IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 March 2024, accepted 21 May 2024, date of publication 29 May 2024, date of current version 14 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3406748

SURVEY

Chronic Diseases Prediction Using Machine Learning With Data Preprocessing Handling: A Critical Review

NUR GHANIAVIYANTO RAMADHAN[1], ADIWIJAYA[2], (Member, IEEE), WARIH MAHARANI[1], AND ALFIAN AKBAR GOZALI[3]

[1]Department of Data Science, School of Computing, Telkom University, Bandung 40257, Indonesia
[2]Department of Informatics, School of Computing, Telkom University, Bandung 40257, Indonesia
[3]Department of Application Software Engineering, School of Applied Science, Telkom University, Bandung 40257, Indonesia

Corresponding author: Adiwijaya (adiwijaya@telkomuniversity.ac.id)

This work was supported in part by the Directorate of Research, Technology, and Community Service; in part by the Directorate General of Higher Education, Research, and Technology; in part by the Ministry of Education, Culture, Research, and Technology, Republic of Indonesia under Grant 331/PNLT2/PPM/2023; and in part by Telkom University.

ABSTRACT According to the World Health Organization (WHO), some chronic diseases such as diabetes mellitus, stroke, cancer, cardiac vascular, kidney failure, and hypertension are essential for early prevention. One of the prevention that can be taken is to predict chronic diseases using machine learning based on personal medical record or general checkup result. The common prediction objective is to minimize the prediction error as low as possible. The most influencing chronic diseases prediction factors are the quality of data and the choice of predictor such as machine learning methods. The five main problems those lower data quality are outliers, missing values, feature selection, normalization, and imbalance. After we ensure the quality of data, the next task is to choose the best machine learning methods. The most influencing factor to consider when we choose the predictor its performance evaluation (accuracy, recall, precision, f1-score). Thus, predicting chronic disease aims to produce increased performance and solve problems in medical data. This paper presents a Systematic Literature Review (SLR) that offers a comprehensive discussion of research on chronic diseases prediction using machine learning and its data preprocessing handling. This paper covers machine learning methods discussion such as supervised learning, ensemble learning, deep learning, and reinforcement learning. The preprocessing handling we discuss includes missing values, outliers, feature selection, normalization, and imbalance. The final discussions of this paper are open issues, and the potential future works in improving the prediction performance for chronic diseases using a data preprocessing handling and machine learning methods.

INDEX TERMS Chronic disease prediction, machine learning, preprocessing data, systematic literature review (SLR).

I. INTRODUCTION

Developments in today's world, a lot of data is collected every day and analyzed for managing businesses [1]. Previously, this paper looked at data using traditional methods like Microsoft Excel. Analyzing data this way takes time and can be frustrating. Plus, it only works well in certain situations, like in healthcare. Figuring out what we can learn from datasets is a big challenge. The goal of knowledge discovery is to find the useful parts of the data. Data mining is one part of knowledge discovery that helps get useful information. It involves finding and pulling out hidden info, patterns, and connections in specific datasets [2].

Today, the healthcare industry gathers a lot of complex data about patients, hospital resources, disease diagnoses, electronic patient records, and medical devices. Having a

The associate editor coordinating the review of this manuscript and approving it for publication was Muammar Muhammad Kabir.

large amount of data is crucial for data mining. Healthcare data mining has great potential, and some of its most important applications include predicting and diagnosing diseases, assessing treatment effectiveness, managing healthcare, and improving the medical device industry [3]. Errors in choosing treatments for patients not only waste time and money but can also lead to serious consequences like patient deaths. That's why accurately diagnosing and selecting the right treatment is extremely important for patients. Data mining can assist in predicting and identifying different diseases in healthy populations.

In the current prediction process, there's something called a machine learning-based approach. Machine learning can be used for data mining in the healthcare sector [4]. Applying machine learning in health data can help predict if a patient might have six chronic diseases: diabetes mellitus [5], [6]; cancer [7], [8]; stroke [9], [10]; hypertension [11], [12]; kidney failure [13], [14]; and heart issues [15], [16].

Machine learning methods used to predict chronic diseases include ensemble tree-based techniques like random forest and CatBoost, fuzzy-based methods such as fuzzy Sugeno and fuzzy Mamdani, and deep learning-based approaches like neural networks and multilayer perceptrons. As mentioned earlier, predictions in the healthcare sector can rely on datasets derived from various sources, including medical examinations conducted by patients themselves, laboratory results, consultations with doctors, and findings from general medical studies or checkups. However, current research on disease prediction using medical data faces several significant challenges, including missing values [17], the impact of features [18], and data imbalance [19].

This paper examines the capability of machine learning in predicting diseases and managing issues with medical data. However, it highlights the absence of a comprehensive survey paper that thoroughly covers chronic diseases and tackles data-related challenges to ensure accurate predictions. Conducting a survey could help identify the most effective machine learning methods and techniques for handling data, ultimately improving prediction accuracy. Filling this research gap could result in significant scientific advancements in disease prediction using machine learning.

A. CURRENT TREND OF CHRONIC DISEASE PREDICTION

This section explores the most recent trends in predicting chronic diseases. Furthermore, the paper addresses the challenges related to managing problematic data. It's crucial to acknowledge that incomplete or inadequate data can impact the accuracy of prediction results, as discussed earlier.

According to Ramadhan and Romadhony [20] in 2021, The predictions derived from laboratory data for diabetes showed an enhancement in precision and recall results by 20-24%. The study pinpointed three factors influencing diabetes prediction accuracy: (i) the quantity of missing values, (ii) the utilization of influential features, and (iii) an imbalance in the number of positive and negative cases. The prediction method employed in the study was an ensemble method called Random Forest.

In 2019, Fitriyani et al. [21] the research focused on predicting hypertension using four distinct types of data. The study outlined three data-related challenges: (i) missing values, (ii) outliers, and (iii) an imbalance in the distribution of data. The prediction methods utilized were based on ensemble learning techniques, including Multi-Layer Perceptron, Support Vector Machine (SVM), Decision Tree, and Logistic Regression.

Muthulakshmi and Parveen [18] in their research, heart disease was predicted using public data from the UCI repository. Muthulakshmi highlighted three factors that impacted the prediction results: (i) missing values, (ii) noise, and (iii) feature selection. However, the study did not specify the predictive model that was utilized.

Kumawat et al. [8] the paper discussed the diagnosis and prognosis of cervical cancer. It faced challenges in handling data, especially in selecting disease-related features that had a significant impact. The study utilized data from the UCI public repository, and the prediction algorithms employed included SVM, Random Tree, Logistic Tree, and Xtreme Gradient Boosting.

Dash et al. [10] in 2022 a discussion revolved around the practical identification and prediction of early stages of stroke. The strategy to tackle the data challenge involved balancing the number of positive and negative classes, which were initially imbalanced. The dataset for the study was obtained from the UCI repository. The prediction method employed was based on ensemble learning, specifically utilizing CatBoost.

Revathy et al. [13] a study was conducted to predict chronic kidney failure, comparing different prediction algorithms to identify the most effective one. The data underwent two processing approaches: (i) data transformation and (ii) imputation of missing values. The dataset used in the study was obtained from the UCI repository. Prediction models utilized in the study included Decision Tree, Random Forests, and Support Vector Machine (SVM).

B. RELATED SURVEY PAPERS

In this section, we will discuss related survey papers on chronic disease prediction. Several survey papers have concentrated on disease prediction, but there's a lack of thorough discussion specifically focusing on chronic diseases, along with concerns regarding the utilized data. A comparison of relevant survey papers can be found in Table 1.

Wadghiri et al. [22] in their work, the authors engaged in a comprehensive discussion covering the following aspects: (i) Analyzing publication trends. (ii) Compiling a list of datasets utilized by researchers. (iii) Exploring individual machine learning techniques. (iv) Summarizing the collective efficacy of ensemble methods and contrasting their accuracy. (v) Identifying existing gaps and proposing suggestions for future research contributions regarding diabetes mellitus.

As a result, the review aimed to compare the predictive performance of ensemble methods with alternative approaches in diabetes prediction.

Abrar et al. [23] conducted comprehensive reviews that encompassed the following domains: (i) exploration of dimensionality reduction techniques commonly applied to manage gene data sets characterized by high dimensions, (ii) examination of feature selection techniques often employed to address high-dimensional gene data, (iii) analysis of frequently used data sets for managing high-dimensional gene data, and (iv) investigation into the models utilized for the identification of crucial gene sequences relevant to cancer disease classification, utilizing high-dimensional DNA gene data. In conclusion, the review critically assesses frequently utilized hybrid algorithms and concisely synthesizes advancements and emerging trends within cancer classification and prediction, all founded on high-dimensional gene data and employing machine learning methodologies.

The review conducted by De Jong et al. [24] encompassed the following aspects: (i) examination of various models for predicting stroke risk, (ii) comparative analysis of different stroke risk prediction models, (iii) evaluation of potential biases in 15 predictive models related to risk assessment, (iv) assessment of the performance of predictive models, and (v) identification of models that are not suitable for use in medically fragile patients. As a result, the main objective of the review paper was to externally assess and evaluate the predictive capabilities of ischemic stroke models.

Silva et al. [25] conducted reviews with the following focuses: (i) selection of the most recent article about hypertension prediction, (ii) comparative assessment of parameters within the selected article, including feature selection, train-test data division, data balancing, result obtained, and performance metrics, and (iii) identification of algorithms that are commonly recognized as exhibiting superior performance, notably Support Vector Machine (SVM), Xtreme Gradient Boosting, and Random Forest. The fundamental aim of this review was to present a comprehensive overview of the literature regarding the application of machine learning algorithms in predicting hypertension.

Sanmarchi et al. [26] conducted a review that encompassed the following aspects: (i) exploration of machine learning methods used for diagnosing chronic kidney failure, (ii) investigation into machine learning methods employed for predicting or prognosing chronic kidney failure, (iii) examination of treatment approaches facilitated by machine learning, and (iv) identification of the frequently utilized algorithm with the best performance, often being Neural Network. The central objective of this review was to identify the machine learning algorithm that exhibits optimal performance in diagnosing, predicting, and treating kidney disease.

Marimuthu et al. [27] conducted a review that revolved around the following objectives: (i) offering an understanding of the algorithms currently employed in predicting heart disease, and (ii) offering a comprehensive summary of the outcomes stemming from algorithmic predictions. Consequently, this review furnished insights limited to applying algorithms such as Artificial Neural Networks, Decision Tree, Fuzzy Logic, K-Nearest Neighbor, Naïve Bayes, and Support Vector Machine (SVM) for predicting heart disease.

C. CONTRIBUTION AND ORGANIZATION

To the authors knowledge, in this paper is the first paper that discusses machine learning algorithms for research on the prediction of chronic disease and problems with the data used. The main contributions of this paper include:

1) Mapping machine learning methods in chronic disease prediction studies from 2020-2022.
2) Mapping solving methods on data problems used for prediction of chronic disease.
3) Discussion on the topic of chronic diseases prediction based on the data used.
4) Identify research gaps in chronic disease prediction studies.

The paper is organized into the following sections: Section II: Systematic Literature Review Method. Section III: Predicting Chronic Disease Using Medical Data. Section IV: Machine Learning Algorithms Used in the Prediction of 6 Conditions. Section V: Discussion of Open Issues and Opportunities for Future Research. Section VI: Conclusion.

II. SLR METHODOLOGY

A. SYSTEMATIC LITERATURE REVIEW (SLR) METHODOLOGY

This section describes the methods used in the preparation of SLR. This paper uses the systematic literature review (SLR) method, which has several stages [28], [29]. These stages are (i) determining research questions, (ii) compiling a search and selection strategy, and (iii) displaying the results of extraction and synthesis data.

B. RESEARCH QUESTIONS

The initial step in SLR is to compile a research question (RQ). RQ is essential to the motivation to collect SLR [30], [31]. This paper has four RQ to start the search; here are the details:

1) RQ1: What are the problems present in the medical dataset?
2) RQ2: What approach is taken to address issues in medical data?
3) RQ3: What are the machine learning models used for predicting chronic diseases?
4) RQ4: What are opportunities for preprocessing data and machine learning models in predicting chronic disease?

C. SEARCH STRATEGY AND SELECTION

This step has two criteria in the paper search: inclusion and exclusion. Inclusion criteria set research boundaries and help researchers find a research gap [32]. The inclusion criteria for this paper are as follows:

TABLE 1. Comparisons of related survey paper.

Survey Paper	Focus Discussion	Scope
Wadghiri, *et al.* [22]	This paper primarily focused on the comparative analysis of accuracy results between ensemble and single machine learning methods. The scope of the study encompassed publication trends within the timeframe of 2000 to 2020. Additionally, the paper presented a comprehensive compilation of datasets utilized in the research, emphasizing openly available ones.	While the paper provides an extensive discussion on diabetes, the utilized datasets, and the resultant outcomes, it regrettably does not delve into the specific challenges or issues inherent within the diabetes mellitus data.
Abrar Yaqoob, *et al.* [23]	This review primarily focused on analyzing nature-inspired algorithms applied to cancer prediction. Additionally, the paper delved into a comprehensive comparison of feature dimension methods designed for high-dimensional data. Within this context, the review also addressed problematic data handling techniques that emerged during the analysis. Furthermore, the review offered an insightful discussion regarding the comparative efficacy of nature-inspired methods.	The scope of this review paper is that it needs to discuss details regarding what methods exist and are used in handling problematic data and only focuses on nature-inspired algorithms.
De Jong, ype, *et al.* [24]	This paper focused on evaluating the performance of a prediction model specifically designed for ischemic strokes. The primary objective was to assess the predictive accuracy in cases involving patients diagnosed with ischemic strokes. The paper thoroughly examined the performance of the prediction method, including its limitations, and identified any shortcomings that may have arisen during the process.	The scope of this paper are that it does not compare other machine learning prediction methods and does not display problems in the data.
Silva, Gabriel, *et al.* [25]	This paper focused on machine learning algorithms applied to the domain of hypertension prediction. Within its scope, the paper meticulously examined a range of performance outcomes to determine the most optimal results achieved. By doing so, the paper provided substantiated evidence that machine learning algorithms indeed contribute to enhancing the accuracy of hypertension prediction outcomes.	The scope of this paper do not discuss problems with the data and are not detailed in the algorithms used.
Sanmarchi, Francesco, *et al.* [26]	This review was dedicated to exploring the role of artificial intelligence in predicting, diagnosing, and treating chronic kidney disease. The main aim was to establish the efficacy of machine learning as a valuable tool for prediction in this context. The review's intent was further underscored by its effort to compare and contrast the various machine-learning methods available to determine the optimal approaches for predicting chronic kidney disease.	The scope of this review do not mention the problems that exist in the data to improve prediction results.
Marimuthu, *et al.* [27]	This review was dedicated to comparing performance outcomes among established machine learning algorithms utilized in heart disease prediction. Additionally, the review highlighted the tools employed for data analysis in this context. The ultimate goal of this review was to determine which machine learning algorithm stood out as the most effective for heart disease prediction based on the accumulated evidence and performance comparisons.	The scope of this review do not explain whether the data used for prediction is problematic and not all machine learning algorithms are compared.

1) Database: Scopus
2) String search: "diabetes mellitus" OR "hypertension" OR "stroke" OR "cancer" OR "kidney failure" OR "heart" AND "disease" AND "prediction" AND "machine learning". The search is carried out including the title, abstract, and keywords sections on the paper.
3) Language: English.
4) Year: 2010-2022
5) Subject area: Computer science
6) Document type: Article and Conference paper
7) Accessibility: Documents available in Google Scholar
8) Document type: PDF

All topics are searched and those that are irrelevant to the RQ are exclude [33]. This paper has the following exclude criteria:

1) Exclude paper that in the title, abstract, and keywords does not discuss things in RQ.
2) Exclude paper that does not have keywords.
3) Exclude paper whose content is irrelevant to RQ.

D. DATA EXTRACTION AND SYNTHESIS

After doing the next stage of inclusion and exclusion, this stage displays the extras data. Extraction data has two stages, namely: (i) collecting paper information based on extraction form (Table 3), and (ii) find more in-depth information on quality assessments (QA) (Table 3) [30].

At the data extraction stage, the author will usually fill out a form that aims to collect data extracted from an article [34]. After data extraction, the QA stage is carried out. QA is necessary because it evaluates the quality of the articles sought [35].

TABLE 2. Data extraction form.

Item	Description
Title	Article title
Year	Year of published article
Author	Author name on the article
Journal	Journal name of published article
Publisher	Journal publisher name
Keywords	Keywords given by the article
Abstract	Abstract on the article

TABLE 3. Quality assessments (QA).

QA Number	Description
QA1	What is the problem with the article being written?
QA2	Does the purpose of the research answer the research problem?
QA3	Does it have relevant sentences for the problem raised?
QA4	How to describe methodology?
QA5	Does the article have results and answer problems?
QA6	Are research gap explained?

E. LITERATURE DEMOGRAPHICS

Based on the inclusion and exclusion criteria outlined in the search strategy and selection chapter, this paper acquired 182 papers related to predicting chronic disease using machine learning and addressing issues with the data utilized. Figure 1 illustrates a bar chart showcasing the number of publications predicting chronic disease, along with a trend line. Over the past three years (2020-2022), research publications in this area have increased by 90%. This significant increase indicates that research on predicting chronic disease using machine learning has been continuously evolving, covering various topics each year.

This study learns more about the topic of each collected issue. The topic of the chronic disease prediction article was grouped based on this issue. This issue is the most important to be seen as the motivation of researchers to conduct research. Issues were grouped based on the keywords in the article. When an article does not have keywords, it is not used.

Based on the results of keyword grouping related to issues, five main issues were identified when researchers researched the prediction of chronic diseases: (i) missing values, (ii) feature data, (iii) data imbalance, (iv) outliers, and (v) data normalization. Furthermore, articles were grouped based on these five issues to determine the number of articles for each issue. Figure 2 displays a diagram showing the number of publications in each issue. Based on Figure 2, it can be observed that the order of issues in medical data starts from the most to the least common, namely imbalance, feature data, missing values, outliers, and normalization.

III. TOPIC ON THE MEDICAL DATA PROBLEM AND ITS HANDLING APPROACHES

In this section, we will address research questions one and two. RQ1 inquiries about the problems present in the medical dataset, while RQ2 seeks to understand the approach taken to address issues in medical data. The explanation will commence with a historical perspective on when the first issues with medical datasets surfaced and the approaches employed to tackle these problems.

A. HISTORY AND DEFINITIONS

In 1991, research was first conducted on three types of diseases: cancer, heart, and diabetes mellitus [36]. However, the paper focused on comparing results using distance metrics with the nearest neighbor algorithm rather than improving prediction results based on data. The initiative to troubleshoot data to enhance accuracy emerged in 2001 [37]. The study utilized a feature subset selection (FSS)-tree approach based on a genetic algorithm. The study pointed out that a significant problem in the medical world is related to the prediction of patient survival, and the application of FSS-Tree was demonstrated to enhance prediction accuracy.

In 2015, the first treatment of null values in medical data, focusing specifically on diabetes and cancer, was conducted [38]. The paper addressed the handling of missing values using k-means clustering algorithms and the most effective MLP machine learning algorithms. The study aimed to predict three diseases: diabetes mellitus, cancer, and hepatitis. Data for the research was sourced from the UCI repository. The outcomes of the study demonstrated that the application of missing value imputation positively impacted increasing accuracy. Subsequent research aimed to further test and enhance models for imbalance classification problems.

Meanwhile, the issue of imbalanced medical data first emerged in 2011, particularly in the context of diabetes prediction [39]. However, in this study, the handling of imbalanced data was carried out without addressing the issue. In 2013, new treatments for handling imbalanced data were implemented to predict heart disease, diabetes, and cancer [40]. The study mentioned several methods to balance imbalanced data, including SMOTE and SVM Weighting. Data for the study was sourced from the UCI open repository. The study highlighted the importance of addressing imbalanced data in the medical domain to increase sensitivity.

B. MEDICAL DATA PROBLEM AND ITS HANDLING APPROACHES

This section discusses the main topic of the issue of predicting chronic disease using machine learning. Figure 3 is a mind map diagram solving techniques related to research on prediction of chronic disease that conducted five main issues. In the mind map, five issues were grouped with methods for solving them in cases of chronic disease prediction.

1) MISSING VALUES

Data missing values are characterized by data that has no value or is denoted as “N/A” or “null”. Table 4 provides an example of a data form with missing values represented

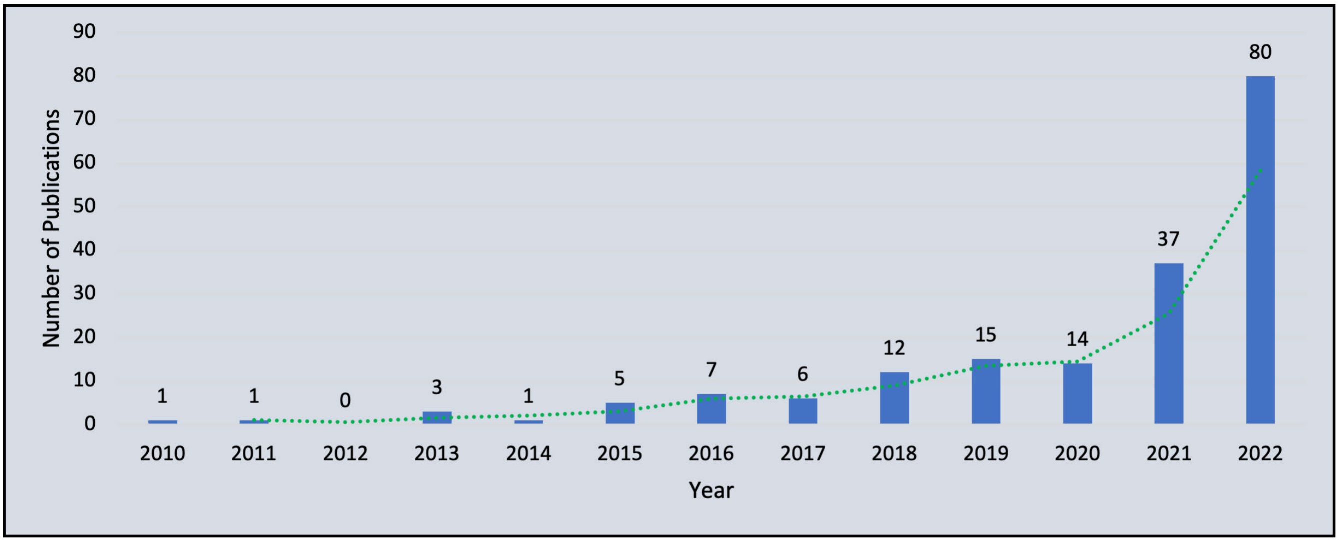

FIGURE 1. A bar chart with a trend line showing growth in chronic disease prediction research.

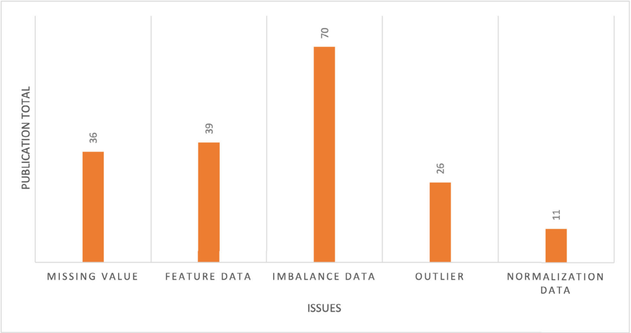

FIGURE 2. A bar diagram showing the number of publications on five issues related to chronic disease prediction.

as "N/A" or "null". Missing value is a problem related to replacing null values in data variables [41]. Missing values in data can be caused by human error when entering data [42], the patient did not provide information [43], and patient data security [44]. In general, missing value imputation techniques can be grouped into two types: statistic-based techniques and machine learning [45], [46], [47]. In this missing value, of course, it has a rate that determines whether the missing value is a little or a lot. In an online tutorial, the 5% rate is the maximum missing data limit for extensive data sets [48]. Some consider the missing value rate small, for example, less than 30%, while others focus on an extensive range of missing values of 50%-80% [49].

In diabetes mellitus prediction research, the imputation process of values to address missing value problems has been shown to increase prediction accuracy by 6% [50]. Other measurement results include a precision of 54.5%, recall of 86.2%, and an f1-score of 66.7%. However, in that study, the discussion of the results is primarily focused on addressing imbalanced data using oversampling techniques. Regarding the missing value process, there are three general techniques to replace null values: (i) In the most commonly used statistical methods, such as mean [51], [52], [53]; median [20], [54], [55]; Standard Deviation [56], and MCMC [57]. (ii) In commonly used machine learning methods, such as KNN [58], [59] and Naïve Bayes [53]. (iii) The balancing method used is Tomek Links [60].

Some studies handle missing values by deleting null data rows instead of imputing values. For instance, in a study on cancer data, missing values were present but were addressed by deleting the corresponding data rows. Consequently, the study achieved an f1-score of 76.06%, accuracy of 79.36%, precision of 80.85%, and recall of 71.81% using the ODNN algorithm [61]. However, it's important to note that this study's dataset contained duplicates and missing values. The missing value handling approach involved deleting rows with missing values, resulting in the utilization of only 400 data points out of an initial 5000.

Basant Abdel et al. [62] a similar study was conducted focusing on heart disease prediction, specifically addressing the handling of null data through deletion. The study's findings indicated an enhancement in prediction accuracy by 8%. Meanwhile, the best precision result was 91%, and the recall was 88%. However, the pivotal factor contributing to this improvement was not the removal of missing values but rather the selection of significant features. It's worth noting that the study should have stated the amount of data deleted due to missing values for transparency and completeness of the research.

In stroke prediction, missing data was handled by deletion, resulting in a prediction accuracy of only 73.52%, an Area Under Curve (AUC) of 83.03%, and a specificity of 73.43% using a logistic regression machine learning algorithm [63]. However, the primary issue in that study was identified as data imbalance. In the research by Kokkotis et al. the

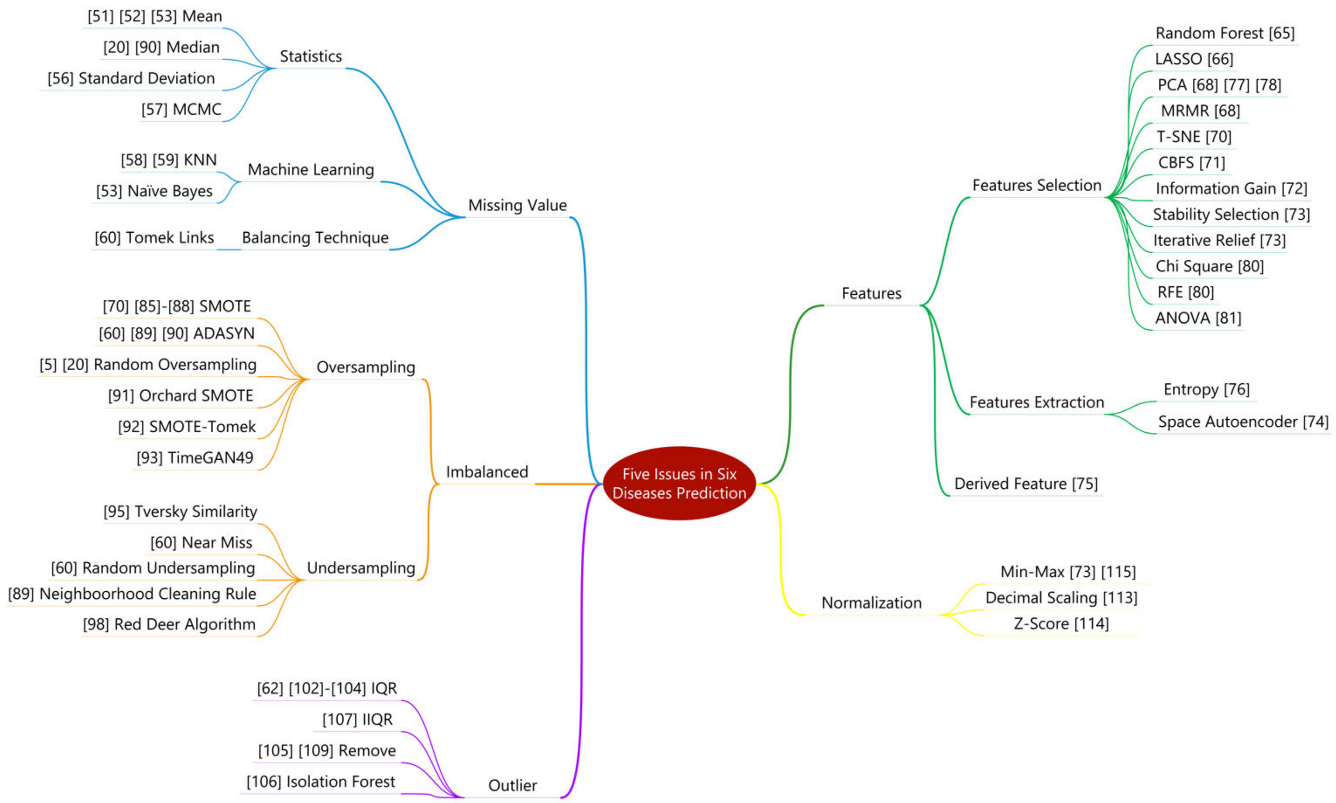

FIGURE 3. Mind map diagram-solving techniques on five issues predicting chronic disease.

total amount of data used was reported as 43,400, but the exact number of data points deleted due to missing values was not specified. Moreover, the study primarily focused on investigating risk factors in the data that contribute to stroke. Consequently, addressing missing values was not the central topic of discussion.

Potharaju et al. [64] conducted a study on predicting kidney failure where missing data was deleted. The reported measurement results include accuracy at 99%, precision at 99%, recall at 100%, f1-score at 99.5%, and a receiver operating characteristic (ROC) of 0.99%. However, the high resulting accuracy is attributed to the influential factor of handling imbalanced data using ensemble-based machine learning algorithms (stacking, bagging, boosting, and voting). The study explicitly mentioned that only 400 data points were used after deletion due to missing values. Additionally, the primary focus of the study was on handling imbalanced data and utilizing ensemble methods.

2) FEATURES DATA

Almustafa et al. conducted the research, [65] examined the utilization of feature selection techniques in predicting kidney disease by implementing ensemble tree-based algorithms (random tree and decision tree). The study outcomes demonstrated that employing feature selection had a discernible effect, leading to a 0.25% increase in accuracy. Other measurement results include ROC 0.978, Mean Absolute Error (MAE) 0.0583, Root Mean Square Error (RMSE) 0.1992, f1-score of 95.8%, recall of 95.8%, and precision of 95.8%. However, the study did not state what kind of feature selection technique was used, whether from a machine learning algorithm or another feature selection technique.

In contrast, the research by Zhang et al. [66] employed the LASSO feature selection method in combination with MLP machine learning algorithms. However, this results in an 8% decrease in prediction performance (sensitivity, recall, and f1-measure). This suggests that the study revealed the presence of other influential factors that hold greater significance in terms of managing prediction outcomes.

According to Khalid et al. [67], data handling related to features is categorized into two main aspects: (i) feature selection and (ii) feature extraction. Within the feature selection category are sub-sections, including filters and wrappers. Meanwhile, the feature extraction category encompasses sub-sections such as transformation, determination of the number of new features, and assessing performance metrics.

Zou et al. [68] in the prediction of diabetes mellitus, the use of feature selection techniques was carried out to see the influence between features. The results showed a difference in accuracy of 2% when applying the selection feature. Meanwhile, the result of Matthews correlation coefficient (MCC) 0.77, specificity 0.81, and sensitivity 0.95. The selection feature methods used in the study were PCA and MRMR with RF, NN, and J48 machine learning algorithms.

TABLE 4. Missing value in dataset.

blood pressure MAP	BMI	stomach circumference	blood sugar fast glucose	cholesterol
N/A	23.4	120	null	89
86.66	28.1	101	166	74
70.21	N/A	91	100	105
55.14	N/A	N/A	115	78
65.37	25.7	87	N/A	null
N/A	N/A	N/A	N/A	N/A
82.7	36.2	80	102	94

The MRMR feature selection technique produced better performance than PCA. The best machine-learning algorithm is RF. The research also states that the results will not be good if only use the most essential features in the dataset for prediction.

Sornsuwit [69] conducted a study involving feature selection by implementing CFS (Correlation-based Feature Selection) in conjunction with AdaBoost machine learning. This was combined with K-Nearest Neighbors (KNN), Naïve Bayes, and Neural Network algorithms. The study's outcome indicated an improvement of 3% in the predictive results. Meanwhile, the result of precision is 87.5%, sensitivity and specificity is 91%, and f1-score is 87%. However, the dataset initially contained eight variables, and the study employed the CFS technique to utilize only four variables for prediction. Moreover, the author compared the performance results for each positive and negative class, even though the data did not balance the proportion of the number of classes, rendering the comparison unfair.

Pokharel et al. [70] investigated the utilization of feature selection and feature extraction based on t-distributed Stochastic Neighbor Embedding (t-SNE) values in combination with deep learning prediction algorithms. This approach led to a 2.5% improvement in the obtained results. Meanwhile, sensitivity score is 40.3% and specificity is 98.3%. However, in that study, it does not mention the number of selected and used features or the number of features extracted from the data. The research only mentions the number of data samples after applying imbalanced data techniques. In another diabetes prediction study addressing data-related challenges, one approach focused on employing importance-based features derived from Gini scores [20]. However, the study should have elaborated on the order of each feature in the data after applying the Gini value. The author primarily focused on experimenting and examining the impact of applying imbalance techniques rather than thoroughly explaining the specific order of features.

Additionally, other methods were applied for diabetes mellitus prediction, such as fast-correlation-based Feature Selection (fast-CBFS) [71], information gain [72], stability selection (SS), and iterative relief [73]. These techniques were used to enhance the prediction accuracy of diabetes mellitus. Kishor and Chakraborty [71] analyzed comparing prediction results after applying the fast-CBFS feature selection technique. Additionally, the author applied the SMOTE technique to handle imbalanced data. The analysis results indicated that the application of the fast-CBFS technique influenced the prediction accuracy value, resulting in an improvement of approximately 10%-15%. Meanwhile, AUC result is 99.35%, sensitivity 99.32%, and specificity 98.86%. However, it is worth noting that the author did not conduct a specific analysis to compare the prediction results after applying the imbalance technique.

Saxena et al. [72] conducted a comparative analysis of the results using various feature selection techniques, including PCA, CFS, and information gain. The dataset initially contained nine features. Through experimentation, the feature selection techniques produced results with four and six features, respectively. The findings indicated that the most compelling feature selection technique was CFS, resulting in six features. The results obtained in this study are accuracy 79.83%, sensitivity 79.83%, specificity 71.4%, and AUC 0.83. However, the proposed technique in this study demonstrated a marginal increase of 1.5% compared to the state-of-the-art technique mentioned in the referenced study.

Akyol and Şen [73] focused on diabetes mellitus classification, and three distinct feature selection techniques were employed: IR, RFE, and SS. Additionally, the study conducted tests utilizing these techniques on three different dataset types. The classification outcomes demonstrated that the SS technique exhibited a 0.28% improvement compared to the IR technique and a 1.28% enhancement over the RFE technique. Meanwhile, measurement results from the proposed method include accuracy 73.88%. This study only focuses on the accuracy level. However, it is noteworthy that the study needed to detail the ranking or identification of the most crucial features for each of the three feature selection techniques.

García-Ordás et al. [74] introduced new features by utilizing a sparse autoencoder to predict diabetes mellitus. A comparison was made between sparse autoencoder techniques and various deep-learning algorithms. The primary purpose of the sparse autoencoder was to predict data dimensions. The prediction results indicated that incorporating the sparse autoencoder led to a 13% improvement compared to utilizing only the deep learning algorithm. This study only focuses on the accuracy level. However, the study did not specify the amount of data utilized, including the reduction or any specifics regarding the dataset size.

Rajkamal et al. [75] introduced an intriguing approach by implementing derivative features, which led to the generation of new features derived from the existing features. The creation of derived features followed established standards by recognized health organizations like the WHO. The results obtained using the proposed method are as follows: accuracy 0.89, sensitivity 0.79, specificity 0.94, and AUC 0.94. However, it is essential to note that not all features were derived; only two out of nine were subjected to this process. In cancer prediction research, the extraction feature is applied based on the entropy value [76]. The results obtained by applying entropy index to features are as follows: accuracy 0.95, precision 0.8, recall 0.83, and f1-score 0.81. However, it was observed that the entropy value did not significantly influence the prediction results. Instead, the prediction outcomes were notably affected by applying a cost-sensitive algorithm.

Qi et al. [77] involved applying PCA for feature selection and Deep Belief Network (DBN) for feature extraction. The findings demonstrated the superiority of the DBN technique over PCA. The results obtained using the best method are as follows: accuracy 98.2%, specificity 98.5%, and sensitivity 62.5%. However, a limitation of this research is the need for more consideration of data imbalance issues. Furthermore, the study utilized different machine learning models, namely PCA-ANN and DBN-ELM-BP, which may impact the comparability of results.

Dev et al. [78] applied PCA techniques with various machine learning algorithms for stroke prediction. The results obtained using PCA technique are as follows: accuracy 0.73, f1-score 0.72, recall 0.68, and precision 0.75. However, specific details still needed to be provided regarding the number of features before they needed to be replicated. Additionally, a limitation of this research is the need for more data imbalance handling, which could impact the robustness and fairness of the predictive model. A study that discusses the prediction of heart disease using the feature optimality criterion selection technique aims to optimize features in the data [79]. The study, in terms of measuring results, focuses on the running time in diagnosing heart disease. The running time result with a dataset size of n = 30 is 24.9 milliseconds.

Feature selection methods for stroke prediction include chi-square and RFE [80]. The accuracy obtained is 85%, and the f1-score is 88%. Almasoud et al. [81] predicted kidney failure using the ANOVA method to identify the smallest subset of features relevant to the prediction task. ANOVA was utilized for two distinct feature categories: numerical and categorical. he results obtained using the proposed feature selection technique with the gradient boosting model are as follows: f1-score of 99.1%, specificity of 99.33%, and sensitivity of 98.8%. However, the analysis and discussion in the paper primarily focus on the machine learning models, especially detailing the parameter tuning process for these models. The specific attention to model parameters and tuning is evident throughout the study.

3) IMBALANCE DATA

Imbalance is a problem that plays a crucial role in machine learning [82]. This problem is defined into two categories: majority class and minority class. Majority class is where the number of one class is more dominant than the other class; the opposite definition is used for minority class. The ratio of data is said to be imbalance into three categories: 1:100 small, 1:1000 medium, and 1:10000 large. This problem is not only found in binary classes but also multi-classes. Some data-related fields have this problem, including the field of health diagnosis and the financial industry [83]. Techniques in handling imbalance are divided into two, namely oversampling and undersampling [84].

Figure 4 is a form of a diagram of two categories of imbalance handling. Based on Figure 4, the two ways of handling imbalance, namely, oversampling and undersampling, are different. Oversampling is performed by adding synthetic data to minority data, whereas undersampling reduces the amount of actual data to the majority data. Based on research on the prediction of chronic disease, here are some names of methods used for handling imbalance data. (1) oversampling techniques: SMOTE [70], [85], [86], [87], [88]; ADASYN [60], [89], [90]; ROS [5], [20]; orchard SMOTE [91]; SMOTE-Tomek [92]; TimeGAN49 [93]; and SVM-SMOTE [94]. (2) undersampling techniques such as: Tversky similarity [95], near miss [60], RUS [60], and NCL [89].

Potharaj et al. [64] adopted a systematic approach to tackle data classification problems characterized by imbalanced datasets. The dataset utilized in this study originated from Apollo Hospital in India. The technique employed to address class imbalance was SMOTE. The matrix of measurement results used includes accuracy at 99%, precision at 99%, recall at 100%, f1-score at 99.5%, and ROC of 0.99%. However, the study does not describe how the balancing process was conducted, whether it was applied to the training data or the dataset before the splitting process.

Zidan et al. [85] focused on diabetes mellitus risk prediction, the authors employed the Synthetic Minority Over-sampling Technique (SMOTE) to balance the training data. The best results were obtained using the random forest model with an accuracy of 79.27%, sensitivity of 50.29%, specificity of 84.31%, and AUC of 76.08%. However, the study does not explicitly mention the quantity of data after the application of SMOTE, nor does it provide a comparison of the results before and after the implementation of SMOTE. Azad et al. [86] various training-testing proportions were employed to observe their impact on prediction results. The imbalanced technique utilized is SMOTE combined with genetic algorithm. The proposed method yields the following values: precision 77.97, sensitivity 85.98, f1-score 81.7, and ROC-AUC 84.9. However, the study did not explicitly present the number of data samples utilized after applying SMOTE.

Ramadhan [90] conducted a comparison of two oversampling techniques, SMOTE and ADASYN, for diabetes

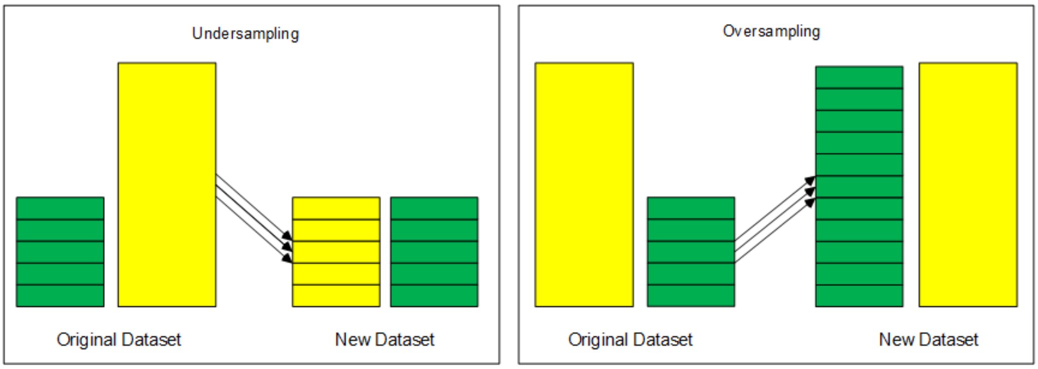

FIGURE 4. Two category of imbalance dataset.

detection. The findings indicated that the ADASYN technique outperformed SMOTE by a margin of 2%. The accuracy obtained by the ADASYN technique is 87.3%, while for SMOTE it is 85.4%. This study focuses solely on the measurement of accuracy as its outcome metric. However, the research needed to explain the fundamental differences between these two oversampling techniques. Moreover, the number of data samples after applying both SMOTE and ADASYN was not specified.

Sreejith et al. [91] the aim of the study was to enhance the Orchard algorithm by incorporating the SMOTE technique. The experiment involved comparing the results before and after this modification. The research revealed that the proposed method led to a 14% improvement. The prediction results obtained by the proposed method are as follows: accuracy 89.04%, sensitivity 89.74%, specificity 88.39 precision 88.17%, f1-score 89%, and MCC 0.78%. However, a critical error was identified in the data balancing process. The proposed method showed that the balancing process was performed on the initial data rather than the training data. This is a significant error, and the balancing process should ideally be conducted on the training data for accurate results.

Roy et al. [92] the study focused on addressing imbalanced data, outliers, and missing values in diabetes mellitus detection. Missing values were handled through KNN imputation, while imbalances were addressed using SMOTETomek, a combined oversampling and undersampling technique. The presented results, considering the impact of data imputation, show precision of 98%, recall of 98%, specificity of 99%, and an f1-score of 98%. However, it is important to note that the choice of machine learning algorithm significantly influenced these results. Additionally, the research predominantly emphasized the handling of outliers rather than addressing imbalances in the dataset.

Ning et al. [93] the study utilized the time series bootstrapping technique to address imbalanced data and aimed to establish relationships between latent variables and clinical features within a latent Dynamic Bayesian Network framework. However, the study did not explicitly present the prediction results obtained from the implemented methodology. Furthermore, the study did not provide a comparison of the data proportions after bootstrapping.

Kamaladevi [95] focused on comparing various undersampling techniques, including ENN, near miss, random undersampling, and Tomeklink. The objective was to analyze and discern the differences between these techniques. The study identified the near-miss technique using the RF machine learning algorithm as superior in accuracy. Testing was conducted on two public datasets, namely Pima Indian Diabetes and Hepatitis. For the diabetes dataset, the prediction results obtained are as follows: precision 80%, recall 78%, f1-score 96%, accuracy 75%, and ROC-AUC 0.94. However, after applying these undersampling techniques, the study needed to provide more information regarding the number of data samples.

Xiao et al. [96] conducted research focused on enhancing cancer prediction performance by addressing imbalance data. The approach adopted was oversampling utilizing the Wasserstein Generative Adversarial Network (WGAN) method. The study demonstrated that the achieved predictive outcome reached a remarkable accuracy of 98.3%, precision 100%, recall 96.67%, f1-score 98.31%, and AUC 0.98. Utilized various cancer datasets for its analysis. It employed imbalanced data handling techniques, including oversampling techniques like ROS and SMOTE, as well as the undersampling technique RUS. However, the study did not provide details regarding the number of data samples before and after the application of these imbalance techniques.

Zhang et al. [97] employed a boosting approach to address imbalance data issues in cancer prediction. The methods employed were SMOTE Boosting and RUS Boosting. As a result, the prediction achieved an impressive accuracy of 98.2%, sensitivity 93.75%, specificity 100%, G-mean 0.96, and AUC 0.96. Cost sensitive imbalance methods were also used to improve the accuracy of cancer predictions, the study proved to increase accuracy by 5% [76]. Meanwhile, the result of precision 0.8, recall 0.83, and f1-score 0.81.

Moghadam and Ahmadi [98] in the studied as a result of their efforts, the accuracy in predicting kidney failure disease was enhanced by 1.6%. imbalance treatment of kidney failure prediction uses a clustering approach using RDA which is based on undersampling. The findings indicated that the decision tree model surpassed other algorithms, achieving 0.96, 0.94, 0.97, 0.95, and 0.95 for accuracy, sensitivity, specificity, F1 score, and AUC, respectively. Furthermore, the results derived from decision trees are more comprehensible and interpretable. However, the min–max technique was used to select the majority of samples to be deleted. This technique can result in the value when a comparison is made with the minority class which will contain all min-max values. In the prediction of other disease, there are also problems with imbalance data using the SMOTE and SMOTE-Tomek methods [21], [99], [100].

López-Martínez et al. [99] predicted hypertension by implementing imbalance management using SMOTE and comparing results with various machine learning algorithms. The results indicated that the application of the SMOTE technique increased prediction results by up to 13%. The accuracy, precision, recall, and f1-score values obtained from the proposed method are 0.73, 0.57, 0.4, and 0.47, respectively. The f1-score value is still relatively low in disease prediction cases, mainly due to the highly imbalanced medical dataset utilized. However, crucial details such as the number of data samples after implementing SMOTE and a clear explanation of the comparison methodology were lacking, potentially causing confusion for readers.

Ramezankhani et al. [100] analyzed the impact of applying SMOTE for diabetes mellitus prediction using three different types of classifiers: Probabilistic Neural Network, Decision Tree (DT), and Naïve Bayes (NB). The experiment involved analyzing the percentage differences in the proportion of the number of major classes from 100% to 700% for balancing. In the study, the outcome measurement focuses on sensitivity, which experienced a 5% increase after applying the SMOTE technique. On the other hand, the accuracy value decreased. Accuracy in imbalanced data cases is a biased measurement tool. However, the accuracy results decreased after implementing SMOTE.

4) OUTLIER

Outlier detection is an essential task in many domains because outliers exhibit abnormal conditions, which can result in decreased performance in the domain [101]. Outlier detection is widely used in several domains, such as fraud detection, time series monitoring, and monitoring medical conditions. There are three approaches to the problem of detecting outliers, namely: (i) unsupervised learning, (ii) supervised learning, and (iii) semi-supervised learning.

In the study of chronic disease prediction, most of the data had outliers removed. Removing outliers was done with IQR techniques [62], [102], [103], [104]. Yuk et al. [102] considered outliers, stipulating that any data point with a pulse rate variable greater than or equal to 300 should be excluded from the analysis. Outlier handling was not the primary focus of this study. Researchers primarily concentrated on comparing the outcomes of various machine learning models. The best results were achieved by an ensemble boosting-based machine learning model with an accuracy of 0.84, AUC of 0.86, sensitivity of 0.69, and specificity of 0.88.

Peñafiel et al. [103] used representative rules to detect and handle outliers in the data. These rules were designed to identify and measure the portion of patients who met specific criteria, thereby addressing the presence of outliers in the dataset. Outliers in the study will be avoided by employing rule-based representations. This research, instead, focuses on the effects of missing values, where skipping missing values yields a better ROC value compared to imputing missing values using the mean technique. The difference in ROC values obtained is 0.16.

Rajendran and Anitha [105] carried out a study focused on the amalgamation of managing missing values and eliminating outliers to predict heart disease. The study's findings demonstrated that the combined application of these preprocessing techniques yielded superior results compared to other combinations of preprocessing techniques. In the research, outlier detection and handling were performed by examining the relationship between healthcare attributes and Mahalanobis distance. The results of this study, by applying the proposed preprocessing techniques (missing value handling, outlier detection, and feature selection) to the best machine learning model, Naive Bayes, show an increase in accuracy by 6.48%, AUC by 1.43%, sensitivity by 5.26%, specificity by 8%, precision by 4.23%, and f1-score by 6.18%. However, it was noted that interpreting the Mahalanobis distance calculation could be challenging and not intuitive.

Yasmeen et al. [106] utilized the isolation forest algorithm to identify outliers within the dataset related to hypertension patient data. In this study, the analysis focuses on the implementation of the isolation forest technique applied to several machine learning models such as MLP, SVM, LR, DT, and Stack ensemble. The research primarily focuses on measuring the AUC values, which are 0.918, 0.879, 0.865, 0.775, and 0.92, respectively. However, the study needed to explain how the isolation forest method, used for handling outliers works. The IIQR method is used to find outliers on ECG data from heart disease [107]. This study solely focuses on accuracy values. The accuracy obtained from the proposed method by applying outlier removal is 99.45%.However, this research should have focused more on discussing outlier handling techniques.

Ifflath et al. [108] employed the isolation forest method for outlier detection, and the study revealed that the utilization of the isolation forest method led to improved results in cancer prediction compared to not applying the isolation forest method. The f1-score increased by 29% after removing the outliers using DT model. As a result, the final values for accuracy, precision, recall, and F1-score are 96%, 99%, 100%, and 98%, respectively. However, this study does not

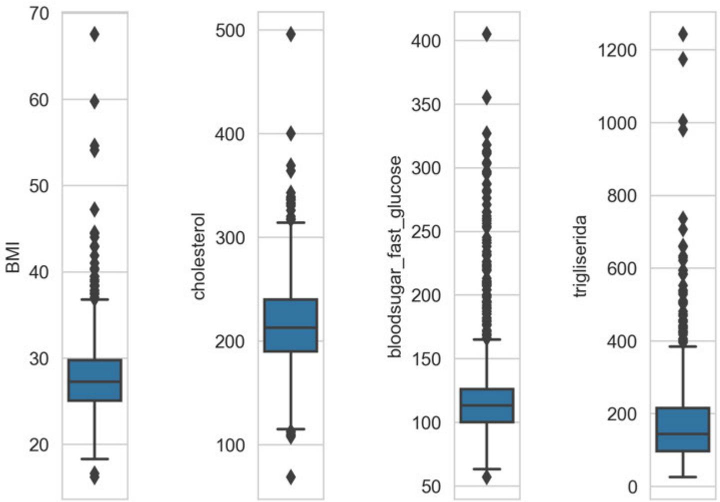

FIGURE 5. Outlier in data.

provide detailed characteristics of the dataset considered as outliers.

In their research on diabetes prediction using big data, Zhang et al. [109] conducted chose to remove outliers. In that study, the focus was solely on accuracy, with a prediction result of 95%. However, it is important to note that outliers in medical data can hold significant value, thus requiring careful consideration in their handling. Figure 5 is an example of visualizing the data form with outliers. Outliers can be seen as black dots far from or close to each other.

5) DATA NORMALIZATION

Data normalization is a mapping or scaling technique at the preprocess stage [110], [111]. In the process researchers can find new ranges from existing ranges, this technique can be very helpful for prediction purposes [112]. Some names of techniques that are currently often used to normalize data on prediction of chronic disease such as: min-max, z-score, and decimal scaling.

Swathi and Kodukula [113] conducted research focused on cancer prediction utilizing gene data. In their study, the researchers employed normalized data along with scaling features to process patient gene data. The accuracy, precision, recall, sensitivity, and f1-score values obtained using the proposed preprocessing technique are 97%, 100%, 95%, 95.2%, and 97.5% respectively. However, in this research, handling normalization was not the primary focus; rather, the main emphasis was on feature reduction using PCA. In another study a researcher used the min-max technique for cancer data [77]. The results obtained using the best method are as follows: accuracy 98.2%, specificity 98.5%, and sensitivity 62.5%. However, the main focus of this research was on handling data dimension reduction.

Fang et al. [114], predicted hypertension occurrences for the upcoming five years by employing the z-score normalization technique. The normalization technique was used to alter the scale of values for the age variable, aiming to achieve a normal data distribution. Additionally, outliers were removed in this study, contributing to the attainment of a normal data distribution, rather than relying solely on the normalization technique. The application of the proposed method yields accuracy, precision, recall, f1-score, and AUC values of 0.861, 0.817, 0.923, 0.867, and 0.951 respectively. However, the study primarily focuses on the hybrid machine learning model used.

Cai et al. [115] the discussion on cancer prediction involved utilizing the min-max normalization technique to process data. Researchers tend to avoid extensively altering data in this field, preferring established normalization techniques. This approach is likely due to the need to re-validate new data ranges with medical professionals. The research results obtained using the proposed preprocessing technique show an AUC of 0.98, accuracy of 0.98, and specificity of 0.97. However, the factor influencing the prediction results is the handling of imbalance using the SMOTE algorithm. Thus, this research primarily emphasizes addressing imbalanced data issues. Moreover, in specific disease prediction research, the application of normalization techniques is limited. This is often because altering values must accurately represent the results of medical examinations, making researchers cautious about introducing new ranges through normalization.

IV. MACHINE LEARNING IN CHRONIC DISEASE PREDICTION

This section will discuss the machine learning algorithms used to predict diseases, addressing RQ3 regarding the machine learning models used for predicting chronic diseases. The paper is categorized into nine algorithmic groups based on the results of an in-depth analysis of the obtained papers. Tables 5 through 7 present detailed information on the machine learning models used for each disease, including the parameters employed in each model.

Table 8 presents a comparison of the advantages and disadvantages of the machine learning algorithms in chronic disease prediction.

A. ENSEMBLE LEARNING

Ensemble learning is a technique that combines multiple methods to make decisions, typically in supervised machine learning scenarios [116]. For instance, Decision Tree, Neural Network, and SVM methods can be combined into an ensemble method. Some advantages of this approach include mitigating overfitting, reducing computational complexity, and providing a straightforward representation of results [117], [118]. Ensemble methods are useful for addressing complex machine learning problems, such as class imbalance, evolving feature distributions, and the curse of dimensionality.

Some familiar ensemble methods are used for the prediction of chronic disease as follows:

AdaBoost is a method that emphasizes previously misclassified instances during training to have a greater impact on subsequent iterations. The value assigned to each instance in the training set is based on weights. Initially, in the first iteration, equal weights are assigned to all instances. Then, with each subsequent iteration, the weight of the misclassified instance increases, while the weight of the correctly classified instance decreases [119].

Azar [120] conducted cancer prediction utilizing various ensemble methods, including AdaBoost, XGBoost, and Random Forest. Interestingly, while AdaBoost did not yield the best prediction results, the random forest method emerged as the most effective for their study. The results obtained using the AdaBoost model are as follows: RMSE (Root Mean Square Error) of 27.76%, time taken is 12.17 seconds, accuracy of 78.25%, sensitivity of 45.66%, specificity of 86.40%, f1-score of 44.76%, and AUC of 66.03%. However, in this research, the primary focus was determining the best parameters using Shapley Additive Explanations (SHAP) for the machine learning method and handling imbalanced data using SMOTE.

Kibria et al. [121] delved into diabetes prediction within explainable AI, employing an ensemble method approach. The AdaBoost model yields the following prediction results for each measurement: precision 0.82, recall 0.85, f1-score 0.83, AUC 0.95, and accuracy 0.83. However, these results are still lower than those obtained using the voting-based ensemble model (XGBoost and Random Forest), with differences ranging from 0.4% to 0.6% for each measurement. In the study, various data issues were addressed, including handling missing values using the mean, addressing imbalanced data using SMOTE, and determining the primary driving factors influencing disease prediction from various perspectives using SHAP.

The AdaBoost method can also be used for high-dimensional data [122]. The AdaBoost model yielded an accuracy of 97%, precision of 96%, recall of 95%, and an f1-score of 95%. However, these results were still lower compared to the ensemble model combining AdaBoost with deep learning, with a difference result only 1%. The study also addressed data issues, particularly feature selection using genetic algorithms.

Bagging is a method that uses a simple but effective approach to produce an independent ensemble model where each inducer is trained using examples taken from the original dataset instead of [123]. To ensure sufficient instances per inducer, each sample typically contains the same number of instances as in the original data set. Bagging methods in several studies were used for cancer data [124], [125].

In the research Syed et al. [124], an accuracy of 0.76, precision of 0.75, recall of 0.94, and f1-score of 0.84 were obtained. Looking at these results, it indicates that the dataset used is indeed imbalanced because of the significant gap between precision and recall values. The study attributes these results to addressing the imbalanced dataset using a GAN (Generative Adversarial Network) model.

In the research Hesham et al. [125], the results showed an accuracy of 94.73%, AUC of 94%, and F1-score of 93%. However, the bagging model's performance was still lower compared to the stacking ensemble model, with an accuracy difference of 4%. This study identified the impact of applying feature selection using the recursive feature elimination technique. The results indicated that there was an impact of 2-5% on the accuracy values due to the application of feature selection. However, in terms of the F1-score, the impact was only 1-2%. The bagging method is also used in heart disease prediction, and the results show this method is superior [126], [127].

Ashfaq et al. [127] compared several ensemble techniques such as bagging, voting, stacking, and boosting. This research handled feature scaling using z-score. The research results for the bagging model yielded an accuracy of 82%, for stacking 84%, and for voting 85%. This indicates that the voting model outperforms the other ensemble models. The precision, recall, and f1-score for the stacking model were 85%, 84%, and 84% respectively. However, this research compared the results with previous research without considering the parameters used in the previous algorithm. So, it could be said that it was not an apples-to-apples comparison.

Stacking is a method that is the most popular ensemble modeling technique in machine learning. Various weak learning models are combined in parallel so that combining them with meta-learning can better predict the future [116]. Kumar et al. [128] optimized a stacking ensemble technique for cancer classification, achieving an impressive accuracy of 99.45%, precision of 99%, recall of 98%, and f1-score 99%. The study also addressed data issues such as feature importance, although the specific technique used was not mentioned, and outlier removal. Optimization was carried out using a genetic algorithm to determine the best classifier. However, in this research, many machine learning algorithms were stacked, potentially causing the results to be affected by overfitting.

Gupta and Gupta [129] addresses the aspect of computational time in cancer prediction through ensemble algorithms. The study demonstrated that the stacking method required a running time of 5.6 seconds for its execution. The AUC value using the stacking model is 0.997. This research focuses on measuring running time and AUC by comparing several other machine learning models. However, in this research, the discussion focused more on handling data issues, namely, missing values using KNN, normalization using z-score, imbalance using ROS, and feature selection using RF.

Dritsas and Trigka [130] predicted strokes using stacking methods involving algorithms such as Random Forest, Naïve Bayes, RepTree, and J48. Their prediction outcomes yielded an impressive accuracy rate of 98%, precision 97.4%, recall 97.4%, f1-score 97.4%, and AUC 0.989. However, this research focused on a comparative analysis of handling problems in the data. The data was handled by removing missing values, detecting important features using RF and information gain, and handling imbalances using SMOTE.

Random forest is a method in which it performs a combination of predictors using tree shapes. Each tree relies on random vector values sampled independently and with the same distribution for all trees in the forest [131]. This method has proven influential in classifying or predicting medical data. It is proven that this method for predictive results is always superior to other methods [20], [132]. This method is famously powerful for disease prediction.

Ramadhan and Romadhony [20] conducted the detection of diabetes using RF and comparing it with the logistic regression algorithm; the results showed that the RF algorithm was superior to logistic regression. The prediction results include precision of 83, recall of 88, and f1-score of 86. However, this research focused on analyzing problems in handling disease data, such as replacing missing values using median, balancing data classes using ROS, and feature importance using entropy value.

Masetic and Subasi [132] detected heart disease using several machine learning algorithms, including RF. In addition, the researchers applied the autoregressive burg for feature selection technique. This study solely focuses on achieving an accuracy of 100%. However, this high accuracy was not re-validated, leading to indications of overfitting to the model. And then, this study did not modify the RF algorithm and focused on comparing the results after handling feature selection.

XGBoost is a method scalable machine learning system for increasing the number of trees in prediction [133]. In 2015, this method showed that 17 of the 29 datasets in Kaggle using the XGBoost method were superior in terms of classification or prediction [112]. This algorithm improves the GBM algorithm by adding several trees to improve scalability. XGBoost was known for its excellent performance across various types of datasets. This algorithm efficiently handled large datasets with many features.

On cancer prediction, using the XGBoost algorithm combined with the feature selection technique resulted in a prediction accuracy [113]. The XGBoost algorithm was superior to the RF algorithm. The accuracy, precision, recall, sensitivity, and f1-score values obtained using the proposed preprocessing technique are 97%, 100%, 95%, 95.2%, and 97.5% respectively. However, the analysis focused more on the results of applying the selection feature using PCA.

Yang and Guan [134] engaged in heart disease prediction utilizing XGBoost in conjunction with imbalance data handling techniques. This study also handled dirty datasets by including replacing missing values with the mean, normalization using min-max, feature selection using information gain, and addressing imbalanced data using SMOTE-ENN. The prediction results obtained were an accuracy of 0.9344, precision of 0.9266, recall of 0.9716, and an f1-score of 0.9486. Important factors in handling dirty medical datasets were imbalanced data and feature selection. However, this research focused on the contribution of handling selection and imbalance features.

B. DEEP LEARNING

Deep learning is the development of a small subset of machine learning [135]. Another article mentions that deep learning is a method that can solve all problems in machine learning [136]. Some advantages of using deep learning are that there are no experts to measure the validation of the results, problems that change over time, solutions that need to be solved with specific cases, and problems that are too big (web page ranking by all parameters). However, the weakness of the deep learning algorithm is that it is a black box, making it difficult to understand the running process. Several deep learning methods, including the following, are used to predict chronic disease.

Artificial Neural Network is the most straightforward and most complex deep learning method. ANN has become a relatively competitive model with conventional regression and statistical models [137]. ANN is said to solve all problems in the medical, economic, energy, and transportation domains [138]. Mridha [139] conducted a comparison between Artificial Neural Network (ANN) methods and conventional machine learning techniques for cancer prediction.

The findings indicated that ANN achieved an accuracy enhancement of 99.73%, sensitivity 100%, specificity 98%. The results indicated the occurrence of overfitting to the prediction model. The preprocessing handling on the dataset involved normalization using z-score, removing rows with missing values, and feature selection using RFE.However, despite this improvement, the results achieved by ANN were still outperformed by ensemble-based methods [139].

Other researchers used ANN for cancer prediction [77]. The results obtained using the best method are as follows: accuracy 98.2%, specificity 98.5%, and sensitivity 62.5%. The study addressed the dirty dataset by imputing missing values using the mean, normalizing using min-max, and removing outliers. However, this research focused on handling several feature selection techniques.

Zhang et al. [66] explored the utility of Artificial Neural Networks (ANN) for predicting kidney failure. The study yielded an accuracy of 0.965, sensitivity of 0.732, specificity of 0.9973, precision of 0.9848, recall of 0.732, and F-measure of 0.7912. It can be observed from the significant difference between the F1-measure and accuracy that the dataset used was imbalanced, and the study did not handle the imbalance well. However, the study revealed decreased accuracy attributed to improper feature selection methods.

López-Martínez et al. [99] employed an Artificial Neural Network (ANN) model to classify patients with kidney disease and diabetes among hypertensive patients. The results indicated that the application of the SMOTE technique increased prediction results by up to 13%. The accuracy, precision, recall, and f1-score values obtained from the proposed method are 0.73, 0.57, 0.4, and 0.47, respectively. The f1-score value is still relatively low in disease prediction cases, mainly due to the highly imbalanced medical dataset utilized. However, this enhancement was attributed to addressing the challenges of imbalanced data.

Convolutional Neural Network is one of the feedforward-type neural network algorithms that can extract features from data with a convolution structure [140]. This CNN method is usually widely used for problems in image detection, face recognition, and intelligent medical treatment. However, in predicting chronic disease, some use this method as done.

Dev et al. [78] conducted research to enhance stroke prediction. The results obtained using PCA technique with CNN are as follows: accuracy 0.73, f1-score 0.72, recall 0.68, and precision 0.75. However, the study's outcomes indicated that this method still needed to catch up when compared to the effectiveness of the ensemble random forest method. Additionally, this research focused more on analyzing the use of features selection. This research also a limitation of this research is the need for more data imbalance handling, which could impact the robustness and fairness of the predictive model.

Ashrafuzzaman et al. [141] predicted strokes through the utilization of Deep Convolutional Neural Network (CNN), yielding a notable prediction accuracy of 95.5%, precision 96%, recall 100%, f1-score 98%. However, this research focused on problem data handling, namely imputation of missing values using the mean, label encoding, feature selection using univariate selection, feature importance using decision trees, not removing outliers, and feature correlation using heatmaps. So, the CNN algorithm only hypermatized several parameters, such as ReLU value and activation.

Gayathri et al. [142] conducted cancer prediction using Mayfly CNN optimization techniques, achieving impressive prediction results with an accuracy of 97%. However, in that study, the focus was solely on the accuracy value, and specific preprocessing data handling was not specifically discussed. Sateesh and Balamanigandan [143] compared the accuracy performance between the decision tree and Convolutional Neural Network (CNN) methods for predicting heart disease. The decision tree algorithm demonstrates superior accuracy at 87.75% compared to CNN, which achieves 84.5% accuracy. Both algorithms exhibit statistical significance, with an independent sample T-Test value of 0.001 ($p<0.05$).

C. BAYES ALGORITHM

This algorithm, also called Bayes Network, has a structural model and a set of conditional probabilities [144]. The structural model is a directional graph in which nodes represent attributes, and arcs represent attribute dependencies. The Bayes method, which researchers often use to predict chronic disease, is Naïve Bayes.

Banerjee [145] conducted a comparison of prediction methods for cancer, explicitly evaluating Naïve Bayes (NB) against K-Nearest Neighbors (KNN) and tree-based (J48) methods. The outcomes demonstrated that the NB method's performance was still slightly lower than the other two methods, with a difference of 1.2% in accuracy. Here are the results from the Naive Bayes model regarding the measurement matrix used: precision 92%, recall 92%, f1-score 92%, ROC 0.97, and MCC 0.83. However, the characteristics of the data were not explained in detail, and the problems in the data needed to be explained.

Patidar et al. [146] compared various machine learning algorithms, including RF, Decision Tree, and K-Nearest Neighbors (KNN), with Gaussian Naïve Bayes for predicting heart disease. The study revealed that the Gaussian Naïve Bayes method achieved the lowest prediction accuracy, with a result of 74.63%. Meanwhile, the precision, recall, and f1-score values are 68.34%, 92.23%, and 78.51%, respectively. The accuracy results of Naive Bayes were still lower than the RF algorithm. The dataset used does not have missing values. Therefore, the preprocessing steps involved one-hot encoding to convert categorical data into numerical form. Additionally, the data was scaled using StandardScaler. However, the analysis of the results was less in-depth, only showing ROC-AUC results.

Chourib et al. [80] categorized stroke patient data using K-Nearest Neighbors (KNN), Naïve Bayes, Decision Tree, and Random Forest algorithms. The study demonstrated that the random forest and decision tree methods achieved the most accurate prediction results compared to naive bayes. The accuracy score of the Naive Bayes model was 82%, while that of the Random Forest model was 85%. Meanwhile, the measurement values of RMSE and f1-score for the Naive Bayes model were 0.91 and 0.9, respectively.

The study handled dataset preprocessing by inputting missing values with a value of 0, removing incorrect data, normalizing the data, and implementing several feature selection techniques such as chi-squared, RFE, and tree-based methods. This algorithm exhibits a weakness, specifically in the assumption of attribute independence. The feature selection technique that yielded the highest prediction measurement values was based on a tree-based method. This assumption introduces a bias that becomes problematic when encountering issues in the dataset, potentially leading to lower prediction accuracy. However, the study focused on examining the influence of implementing various feature selection techniques.

D. FUZZY ALGORITHM

The Fuzzy theory can encompass inaccuracies or linguistic ambiguities in statements [147]. Fuzzy algorithms typically manifest as IF-ELSE constructs to ascertain the degrees of membership to specific linguistic categories. For instance, when a resulting value of 36.54 falls within a medium or heavy classification, fuzzy algorithms can ascertain how much the value leans towards the medium or heavy category.

Jaiswal Sushima performed detection of diabetic patients using fuzzy Mamdani [148]. The fuzzy algorithm was superior to several algorithms, such as SVM, MLP, and NB. However, the fuzzy mamdani F-Measure results were still lower than J48.

Some other studies that predict diabetes using fuzzy do not fully use fuzzy but are combined with other machine learning algorithms such as SVM [149], neural network [150], and optimization [56]. Praveen et al. [151], predicted kidney failure by utilizing a Neuro-fuzzy approach, achieving a predictive result with an accuracy of 97%, 94% of precision, 96% of specificity, 94% of recall, and 96% of F1-score. In that study, the preprocessing of the data only involved feature selection using the Random Forest model. However, this research did not explain the fuzzy rules created.

Manur et al. [152] harnessed big data for heart disease prediction through a hybrid approach combining fuzzy logic with Deep CNN, achieving a commendable accuracy rate of 95.26% and 92% of f1-score. The preprocessing steps to clean the dataset included removing missing values and noise, and normalization using min-max. However, this study did not mention the fuzzy rules to predict heart disease. This algorithm is seldom employed in isolation for disease prediction, as its standalone usage often yields shallow prediction outcomes.

Nguyen et al. [153] Proposed an attentive hierarchical adaptive neuro-fuzzy inference system (AH-ANFIS) that combined fuzzy inference in a hierarchical architecture with attention for selecting the essential rules. The proposed model benefited from the rule-based structure of ANFIS that enabled the user to interpret the abstractions of hidden layers. The accuracy results obtained were 79.69%. The research also compares with several algorithms, such as SVM and CNN. AH-ANFIS results are superior to the two comparison algorithms. However, the fuzzy rules used are still native to the fuzzy algorithm.

E. REGRESSION

Regression methods are tailored to forecast continuous numerical outputs by establishing sequence relationships [154]. The domain of statistics has extensively explored regression techniques. Within the realm of disease prediction, familiar methodologies encompass logistic regression and linear regression, commonly employed for forecasting the occurrence of chronic disease. This algorithm exhibits a vulnerability when confronted with imbalanced data. In such cases, underfitting can transpire, leading to diminished prediction accuracy.

Joshi and Dhakal [155] predicted diabetes using the logistic regression method and achieved a prediction result with a accuracy of 78.26% and a cross-validation error rate of 21.74% was observed. The data preprocessing involved imputing missing values using the median value. However, in some studies related to chronic disease prediction, regression methods are still inferior to the accuracy produced compared to ensemble methods (random forest and XGBoost) [71], [156], [157].

Mienye and Sun [156] analyzed an imbalanced cost-sensitive method using several supervised learning algorithms such as logistic regression and XGBoost. The results obtained by the logistic regression algorithm were 9% lower than the XGBoost algorithm. In this study, the proposed model was applied to several datasets including the Pima Indian Diabetes dataset, Chronic Kidney Disease dataset, and Breast Cancer dataset. The ROC results obtained using the cost-sensitive logistic regression model on the datasets used were as follows: for the Pima Indian dataset, it was 0.8; for the Chronic Kidney dataset, it was 1.0; for the Breast Cancer dataset, it was 0.83; and for the Cervical Cancer dataset, it was 0.98. This research focused more on the application of the imbalanced data method using cost sensitive.

Assegie et al. [157] applied several classification algorithms to see the performance of the most superior algorithm. The classification algorithms used were SVM, DT, RF, and logistic regression. The results showed that the logistics algorithm still needed to be improved for DT and RF. The logistic regression model yielded an accuracy of 85.8%, an AUC of 0.88, and a precision of 94%. Meanwhile, the Random Forest model resulted in an accuracy of 99%, an AUC of 1.0, and a precision of 100%. However, the results were analyzed by looking at the application of the RFE for feature selection technique.

F. SVM

SVM is one of the best machine learning algorithms primarily used for pattern recognition [158]. This algorithm works based on the hyperplane in the data, where the best results are obtained to find the optimal hyperplane. Tyagi and Mittal [89] employed SVM to classify diabetic patients, achieving an impressive accuracy of 89%, sensitivity of 99%, specificity 73%, precision 98%, and negative predictive value (NPV) 82%. However, it is essential to note that the high accuracy was influenced by how imbalanced data was handled in the study. Among the various imbalanced techniques applied, ADASYN yielded the best results in terms of improving the imbalanced ratio.

In the study, SVM kidney disease prediction for performance results was still inferior to random forest [159]. SVM with Recursive Feature Elimination with Cross-Validation (RFECV) using 9 features resulted in an accuracy of 95.5%, precision of 98.7%, recall of 92.2%, f1-score of 95.3%, sensitivity of 92.2%, and specificity of 98.8% for binary classification cases. However, this study focused on the impact of using two feature selection techniques: RFECV and univariate feature selection (UFS). The best feature selection technique yielded was RFECV.

Kumar et al. [160] utilized SVM for cancer prediction, particularly in high-dimensional data. Several kernels available in the SVM model were used to determine their best performance. The results indicated that the best kernels for predicting cancer cases were the radial basis function (RBF), polynomial, and linear kernels. However, the sigmoid kernel was less suitable for this case. The predictive values for the three best kernels were accuracy 74%, precision 55%, recall 74%, and f1-score 63%. This study implemented feature selection but did not mention which feature selection technique was used and what the results were.

In heart disease prediction, the results of the SVM method are still below the random forest and decision tree [146]. The prediction results using the SVM model were as follows: accuracy 87.31%, precision 84.07%, recall 92.23%, and f1-score 87.96%. The accuracy results of Naive Bayes were still lower than the RF algorithm. The dataset used does not have missing values. Therefore, the preprocessing steps involved one-hot encoding to convert categorical data into numerical form. Additionally, the data was scaled using StandardScaler. Examining prediction outcomes from prior studies reveals a recurring trend wherein SVM consistently lags behind other algorithms. This algorithm's inherent limitation becomes evident when tasked with predictions involving datasets characterized by abundant features or high-dimensional spaces.

G. KNN

KNN is an algorithm for identifying unknown categories of data points based on their nearest neighbors of known class [161]. These rules are commonly used for pattern recognition [162], text categorization [163], object detection [164], and event recognition [165].

Syed et al. [124] conducted cancer diagnosis using KNN compared to other machine learning algorithms such as SVM and MLP. The result accuracy of 0.76, precision of 0.75, recall of 0.94, and f1-score of 0.84 were obtained. However, the diagnostic accuracy achieved by KNN was comparatively lower than that of other algorithms. This discrepancy was addressed by applying imbalance oversampling techniques, leading to an improvement in accuracy. This is because the KNN algorithm has difficulties determining the relevant features in the data and determining the biased value of K.

Thummala et al. [166] compared the KNN and Naïve Bayes methods to determine their effectiveness for heart disease prediction. The Naïve Bayes method outperformed the KNN method with a superiority of 3% in accuracy. The precision values generated by the KNN model were 79%, while the Naive Bayes model achieved 82%. The study focused on measuring precision results. The precision results indicate the presence of issues within the dataset used. However, this study did not address the dataset issues. Additionally, this research only compared the two machine learning algorithms used.

Gupta and Goel [167] conducted a study where KNN was utilized to predict diabetes mellitus. The research explored the optimal K value within the range of 1-200. The optimal K value was found to be within the range of 72-73. The performance metrics obtained from the optimal value of K were as follows: Accuracy 81.17%, Precision 84.21%, Recall 58.18%, Specificity 93.94%, F1 Score 68.82%, and Error Rate 18.83%. However, this research mainly emphasized data handling, specifically removing irrelevant features using ANOVA.

Majid and Utomo et al. [168] also compared the KNN method with the AdaBoost ensemble technique in diabetes prediction. The accuracy result obtained for the method KNN + Discretization + Adaboost was 83.18%, while the accuracy value generated by using only the KNN method was 73.2%. However, the results obtained for the KNN algorithm are still 3% lower than the Decision Tree algorithm. In addition, this research applies discretization and the AdaBoost algorithm to combine with KNN. The results indicated that the AdaBoost method achieved higher accuracy compared to KNN.

H. GENETIC ALGORITHM

Genetic algorithm (GA) is a method based on population genetics using adaptive heuristic search [172]. The GA algorithm performs probabilistic searches based on natural selection and natural genetics. Komal [173] employed a combination of the Genetic Algorithm (GA) and the Random Forest algorithm for diabetes classification, achieving a significant accuracy of 94.5%, specificity of 92.4%, sensitivity of 90%, MCC of 0.886, and ROC of 0.874. However, the analysis in this study was limited to comparison without an in-depth analysis.

In predicting chronic disease, the GA algorithm is mostly hybrid with other methods such as ANN with a result accuracy of 80% [104]. The study focused solely on measuring accuracy results. Azad et al. [86] The prediction of diabetes mellitus was conducted using a hybrid model consisting of decision trees, SMOTE for handling imbalanced data, and GA for dimensionality reduction. The prediction results were an accuracy of 80.19%, error rate of 19.80%, precision of 77%, sensitivity of 85%, f1-score of 81%, and AUROC of 0.84. This study handling missing values and outliers, but did not specify the techniques used.

I. OPTIMIZATION

An optimization algorithm finds the value of x such that it produces the smallest or largest possible value of f(x) for a given function [174]. The optimization method researchers often use to predict chronic disease is PSO. Most researchers also perform optimizations based on single machine learning algorithms, such as Optimized

Gaussian Naïve Bayes (OGNB) [175]. The OGNB classifier blends Gaussian naive Bayes, Adaboost, and the random search method to construct a proficient classifier that optimizes prediction scores while mitigating overfitting concerns.

Kumari and Ahlawat [175] study also handled data preprocessing, including imputing missing values using regression, normalization using min-max, and Sequential Backward Feature Elimination (SBFE). The study analyzed the implementation of regression values as missing value imputation, where the application of regression values resulted in an increase in accuracy by 1.6% to 78.74%. However, sensitivity decreased by 6% to 67.87%, precision decreased by 4.4% to 64.89%, and f1-score decreased by 1% to 65.17%. The study also applied SBFE technique, resulting in an accuracy of 81.85%, sensitivity of 81.17%, precision of 81.46%, and f1-score of 72.47%. However, another factor in the dataset, namely imbalanced data, was not addressed in the study.

Ahlawat [176] conducted a study predicting diabetes using KNN optimized by searching for the optimal K value within the range of 1 to 100. Preprocessing of the data included removing outliers using z-score, imputing missing values using regression, normalization using min-max, and feature selection using correlation score. The prediction results obtained with the optimal K value of 23 include an accuracy of 92.28%, precision of 92.38%, recall of 92.36%, and f1-score of 92.36%. However, the study's process of finding the optimal K value was conducted through trial and error.

Ramalingaswamy et al. [177] classified diabetes using the model optimized radial basis function NN. The optimization was done by tuning the hyperparameters of the NN model such as setting activation functions, the number of hidden layers, and the weight between the hidden layer and the output layer. The prediction results yielded an accuracy, sensitivity, and specificity of 73.50%, 64.44%, and 78.75%, respectively. However, the study did not conduct an analysis regarding the issues arising from the medical dataset used.

Reddy et al. [178] introduced a unique approach to diabetes prediction by employing a hybrid method combining PSO with an Artificial Fish Swarm Algorithm. This study also performed data preprocessing, including missing value imputation using the mean and normalization using min-max. The outcome of this approach resulted in an impressive accuracy of 98.5%, sensitivity of 98.7%, specificity 83.2%, MCC 93.8%, and kappa statistics 0.967. However, this research focused on analyzing disease risk factors in the data.

El-Shafiey et al. [179] conducted heart disease prediction using a hybrid approach combining GA and PSO algorithms in conjunction with Random Forests. The predictive outcome of this hybrid method achieved an accuracy of 95.6%, precision of 97.44%, recall of 92.68%, and AUC 0.94. The preprocessing conducted in this study only involved normalization using min-max. However, GA and PSO algorithms focused more on selecting the best features. So, in this research, the main contribution was determining the best features.

V. ISSUES ON PREDICTION RESULT

Based on the discussions from RQ2 about what approaches to use for dealing with issues in medical data and RQ3 about which machine learning models are used in predicting chronic diseases, this section will discuss which data problems impact prediction results. Additionally, it will explore what types of machine learning models can improve predictions for chronic diseases.

A. ISSUES RELATED TO DATA

Research into predicting long-term health issues often relies on medical information that's available to the public. This data typically comes from places like the UCI repository and includes sets like the Pima Indian data for diabetes, NHANES for hypertension dataset, microarray for cancer, breast cancer wisconsin dataset, cleveland and statlog dataset for heart disease, stroke event dataset, and chronic kidney dataset. Some of these datasets are easy to find on a popular website called Kaggle (https://www.kaggle.com/datasets). Additionally, there are private medical sets gathered in labs or hospitals that can also be used for predictions. Both public and private medical data share common challenges [20], [77], like missing data, unusual values, choosing relevant information, making the data comparable, and dealing with data that's not evenly distributed. It's important to choose the right methods to address these issues when predicting diseases.

Based on the papers reviewed, data problems significantly impact prediction outcomes. Out of the five data issues discussed earlier, only normalizing the data doesn't affect prediction results. The four data problems that do impact predictions are imbalanced class, missing values, feature selection, and outliers. Among these, imbalanced data has the most significant influence [20], [97]. It disrupts the balance between training and testing data for machine learning algorithms, and addressing it can improve prediction results by 30% [99]. However, feature selection typically only slightly enhances predictions, usually by 2%-3% [68], [69]. Interestingly, using the principal component analysis model for feature selection decreased prediction results by 2% [76]. Regarding outliers, some studies handle them while others do not. Outliers in medical data are important because they might represent accurate values. Hence, dealing with outliers requires careful consideration.

Meanwhile, addressing the missing value problem can involve replacing the value or deleting the null row [63]. In some studies, the ratio of null values is relatively low (2%-5%), thus having minimal impact on prediction results [205]. However, when the ratio of missing values exceeds 80%, there hasn't been much research discussed on this issue. Therefore, in the future, if medical data with a missing value

TABLE 5. Detailed of machine learning algorithms for predicting each chronic disease.

Disease	Machine Learning Algorithm	Parameter Used	Reasons the Chosen Method
Diabetes Mellitus	Decision Tree and J48 [68], [86], [180], [181]	criterion: entropy	Providing information about the importance of each feature or variable in decision-making
	C4.5 [86], [181]	criterion: entropy	Reducing the risk of overfitting
	ID3 and CART [121], [180], [182]	criterion: gini	The ability to effectively handle categorical variables
	Random Forest [20], [53], [68], [71], [121], [173], [182], [183]	n_estimators: 300, criterion: entropy	Successfully applied in various disciplines for classification purposes
	AdaBoost [168]	n_estimators: 50	The generated predictions are accurate, and their implementation is straightforward
	XGBoost [121], [156]	n_estimators: 100, scale_pos_weight: 1	Having parameters scale_pos_weight can assist in addressing this class imbalance
	Fuzzy Logic [149], [150]	Mamdani Technique	Providing flexibility in describing complex relationships among variables
	SVM [71], [89], [149], [180], [181], [184], [185]	kernel: sigmoid	Providing a viable solution
	KNN [71], [167], [168], [181]	K: 1 to 200	The simplest classification method
	Logistic Regression [?], [71], [181], [186], [187]	solver: lbfgs	Modeling the relationship between a categorical response variable and covariates
	Deep CNN [188]	Optimizer: Adam, learning rate: 0.0001 to 0.01, max pooling	The nature of its architecture is compact and performing only 1D convolutions
	DNN [189]	Learning rate: 0.01, Epochs: 200, K-fold: 5, Hidden layer 1: units 10, Hidden layer 2: units 3, Optimizer: Adam, SGD, RMSprop, and Adagrad	Two hidden layer
	ANN [70], [190], [193]	Learning rate: 0.1, Layer: 4, Regularization: Lasso and Ridge	The architecture of ANN is highly flexible
	Naïve Bayes [71], [181], [193]	Posterior distribution	The most efficient and effective inductive learning for machine learning
	Genetic Algorithm [104], [173], [192]	Population: 50, crossover probability: 0.7, mutation probability: 0.1	Optimal parameter combination for predictive model
	Optimized NN [176], [177]	epoch: 400, learning rate: 0.01, hidden layer: 2	Flexible in tuning parameters
	Optimized Gaussian Naive Bayes [175]	Number of weak learners: 50, learning rate: 0.1, random state: 50	Chosen risk factors for the development of an improved diagnostic system
Cancer	Decision Tree [120], [196], [197]	Spliter: best, max_depth: none, criterion: gini	Handling a combination of numerical and categorical data, often encountered in medical data
	Random Forest [115], [120], [128], [196], [197]	Criterion: gini, max_depth: none, n_estimators: 150	Effectively addressing class imbalance through techniques such as weighted classes
	AdaBoost [120]	n_estimators: 50, learning_rate: 0.5, algorithm: SAMME	Addressing the issue of class imbalance by assigning greater weight to prediction errors on samples from the positive class
	XGBoost [113], [115], [120], [128], [195], [196]	n_estimators: 300, scale_pos_weight: 1, max_depth: 4, Sampling method: uniform, eta: 0.3, booster: gbtree	Provides numerous parameters that can be optimized
	Bagging [125]	n_estimators: 20, base_estimator: decision tree	Becoming more robust to outliers
	SVM [7], [40], [108], [120], [160], [171]	Kernel: rbf, gamma: 1, C: 10	Effective in high-dimensional feature space
	KNN [76], [108], [120], [124], [145], [198]	Algorithm: ball_tree, p:2, n_neighbors: 14	Capturing non-linear and complex patterns in data
	MLP [115]	Optimizer: Adam, learning rate: 0.001	Identifying complex patterns that may be associated with cancer
	CNN [96]	Learning rate: 0.01, Epochs: 100, Optimizer: Adam	Recognizing patterns and relationships among structures
	ANN [77]	Learning rate: 0.01, Regularization: Lasso and Ridge	The ability to automatically extract relevant features from data
	Naïve Bayes [71], [181], [193]	Posterior distribution	Requires minimal training data and exhibits high prediction speed
Stroke	Decision Tree [80], [199], [200]	max_depth: none, criterion: gini	Capable of handling mixed data types
	Random Forest [9], [63], [103]	Criterion: entropy, max_depth: 50, n_estimators: 150	More tolerant to overfitting compared to a single decision tree
	XGBoost [63]	max_depth: 1-10, gamma: [0, 0.4-1], min_child_weight: [1-6,8,10]	Boosting for accuracy prediction
	C4.5 [205]	Criterion: gini, max_depth: none, n_estimators: 150	Automatically identifying key risk factors associated with stroke
	CatBoost [10]	learning_rate: 0.03, class_weight: 1, iterations: 100, depth: 6	Utilizing the class weights parameter for handling imbalanced classes
	Logistic Regression [63]	C: [0.01, 0.1, 1, 10, 100], penalty: l1, l2	Modeling the relationship between a categorical response variable and covariates

TABLE 6. Detailed of machine learning algorithms for predicting each chronic disease (continued).

Disease	Machine Learning Algorithm	Parameter Used	Reasons the Chosen Method
Stroke	MLP [63], [202]	Optimizer: [sgd,Adam], learning rate: [constant,adaptive], activation: [tanh,relu], alpha: [0.0001,0.05], hidden layer size: [(2,5,10),(5,10,20),(10,20,50)]	have flexible architectures with multiple hidden layers and neurons
	CNN [78]	kernel: 3, activation: relu, convolution layer: 2, linear layer: 2, Features In/Out: 32/16 and 16/1	Capable of processing and leveraging large-scale data
	Deep CNN [141]	activation: relu, dropout: [32,32], learning rate: 0.5, batch size: 32	The ability to build complex feature hierarchies
	LSTM-RNN [203]	epoch: 10, activation: relu, learning rate: 0.1	The model enables the retention and comprehension of temporal patterns and relationships among data in a time sequence
	DNN [204], [206]	Learning rate: 0.001, loss function: entropy, neurons: 12, layer: 9, Optimizer: Adam	Effectively handling the presence of missing values in the data
	KNN [63], [80], [201]	algorithm: [auto, ball tree, kd tree], leaf size: [1,2,3,5], n_neighbor: 3,4,5,7,9,12,14,15,16,17, weight: [uniform, distance]	The simplest classification method
	Naive Bayes [80], [130], [201], [202]	var_smoothing: 0.0533	Having low time complexity and simplicity in terms of implementation
	SVM [63], [200]– [202]	Kernel: [rbf,linear,sigmoid], gamma: [0.01,0.001,0.0001], C: [0.001,0.01,0.1,10,25,50,100,1000]	Effective in high-dimensional feature space
Hypertension	Decision Tree [12], [21], [106], [209]	max_depth: none, number of tree: 100, criterion: gini	providing information about the importance of each variable in making predictions
	Random Forest [11], [102], [207]	Criterion: gini, max_depth: 5, n_estimators: 50	More tolerant to overfitting compared to a single decision tree
	XGBoost [11], [207]	max_depth: 5, number trees: 50	Handling correlated variables effectively and even identifying interactions among these variables adeptly
	C4.5 [11], [12]	Criterion: gini, max_depth: 5, n_estimators: 50	Equipped with built-in pruning strategies to prevent overfitting
	LightGBM [102], [114]	max_depth: 15, num leaves: 48, min child samples: 299, min child weight: 10, reg alpha: 2, reg lambda: 1, n_estimators: 1600, learning rate: 0.1	LightGBM efficiently handles categorical features without the need for one-hot encoding
	ANN [99], [210]	learning rate: 0.01, activation: relu, loss function: entropy, input nodes: 11, hidden nodes: 7, output nodes: 2	The ability to automatically extract relevant features from data
	MLP [12]	num of epoch: 500, learning rate: 0.3, momentum: 0.2	have flexible architectures with multiple hidden layers and neurons
	KNN [114], [207], [209]	weight: [uniform,distance], p:1-20, n_neighbors: 1-20	Capturing non-linear and complex patterns in data
	SVM [11], [207], [209]	Kernel: [linear,rbf], gamma: 1	Can handle complex and unstructured data effectively
	Naive Bayes [102], [207]	posterior distribution	Providing good results without requiring a very large amount of training data
	Logistic Regression [102], [207]–[209]	C: [10,100], penalty: l1, l2	The logistic regression exhibits resistance to outliers
Cardiovascular	Decision Tree [15], [127], [146], [157], [211]	max_depth: none, criterion: gini	Explaining the relationship between predictor variables and the target
	Random Forest [15], [127], [132], [146], [157], [211], [212]	Criterion: gini, n_estimators: 100	Effectively addressing the common issue of class imbalance often encountered in medical datasets
	AdaBoost [105], [126]	n_estimators: [10,50,100,200, learning_rate: 0.5, algorithm: SAMME	Adaboost can utilize relatively simple base models
	XGBoost [105], [127], [134], [211]	n_estimators: 300, scale_pos_weight: 1, max_depth: 4, min child weight: 5, gamma: 1.5, subsample: 1, colsample_bytree: 0.6	XGBoost has various parameters that can be optimized to enhance the models performance, such as the learning rate
	NN [15]	hidden layers: 100, activation: truncated linear transform, optimization: adam, epoch: 200	ANN can undergo automatic exploration of architecture and model parameter
	MLP [16], [211]	number of layers: [1,3,5], max units: [30,50,100], dropout: [0,0.5]	MLP can handle missing data effectively
	RNN [16]	hidden size: [30,50,100], num layers: [1,2], bidirectional: [true,false], dropout: [0,0.3], model: [GRU,LSTM,RNN]	RNN have a memory mechanism that allows them to remember past information and use it to make predictions

TABLE 7. Detailed of machine learning algorithms for predicting each chronic disease (continued).

Disease	Machine Learning Algorithm	Parameter Used	Reasons the Chosen Method
Cardiovascular	CNN [143]	kernel: 5, activation: relu	Capable of processing and leveraging large-scale data
	DNN [61]	Learning rate: 0.01, epoch: 40, Optimizer: Adam	Effectively handling the presence of missing values in the data
	KNN [15], [146], [166], [211], [212]	metric: minkowski distance, n_neighbor: 5, weight: distance	The strength to handle outliers is due to its non-involvement in assumptions about a specific data distribution
	SVM [15], [16], [146], [157]	Kernel: rbf, degree of polynomial: 3, cache size: 200mb	Effective in handling datasets with a smaller number of samples
	Logistic Regression [15], [16], [146], [157], [211], [212]	C: 100, optimization algorithm: broyden–Fletcher–Goldfarb–Shanno algorithm with limited memory, penalty: ridge	The logistic regression exhibits resistance to outliers
	Naive Bayes [15], [146], [166]	var_smoothing: 10^9	The Naive Bayes model is relatively simple and quick to train
	Genetic Algorithm [62], [179], [212]	number of generations: 100, crossover rate: 0.5, mutation rate: 0.07-0.08, population size: 50	Genetic Algorithm can assist in addressing the challenge of class imbalance by optimizing the selection of parameters and model structures
Kidney Failure	Decision Tree [13], [14], [26], [98], [217]	criterion: entropy	Decision tree can provide insights into the significance of each feature in the decision-making process
	Random Forest [13], [26], [81], [214], [217]	Criterion: entropy, n_estimators: 50, max depth: 2	Capable of handling imbalanced datasets effectively
	AdaBoost [26], [64]	n_estimators: 100, learning_rate: 0.5, algorithm: SAMME	Adaboost has an inherent property to reduce overfitting
	XGBoost [26], [81]	n_estimators: 100, scale_pos_weight: 1, max_depth: none	XGBoost is designed for handling large-scale data and parallelization
	J48 [65]	n_estimators: 100	J48 provides information about the importance of each feature in the decision-making process
	ANN [66], [98]	learning rate: 0.001, activation: relu, epoch: 200, momentum: 0.9	ANN are adaptive and can model complex data distributions
	MLP [26]	Optimizer: Adam, activation: relu, hidden layer: 28	MLP can exhibit robustness to noisy data, which might be present in medical datasets
	SVM [13], [26], [81], [98], [159], [214], [215]	Kernel: rbf, gamma: 3, C: 1	SVM are suitable for datasets with a relatively small to medium-sized number of samples
	Naive Bayes [14], [65]	posterior distribution	Naive Bayes can perform well even with a large number of features
	KNN [65], [98], [215]	n_neighbor: 1-9, weight: distance	KNN adapts well to changes in data or distribution
	Logistic Regression [81]	C: 1000, penalty: ridge	Providing output in the form of predictive probability
	Fuzzy Logic [151]	Mamdani Technique	Easily interpretable by humans as fuzzy rules can be formulated in natural language

ratio of over 80% emerges, it will require a different approach than current techniques.

B. ISSUES RELATED TO MACHINE LEARNING ALGORITHMS

In addition to data-related challenges, other studies suggest that optimizing or modifying the machine learning algorithms used can also improve prediction outcomes. For example, Cai et al. [115] predicted cancer using an optimized naïve Bayes algorithm, achieving an AUC of 0.98, accuracy of 0.98, and specificity of 0.97. Kumar et al. [128] optimized an ensemble stacking model for cancer prediction, resulting in an impressive accuracy of 99.45%, precision of 99%, recall of 98%, and F1-score of 99%. Manur et al. [152] explored heart disease prediction by combining fuzzy logic and Deep CNN models, achieving a commendable accuracy rate of 95.26% and an F1-score of 92%.

Selvi et al. [170] conducted diabetes prediction research by employing a hybrid model that combined the improved k-means clustering technique with an ensemble method (gradient boosting tree). Improved K-means clustering introduces a new mechanism where the seed value is determined based on the minimal clustering error (CE). The gradient boosting tree model utilizes functions as parameters, making it challenging to enhance using conventional optimization techniques in Euclidean space. The model consistently outperforms other methods, achieving impressive precision of 99.23%, recall of 97.48%, accuracy of 97.79%, F1-score of 98.34%, and kappa value of 95.02%.

El-Shafiey et al. [179] proposed a hybrid model combining genetic algorithm with particle swarm optimization was employed. Both algorithms were utilized to determine the best features for classification. Following the application of these algorithms, a set of 9 features out of the original 13 was selected for use. Additionally, the classification model employed in this study was random forest. The predictive performance of this hybrid approach yielded an accuracy of 95.6%, precision of 97.44%, recall of 92.68%, and an AUC of 0.94.

Shrestha et al. [185] conducted the SVM model was modified to enhance its optimally, resulting in a model called Sparse Balance (SB)-SVM. The proposed solution

TABLE 8. Comparison of strength and weakness machine learning algorithms in chronic disease prediction.

Category	Method	Strength	Weakness	Paper
Supervised Learning	Decision Tree	The data preprocessing was simple	Overfitting on complex training data	[180]
	KNN	Simple and easy to implement	Susceptible to noise and outliers	[181]
	Naive Bayes	Resistant to data loss	Attributes are mutually independent	[184]
	SVM	Effective in high-dimensional features	Determining the correct kernel	[185]
	Fuzzy Logic	Modeling is intuitive	Overfitting when the rules defined are too complex	[150]
	Logistic Regression	Suitable for binary classification	Sensitive to outliers	[186], [187]
Ensemble Learning	Random Forest	Overcomes overfitting and can handle imbalance data	High time computing	[182], [183]
	Bagging	Prediction results are more stable	Higher computing	[125]
	Stacking	Effective in handling datasets with complex characteristics and patterns	High computation and overfitting	[128]
	Voting	Stable and robust	Giving equal weight to the base model	[124], [127], [197]
	XGBoost	Ability to handle missing values and imbalanced data	Hyperparameter setting and overfitting	[134]
	AdaBoost	Effective in handling imbalanced data	Sensitive to noise and outliers	[168]
Deep Learning	ANN	Able to overcome noise and incomplete data	Large amounts of data and hyperparameter settings	[139]
	CNN	Automatic feature extraction	Requires large training data and computational time	[143]
	DNN	Good understanding of patterns	Requires large training data and computational time	[189]
	MLP	Able to overcome classification problems in data processing	Requires large data and non-trivial hyperparameter settings	[160]
Reinforcement Learning	Genetic Algorithm	Search a wide solution space	Computational time and does not guarantee an optimal solution	[173]
	PSO	Ability to converge to an optimal solution	Sensitive to initial configuration	[173]
	AFS	Efficient in optimization problems	Does not guarantee an optimal solution	[178]

addresses the shortcomings of longitudinal and sparse data. Additionally, mean imputation will be employed to handle missing data by replacing missing values with the mean of each variable. Researcher utilize SB-SVM to enhance the interpretability of the model and address imbalanced data. The predictive performance metrics for the model are as follows: accuracy of 76.36%, precision of 66.86%, recall of 76.74%, AUC of 0.85. This research enhances the accuracy by 9.14%, precision by 3.93%, recall by 6.78%, and AUC by 0.16 compared to the state-of-the-art solution.

Chen et al. [204] proposed a novel Hybrid Deep Transfer Learning-based Stroke Risk Prediction (HDTL-SRP) scheme to exploit the knowledge structure from multiple correlated sources (i.e., external stroke data, chronic diseases data, such as hypertension and diabetes). The proposed HDTL-SRP framework operated in a distributed manner, eliminating the need to directly share patients' records between hospitals. It consist of four components:

1) Generative Instance Transfer (GIT): Utilizing GAN in external data to generate synthetic instances for model training.
2) Network Weight Transfer (NWT): Utilize data from highly correlated diseases such as hypertension or diabetes.
3) Bayesian Optimization (BO): Employed to identify the best transferred parameters.
4) Active Instance Transfer (AIT): Selecting informative synthetic stroke instances to create a balanced stroke dataset, subsequently used to fine-tune the SRP model.

The predicted outcomes of the proposed model were an accuracy of 78.7%, recall of 67.8%, f1-score of 77.2%, and an AUC of 0.84.

Chang et al. [218] conducted hybrid XGB-SVM model was proposed to predict whether hypertensive patients would develop hypertensive heart disease within three years. The hybrid model utilizes the XGB method as a feature converter to construct new feature combinations for training the SVM model. This study employs the AUC and INE (Improved Normalized Entropy) metrics for measurement. The AUC value is 0.93 and the INE value is 0.895. Both values are superior compared to other single models such as RF, XGBoost, RF+SVM, and Gradient Boosting DT. In model evaluation, a smaller INE value and a larger AUC value indicate better model performance. This approach can help alleviate the psychological, physiological, and economic burden associated with the disease.

Research that explores hybrid or optimized machine learning algorithms aims to develop more robust models for chronic disease prediction. However, only a few studies have focused on hybrid or optimized models, specifically addressing significant factors that affect prediction outcomes when handling preprocessing datasets [77], [219]. This highlights the importance of preparing datasets to be clean and suitable for analysis, as it significantly impacts disease prediction outcomes, particularly in managing feature selection and imbalanced data [20].

VI. IMPLICATIONS AND APPLICATIONS

In this section, we will discuss the ethical implications of machine learning for use in healthcare services, which include several topics such as privacy of patient data, potential algorithmic bias, the relationship between prediction results and prevention of chronic diseases, and the practical application of machine learning in healthcare services for early intervention. To address these ethical implications, collaboration among experts in ethics, law, technology, and healthcare is crucial. By considering these aspects, the use of machine learning in healthcare services can provide significant benefits to patients while ensuring that ethical principles and fairness are thoroughly maintained.

A. PRIVACY OF PATIENT DATA

Here are some key points in discussing the ethical implications in the privacy of patient data section:

1) PATIENT DATA PRIVACY PROTECTION

ensuring patient data privacy is crucial when using machine learning in healthcare services. Patient medical data is highly sensitive and must be tightly guarded. This includes information like medical history, lab test results, medical records, and personal identification information. Misuse or violation of patient data privacy can have serious consequences for patients, including personal information misuse, discrimination, or even physical security risks.

2) TRANSPARENCY AND CLARITY

the use of machine learning in healthcare services should be supported by adequate transparency and clarity for patients. This means patients should be given clear information about the types of data collected, how the data will be used, who will access the data, and how the data will be stored and protected. Patients should give appropriate consent before their medical data is used in analysis or machine learning model development.

3) ANONYMIZATION

in situations where patient data needs to be used to train or test machine learning models, anonymization steps should be taken to protect patient privacy. This involves removing or replacing personal identification information so individuals cannot be directly identified from the data. However, it's important to note that these techniques may not always guarantee data security perfectly, especially when data is combined with other data sources or when there is indirectly identifiable sensitive information.

4) STRONG DATA SECURITY

implementing strong data security measures is essential to protect patient data from unauthorized access or security breaches. This includes data encryption, limited access only to authorized personnel, strong authentication systems, data access monitoring, and other technical security measures. Additionally, clear protocols should be in place to handle data security breaches if they occur, including notifying affected patients and complying with applicable data privacy regulations.

5) MONITORING AND AUDITING

the use of machine learning in healthcare services should be continuously monitored and audited to ensure compliance with data privacy principles. This involves routine evaluation of the algorithms and processes used, as well as identifying and addressing potential privacy breaches or algorithmic biases.

B. POTENTIAL ALGORITHMIC BIAS

Algorithmic bias can occur when machine learning algorithms make inaccurate or unfair decisions or recommendations because they are based on unrepresentative data or contaminated by certain assumptions or preferences. Here are some potential implications of algorithmic bias in machine learning for healthcare services:

1) INCORRECT DIAGNOSIS

if machine learning algorithms are used to support disease diagnosis, algorithmic bias can lead to errors in recognizing symptoms or specific risk factors in patients. For example, if the algorithm is only trained using data from a specific population that does not reflect demographic or genetic diversity comprehensively, then the algorithm may not be able to recognize relevant symptoms or risk factors in patients from different backgrounds.

2) SUB-OPTIMAL TREATMENT

machine learning algorithms can also be used to provide recommendations related to necessary treatments or interventions for patients. However, if the algorithm does not consider the individual needs of the patient accurately, then the recommendations provided can be sub-optimal or even counterproductive. For instance, if the algorithm tends to provide treatment recommendations based on data from a specific population without considering the health conditions or preferences of individual patients, then this can result in treatments that do not meet the patient's needs.

3) INEQUALITY IN ACCESS TO CARE

Algorithmic bias can also result in inequality in access to healthcare. If machine learning algorithms tend to provide recommendations or make decisions that benefit one demographic group while neglecting the needs or preferences of other groups, then this can lead to inequality in access to healthcare. For example, if the algorithm tends to prioritize treatment or interventions for patients from a certain group while neglecting patients from other groups, then this can exacerbate disparities in healthcare.

To address the potential implications of algorithmic bias in machine learning in healthcare services, several steps can be taken, including:

1) Use of Representative Data: Ensuring that the data used to train machine learning algorithms reflects population diversity comprehensively, including demographic, ethnic, and geographic diversity.
2) Regular Evaluation: Conducting regular evaluations of the performance of machine learning algorithms to identify and address biases that may arise.
3) Transparency and Accountability: Ensuring transparency in the use of machine learning algorithms in healthcare services and being accountable for the consequences of the decisions or recommendations generated by these algorithms.
4) Education and Awareness: Increasing awareness and understanding of the potential algorithmic bias among healthcare professionals and other relevant stakeholders to ensure that decisions are based on accurate and reliable information.

C. THE RELATIONSHIP BETWEEN PREDICTION RESULT AND PREVENTION OF CHRONIC DISEASES

The use of machine learning in healthcare services also has a significant impact on preventing chronic diseases. By utilizing algorithms that can quickly and accurately analyze patient data, healthcare professionals can identify risk factors for chronic diseases earlier and provide timely interventions to patients. However, it is important to note that accuracy in predictions does not always equate to effectiveness in disease prevention. There are ethical aspects to consider regarding how these prediction results are used. For example, if prediction results indicate that someone is at high risk of developing a chronic disease, but the recommended interventions are financially unaffordable for the patient, this can lead to issues of unequal access to healthcare. Therefore, it is important for developers and users of machine learning systems to consider not only the accuracy of predictions but also the social and economic implications of the recommended interventions.

D. THE PRACTICAL APPLICATION OF MACHINE LEARNING IN HEALTHCARE SERVICES FOR EARLY INTERVENTION

One of the main benefits of machine learning in healthcare services is its ability to detect symptoms of diseases or health risks early on. For instance, algorithms can be used to analyze patterns of symptoms or biomarkers that may indicate the development of certain diseases in patients. By detecting symptoms or health risks early, healthcare professionals can provide early interventions, which in turn can improve patient prognosis and reduce the burden of chronic diseases in the long term. However, in implementing machine learning for early interventions, it is important to strike a balance between clinical benefits and adherence to patient ethics and privacy. It is important to ensure that patients provide appropriate consent for the use of their data in analysis and that adequate data security measures are implemented to protect their privacy.

VII. OPEN ISSUES AND FUTURE WORK

Based on the review of articles related to chronic disease prediction, a gap opportunity was identified for future work. This gap opportunity consists of two aspects: preprocessing data handling and machine learning model. In this section, we will discuss opportunities related to both aspects, this also addresses RQ4.

A. PREPROCESSING DATA HANDLING

Some future work on the data issues that can be developed is applying a collection of other disease datasets with broader features or variables and applying the other techniques for outlier detection [21]. Improving the missing value algorithm to predict different types of disease [53]. Applying SMOTE and ensemble techniques to big-data analysis [64]. Systematically set hyperparameters and address the problem of imbalanced datasets [66].

Creating new datasets with more valuable characteristics and more individuals allows for better generalization of learning [74]. The derived feature techniques are used for other diagnoses in the health domain [75]. Integrating electronic record datasets with background knowledge about various diseases [78].

The data used cover different geographical areas and origins. It would be interesting to examine how the model behaves with a mixed group of men and women [92]. Learning more factors influencing hypertension. According to the literature, family history and risk factors such as smoking and alcohol abuse are directly related to hypertension. Additionally, it uses population health data to update this information [114]. Check whether the base classifier is optimal and attempt feature engineering [115].

Extending the model to handle different types of data using other feature selection methods, investigating the role of machine learning models in handling time series data, and applying classification with deep learning algorithms to study improving classification accuracy [125]. Applying alternative methods to handling imbalanced data [129]. Applying semantic similarity measures to include similarity information between features (for example, using latent semantic indexing [129].

Collecting new CT scan data [130]. It was conducting hyper-parameter tuning and feature selection [146]. Applying other combinations of feature selection, re-sampling, and cost-sensitive learning methods [156]. Explore the side effects that can cause diabetes [178]. Evaluate the model using other datasets and risk factors that influence chronic disease [182]. Studying the power of genetic programming on incomplete dataset [199]. Collecting a new data set originating from hospitals [202]. Carry out feature importance analysis to determine control indicators [206].

TABLE 9. List of glossary on prediction of chronic disease.

Abbreviation	Alias	Description
AdaBoost	Adaptive Boosting	Popular ensemble learning technique in machine learning that combines multiple weak classifiers to create a strong classifier
AdaGrad	Adaptive Gradient Algorithm	Gradient-based optimization algorithm commonly used in machine learning for training models
Adam	Adaptive Moment Estimation	Optimization algorithm commonly used in machine learning for training deep neural networks
ANN	Artificial Neural Network	Computational model inspired by the structure and function of the human brain, consisting of interconnected nodes called neurons that process and transmit information
AFS	Artificial Fish Swarm	Computational algorithm inspired by the collective behavior of fish schools for solving optimization problems in various domains
ANOVA	Analysis of Variance	Statistical method in feature selection used to compare means of two or more groups to determine whether there are statistically significant differences between them
ADASYN	Adaptive Synthetic Sampling	Generates synthetic samples for the minority class by adaptively adjusting the balance between the classes based on their densities
Bagging	Bootstrap Aggregating	Ensemble learning technique in machine learning that improves the stability and accuracy of models by training multiple base models on different subsets of the training data using bootstrapping and then combining their predictions through averaging (for regression) or voting (for classification)
CatBoost	Category Boosting	Typically involve iteratively training weak models (e.g., decision trees) that focus on predicting specific categories or levels of the target variable
CART	Classification and Regression Tree	Predictive modeling technique that uses binary tree structures to recursively partition the dataset into subsets based on the value of input features, ultimately resulting in either classification or regression outcomes for each leaf node
CBFS	Correlation Based Feature Selection	Method used in machine learning to identify and select the most relevant features for improving the performance of predictive models
CNN	Convolutional Neural Network	Deep learning model commonly used for image recognition and classification tasks
DNN	Deep Neural Network	Type of artificial neural network that consists of multiple hidden layers, enabling it to learn complex patterns and relationships in data
DNA	Deoxyribose Nucleic Acid	Molecule that carries the genetic instructions for the development, functioning, growth, and reproduction of all known living organisms and many viruses

TABLE 10. List of glossary on prediction of chronic disease (continued).

Abbreviation	Alias	Description
GBM	Gradient Boosting Machine	Machine learning algorithm that builds predictive models by combining the predictions of multiple weak learners, typically decision trees, in a sequential manner while minimizing the loss function using gradient descent
GRU	Gate Recurrent Units	Recurrent neural network (RNN) architecture that is designed to efficiently capture long-range dependencies in sequential data by utilizing gating mechanisms to control the flow of information within the network
IQR	Inter Quartile Range	statistical measure that represents the range between the first quartile (25th percentile) and the third quartile (75th percentile) of a dataset
IIQR	Improve Inter Quartile Range	Increasing the sample size or using more robust statistical methods may also help improve the accuracy and reliability of the interquartile range estimation
KNN	K-Nearest Neighbors	Non-parametric machine learning algorithm used for classification and regression tasks
LASSO	Least Absolute Shrinkage and Selection Operator	Regression analysis method that performs both variable selection and regularization to improve the prediction accuracy and interpretability of the statistical model
LR	Logistic Regression	Statistical model used for binary classification tasks, where the output variable takes on two possible outcomes
MCC	Matthew Correlation Coefficient	Measure of the quality of binary classifications, taking into account true and false positives and negatives. It ranges from -1 to 1, where 1 indicates perfect prediction, 0 indicates random prediction, and -1 indicates total disagreement between prediction and observation
MLP	Multi-Layer Perceptron	Artificial neural network consisting of multiple layers of nodes, each connected to the nodes in the adjacent layers
MRMR	Minimum Redundancy Maximum Relevance	Select a subset of features that have the highest relevance to the target variable while minimizing redundancy among the selected features
NB	Naïve Bayes	Probabilistic machine learning algorithm based on Bayes theorem, often used for classification tasks, assuming independence among features
MCMC	Markov Chain Monte Carlo	Computational technique that uses Markov chains to sample from complex probability distributions, commonly employed in Bayesian statistics and machine learning

B. MACHINE LEARNING MODEL

Some future works on machine learning algorithm issues include combining various prediction algorithms to increase prediction accuracy [61]. Predicting survival duration using the regression model [66]. The other research is developing real-time and precise AdaBoost models for other disease detection [69]. Predict whether someone's with risk factors

TABLE 11. List of glossary on prediction of chronic disease (continued).

Abbreviation	Alias	Description
NCL	Neighborhood Cleaning Rule	Data preprocessing technique that removes noisy instances from imbalanced datasets by considering the majority class instances in the neighborhood of minority class examples
ODNN	Optimal Deep Neural Network	Model architecture and parameters that achieve the best performance for a specific task, balancing complexity and accuracy based on the available data and computational resources
PSO	Particle Swarm Optimization	Population-based optimization algorithm inspired by the social behavior of birds flocking or fish schooling
PCA	Principal Component Analysis	Dimensionality reduction technique used to transform high-dimensional data into a lower-dimensional space while preserving most of the variability in the data
RDA	Red Deer Algorithm	The algorithm simulates the mating behavior of red deer, where stags compete to find mates by balancing between exploration and exploitation of the search space
RepTree	Reduced Error Pruning-Tree	Technique used in decision tree algorithms to improve their generalization performance and avoid overfitting
RF	Random Forest	Improves on decision tree performance by introducing randomness during the construction of each tree and combining the predictions of multiple trees to reduce overfitting and improve accuracy
RFE	Recursive Feature Elimination	Feature selection technique used in machine learning to select the most important features from a given dataset
RMSprop	Root Mean Squared Propagation	Mathematical calculation used to measure the average magnitude of a set of values
ROS	Random Oversampling	Instances from the minority class are randomly duplicated or replicated to increase their representation in the dataset and balance the class distribution
RUS	Random Undersampling	Instances from the majority class are randomly removed or downsampled to reduce their representation in the dataset and balance the class distribution
SGD	Stochastic Gradient Descent	Particularly useful when dealing with large datasets or complex models, as it allows for efficient training by processing data in mini-batches rather than all at once
SVM	Support Vector Machine	Handle both linear and non-linear classification problems by using different kernel functions
SMOTE	Synthetic Minority Oversampling Technique	Specifically, for each minority class instance, SMOTE identifies its k nearest neighbors in feature space and creates synthetic instances by randomly selecting one or more of these neighbors and interpolating between them

TABLE 12. List of glossary on prediction of chronic disease (continued).

Abbreviation	Alias	Description
SNE	Stochastic Neighbor Embedding	Dimensional reduction technique used for visualizing high-dimensional data in a lower-dimensional space, typically two or three dimensions, while preserving the local structure of the data as much as possible
SS	Stability Selection	This technique is particularly useful in high-dimensional datasets where the number of features is much larger than the number of samples, and it can be applied with various machine learning algorithms and model types
UCI	Repository Irvine	This repository is a collection of datasets that are widely used for machine learning research and experimentation
WGAN	Wasserstein Generative Adversarial Networks	Variant of Generative Adversarial Networks (GANs), which are a class of deep learning models used for generating synthetic data that resembles real data
XGBoost	Xtreme Gradient Boosting	XGBoost builds an ensemble of decision trees sequentially, where each tree corrects the errors made by the previous trees. It uses gradient boosting techniques to optimize a specific objective function, such as mean squared error for regression problems or log loss for classification problems

for chronic kidney failure such as diabetes, hypertension, and a family history of kidney failure suffers from chronic kidney failure in the future or not using the appropriate dataset [81].

Ensemble model (NB + LR) to be integrated with the web application. The web application (user interface) will be developed so that the user can input the attributes required for heart disease prediction to help doctors as well [105]. Evaluating deep learning-based models for stroke [130]. Build a real-time cardiac diagnosis model by combining several machine learning models and incorporating the latest data preprocessing techniques [134]. Using the ensemble-based method [146].

Implement a deeper hierarchical fuzzy inference system that integrates expert knowledge for expert systems [154]. Applying ANOVA and KNN models for bigger data sizes [167]. Combination of other classification algorithms with genetic algorithms to produce high accuracy [173]. Make comparisons with other prediction models [182]. Adopt embedded method to extract intelligent patterns [186]. It is building a model to imagine a stroke in one day with the aim of doctors being able to treat their patients as early as possible [205].

Expanding the prediction model by collecting data that includes more influencing factors such as family history, smoking, and alcohol [210]. Shows the possibility of someone's suffering from the disease using a machine learning algorithm [212]. Shows the stage of the disease that has been reached [216]. Uses predictive proposed models (hyrbid XGB-SVM) for other problems in disease [218].

C. EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI)

On the other hand, XAI is also an issue that is currently the subject of much research, particularly in the field of chronic disease prediction. One of the goals of XAI is to understand the mechanisms behind a machine learning prediction model to create reliable, non-invasive, and advanced prediction tools for doctors [63]. Some future work on XAI issues includes using techniques other than Shapley Additive Explanations (SHAP) to measure the significance of individual features by assessing their average marginal contribution across all potential combinations [63].

Evaluating the performance characteristics of XAI approaches such as Local Interpretable Model-Agnostic Explanations (LIME) and SHAP for other relevant datasets [121]. Additionally, enhancing model explainability can be attempted using different machine learning algorithms to develop various types of ensemble models [121]. Further independent studies should be conducted to test the performance of the Bayesian Optimization-TabNet model, which can be generally explained in predicting diabetes.

Using XAI for attribute reduction and interpreting the final prediction system. Addressing issues in finding samples and health standards for real-time datasets, and also evaluating ensemble models using the XAI framework. XAI can be valuable in medical practice, helping stakeholders accept recommendations to reduce patient readmission and save public healthcare costs in the future. Other researchers can use the proposed XAI framework, such as feature importance and LIME, for predicting other diseases like heart disease. Applying the XAI frameworks Eli5 and LIME for predicting diseases other than stroke.

Based on the future work outlined in previous research, there are still significant opportunities for conducting further research related to the prediction of chronic diseases. This includes creating new datasets directly from hospitals, exploring alternative approaches to handling data problems, predicting the likelihood of individuals developing specific diseases, developing hybrid machine learning models, and exploring additional strategies to address data issues when the ratio of missing values and imbalance exceeds 80%.

VIII. CONCLUSION

This study offers a detailed overview of predicting chronic diseases. It suggests using SLR to explore how machine learning can improve prediction result by addressing issues in disease data. Properly handling data problems is crucial for achieving accurate predictions. Additionally, selecting the most appropriate machine learning algorithm is vital as each algorithm has unique characteristics. Thus, choosing the right algorithm based on the data characteristics is key to successful disease prediction.

Problems in medical data such as outliers, missing values, feature selection, normalization, and imbalance require appropriate handling. The accuracy of prediction outcomes is influenced by the choice of approach used. Based on the results of the SLR, the most influential factors in data are feature selection and imbalance, which can lead to a difference of up to 15% in prediction outcomes. On the other hand, missing values, outliers, and normalization have less significant impact. Missing values become influential when the percentage of null values reaches 40% of the total data. Outliers in medical data are generally not a major concern as they often represent genuine values.

Discussing the use of machine learning models in predicting chronic diseases is crucial and deserves attention. In chronic disease prediction, machine learning models commonly used fall into categories such as supervised learning, ensemble learning, deep learning, and reinforcement learning. Each of these model categories has distinct characteristics. Selecting the appropriate model based on the characteristics of the data used is essential for achieving the best prediction outcomes. The results of the SLR indicate that ensemble learning-based models are the most powerful. Ensemble learning models are powerful because they combine multiple machine learning models into one for prediction. Therefore, when selecting a model, the use of ensemble learning requires special attention, taking into account the strengths and weaknesses of the models involved.

The results of this SLR also indicate that the best evaluation metrics for chronic disease prediction are F1-score, AUC-ROC, and MSE. The use of these evaluation metrics heavily depends on the characteristics of the dataset, the issues within the data, and the type of prediction model used. Finally, this paper addresses future challenges in chronic disease prediction. It discusses the challenges of enhancing prediction accuracy through hybrid machine learning algorithms and integrating data troubleshooting techniques. Another future challenge is acquiring new datasets from hospitals with more prevalent variables for each disease. Additionally, accurately predicting the likelihood of an individual being affected by certain diseases remains a significant challenge.

APPENDIX

Table 9 until Table 12 provides a list of glosarium related to predicting chronic diseases using machine learning for readers reference.

REFERENCES

[1] R. Ghorbani and R. Ghousi, "Predictive data mining approaches in medical diagnosis: A review of some diseases prediction," *Int. J. Data Netw. Sci.*, vol. 3, no. 2, pp. 47–70, 2019.

[2] F. Gorunescu, *Data Mining: Concepts, Models and Techniques*. India: Springer, 2011.

[3] H. C. Koh and G. Tan, "Data mining applications in healthcare," *J. Healthc. Inf. Manag.*, vol. 19, no. 2, p. 65, 2011.

[4] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," *Comput. Struct. Biotechnol. J.*, vol. 15, pp. 104–116, Jan. 2017.

[5] B. S. Ahamed, M. S. Arya, and A. O. V. Nancy, "Diabetes mellitus disease prediction using machine learning classifiers with oversampling and feature augmentation," *Adv. Hum.-Comput. Interact.*, vol. 2022, pp. 1–14, Sep. 2022.

[6] P. Theerthagiri, A. U. Ruby, and J. Vidya, "Diagnosis and classification of the diabetes using machine learning algorithms," *Social Netw. Comput. Sci.*, vol. 4, no. 1, p. 72, Nov. 2022.

[7] R. R. Kadhim and M. Y. Kamil, "Comparison of machine learning models for breast cancer diagnosis," *IAES Int. J. Artif. Intell. (IJ-AI)*, vol. 12, no. 1, p. 415, Mar. 2023.

[8] G. Kumawat, S. K. Vishwakarma, P. Chakrabarti, P. Chittora, T. Chakrabarti, and J. C.-W. Lin, "Prognosis of cervical cancer disease by applying machine learning techniques," *J. Circuits, Syst. Comput.*, vol. 32, no. 1, Jan. 2023, Art. no. 2350019.

[9] R. Huang, J. Liu, T. K. Wan, D. Siriwanna, Y. M. P. Woo, A. Vodencarevic, C. W. Wong, and K. H. K. Chan, "Stroke mortality prediction based on ensemble learning and the combination of structured and textual data," *Comput. Biol. Med.*, vol. 155, Mar. 2023, Art. no. 106176.

[10] P. B. Dash, "Efficient ensemble learning based CatBoost approach for early-stage stroke risk prediction," in *Ambient Intelligence in Health Care: Proceedings of ICAIHC 2022*. Singapore: Springer, 2022, pp. 475–483.

[11] W. Chang, Y. Liu, Y. Xiao, X. Yuan, X. Xu, S. Zhang, and S. Zhou, "A machine-learning-based prediction method for hypertension outcomes based on medical data," *Diagnostics*, vol. 9, no. 4, p. 178, Nov. 2019.

[12] M. A. J. Tengnah, R. Sooklall, and S. D. Nagowah, "A predictive model for hypertension diagnosis using machine learning techniques," in *Telemedicine Technologies*. Mauritius: Academic, 2019, pp. 139–152.

[13] S. Revathy, "Chronic kidney disease prediction using machine learning models," *Int. J. Eng. Adv. Technol.*, vol. 9, no. 1, pp. 6364–6367, 2019.

[14] K. R. A. Padmanaban and G. Parthiban, "Applying machine learning techniques for predicting the risk of chronic kidney disease," *Indian J. Sci. Technol.*, vol. 9, no. 29, pp. 1–6, Aug. 2016.

[15] I. V. Stepanyan, "Comparative analysis of machine learning methods for prediction of heart disease," *J. Mach. Manuf. Reliab.*, vol. 51, no. 8, pp. 789–799, 2022.

[16] K. M. Alfadli and A. O. Almagrabi, "Feature-limited prediction on the UCI heart disease dataset," *Comput., Mater. Continua*, vol. 74, no. 3, pp. 5871–5883, 2023.

[17] X. Lu, L. Yuan, R. Li, Z. Xing, N. Yao, and Y. Yu, "An improved Bi-LSTM-based missing value imputation approach for pregnancy examination data," *Algorithms*, vol. 16, no. 1, p. 12, Dec. 2022.

[18] P. Muthulakshmi and M. Parveen, "Z-score normalized feature selection and iterative African buffalo optimization for effective heart disease prediction," *Int. J. Intell. Eng. Syst.*, vol. 16, no. 1, pp. 25–37, 2022.

[19] F. Yang, K. Wang, L. Sun, M. Zhai, J. Song, and H. Wang, "A hybrid sampling algorithm combining synthetic minority oversampling technique and edited nearest neighbor for missed abortion diagnosis," *BMC Med. Informat. Decis. Making*, vol. 22, no. 1, p. 344, Dec. 2022.

[20] N. G. Ramadhan and A. Romadhony, "Preprocessing handling to enhance detection of type 2 diabetes mellitus based on random forest," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 7, pp. 223–228, 2021.

[21] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "Development of disease prediction model based on ensemble learning approach for diabetes and hypertension," *IEEE Access*, vol. 7, pp. 144777–144789, 2019.

[22] M. Z. Wadghiri, A. Idri, T. El Idrissi, and H. Hakkoum, "Ensemble blood glucose prediction in diabetes mellitus: A review," *Comput. Biol. Med.*, vol. 147, Aug. 2022, Art. no. 105674.

[23] A. Yaqoob, R. M. Aziz, N. K. Verma, P. Lalwani, A. Makrariya, and P. Kumar, "A review on nature-inspired algorithms for cancer disease prediction and classification," *Mathematics*, vol. 11, no. 5, p. 1081, Feb. 2023.

[24] Y. de Jong, C. L. Ramspek, V. H. W. van der Endt, M. B. Rookmaaker, P. J. Blankestijn, R. W. M. Vernooij, M. C. Verhaar, W. J. W. Bos, F. W. Dekker, G. Ocak, and M. van Diepen, "A systematic review and external validation of stroke prediction models demonstrates poor performance in dialysis patients," *J. Clin. Epidemiol.*, vol. 123, pp. 69–79, Jul. 2020.

[25] G. F. S. Silva, T. P. Fagundes, B. C. Teixeira, and A. D. P. Chiavegatto Filho, "Machine learning for hypertension prediction: A systematic review," *Current Hypertension Rep.*, vol. 24, no. 11, pp. 523–533, Nov. 2022.

[26] F. Sanmarchi, C. Fanconi, D. Golinelli, D. Gori, T. Hernandez-Boussard, and A. Capodici, "Predict, diagnose, and treat chronic kidney disease with machine learning: A systematic literature review," *J. Nephrol.*, vol. 36, no. 4, pp. 1101–1117, Feb. 2023.

[27] M. Marimuthu, M. Abinaya, K. S., K. Madhankumar, and V. Pavithra, "A review on heart disease prediction using machine learning and data analytics approach," *Int. J. Comput. Appl.*, vol. 181, no. 18, pp. 20–25, Sep. 2018.

[28] S. Nazah, S. Huda, J. Abawajy, and M. M. Hassan, "Evolution of dark web threat analysis and detection: A systematic approach," *IEEE Access*, vol. 8, pp. 171796–171819, 2020.

[29] P. Saint-Louis, M. C. Morency, and J. Lapalme, "Defining enterprise architecture: A systematic literature review," in *Proc. IEEE 21st Int. Enterprise Distrib. Object Comput. Workshop (EDOCW)*, Oct. 2017, pp. 41–49.

[30] A. G. Putrada, M. Abdurohman, D. Perdana, and H. H. Nuha, "Machine learning methods in smart lighting toward achieving user comfort: A survey," *IEEE Access*, vol. 10, pp. 45137–45178, 2022.

[31] C. Fisch and J. Block, "Six tips for your (systematic) literature review in business and management research," *Manage. Rev. Quart.*, vol. 68, no. 2, pp. 103–106, Apr. 2018.

[32] W. R. Clark, "Extending Fisch and block's (2018) tips for a systematic review in management and business literature," *Manag. Rev. Quart.*, vol. 71, no. 1, pp. 215–231, 2018.

[33] Y. Xiao and M. Watson, "Guidance on conducting a systematic literature review," *J. Planning Educ. Res.*, vol. 39, no. 1, pp. 93–112, Mar. 2019.

[34] F. Shokraneh and C. E. Adams, "Increasing value and reducing waste in data extraction for systematic reviews: Tracking data in data extraction forms," *Systematic Rev.*, vol. 6, no. 1, pp. 1–3, Dec. 2017.

[35] L. Yang, H. Zhang, H. Shen, X. Huang, X. Zhou, G. Rong, and D. Shao, "Quality assessment in systematic literature reviews: A software engineering perspective," *Inf. Softw. Technol.*, vol. 130, Feb. 2021, Art. no. 106397.

[36] S. Salzberg, "Distance metrics for instance-based learning," in *Proc. Methodolog Intell. Syst., 6th Int. Symp. (ISMIS)*, Charlotte, NC, USA, 1991, pp. 16–19.

[37] I. Inza, "Feature subset selection by genetic algorithms and estimation of distribution algorithms: A case study in the survival of cirrhotic patients treated with TIPS," *Artif. Intell. Med.*, vol. 23, no. 2, pp. 187–205, 2001.

[38] A. Purwar and S. K. Singh, "Hybrid prediction model with missing value imputation for medical data," *Expert Syst. Appl.*, vol. 42, no. 13, pp. 5621–5631, Aug. 2015.

[39] H.-C. Lin, C.-T. Su, and P.-C. Wang, "An application of artificial immune recognition system for prediction of diabetes following gestational diabetes," *J. Med. Syst.*, vol. 35, no. 3, pp. 283–289, Jun. 2011.

[40] T. R. Tavares, A. L. I. Oliveira, G. G. Cabral, S. S. Mattos, and R. Grigorio, "Preprocessing unbalanced data using weighted support vector machines for prediction of heart disease in children," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Aug. 2013, pp. 1–8.

[41] K. Mahboob, S. A. Ali, and U. Laila, "Investigating learning outcomes in engineering education with data mining," *Comput. Appl. Eng. Educ.*, vol. 28, no. 6, pp. 1652–1670, Nov. 2020.

[42] V. Bhat, "An efficient framework for prediction in healthcare data using soft computing techniques," in *Proc. Int. Conf. Adv. Comput. Commun.* Berlin, Germany: Springer, 2011, pp. 522–532.

[43] I. Kadi, A. Idri, and J. L. Fernandez-Aleman, "Knowledge discovery in cardiology: A systematic literature review," *Int. J. Med. Informat.*, vol. 97, pp. 12–32, Jan. 2017.

[44] M. Albayrak, K. Turhan, and B. Kurt, "A missing data imputation approach using clustering and maximum likelihood estimation," in *Proc. Med. Technol. Nat. Congr. (TIPTEKNO)*, Oct. 2017, pp. 1–4.

[45] W.-C. Lin and C.-F. Tsai, "Missing value imputation: A review and analysis of the literature (2006–2017)," *Artif. Intell. Rev.*, vol. 53, no. 2, pp. 1487–1509, Feb. 2020.

[46] T. Aittokallio, "Dealing with missing values in large-scale studies: Microarray data imputation and beyond," *Briefings Bioinf.*, vol. 11, no. 2, pp. 253–264, Mar. 2010.

[47] P. J. García-Laencina, J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal, "Pattern classification with missing data: A review," *Neural Comput. Appl.*, vol. 19, no. 2, pp. 263–282, Mar. 2010.

[48] A. M. Robitzsch. (2015). *Imputing Missing Data With R; MICE Package 2015*. [Online]. Available: https://datascienceplus.com/imputing-missing-datawith-r-mice-package

[49] P. Madley-Dowd, R. Hughes, K. Tilling, and J. Heron, "The proportion of missing data should not be used to guide decisions on multiple imputation," *J. Clin. Epidemiol.*, vol. 110, pp. 63–73, Jun. 2019.

[50] S. Balasubramanian, R. Kashyap, S. T. CVN, and M. Anuradha, "Hybrid prediction model for type-2 diabetes with class imbalance," in *Proc. IEEE Int. Conf. Mach. Learn. Appl. Netw. Technol. (ICMLANT)*, Dec. 2020, pp. 1–6.

[51] A. Azrar, Y. Ali, M. Awais, and K. Zaheer, "Data mining models comparison for diabetes prediction," *Int. J. Adv. Comput. Sci. Appl.*, vol. 9, no. 8, pp. 320–323, 2018.

[52] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, "Type 2 diabetes mellitus prediction model based on data mining," *Informat. Med. Unlocked*, vol. 10, pp. 100–107, Jan. 2018.

[53] Q. Wang, W. Cao, J. Guo, J. Ren, Y. Cheng, and D. N. Davis, "DMP_MI: An effective diabetes mellitus classification algorithm on imbalanced data with missing values," *IEEE Access*, vol. 7, pp. 102232–102238, 2019.

[54] M. Maniruzzaman, M. J. Rahman, M. Al-MehediHasan, H. S. Suri, M. M. Abedin, A. El-Baz, and J. S. Suri, "Accurate diabetes risk stratification using machine learning: Role of missing value and outliers," *J. Med. Syst.*, vol. 42, no. 5, pp. 1–17, May 2018.

[55] U. M. Faustin and B. Zou, "An improved homogeneous ensemble technique for early accurate detection of type 2 diabetes mellitus (T2DM)," *Computation*, vol. 10, no. 7, p. 104, Jun. 2022.

[56] J. B. Raja and S. C. Pandian, "PSO-FCM based data mining model to predict diabetic disease," *Comput. Methods Programs Biomed.*, vol. 196, Nov. 2020, Art. no. 105659.

[57] J. J. Khanam and S. Y. Foo, "A comparison of machine learning algorithms for diabetes prediction," *ICT Exp.*, vol. 7, no. 4, pp. 432–439, Dec. 2021.

[58] P. Ghosh, S. Azam, A. Karim, M. Hassan, K. Roy, and M. Jonkman, "A comparative study of different machine learning tools in detecting diabetes," *Proc. Comput. Sci.*, vol. 192, pp. 467–477, Jan. 2021.

[59] V. R. E. Christo, H. K. Nehemiah, J. Brighty, and A. Kannan, "Feature selection and instance selection from clinical datasets using co-operative co-evolution and classification using random forest," *IETE J. Res.*, vol. 68, no. 4, pp. 2508–2521, Jul. 2022.

[60] S. Mundra, "Classification of imbalance medical data: An empirical study of machine learning approaches," *J. Intell. Fuzzy Syst.*, vol. 43, no. 2, pp. 1933–1946, 2022.

[61] V. Thambusamy and L. Umasankar, "Prediction of heart disease using name entity recognition based on back propagation and whale optimization algorithms," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 10, pp. 437–443, 2019.

[62] B. A. M. Metwally, N. E. Mekky, and I. M. Elhenawy, "Heart disease prediction using genetic algorithm with machine learning classifiers," *Int. J. Adv. Sci., Eng. Inf. Technol.*, vol. 12, no. 5, pp. 1887–1894, Sep. 2022.

[63] C. Kokkotis, G. Giarmatzis, E. Giannakou, S. Moustakidis, T. Tsatalas, D. Tsiptsios, K. Vadikolias, and N. Aggelousis, "An explainable machine learning pipeline for stroke prediction on imbalanced data," *Diagnostics*, vol. 12, no. 10, p. 2392, Oct. 2022.

[64] S. P. Potharaju and M. Sreedevi, "Ensembled rule based classification algorithms for predicting imbalanced kidney disease data," *J. Eng. Sci. Technol. Rev.*, vol. 9, no. 5, pp. 201–207, Oct. 2016.

[65] K. M. Almustafa, "Prediction of chronic kidney disease using different classification algorithms," *Informat. Med. Unlocked*, vol. 24, Jan. 2021, Art. no. 100631.

[66] H. Zhang, C.-L. Hung, W. C. Chu, P.-F. Chiu, and C. Y. Tang, "Chronic kidney disease survival prediction with artificial neural networks," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Dec. 2018, pp. 1351–1356.

[67] S. Khalid, T. Khalil, and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," in *Proc. Sci. Inf. Conf.*, Aug. 2014, pp. 372–378.

[68] H. Zou, "Predicting diabetes mellitus with machine learning techniques," *Frontiers Genet.*, vol. 6, no. 9, p. 515, 2018.

[69] P. Sornsuwit, "Enhance weak learner model of adaboost (EWDM) for diabetes mellitus classification," *Int. J. Innov. Comput. Inf. Control*, vol. 18, no. 4, pp. 1117–1132, 2022.

[70] M. Pokharel, A. Alsadoon, T. Q. V. Nguyen, T. Al-Dala'in, D. T. H. Pham, P. W. C. Prasad, and H. T. Mai, "Deep learning for predicting the onset of type 2 diabetes: Enhanced ensemble classifier using modified t-SNE," *Multimedia Tools Appl.*, vol. 81, no. 19, pp. 27837–27852, Aug. 2022.

[71] A. Kishor and C. Chakraborty, "Early and accurate prediction of diabetics based on FCBF feature selection and SMOTE," *Int. J. Syst. Assurance Eng. Manage.*, vol. 12, no. 4, pp. 1–9, Jun. 2021.

[72] R. Saxena, S. K. Sharma, M. Gupta, and G. C. Sampada, "A novel approach for feature selection and classification of diabetes mellitus: Machine learning methods," *Comput. Intell. Neurosci.*, vol. 2022, pp. 1–11, Apr. 2022.

[73] K. Akyol and B. Sen, "Diabetes mellitus data classification by cascading of feature selection methods and ensemble learning algorithms," *Int. J. Mod. Educ. Comput. Sci.*, vol. 10, no. 6, pp. 10–16, 2018.

[74] M. T. García-Ordás, C. Benavides, J. A. Benítez-Andrades, H. Alaiz-Moretón, and I. García-Rodríguez, "Diabetes detection using deep learning techniques with oversampling and feature augmentation," *Comput. Methods Programs Biomed.*, vol. 202, Apr. 2021, Art. no. 105968.

[75] R. Rajkamal, A. Karthi, and X.-Z. Gao, "Diabetes prediction using derived features and ensembling of boosting classifiers," *Comput., Mater. Continua*, vol. 73, no. 1, pp. 2013–2033, 2022.

[76] C. Song and X. Li, "Cost-sensitive KNN algorithm for cancer prediction based on entropy analysis," *Entropy*, vol. 24, no. 2, p. 253, Feb. 2022.

[77] H. Qi, S. Xie, Y. Chen, C. Wang, T. Wang, B. Sun, and M. Sun, "Prediction methods of common cancers in China using PCA-ANN and DBN-ELM-BP," *IEEE Access*, vol. 10, pp. 113397–113409, 2022.

[78] S. Dev, H. Wang, C. S. Nwosu, N. Jain, B. Veeravalli, and D. John, "A predictive analytics approach for stroke prediction using machine learning and neural networks," *Healthcare Anal.*, vol. 2, Nov. 2022, Art. no. 100032.

[79] S. Prakash, K. Sangeetha, and N. Ramkumar, "An optimal criterion feature selection method for prediction and effective analysis of heart disease," *Cluster Comput.*, vol. 22, no. S5, pp. 11957–11963, Sep. 2019.

[80] I. Chourib, G. Guillard, I. R. Farah, and B. Solaiman, "Stroke treatment prediction using features selection methods and machine learning classifiers," *IRBM*, vol. 43, no. 6, pp. 678–686, Dec. 2022.

[81] M. Almasoud and T. E, "Detection of chronic kidney disease using machine learning algorithms with least number of predictors," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 8, pp. 1–9, 2019.

[82] N. Rout, D. Mishra, and M. K. Mallick, "Handling imbalance data: A survey," in *International Proceedings on Advances in Soft Computing, Intelligent Systems and Applications*. Singapore: Springer, 2018, pp. 431–443.

[83] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," *IEEE Trans. Syst. Man, Cybern. C, Appl. Rev.*, vol. 42, no. 4, pp. 463–484, Jul. 2012.

[84] B. Krawczyk, "Learning from imbalanced data: Open challenges and future directions," *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221–232, Nov. 2016.

[85] Z. Xie, O. Nikolayeva, J. Luo, and D. Li, "Building risk prediction models for type 2 diabetes using machine learning techniques," *Preventing Chronic Disease*, vol. 16, pp. 1–9, Sep. 2019.

[86] C. Azad, B. Bhushan, R. Sharma, A. Shankar, K. K. Singh, and A. Khamparia, "Prediction model using SMOTE, genetic algorithm and decision tree (PMSGD) for classification of diabetes mellitus," *Multimedia Syst.*, vol. 28, no. 4, pp. 1289–1307, Aug. 2022.

[87] M. S. Roobini and M. Lakshmi, "Autonomous prediction of type 2 diabetes with high impact of glucose level," *Comput. Electr. Eng.*, vol. 101, Jul. 2022, Art. no. 108082.

[88] P. Madan, V. Singh, V. Chaudhari, Y. Albagory, A. Dumka, R. Singh, A. Gehlot, M. Rashid, S. S. Alshamrani, and A. S. AlGhamdi, "An optimization-based diabetes prediction model using CNN and bi-directional LSTM in real-time environment," *Appl. Sci.*, vol. 12, no. 8, p. 3989, Apr. 2022.

[89] S. Tyagi and S. Mittal, "Sampling approaches for imbalanced data classification problem in machine learning," in *Proc. ICRIC*, 2020, pp. 209–221.

[90] N. G. Ramadhan, "Comparative analysis of ADASYN-SVM and SMOTE-SVM methods on the detection of type 2 diabetes mellitus," *Sci. J. Informat.*, vol. 8, no. 2, pp. 276–282, Nov. 2021.

[91] S. Sreejith, H. K. Nehemiah, and A. Kannan, "Clinical data classification using an enhanced SMOTE and chaotic evolutionary feature selection," *Comput. Biol. Med.*, vol. 126, Nov. 2020, Art. no. 103991.

[92] K. Roy, M. Ahmad, K. Waqar, K. Priyaah, J. Nebhen, S. S. Alshamrani, M. A. Raza, and I. Ali, "An enhanced machine learning framework for type 2 diabetes classification using imbalanced data with missing values," *Complexity*, vol. 2021, pp. 1–21, Jul. 2021.

[93] Z. Ning, Z. Ye, Z. Jiang, and D. Zhang, "BESS: Balanced evolutionary semi-stacking for disease detection using partially labeled imbalanced data," *Inf. Sci.*, vol. 594, pp. 233–248, May 2022.

[94] L. Yousefi and A. Tucker, "Identifying latent variables in dynamic Bayesian networks with bootstrapping applied to type 2 diabetes complication prediction," *Intell. Data Anal.*, vol. 26, no. 2, pp. 501–524, Mar. 2022.

[95] M. Kamaladevi and V. Venkatraman, "Tversky similarity based Under-Sampling with Gaussian kernelized decision stump AdaBoost algorithm for imbalanced medical data classification," *Int. J. Comput. Commun. Control*, vol. 16, no. 6, p. 4291, Nov. 2021.

[96] Y. Xiao, J. Wu, and Z. Lin, "Cancer diagnosis using generative adversarial networks based on deep learning from imbalanced data," *Comput. Biol. Med.*, vol. 135, Aug. 2021, Art. no. 104540.

[97] J. Zhang, L. Chen, and F. Abid, "Prediction of breast cancer from imbalance respect using cluster-based undersampling method," *J. Healthcare Eng.*, vol. 2019, pp. 1–10, Oct. 2019.

[98] P. Moghadam and A. Ahmadi, "A machine learning framework to predict kidney graft failure with class imbalance using red deer algorithm," *Expert Syst. Appl.*, vol. 210, Dec. 2022, Art. no. 118515.

[99] F. López-Martínez, E. R. Nuñez-Valdez, R. G. Crespo, and V. García-Díaz, "An artificial neural network approach for predicting hypertension using NHANES data," *Sci. Rep.*, vol. 10, no. 1, pp. 1–14, Jun. 2020.

[100] A. Ramezankhani, O. Pournik, J. Shahrabi, F. Azizi, F. Hadaegh, and D. Khalili, "The impact of oversampling with SMOTE on the performance of 3 classifiers in prediction of type 2 diabetes," *Med. Decis. Making*, vol. 36, no. 1, pp. 137–144, Jan. 2016.

[101] V. Hodge and J. Austin, "A survey of outlier detection methodologies," *Artif. Intell. Rev.*, vol. 22, no. 2, pp. 85–126, Oct. 2004.

[102] H. Yuk, J. Gim, J. K. Min, J. Yun, and T.-Y. Heo, "Artificial intelligence-based prediction of diabetes and prediabetes using health checkup data in Korea," *Appl. Artif. Intell.*, vol. 36, no. 1, Dec. 2022, Art. no. 2145644.

[103] S. Peñafiel, N. Baloian, J. A. Pino, J. Quinteros, Á. Riquelme, H. Sanson, and D. Teoh, "Associating risks of getting strokes with data from health checkup records using Dempster-Shafer theory," in *Proc. 20th Int. Conf. Adv. Commun. Technol. (ICACT)*, Feb. 2018, pp. 239–246.

[104] A. Rajagopal, S. Jha, R. Alagarsamy, S. G. Quek, and G. Selvachandran, "A novel hybrid machine learning framework for the prediction of diabetes with context-customized regularization and prediction procedures," *Math. Comput. Simul.*, vol. 198, pp. 388–406, Aug. 2022.

[105] R. Rajendran and A. Karthi, "Heart disease prediction using entropy based feature engineering and ensembling of machine learning classifiers," *Expert Syst. Appl.*, vol. 207, Nov. 2022, Art. no. 117882.

[106] Y. Shaikh, V. Parvati, and S. R. Biradar, "Early disease predictionalgorithm for hypertension-based disease using data aware algorithms," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 27, no. 2, pp. 1100–1108, 2022.

[107] V. A. Ardeti, V. R. Kolluru, G. T. Varghese, and R. K. Patjoshi, "An outlier detection and feature ranking based ensemble learning for ECG analysis," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 6, pp. 727–737, 2022.

[108] F. Iffath, S. J. Maisha, and M. Rashida, "Comparative analysis of machine learning techniques in classification cervical cancer using isolation forest with ADASYN," in *Proc. Int. Conf. Big Data, IoT, Mach. Learn. (BIM)*, Singapore. Springer, 2021, pp. 15–26.

[109] Y. Zhang, J. Li, P. Qu, W. Yao, and C. Lu, "Analysis and prediction of diabetes incidence based on big data," in *Proc. 5th Int. Conf. Big Data Technol.*, Sep. 2022, pp. 1–7.

[110] S. G. K. Patro and K. K. Sahu, "Normalization: A preprocessing stage," 2015, *arXiv:1503.06462*.

[111] L. A. Shalabi, Z. Shaaban, and B. Kasasbeh, "Data mining: A preprocessing engine," *J. Comput. Sci.*, vol. 2, no. 9, pp. 735–739, Sep. 2006.

[112] S. G. K. Patro, P. P. Sahoo, I. Panda, and K. K. Sahu, "Technical analysis on financial forecasting," 2015, *arXiv:1503.03011*.

[113] K. Swathi and S. Kodukula, "XGBoost classifier with hyperband optimization for cancer prediction based on geneselection by using machine learning techniques," *Revue d'Intell. Artificielle*, vol. 36, no. 5, pp. 665–670, Dec. 2022.

[114] M. Fang, Y. Chen, R. Xue, H. Wang, N. Chakraborty, T. Su, and Y. Dai, "A hybrid machine learning approach for hypertension risk prediction," *Neural Comput. Appl.*, vol. 35, no. 20, pp. 14487–14497, Jul. 2023.

[115] T. Cai, H. He, and W. Zhang, "Breast cancer diagnosis using imbalance learning and ensemble method," *Appl. Comput. Math.*, vol. 7, no. 3, pp. 146–154, 2018.

[116] O. Sagi and L. Rokach, "Ensemble learning: A survey," *WIREs Data Mining Knowl. Discovery*, vol. 8, no. 4, Jul. 2018, Art. no. e1249.

[117] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits Syst. Mag.*, vol. 6, no. 3, pp. 21–45, Jun. 2006.

[118] T. G. Dietterich, "Ensemble learning," in *The Handbook of Brain Theory and Neural Networks*. Cambridge, MA, USA: MIT Press, 2002, pp. 110–125.

[119] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *Proc. Comput. Learn. Theory, 2nd Eur. Conf. (EuroCOLT)*, Barcelona, Spain, Cham, Switzerland: Springer, Mar. 1995, pp. 23–37.

[120] A. S. Azar, "Application of machinelearning techniques for predicting survival in ovarian cancer," *BMC Med. Inform. Decis. Mak.*, vol. 22, no. 1, p. 345, 2022.

[121] H. B. Kibria, M. Nahiduzzaman, M. O. F. Goni, M. Ahsan, and J. Haider, "An ensemble approach for the prediction of diabetes mellitus using a soft voting classifier with an explainable AI," *Sensors*, vol. 22, no. 19, p. 7268, Sep. 2022.

[122] P. R. Krishna and P. Rajarajeswari, "EapGAFS: Microarray dataset for ensemble classification for diseases prediction," *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 10, no. 8, pp. 1–15, Aug. 2022.

[123] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996.

[124] A. R. P. Syed, R. Anbalagan, A. S. Setlur, C. Karunakaran, J. Shetty, J. Kumar, and V. Niranjan, "Implementation of ensemble machine learning algorithms on exome datasets for predicting early diagnosis of cancers," *BMC Bioinf.*, vol. 23, no. 1, pp. 1–24, Nov. 2022.

[125] A. Hesham, N. El-Rashidy, A. Rezk, and N. A. Hikal, "Towards an accurate breast cancer classification model based on ensemble learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 12, pp. 590–602, 2022.

[126] T. A. Shilpa, "Applying ensemble techniques of machine learning to predict heart disease," in *Proc. Int. Conf. Cogn. Intell. Comput. (ICCIC)*, 2021, pp. 775–783.

[127] A. Ashfaq, A. Imran, I. Ullah, A. Alzahrani, K. M. A. Alheeti, and A. Yasin, "Multi-model ensemble based approach for heart disease diagnosis," in *Proc. Int. Conf. Recent Adv. Electr. Eng. Comput. Sci. (RAEE CS)*, Oct. 2022, pp. 1–8.

[128] M. Kumar, S. Singhal, S. Shekhar, B. Sharma, and G. Srivastava, "Optimized stacking ensemble learning model for breast cancer detection and classification using machine learning," *Sustainability*, vol. 14, no. 21, p. 13998, Oct. 2022.

[129] S. Gupta and M. K. Gupta, "Computational prediction of cervical cancer diagnosis using ensemble-based classification algorithm," *Comput. J.*, vol. 65, no. 6, pp. 1527–1539, Jun. 2022.

[130] E. Dritsas and M. Trigka, "Stroke risk prediction with machine learning techniques," *Sensors*, vol. 22, no. 13, p. 4670, Jun. 2022.

[131] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, Oct. 2001.

[132] Z. Masetic and A. Subasi, "Congestive heart failure detection using random forest classifier," *Comput. Methods Programs Biomed.*, vol. 130, pp. 54–64, Jul. 2016.

[133] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[134] J. Yang and J. Guan, "A heart disease prediction model based on feature optimization and smote-xgboost algorithm," *Information*, vol. 13, no. 10, p. 475, Oct. 2022.

[135] M. Z. Alom, T. M. Taha, C. Yakopcic, S. Westberg, P. Sidike, M. S. Nasrin, M. Hasan, B. C. Van Essen, A. A. S. Awwal, and V. K. Asari, "A state-of-the-art survey on deep learning theory and architectures," *Electronics*, vol. 8, no. 3, p. 292, Mar. 2019.

[136] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.

[137] K. D. Dave and S. Vachik, "Neural network-based models for software effort estimation: A review," *Artif. Intell. Rev.*, vol. 42, no. 4, pp. 295–307, 2014.

[138] O. I. Abiodun, A. Jantan, A. E. Omolara, K. V. Dada, N. A. Mohamed, and H. Arshad, "State-of-the-art in artificial neural network applications: A survey," *Heliyon*, vol. 4, no. 11, Nov. 2018, Art. no. e00938.

[139] K. Mridha, "Early prediction of breast cancer by using artificial neural network and machine learning techniques," in *Proc. 10th IEEE Int. Conf. Commun. Syst. Netw. Technol. (CSNT)*, Jun. 2021, pp. 582–587.

[140] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, "A survey of convolutional neural networks: Analysis, applications, and prospects," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 12, pp. 6999–7019, Dec. 2022.

[141] M. Ashrafuzzaman, S. Saha, and K. Nur, "Prediction of stroke disease using deep CNN based approach," *J. Adv. Inf. Technol.*, vol. 13, no. 6, pp. 604–613, 2022.

[142] T. Gayathri, T. Madhavi, and K. R. Kumari, "A prediction of breast cancer based on mayfly optimized CNN," in *Proc. Int. Conf. Comput., Commun. Power Technol. (IC3P)*, Jan. 2022, pp. 176–180.

[143] C. Sateesh and R. Balamanigandan, "Heart disease prediction using innovative decision tree technique for increasing the accuracy compared with convolutional neural networks," in *Proc. 2nd Int. Conf. Innov. Practices Technol. Manage. (ICIPTM)*, vol. 2, Feb. 2022, pp. 583–587.

[144] L. Jiang, "Survey of improving naive Bayes for classification," in *Proc. Int. Conf. Adv. Data Mining Appl.*, Harbin, China, 2007, pp. 134–145.

[145] G. G. Banerjee Trishit, "Breast cancer prediction by machine learning algorithms—A comparative study of Naive Bayes, KNN and J48 in Weka environment," in *Proc. Workshop Comput. Vis. Mach. Learn. Healthcare Workshop Technolog Innov. Educ. Knowl. Dissemination, CVMLHWTEK*, 2022, pp. 47–53.

[146] S. Patidar, A. Jain, and A. Gupta, "Comparative analysis of machine learning algorithms for heart disease predictions," in *Proc. 6th Int. Conf. Intell. Comput. Control Syst. (ICICCS)*, May 2022, pp. 1340–1344.

[147] A. Q. Ansari, "The basics of fuzzy logic: A tutorial review," *Comput. Educ.-Stafford-Comput. Educ. Group*, vol. 88, pp. 5–8, Feb. 1998.

[148] T. Jaiswal and S. Jaiswal, "Machine learning-based classification models for diagnosis of diabetes," *Recent Adv. Comput. Sci. Commun.*, vol. 15, no. 6, pp. 813–821, Jul. 2022.

[149] J. Hao, S. Luo, and L. Pan, "Rule extraction from biased random forest and fuzzy support vector machine for early diagnosis of diabetes," *Sci. Rep.*, vol. 12, no. 1, pp. 1–12, Jun. 2022.

[150] A. K. Dehariya and P. Shukla, "Medical data classification using fuzzy main max neural network preceded by feature selection through moth flame optimization," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 12, pp. 655–662, 2020.

[151] S. P. Praveen, V. E. Jyothi, C. Anuradha, K. VenuGopal, V. Shariff, and S. Sindhura, "Chronic kidney disease prediction using ML-based neuro-fuzzy model," *Int. J. Image Graph.*, vol. 22, no. 5, Dec. 2022, Art. no. 2340013.

[152] M. Manur, A. Pani, and P. Kumar, "A big data analysis using fuzzy deep convolution network based model for heart disease classification," *Int. J. Intell. Eng. Syst.*, vol. 14, no. 2, pp. 147–156, Apr. 2021.

[153] T.-L. Nguyen, S. Kavuri, S.-Y. Park, and M. Lee, "Attentive hierarchical ANFIS with interpretability for cancer diagnostic," *Expert Syst. Appl.*, vol. 201, Sep. 2022, Art. no. 117099.

[154] M. Fernández-Delgado, M. S. Sirsat, E. Cernadas, S. Alawadi, S. Barro, and M. Febrero-Bande, "An extensive experimental survey of regression methods," *Neural Netw.*, vol. 111, pp. 11–34, Mar. 2019.

[155] R. D. Joshi and C. K. Dhakal, "Predicting type 2 diabetes using logistic regression and machine learning approaches," *Int. J. Environ. Res. Public Health*, vol. 18, no. 14, p. 7346, Jul. 2021.

[156] I. D. Mienye and Y. Sun, "Performance analysis of cost-sensitive learning methods with application to imbalanced medical data," *Informat. Med. Unlocked*, vol. 25, Jun. 2021, Art. no. 100690.

[157] T. A. Assegie, P. K. Rangarajan, N. K. Kumar, and D. Vigneswari, "An empirical study on machine learning algorithms for heart disease prediction," *IAES Int. J. Artif. Intell. (IJ-AI)*, vol. 11, no. 3, p. 1066, Sep. 2022.

[158] A. Pradhan, "Support vector machine—A survey," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 2, no. 8, pp. 82–85, 2012.

[159] S. Pal, "Chronic kidney disease prediction using machine learning techniques," *Biomed. Mater. Devices*, vol. 1, no. 1, pp. 534–540, Mar. 2023.

[160] M. Kumar, S. K. Khatri, and M. Mohammadian, "Predicting cancer survival using multilayer perceptron and high-dimensional SVM kernel space," *Ingénierie des Systèmes d Inf.*, vol. 27, no. 5, pp. 829–834, Oct. 2022.

[161] N. Bhatia, "Survey of nearest neighbor techniques," *Int. J. Comput. Sci. Inf. Secur.*, vol. 8, no. 2, pp. 302–305, 2010.

[162] V. Vaidehi, "Person authentication using face recognition," in *Proc. World Congr. Eng. Comput. Sci.*, 2008, pp. 2327–2338.

[163] O. K. Toker, "Text categorization using k nearest neighbor classification," Survey Paper, Middle East Tech. Univ., Türkiye, Tech. Rep. 2103101, 2013.

[164] F. Bajramovic, "A comparison of nearest neighbor search algorithms for generic object recognition," in *Proc. Adv. Concepts Intell. Vis. Syst., 8th Int. Conf. (ACIVS)*, 2006, pp. 18–21.

[165] Y. Yang, T. Ault, T. Pierce, and C. W. Lattimer, "Improving text categorization methods for event tracking," in *Proc. 23rd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2000, pp. 65–72.

[166] G. R. Thummala, R. Baskar, and N. Thiyaneswaran, "Prediction of heart disease using naive Bayes in comparison with KNN based on accuracy," in *Proc. Int. Conf. Cyber Resilience (ICCR)*, Oct. 2022, pp. 1–4.

[167] S. C. Gupta and N. Goel, "Selection of best K of K-nearest neighbors classifier for enhancement of performance for the prediction of diabetes," in *Progress in Advanced Computing and Intelligent Engineering*. Singapore: Springer, 2021, pp. 135–142.

[168] A. M. Majid and W. H. Utomo, "Application of discretization and adaboost method to improve accuracy of classification algorithms in predicting diabetes mellitus," *ICIC Exp. Lett.*, vol. 12, no. 12, pp. 1177–1184, 2021.

[169] M. Ahmed, R. Seraj, and S. M. S. Islam, "The k-means algorithm: A comprehensive survey and performance evaluation," *Electronics*, vol. 9, no. 8, p. 1295, Aug. 2020.

[170] R. T. Selvi and I. Muthulakshmi, "Modelling the map reduce based optimal gradient boosted tree classification algorithm for diabetes mellitus diagnosis system," *J. Ambient Intell. Humanized Comput.*, vol. 12, no. 2, pp. 1717–1730, Feb. 2021.

[171] R. Haripriya, "Predicitive analysis of breast cancer using multiple classification algorithms," in *Proc. Smart Technol., Commun. Robot. (STCR)*, Oct. 2021, pp. 1–5.

[172] M. Kumar, "Genetic algorithm: Review and application," *Int. J. Inf. Technol. Knowl. Manag.*, vol. 2, no. 2, pp. 451–454, 2010.

[173] N. K. Kumar, "An optimized random forest classifier for diabetes mellitus," in *Emerging Technologies in Data Mining and Information Security*, vol. 813. Singapore: Springer, 2019, pp. 765–773.

[174] T. H. Cormen, *Introduction to Algorithms*. Singapore: Springer, 2022.

[175] D. Mohideen, D. F. Mohammed, J. S. S. Raj, and R. S. P. Raj, "Regression imputation and optimized Gaussian Naive Bayes algorithm for an enhanced diabetes mellitus prediction model," *Brazilian Arch. Biol. Technol.*, vol. 64, Apr. 2022, Art. no. e21210181.

[176] M. Kumari and P. Ahlawat, "DCPM: An effective and robust approach for diabetes classification and prediction," *Int. J. Inf. Technol.*, vol. 13, no. 3, pp. 1079–1088, Jun. 2021.

[177] R. Cheruku, D. Edla, and V. Kuppili, "An optimized and efficient radial basis neural network using cluster validity index for diabetes classification," *Int. Arab J. Inf. Technol.*, vol. 16, no. 5, pp. 816–826, 2019.

[178] G. M. Reddy, "Risk assessment of type 2 diabetes mellitus prediction using an improved combination of NELM-PSO," *EAI Endorsed Trans. Scalable Inf. Syst.*, vol. 8, no. 32, pp. 1–26, 2021.

[179] M. G. El-Shafiey, A. Hagag, E.-S.-A. El-Dahshan, and M. A. Ismail, "A hybrid GA and PSO optimized approach for heart-disease prediction based on random forest," *Multimedia Tools Appl.*, vol. 81, no. 13, pp. 18155–18179, May 2022.

[180] M. Shuja, S. Mittal, and M. Zaman, "Effective prediction of type II diabetes mellitus using data mining classifiers and SMOTE," in *Advances in Computing and Intelligent Systems*. Singapore: Springer, 2020, pp. 195–211.

[181] M. F. U. Bhuiyan, M. T. Rahman, M. A. Anik, and M. Khan, "A framework for type-II diabetes prediction using machine learning approaches," in *Proc. 12th Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT)*, Jul. 2021, pp. 1–6.

[182] M. Ijaz, G. Alfian, M. Syafrudin, and J. Rhee, "Hybrid prediction model for type 2 diabetes and hypertension using DBSCAN-based outlier detection, synthetic minority over sampling technique (SMOTE), and random forest," *Appl. Sci.*, vol. 8, no. 8, p. 1325, Aug. 2018.

[183] K. Azbeg, M. Boudhane, O. Ouchetto, and S. Jai Andaloussi, "Diabetes emergency cases identification based on a statistical predictive model," *J. Big Data*, vol. 9, no. 1, pp. 1–25, Dec. 2022.

[184] Z. Tafa, N. Pervetica, and B. Karahoda, "An intelligent system for diabetes prediction," in *Proc. 4th Medit. Conf. Embedded Comput. (MECO)*, Jun. 2015, pp. 378–382.

[185] B. Shrestha, A. Alsadoon, P. W. C. Prasad, G. Al-Naymat, T. Al-Dala'in, T. A. Rashid, and O. H. Alsadoon, "Enhancing the prediction of type 2 diabetes mellitus using sparse balanced SVM," *Multimedia Tools Appl.*, vol. 81, no. 27, pp. 38945–38969, Nov. 2022.

[186] S. K. Kalagotla, S. V. Gangashetty, and K. Giridhar, "A novel stacking technique for prediction of diabetes," *Comput. Biol. Med.*, vol. 135, Aug. 2021, Art. no. 104554.

[187] J. Ramesh, R. Aburukba, and A. Sagahyroon, "A remote healthcare monitoring framework for diabetes prediction using machine learning," *Healthcare Technol. Lett.*, vol. 8, no. 3, pp. 45–57, Jun. 2021.

[188] S. A. Alex, J. J. V. Nayahi, H. Shine, and V. Gopirekha, "Deep convolutional neural network for diabetes mellitus prediction," *Neural Comput. Appl.*, vol. 34, no. 2, pp. 1319–1327, Jan. 2022.

[189] E. S. K. Chandrasekara, W. K. T. Kanchana, and E. J. K. P. Nandani, "Comprehensive study for diabetes identification ability of various optimizers in deep learning neural network," in *Proc. 5th SLAAI Int. Conf. Artif. Intell. (SLAAI-ICAI)*, Dec. 2021, pp. 1–6.

[190] H. Naz and S. Ahuja, "Deep learning approach for diabetes prediction using PIMA Indian dataset," *J. Diabetes Metabolic Disorders*, vol. 19, no. 1, pp. 391–403, Jun. 2020.

[191] S. T. Nivetha, B. Valarmathi, K. Santhi, and Chellatamilan, "Detection of type 2 diabetes using clustering methods—Balanced and imbalance pima Indian extended dataset," in *Proc. Int. Conf. Comput. Netw., Big Data (IoT)*, Cham, Switzerland. Springer, 2019, pp. 610–619.

[192] R. Vaishali, R. Sasikala, S. Ramasubbareddy, S. Remya, and S. Nalluri, "Genetic algorithm based feature selection and MOE fuzzy classification algorithm on pima Indians diabetes dataset," in *Proc. Int. Conf. Comput. Netw. Informat. (ICCNI)*, Oct. 2017, pp. 1–5.

[193] A. Sarwar and V. Sharma, "Comparative analysis of machine learning techniques in prognosis of type II diabetes," *AI Soc.*, vol. 29, no. 1, pp. 123–129, Feb. 2014.

[194] A. I. Pritom, Md. A. R. Munshi, S. A. Sabab, and S. Shihab, "Predicting breast cancer recurrence using effective classification and feature selection technique," in *Proc. 19th Int. Conf. Comput. Inf. Technol. (ICCIT)*, Dec. 2016, pp. 310–314.

[195] T. K. Avramov and D. Si, "Comparison of feature reduction methods and machine learning models for breast cancer diagnosis," in *Proc. Int. Conf. Compute Data Anal.*, May 2017, pp. 69–74.

[196] K. S. Bhangu, J. K. Sandhu, and L. Sapra, "Improving diagnostic accuracy for breast cancer using prediction-based approaches," in *Proc. 6th Int. Conf. Parallel, Distrib. Grid Comput. (PDGC)*, Nov. 2020, pp. 438–441.

[197] P. Sivakumar, T. U. Lakshmi, N. S. Reddy, R. Pavani, and V. Chaitanya, "Breast cancer prediction system: A novel approach to predict the accuracy using majority-voting based hybrid classifier (MBHC)," in *Proc. IEEE India Council Int. Subsections Conf. (INDISCON)*, Oct. 2020, pp. 57–62.

[198] A. Islam, M. M. Rahman, E. Ahmed, F. Arafat, and M. F. Rabby, "Adaptive feature selection and classification of colon cancer from gene expression data: An ensemble learning approach," in *Proc. Int. Conf. Comput. Advancements*, Jan. 2020, pp. 1–7.

[199] L. I. Santos, M. O. Camargos, M. F. S. V. D'Angelo, J. B. Mendes, E. E. C. D. Medeiros, A. L. S. Guimarães, and R. M. Palhares, "Decision tree and artificial immune systems for stroke prediction in imbalanced data," *Expert Syst. Appl.*, vol. 191, Apr. 2022, Art. no. 116221.

[200] Z. Hadianfard, H. L. Afshar, S. Nazarbaghi, B. Rahimi, and T. Timpka, "Predicting mortality in patients with stroke using data mining techniques," *Acta Inf. Pragensia*, vol. 11, no. 1, pp. 36–47, Mar. 2022.

[201] S. A. Mostafa, D. S. Elzanfaly, and A. E. Yakoub, "A machine learning ensemble classifier for prediction of brain strokes," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 12, pp. 258–266, 2022.

[202] S. Peñafiel, N. Baloian, H. Sanson, and J. A. Pino, "Predicting stroke risk with an interpretable classifier," *IEEE Access*, vol. 9, pp. 1154–1166, 2021.

[203] P. Chantamit-O-Pas and M. Goyal, "Long short-term memory recurrent neural network for stroke prediction," in *Machine Learning and Data Mining in Pattern Recognition*, vol. 10934, P. Perner, Ed. Cham, Switzerland: Springer, 2018, pp. 312–323.

[204] J. Chen, Y. Chen, J. Li, J. Wang, Z. Lin, and A. K. Nandi, "Stroke risk prediction with hybrid deep transfer learning framework," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 1, pp. 411–422, Jan. 2022.

[205] Y.-C. Chen, T. Suzuki, M. Suzuki, H. Takao, Y. Murayama, and H. Ohwada, "Building a classifier of onset stroke prediction using random tree algorithm," *Int. J. Mach. Learn. Comput.*, vol. 7, no. 4, pp. 61–66, Oct. 2017.

[206] T. Liu, W. Fan, and C. Wu, "A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset," *Artif. Intell. Med.*, vol. 101, Nov. 2019, Art. no. 101723.

[207] T. T. Oanh and N. T. Tung, "Predicting hypertension based on machine learning methods: A case study in Northwest Vietnam," *Mobile Netw. Appl.*, vol. 27, no. 5, pp. 2013–2023, Oct. 2022.

[208] F. V. Ferdinand, J. Sebastian, and F. Natalia, "Predicting stroke, hypertension, and diabetes disease based on individual characteristics," *ICIC Exp. Lett. B, Appl.*, vol. 12, no. 8, pp. 723–731, 2021.

[209] I. Arefa, M. S. Alam, I. Siddiquee, and N. Siddique, "Performance analysis of machine learning algorithms for hypertension decision support system," in *Proc. IEEE Int. Conf. Robot., Autom., Artif.-Intell. Internet Things (RAAICON)*, Nov. 2019, pp. 15–20.

[210] D. LaFreniere, F. Zulkernine, D. Barber, and K. Martin, "Using machine learning to predict hypertension from a clinical dataset," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Dec. 2016, pp. 1–7.

[211] C. A. U. Hassan, J. Iqbal, R. Irfan, S. Hussain, A. D. Algarni, S. S. H. Bukhari, N. Alturki, and S. S. Ullah, "Effectively predicting the presence of coronary heart disease using machine learning classifiers," *Sensors*, vol. 22, no. 19, p. 7227, Sep. 2022.

[212] B. Kaur and G. Kaur, "Heart disease prediction using modified machine learning algorithm," in *Proc. Int. Conf. Innov. Comput. Commun.*, vol. 473, D. Gupta, A. Khanna, S. Bhattacharyya, A. E. Hassanien, S. Anand, and A. Jaiswal, Eds. Singapore: Springer, 2022, pp. 189–201.

[213] M. Rizwan, S. Arshad, H. Aijaz, R. A. Khan, and M. Z. U. Haque, "Heart attack prediction using machine learning approach," in *Proc. 3rd Int. Conf. Latest Trends Electr. Eng. Comput. Technol. (INTELLECT)*, Nov. 2022, pp. 1–8.

[214] M. Pavithra and B. T. Geetha, "Prediction of chronic kidney cancer using RBF support vector machine compared with random forest for better accuracy," in *Proc. Int. Conf. Innov. Comput., Intell. Commun. Smart Electr. Syst. (ICSES)*, Jul. 2022, pp. 1–5.

[215] A. Stella and P. V. Kumari, "Forecasting of chronic kidney disease and analysis of the classifiers using ML based classification approaches," in *Proc. Int. Conf. Advancements Electr., Electron., Commun., Comput. Autom. (ICAECA)*, Oct. 2021, pp. 1–5.

[216] L. Antony, S. Azam, E. Ignatious, R. Quadir, A. R. Beeravolu, M. Jonkman, and F. De Boer, "A comprehensive unsupervised framework for chronic kidney disease prediction," *IEEE Access*, vol. 9, pp. 126481–126501, 2021.

[217] Rajeshwari and H. K. Yogish, "Prediction of chronic kidney disease using machine learning technique," in *Proc. 4th Int. Conf. Cognit. Comput. Inf. Process. (CCIP)*, Dec. 2022, pp. 1–6.

[218] W. Chang, Y. Liu, X. Wu, Y. Xiao, S. Zhou, and W. Cao, "A new hybrid XGBSVM model: Application for hypertensive heart disease," *IEEE Access*, vol. 7, pp. 175248–175258, 2019.

[219] H. Wu, X. Song, L. Zhu, X. Feng, Y. Li, and J. Chang, "A hypertension risk prediction model based on improve random forest," in *Proc. 2nd Int. Conf. Bioinf. Intell. Comput.*, Jan. 2022, pp. 429–436.

NUR GHANIAVIYANTO RAMADHAN was born in Surabaya, East Java, Indonesia, in February 1996. He received the bachelor's and master's degrees in informatics from Telkom University, Bandung, in 2019 and 2021, respectively, where he is currently pursuing the Ph.D. degree. He is engaged in research in data science and machine learning. He is a Lecturer and a Researcher with Telkom University. His research interests include data analytics and machine learning. His research dissertation focuses on the topic of the prediction of chronic disease using a preprocessing approach.

ADIWIJAYA (Member, IEEE) received the Ph.D. degree from Bandung Institute of Technology, Indonesia, in 2012. He has been the President of Telkom University, Indonesia, since 2018. He is currently a Professor of mathematics with the School of Computing, Telkom University. He published more than 153 papers in journals and conference proceedings with a Scopus H-index of 18. He was appointed as a keynote or invited speaker for several conferences. His research interests include mathematics, graph theory, data science, data mining, and data analytics. He has received several national research grants from the Ministry of Education, Culture, Research, and Technology, Republic of Indonesia, in the last five years.

WARIH MAHARANI received the bachelor's degree in informatics from STT Telkom, currently Telkom University, Bandung, and the master's and Ph.D. degrees in information technology from ITB Bandung. She is currently a full-time Lecturer and a Researcher with Telkom University. She is also an Associate Professor with Telkom University. She is also the Head of the Data Science Undergraduate Study Program. Her research interests include information extraction and data mining.

ALFIAN AKBAR GOZALI received the bachelor's and master's degrees in informatics from STT Telkom, currently Telkom University, Bandung, and the Ph.D. degree from Waseda University, Japan, in 2020. He is currently a Lecturer and a Researcher with Telkom University. He is also the Head of the Information Technology Product Development Section. He is also an Assistant Professor with Telkom University. His research interests include data analytics, machine learning, and mobile device development.