/" is used as a 3-tuple buzzword. Then, the threshold is set at 94.07% classification accuracy to form the second phoneme group of "/t/-/r/-/n/-/g/-/l/." The best performing five phonemes are then combined. In these combinations, the phoneme classification accuracies are taken in descending order where the /t/ is having the highest classification accuracy and /l/ is having the lowest classification accuracy among the group. Audacity software is used to combine the individual phonemes to form 104 speech samples in both the categories of COVID-19 positive and healthy for the phoneme group of "/t/-/r/-/n/" and "/t/-/r/-/n/-/g/-/l/." The same 78-dimensional feature vectors are extracted and applied to the GAM classifier and the results are listed in Table 5. The ROC-AUC curve is plotted for the phoneme group /t//r//n//g//l/ in Figure 4 and the comparison between spectrogram of COVID-19 positive sample and healthy sample is plotted in Figure 5.

It is observed that the phoneme group with the buzzword "/t//r//n//g//l/" performs better as compared to /t//r//n/. The spectrograms of the buzzword "/t//r//n//g//l/" are plotted for

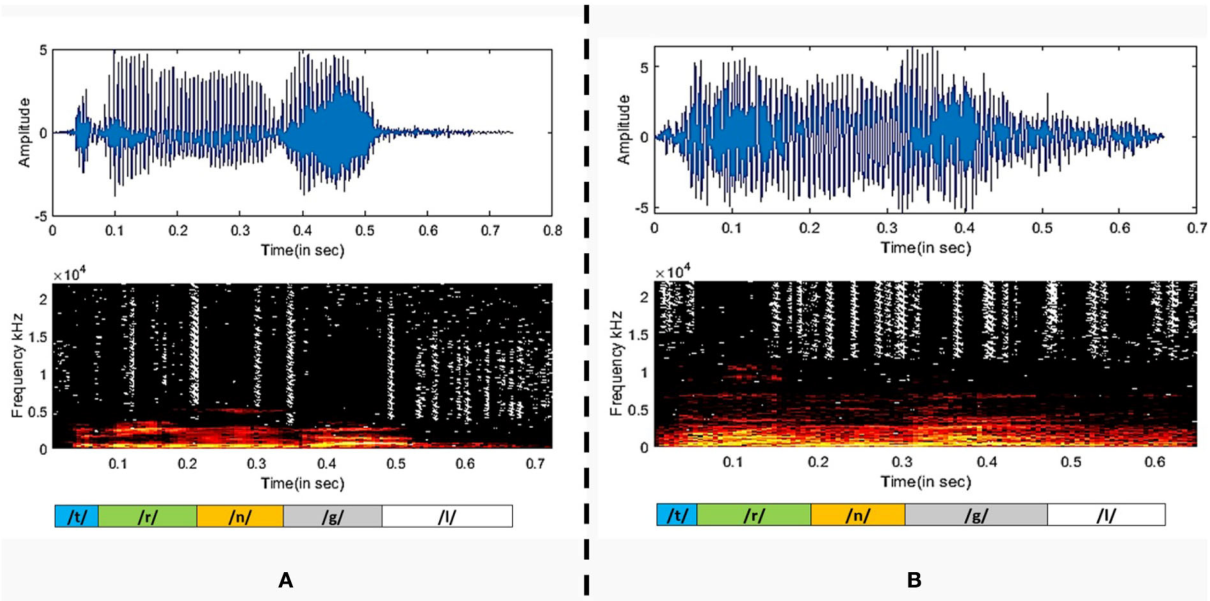


FIGURE 5
The plot of the Spectrogram of **(A)** COVID-19 positive sample and **(B)** a healthy sample.

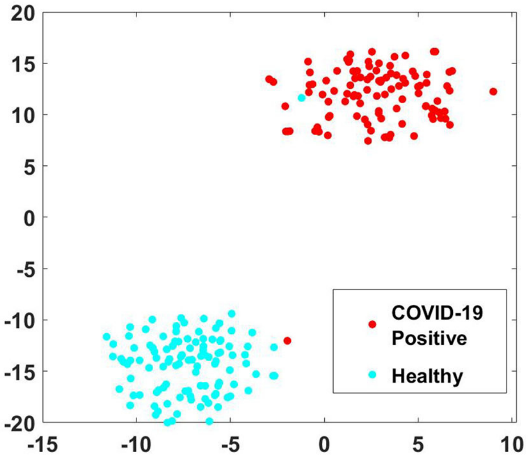


FIGURE 6
t-SNE plots of phoneme group buzzword "/t//r//n//g//l/" for MFCC and GTCC features.

COVID-19 positive and healthy speech samples are plotted in (Narlı, 2021).

A person affected by COVID-19 may lack in energy to produce sound, thus disrupting the normal speech production phenomena. In the stage of sound phonation, the sub-glottal thrust must cross a certain threshold to set the vocal folds in vibration. If the respiration stage of speech production is interrupted, the phonation of the larynx will be accordingly compromised (Asiaee et al., 2020). Therefore, the audio waveform of the plosive /t/ in healthy candidate exhibits strong energy compaction due to sufficient sub glottal pressure as compared to the diseased case. The healthy vocal folds exhibit glottal closures with a trail of strong impulses due to the quick closure of vocal folds, whereas a disordered vocal fold produces a weak impulse due to the incomplete closure of vocal folds (Mandal and Rao, 2018). The ability to increase or decrease vocal cord length and tension governs the frequency at which the cord vibrates and, consequently, the pitch of the sound produced. As the mass of the vocal cords increases, the vibrating frequency and pitch decrease (Dettelbach et al., 1994). In the above spectrograms, the healthy waveform depicts equivalent variation for all phonemes, whereas, in the case of COVID-19 affected sample, certain phonemes are subdued as compared to others. To further evaluate the effectiveness of the extracted MFCC and GTCC features for phoneme group buzzword "/t//r//n//g//l/," the t-SNE plot is shown in Figure 6 (der Maaten and Hinton, 2008). It is observed that in the input space, the pattern of the extracted features is linearly separable which improves the performance of the classification especially the phoneme group buzzword "/t//r//n//g//l/."

This approach to phoneme grouping has the advantage of designing a low computational complexity based COVID-19 detection model as the individual phonemes are not recorded and the group has a higher classification accuracy as compared to individual phonemes.

4. Conclusion

In this study, a hybrid model is designed for the detection of COVID-19 from speech signals by combining phoneme-based signal analysis and spectral smearing. The performance of the detection model is evaluated for 19 individual phonemes and two phoneme groupings using five ML-based classifiers. It is observed that the GAM model performs appreciably better for most pathological phoneme detection. These methods are expected to perform well among suspected COVID-19 patients with minimal or no cough and shortness of breath. Due to insufficient audio samples present in the corpus and to avoid the issues of imbalanced data, the final dataset has been created with the help of data augmentation prior to further processing. In the future, a phone or a web application may be developed for detection based on this buzzword. This proposed methodology needs to be clinically validated in hospitals with large speech datasets.

Data availability statement

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

Author contributions

SM formulated the problem statement and simulated the experiment. TD contributed in drafting the manuscript. GP revised and modified the manuscript. All authors contributed to the article and approved the submitted version.

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

Afshar, P., Heidarian, S., Enshaei, N., Naderkhani, F., Rafiee, M. J., Oikonomou, A., et al. (2021). COVID-CT-MD, COVID-19 computed tomography scan dataset applicable in machine learning and deep learning. *Sci. Data* 8, 1–8. doi: 10.1038/s41597-021-00900-3

Akbari, A., and Arjmandi, M. K. (2014). An efficient voice pathology classification scheme based on applying multi-layer linear discriminant analysis to wavelet packet-based features. *Biomed. Signal Process. Control* 10, 209–232. doi: 10.1016/j.bspc.2013.11.002

Alsmadi, S., and Kahya, Y. P. (2008). Design of a DSP-based instrument for real-time classification of pulmonary sounds. *Comput. Biol. Med.* 38, 53–61. doi: 10.1016/j.compbiomed.2007.07.001

Altan, G. (2021). SecureDeepNet-IoT: a deep learning application for invasion detection in industrial Internet of things sensing systems. *Trans. Emerg. Telecommun. Technol.* 32, e4228. doi: 10.1002/ett.4228

Altan, G. (2022). DeepOCT: An explainable deep learning architecture to analyze macular edema on OCT images. *Eng. Sci. Technol. Int. J.* 34, 101091. doi: 10.1016/j.jestch.2021.101091

Asiaee, M., Vahedian-Azimi, A., Atashi, S. S., Keramatfar, A., and Nourbakhsh, M. (2020). Voice quality evaluation in patients with COVID-19: An acoustic analysis. *J. Voice*. 36, 879.e13–879.e19. doi: 10.1016/j.jvoice.2020.09.024

Boothroyd, A., Mulhearn, B., Gong, J., and Ostroff, J. (1996). Effects of spectral smearing on phoneme and word recognition. *J. Acoust. Soc. Am.* 100, 1807–1818. doi: 10.1121/1.416000

Cheng, O., Abdulla, W., and Salcic, Z. (2005). "Performance evaluation of front-end algorithms for robust speech recognition," in *Proceedings of the Eighth International Symposium on Signal Processing and Its Applications, 2005, Vol. 2* (IEEE), 711–714.

Croux, C., Filzmoser, P., and Joossens, K. (2008). Classification efficiencies for robust linear discriminant analysis. *Statist. Sin.* 18, 581–599. doi: 10.2139/ssrn.1024151

Dash, T. K., Mishra, S., Panda, G., and Satapathy, S. C. (2021a). Detection of COVID-19 from speech signal using bio-inspired based cepstral features. *Pattern Recognit*. 117, 107999. doi: 10.1016/j.patcog.2021.107999

Dash, T. K., and Solanki, S. S. (2019). Investigation on the effect of the input features in the noise level classification of noisy speech. *J. Sci. Ind. Res.* 78, 868–872. Available online at: http://nopr.niscpr.res.in/handle/123456789/52213

Dash, T. K., Solanki, S. S., and Panda, G. (2020). Improved phase aware speech enhancement using bio-inspired and ANN techniques. *Analog Integr. Circ. Signal Process*. 102, 465–477. doi: 10.1007/s10470-019-01566-z

Dash, T. K., Solanki, S. S., and Panda, G. (2021b). Multi-objective approach to speech enhancement using tunable Q-factor-based wavelet transform and ANN techniques. *Circ. Syst. Signal Process.* 40, 6067–6097. doi: 10.1007/s00034-021-01753-2

der Maaten, L. V., and Hinton, G. (2008). Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, 2579–2605.

Dettelbach, M., Eibling, D. E., and Johnson, J. T. (1994). Hoarseness: from viral laryngitis to glottic cancer. *Postgrad. Med.* 95, 143–162. doi: 10.1080/00325481.1994.11945836

Fredouille, C., Pouchoulin, G., Ghio, A., Revis, J., Bonastre, J. F., and, A., et al. (2009). Back-and-forth methodology for objective voice quality assessment: from/to expert knowledge to/from automatic classification of dysphonia. *EURASIP J. Adv. Signal Process*. 2009, 982102. doi: 10.1155/2009/982102

Gallo, O. (2021). The central role of the nasal microenvironment in the transmission, modulation, and clinical progression of SARS-CoV-2 infection. *Mucosal Immunol.* 14, 305–316. doi: 10.1038/s41385-020-00359-2

Goldsworthy, R. L., Delhorne, L. A., Braida, L. D., and Reed, C. M. (2013). Psychoacoustic and phoneme identification measures in cochlear-implant and normal-hearing listeners. *Trends Amplif.* 17, 27–44. doi: 10.1177/1084713813477244

Golestani, N., Rosen, S., and Scott, S. K. (2009). Native-language benefit for understanding speech-in-noise: The contribution of semantics. *Biling. Lang. Cogn.* 12, 385–392. doi: 10.1017/s1366728909990150

Han, J., Brown, C., Chauhan, J., Grammenos, A., Hasthanasombat, A., Spathis, D., et al. (2021). "Exploring automatic COVID-19 diagnosis via voice and symptoms from crowdsourced data," in *InICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE), 8328–8332.

He, L., Zhang, J., Liu, Q., Zhang, J., Yin, H., and Lech, M. (2018). Automatic detection of glottal stop in cleft palate speech. *Biomed. Signal Process. Control*. 39, 230–236. doi: 10.1016/j.bspc.2017.07.027

Hui, Y. U., Juzhi, C., Hui, C., Xiao, W., Xianxiang, Zhiyong, Z., et al. (2019). Three-dimensional magnetotelluric inversion under topographic relief based on the limited-memory quasi-Newton algorithm (L-BFGS). *Chin. J. Geophys*. 62, 3175–3188. doi: 10.1016/j.ijid.2020.01.009

Jax, P., and Vary, P. (2004). "Feature selection for improved bandwidth extension of speech signals," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 1* (Montreal, QC: IEEE), 697.

Kamath, S., and Loizou, P. (2002). "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proceedings of International Conference on Acoustics, Speech, and Signal Processing, Vol. 4* (Orlando, FL), 44164–44164.

Katamba, F. (1989). *An Introduction to Phonology, Vol. 48*. London: Longman.

Kiamanesh, O., Harper, L., Wiskar, K., Luksun, W., McDonald, M., Ross, H., et al. (2020). Lung ultrasound for cardiologists in the time of COVID-19. *Can. J. Cardiol.* 36, 1144–1147. doi: 10.1016/j.cjca.2020.05.008

Kumar, L. K., and Alphonse, P. J. (2021). Automatic diagnosis of COVID-19 disease using deep convolutional neural network with multi-feature channel from respiratory sound data: cough, voice, and breath. *Alexandria Eng. J.* 61, 1319–1334. doi: 10.1007/s00392-020-01730-w

Lamba, R., Gulati, T., Alharbi, H. F., and Jain, A. (2021). A hybrid system for Parkinson's disease diagnosis using machine learning techniques. *Int. J. Speech Technol.* 25, 583–593. doi: 10.1007/s10772-021-09837-9

Lee, S. J., Kang, B. O., Chung, H., and Lee, Y. (2014). Intra-and inter-frame features for automatic speech recognition. *ETRI J.* 36, 514–521. doi: 10.4218/etrij.14.0213.0181

Liu, H. (2008). *Generalized Additive Model, Vol. 55812*. Duluth, MN.

Lopez-Moreno, I., Gonzalez-Dominguez, J., Martinez, D., Plchot, O., Gonzalez-Rodriguez, J., and Moreno, P. J. (2016). On the use of deep feedforward neural networks for automatic language identification. *Comput. Speech Lang.* 40, 46–59. doi: 10.1016/j.csl.2016.03.001

Mandal, T., and Rao, K. S. (2018). Glottal closure instants detection from pathological acoustic speech signal using deep learning. *arXiv preprint arXiv, 1811.09956*. doi: 10.48550/arXiv.1811.09956

Meng, A., Ahrendt, P., Larsen, J., and Hansen, L. K. (2007). Temporal feature integration for music genre classification. *IEEE Trans. Audio Speech Lang. Process*. 15, 1654–1664. doi: 10.1109/TASL.2007.899293

Milner, B. (2002). "A comparison of front-end configurations for robust speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing* (Orlando, FL: IEEE).

Moro-Velazquez, L., Gomez-Garcia, J. A., Godino-Llorente, J. I., Grandas-Perez, F., Shattuck-Hufnagel, S., Yagüe-Jimenez, V., et al. (2019). Phonetic relevance and phonemic grouping of speech in the automatic detection of parkinson's disease. *Sci. Rep*. 9, 1–16. doi: 10.1038/s41598-019-55271-y

Muthusamy, H., Polat, K., and Yaacob, S. (2015). Improved emotion recognition using gaussian mixture model and extreme learning machine in speech and glottal signals. *Math. Problems Eng.* 2015, 394083. doi: 10.1155/2015/394083

Narlı, S. S. (2021). Impact of local histogram equalization on deep learning architectures for diagnosis of COVID-19 on chest X-rays. *Manchester J. Artif. Intell. Appl. Sci.* 2.

Narli, S. S., and Altan, G. (2022) "CLAHE based enhancement to transfer learning in COVID-19 detection," in *Gazi M¨u* hendislik Bilimleri Dergisi, 1–11.

Nocedal, J., and Wright, S. J. (2006). *Numerical Optimization*. New York, NY: Springer.

Pancaldi, F., Pezzuto, G. S., Cassone, G., Morelli, M., Manfredi, A., D'Arienzo, M., et al. (2022). VECTOR: An algorithm for the detection of COVID-19 pneumonia from velcro-like lung sounds. *Comput. Biol. Med.* 142, 105220. doi: 10.1016/j.compbiomed.2022.105220

Peng, M. (2020). Outbreak of COVID-19: an emerging global pandemic threat. *Biomed. Pharmacother*. 129, 110499–110499. doi: 10.1016/j.biopha.2020.110499

Quatieri, T. F. (2002). *Discrete-Time Speech Signal Processing: Principles and Practice*. Prentice Hall, NJ.

Ritwik, K. V. S., Kalluri, S. B., and Vijayasenan, D. (2020). COVID-19 patient detection from telephone quality speech data. *arXiv preprint arXiv, 2011.04299*. doi: 10.48550/arXiv.2011.04299

Salamon, J., and Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Process. Lett*. 24, 2657381. doi: 10.1109/LSP.2017.2657381

Shannon, R. V. (2005). Speech and music have different requirements for spectral resolution. *Int. Rev. Neurobiol*. 70, 121–155. doi: 10.1016/S0074-7742(05)70004-0

Sharma, G., Umapathy, K., and Krishnan, S. (2020). Trends in audio signal feature extraction methods. *Appl. Acoust*. 158, 107020–107020. doi: 10.1016/j.apacoust.2019.107020

Sharma, N., Krishnan, P., Kumar, R., Ramoji, S., Chetupalli, S. R., Nirmala, R., et al. (2020). Coswara-A database of breathing, cough, and voice sounds for COVID-19 diagnosis. *Proc. Ann. Conf. Int. Speech Commun. Assoc. Interspeech* 2020, 4811–4815. doi: 10.21437/Interspeech.2020-2768

Soumaya, Z., Taoufiq, B. D., Benayad, N., Yunus, K., and Abdelkrim, A. (2021). The detection of Parkinson disease using the genetic algorithm and SVM classifier. *Appl. Acoust*. 171, 107528–107528. doi: 10.1016/j.apacoust.2020.107528

Wielgat, R. (2008). Automatic recognition of pathological phoneme production. *Folia Phoniatr. Logopaedica* 60, 323–331. doi: 10.1159/000170083

Xu, L., Thompson, C. S., and Pfingst, B. E. (2005). Relative contributions of spectral and temporal cues for phoneme recognition. *J. Acoust. Soc. Am*. 117, 3255–3267. doi: 10.1121/1.1886405

Zhang, D., and Wu, K. (2020). *Pathological Voice Analysis*. Singapore: Springer Nature Singapore Pte Ltd.

Frontiers in
Big Data

Explores the potential for big data to address global challenges

This innovative journal focuses on the power of big data - its role in machine learning, AI, and data mining, and its practical application from cybersecurity to climate science and public health.

Discover the latest Research Topics

See more →

frontiers

Frontiers in
Big Data

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact