IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 June 2024, accepted 25 July 2024, date of publication 30 July 2024, date of current version 13 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3435948

RESEARCH ARTICLE

Enhancing Coronary Artery Disease Prognosis: A Novel Dual-Class Boosted Decision Trees Strategy for Robust Optimization

TARIQ MAHMOOD[1,2], AMJAD REHMAN[1], (Senior Member, IEEE), TANZILA SABA[1], (Senior Member, IEEE), TAHANI JASER ALAHMADI[3], MUHAMMAD TUFAIL[4], SAEED ALI OMER BAHAJ[5], AND ZOHAIB AHMAD[6]

[1]Artificial Intelligence and Data Analytics (AIDA) Laboratory, College of Computer and Information Science (CCIS), Prince Sultan University, Riyadh 11586, Saudi Arabia
[2]Faculty of Information Sciences, University of Education, Vehari Campus, Vehari 61100, Pakistan
[3]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia
[4]Department of Computer Science, Government Post Graduate College Nowshera, Nowshera 24100, Pakistan
[5]MIS Department, College of Business Administration, Prince Sattam Bin Abdulaziz University, AlKharj 11942, Saudi Arabia
[6]Department of Criminology and Forensics Sciences, Lahore Garrison University, Lahore 54000, Pakistan

Corresponding author: Tahani Jaser Alahmadi (tjalahmadi@pnu.edu.sa)

This research is funded by Princess Nourah Bint Abdulrahman University Researchers Supporting Project number (PNURSP2024R513), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

ABSTRACT The rise in stable coronary artery disease (CAD) due to improved survival rates and population growth has increased patient numbers, straining healthcare systems. Machine learning (ML) models are being developed to predict and identify individual risk factors for early treatment, reducing harm to individuals and families. These models can predict hospitalizations, enable close monitoring of high-risk patients, and optimize medical care. Researchers are developing robust models based on ML algorithms and real-world clinical data to aid in early detection, contributing to AI research in healthcare. Advanced ML models analyze medical imaging, genetic markers, lifestyle, and environmental factors to accurately predict coronary heart disease (CHD) start and progression. Our research introduces four novel models based on two-class Logistic Regression (two-class LR), two-class Neural Network (two-class NN), two-class Decision Jungle (two-class DJ), and two-class Boosted DT (two-class BDT). Our comparative analysis reveals that the two-class Boosted DT model is the most effective, achieving an AUC score of 0.991. This model excels in real-time monitoring by predicting minor changes in patient's health markers, allowing for timely adjustments in treatment plans. It optimizes medication selection, dosing, and intervention timing based on patient characteristics, improving therapeutic efficacy and reducing side effects. The study reveals the transformative potential of these advanced ML models in CAD prediction and management. By focusing on feature selection, algorithm improvement, and integration, our models analyze medical imaging, genetic markers, lifestyle, and environmental factors to accurately predict the onset and progression of CHD. This research proposes valuable insights into the capabilities of these models to revolutionize disease detection and management, ensuring reliable and timely healthcare interventions across various datasets.

INDEX TERMS Health issue, coronary heart disease, two-class LR, two-class NN, two-class DJ, two-class BDT.

I. INTRODUCTION

The human heart, a pivotal organ within the human body, propels blood throughout the circulatory system. Comprising four chambers, the heart primarily manages coronary arteries,

The associate editor coordinating the review of this manuscript and approving it for publication was Binit Lukose.

intricately connected to the aorta, the body's principal artery, residing within the heart's left ventricle. These coronary arteries further branch into the right and left coronary arteries as they emanate from the aorta. On a global scale, the accumulation of plaque in blood vessels leading to arterial constriction and inadequate perfusion, resulting in the obstruction of blood circulation, is a common causative factor for cardiovascular disease, often associated with heart attacks [1], [2].

Among cardiovascular illnesses, CHD stands as the most prevalent. Cardiovascular diseases rank prominently as a leading cause of death, claiming over 17 million lives annually [3]. In Europe alone, CHD contributes to more than 240,000 male and 76,000 female deaths under the age of 65 [3], with a staggering annual prediction of approximately 3.8 million male and 3.4 million female fatalities attributed to CHD [4]. The substantial economic burden, with heart disease-related expenses totaling $219.6 billion in 2017, encompassing healthcare services, medications, and productivity loss [5].

In recent decades, Artificial Intelligence (AI) has found diverse applications, including autonomous driving, gaming [6], and healthcare [7]. Machine learning (ML) algorithms commonly drive AI, offering a plethora of techniques such as support vector machine (SVMs) [8], ANN [9], DTs [10], NB [11], k-nearest neighbors (KNN) [12], and K-means [13]. Each of these methods bears distinct strengths and weaknesses and has undergone extensive investigation across various domains, including medical applications like liver disease [14], electrocardiogram (ECG) analysis [15], Parkinson's disease [16], and skin disease [17], for tasks such as screening, risk assessment, prediction, and decision support [18]. The volume of research varies based on disease prevalence, with cardiovascular diseases recognized by the World Health Organization (WHO) as a global leading cause of death [19], receiving considerable attention [20]. Precise diagnostics and preventative measures hold great potential for reducing CAD-related mortality. Early CAD detection is critical, as the definitive diagnosis involves invasive coronary angiography, which poses risks [20].

In this era of technological advancement and digitalization, healthcare organizations possess substantial patient data, including medical histories, symptoms, diagnoses, and treatment outcomes, all stored in electronic formats [21]. This burgeoning data resource, if effectively leveraged, can enhance patient care, reduce medical personnel burden, enable early detection, and facilitate preventive programs, especially for CHD, based on medical and family histories [21]. Medical professionals and researchers have begun to recognize the wealth of information within these datasets, extending to diseases like Dementia [22], Alzheimer's [22], Tuberculosis screening [23], Autism [24], and Cancer [25], with CHD [26] being a predominant focus. CHD, stemming from the gradual build-up of fat or bad cholesterol in coronary artery walls, poses severe health risks [27]. Risk factors encompass lifestyle choices and uncontrollable factors such as age, ethnicity, and family history [27]. Early CHD symptom detection enables lifestyle modifications and medications to prevent its progression and potentially fatal outcomes.

ML's role in medical diagnosis has expanded significantly, reducing manual errors, enhancing accuracy, and enabling reliable disease diagnoses [29]. Heart disease, a prevalent and often asymptomatic condition [30], remains a leading cause of mortality worldwide. Diagnosis traditionally relies on clinical history, physical exams, and medical tests, yet misdiagnosis rates remain a concern, given overlapping symptoms [31]. Various disease management approaches, including nursing-based and technological interventions, have been explored but face complexities and cost constraints [32]. Cardiac biomarkers offer cost-effective diagnostic potential, though their ambiguity sometimes necessitates additional tests [33]. Leveraging ML, particularly supervised and unsupervised techniques, has emerged as an effective diagnostic strategy. Clinical Decision Support Systems (CDSS) combining knowledge-based and non-knowledge-based approaches enhance diagnosis accuracy [34]. DL, an AI subset, excels in complex data analysis, such as medical datasets, and offers the potential for accurate heart disease diagnosis [35].

The prevalence of heart disease and its impact on global health underscore the need for improved diagnostic tools and predictive models. By harnessing the power of data analysis, ML offers promising avenues for early detection and precise diagnosis. Risk factors, such as high cholesterol, obesity, and hypertension, contribute to heart disease [36], yet diagnosis is challenging due to overlapping symptoms [37]. Access to extensive patient records and research data has led to the development of ML and DL models for accurate disease classification and prediction [38]. Various studies have showcased the potential of ML algorithms, such as RF, SVM, and NN, in diagnosing heart disease [39]. Dimensionality reduction techniques address the challenge of high-dimensional data, improving processing efficiency [40]. Feature engineering and selection methods further streamline data for more effective analysis.

The comparison of several ML algorithms utilized in the prediction of CHD risk is presented in Table 3. The evaluation of techniques such as SVM, NN, RF, and DT involves an assessment of their respective data sources, as well as an analysis of their strengths and limitations. SVMs are known for achieving high accuracy levels in many tasks. NN demonstrates exceptional proficiency in processing intricate data, albeit with a susceptibility to overfitting and a demand for substantial computer resources. RF has been shown to enhance generalization performance, yet they may encounter difficulties when dealing with datasets with many dimensions. Their simplicity in interpretation characterizes DT, although they are susceptible to fluctuations in data and have the potential to exhibit overfitting. This overview is valuable for researchers and healthcare practitioners in

discerning appropriate ML methodologies for assessing CHD risk.

In addressing healthcare challenges, AI is a transformative tool, particularly in the Information and Communication Technology era (*ICT*) [41]. AI's data processing capabilities surpass human capabilities, enhancing diagnostic and healthcare procedures [42]. AI-driven technologies exhibit common-sense reasoning, data extraction from raw data, adaptability, and the ability to act on knowledge [43], enabling precise clinical pattern recognition and diagnostics [44]. Combining AI with human intelligence, often referred to as augmented intelligence, can significantly improve healthcare services [45].

This study addresses the intricate issue of predicting the risk of CHD. This research work provides novel machine-learning techniques to enhance CHD risk prediction. By leveraging diverse patient data sources and employing cutting-edge algorithms, our methodology strives to provide a more thorough and precise evaluation of CHD risk. Moreover, our study aims to fulfill a significant research need by comprehensively evaluating diverse ML techniques. This study aims to provide significant information on the distinct advantages and limitations associated with different techniques in predicting CHD risk. Through this collaborative endeavor, This study aims to advance the current knowledge in CHD risk assessment and equip healthcare professionals with a reliable instrument for timely identification and intervention, ultimately leading to enhanced patient outcomes.

A. DEFICIENCIES IN CONVENTIONAL APPROACHES

Conventional CHD control methods have severe flaws that require modification. Often, these methods focus on symptom reduction and prompt procedures like medication and surgery to address critical conditions. While these treatments are essential for controlling symptoms and acute diseases, they may not address the underlying causes of CHD, which commonly originate from lifestyle factors. CHD development is linked to poor diet, inactivity, stress, and smoking. Pharmaceutical treatments can manage symptoms, but lifestyle changes and patient education are needed to promote healthy habits and lower risk factors. Traditional approaches focus on reactive treatment rather than proactive initiatives to prevent sickness. A comprehensive approach to CHD care improves results. This strategy should include lifestyle changes, preventative action, and patient education on risk factors. Researchers are using statistical analysis and data mining to create early-stage diagnostic tools for healthcare professionals, addressing issues like data dimensionality and missing features. They are also focusing on predicting CHD using publicly available datasets that don't accurately reflect the imbalanced distribution of real CHD samples, aiming to improve, integrate, and heterogeneity models for higher accuracy or AUC values. ML algorithms and methodologies offer significant advantages over conventional CHD control methods. Unlike traditional methods that use predetermined rules and limited data analysis, ML can analyze medical imaging, genetic markers, lifestyle, and environmental factors. ML may accurately predict CHD start and progression using advanced algorithms and deep learning models by identifying complex patterns and relationships in this diverse data. Depending on patient characteristics, these algorithms also optimize medicine selection, doses, and intervention time. This specific strategy improves therapeutic efficacy and reduces side effects. Machine learning-enabled real-time monitoring detects minor changes in a patient's health markers, allowing treatment plan revisions. ML has excellent potential to make CHD management data-driven and patient-centric, improving patient outcomes and public health.

B. RESEARCH MOTIVATION

CHD is a significant and complex problem in the current healthcare environment, with extensive consequences for patient welfare and global healthcare systems [1], [2]. The impetus for additional investigation in this field is firmly grounded in the imperative to augment our comprehension of CHD and to better approaches for early identification. The dataset utilized in this research encompasses a range of patient characteristics, including age, gender, cholesterol levels, chest pain type, resting blood pressure, resting ECG [15], maximal heart rate, and fasting blood sugar [27]. This dataset[1] exemplifies the considerable capacity for deriving valuable insights through data analysis. The study demonstrates the impressive performance of ML-based models, specifically the two-class Decision Jungle and two-class Boosted Decision Tree models. This confirms their effectiveness and highlights the potential for using advanced technologies that have yet to be fully explored. The models above demonstrated remarkable performance in terms of AUC scores, surpassing the forecasts made by earlier research. The impetus for this research stems from the assumption that there is significant potential for progress in CHD prediction. The potential to enhance early detection techniques for a grievous illness has significant implications for the improvement of patient outcomes and the reduction of the burden on healthcare systems. In addition, the utilization of unsupervised machine learning and deep learning methodologies, as outlined in this research, presents intriguing opportunities for further investigation. As the exploration of data science and artificial intelligence progresses, a substantial amount of data remains untapped, presenting significant opportunities for enhancing healthcare outcomes. The findings from this research highlight the possibility for future investigations to advance the limits of CHD prediction, ultimately resulting in the development of more precise and easily available resources for healthcare practitioners. By employing a data-driven approach, we can effectively tackle the obstacles associated with CHD and

[1] Downloaded from https:// ieee- dataport. org/ open- access/ heart- disease dataset- comprehensive

TABLE 1. Comparison of ML techniques in coronary heart disease risk prediction.

Ref.	ML Technique	Data Sources	Strengths	Weaknesses
[1]	SVM	Electronic Health Records	High accuracy, robustness to noise	Limited interpretability, requires large datasets
[2]	Neural Networks	Health Records, Genetics	DL capabilities, handles complex data	Prone to overfitting, computationally intensive
[5]	DTs	Health Records, Biomarkers	Simple to interpret, suitable for rule extraction	Sensitive to small variations in data, may lead to overfitting
[26]	RF	Medical Histories	Ensemble method for improved generalization	May struggle with high-dimensional data
[46]	RF	Clinical Data	High accuracy, ensemble method	Data preprocessing required, may struggle with imbalanced data
[47]	SVM	ECG Data	Effective with feature selection	Requires expert domain knowledge
[48]	Neural Networks	Patient Records	Handles non-linearity	Sensitive to noisy data

work towards a future where early intervention is the prevailing practice, substantially reducing the impact of this debilitating condition. This study aims to extend an invitation to the scientific community to engage in a collective endeavor of exploration, advancement, and cooperation to enhance our comprehension of CHD, improve its management, and ultimately reduce its prevalence in the future.

C. THE INNOVATIVE CONTRIBUTION

Advanced machine learning algorithms are a more suitable choice for assessing CHD risk than deep learning models. These algorithms offer better interpretability, and better clinical adoption, and are practical and resource-efficient in healthcare settings with limited labeled data. They can handle diverse data sources like medical history, imaging, and lifestyle factors in a structured and interpretable manner. The choice to use advanced machine learning algorithms is driven by the need for domain-specific customization, interpretability, resource efficiency, and the effective integration of diverse data types. This makes them a more practical and efficient choice for CHD risk assessment. This makes them a more practical and efficient choice for CHD risk assessment.

The paper's primary contributions are as follows:

- To propose ML-based fusion algorithms to enhance prediction accuracy and model robustness, potentially reducing CHD mortality rates.
- To boost the preciseness of CAD predictive models using novel data preprocessing methods, ensuring an equitable distribution of dataset instances and aiding prompt detection and proactive intervention for strengthened patient care.
- To employ feature scaling techniques to discern and assess the significance of features, conducting thorough analyses on both the original imbalanced dataset and its counterpart, meticulously balanced for class equality.
- To boost the perfection and precision of CHD risk analysis by the integration of diverse patient data sources, including medical history, imaging, and lifestyle factors.
- To compare the strengths and weaknesses of proposed models in predicting CHD risks and improve risk assessment through a robust approach.
- The efficacy of suggested algorithms was compared with cutting-edge models, exhibiting superior efficacy in detecting CAD using benchmark evaluation parameters.

D. ORGANIZATION OF PAPER

Table 2 presents an organized list of abbreviations in alphabetical order and their corresponding full versions to enhance the readability of the proposed study. This table aims to enhance reader comprehension by providing a convenient reference for the numerous acronyms and words in the text, enabling readers to quickly and easily acquaint themselves with them. This table seeks to enhance the overall readability and accessibility of the document by offering a concise reference guide. It ensures that readers can navigate and comprehend the text more efficiently.

In the rest of the paper, the literature review is presented in Section II. The suggested methodology is explained in Section III. Section IV evaluates the proposed technique's results and discusses the observed outcomes. In Section V, experimental work is presented. Finally, Section VI highlights the conclusion and future research.

II. LITERATURE REVIEW

The healthcare sector has significantly transitioned in recent years, primarily due to the rapid growth of data mining and machine learning in the field [49]. These advanced technologies have been applied in different healthcare fields, greatly influencing the field of medical cardiology. The rapid and significant increase in medical data has created unique prospects for researchers to create and evaluate cutting-edge algorithms to transform the early identification and prevention of cardiac disease. This is particularly important because heart disease continues to be a significant cause of death, especially in developing countries.

Waigi et al. [50] utilized ML to identify risk factors for cardiovascular disease in individuals with Metabolic-Associated Fatty Liver Disease (MAFLD). The model accurately

TABLE 2. List of abbreviations: Enhancing readability in cardiovascular disease prediction research.

Abbreviation	Full Form
AUC	Area Under the Curve
ANW	Artificial Neural Networks
CAD	Coronary Artery Disease
CDSS	Clinical Decision Support Systems
CHD	Coronary Heart Disease
CNN	Convolutional Neural Network
CVD	Cardiovascular Disease
DL	Deep Learning
DT	Decision Trees
EHR	Electronic Health Records
ECG	Electrocardiogram
FRS	Framingham Risk Score
ICT	Information and Communication Technology
KNN	K-Nearest Neighbors
LR	Logistic Regression
MAFLD	Metabolic-Associated Fatty Liver Disease
ML	Machine Learning
NB	Naive Bayes
NN	Neural Networks
ANN	Artificial Neural Networks
PCA	Principal Component Analysis
RF	Random Forests
RETAIN	Reverse Time Attention
SGD	Stochastic Gradient Descent
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
SVMs	Support Vector Machines
two-class BDT	Two-Class Boosted Decision Tree
two-class DJ	Two-Class Decision Jungle
two-class LR	Two-Class Logistic Regression
two-class NN	Two-Class Neural Network

identified high cholesterol, plaque scores, and diabetes duration, demonstrating the effectiveness of machine learning in assessing CVD risk. Zhou et al. [51] developed a DL-based prediction model for heart failure using risk factors and innovative techniques. The model, tested on real-world data from the HeartCarer project, achieved an impressive 98.5% accuracy. Zwack et al. [52] analyzed 16,000 cardiovascular disease articles and found that digital technologies have notably enhanced morbidity and mortality outcomes, especially in cardiovascular rehabilitation, out-of-hospital cardiac arrest, and arrhythmia management. Narin et al. [53] present a significant advancement in this area by proposing an ML-based method to forecast the risk of CVD. The objective of this system was to improve the precision of the conventional Framingham Risk Score (FRS) by utilizing a quantum neural network, resulting in an amazing accuracy rate of 98.57%. Shah et al. [54] created an ML model to forecast cardiovascular disorders. Among the models tested, the KNN model obtained an outstanding accuracy rate of 90.8%. This research emphasizes the need to choose suitable machine learning models for accurate and dependable cardiovascular disease predictions. Gokhale et al. [55] improved the QRISK cardiovascular disease risk analysis approach for type 2 diabetes patients using data from 200,000 individuals aged 25-85. The study used Cox's hazards model to calculate a 10-year risk score, predicting survival probabilities of 0.87 for females and 0.84 and 0.83 for males. Adekkanattu et al. [56] examined EHR from three academic medical centers to understand baseline characteristics and ejection fraction changes in heart failure patients. The ML models, particularly XGBoost, effectively predicted EF changes, achieving high F1-scores for a 30% increase and 95.0 for a 30% decrease, based on baseline EF, age, gender, and heart disease. Additionally, the literature features studies focused on specific models and approaches. For instance, the Reverse Time Attention (RETAIN) model, discussed

in [57], was evaluated for its ability to predict heart failure risks across various hospital settings and patient groups. This model demonstrated remarkable efficiency in electronic health record (EHR) prediction modeling, achieving an AUC of 82%, outperforming traditional logistic regression when trained on larger datasets. Parveen and Hiremath [58] proposed a method to predict cardiovascular disease from retinal images, using an Improved GLCM technique to detect microvascular changes based on factors like age, gender, smoking status, and blood pressure, analyzed in MATLAB with data from the UCI Machine Learning Repository.

Alotaibi et al. [59] explored predicting heart failure using machine learning, utilizing data from the Cleveland Clinic Foundation. The DT algorithm achieved an accuracy of 93.19%, showcasing the potential of ML in heart failure prediction and proactive patient care. Chunduru et al. [60] developed a flask web application using ML and DL to predict seven diseases, including diabetes, breast cancer, heart, kidney, liver, malaria, and pneumonia, to improve prediction accuracy to reduce misdiagnoses. Ozsahin et al. [61] introduce an ML meta-model that predicts heart failure with 87% accuracy using clinical data. Employing Random Forest, GNB, DT, and kNN estimators, it's tested across five datasets featuring 11 key attributes. This method aims to preempt critical conditions stemming from heart diseases and related risk factors. Sevket et al. [62] present an ML model for predicting heart disease and heart failure, utilizing advanced optimization algorithms. On Cleveland datasets, it achieves an 88% F-score for heart disease with KNN and 70% for heart failure. Notably, the flower pollination algorithm attains a 99.72% F-score for heart disease for certain population sizes.

Hasan and Bao [63] compared models like RF, SVC, kNN, NB, XGBoost, and ANN for CVD prediction, highlighting XGBoost with the wrapper approach as the top performer at a 73.74% accuracy rate, underlining feature selection's significance in CVD prediction accuracy. Balasubramaniam et al. [64] present the GSSA-DMN technique for detecting heart disease, which combines data preprocessing, ReliefF feature selection, and GSSA optimization. This method achieves approximately 93.2% accuracy, 93% sensitivity, and 91.5% specificity, outperforming conventional ECG approaches in effectiveness. Dritsas and Trigka et al. [65] presents a supervised ML-based methodology for designing efficient prediction models for CVD manifestation, focusing on risk factors. The Stacking ensemble model achieved an accuracy of 87.8%, recall of 88.3%, precision of 88%, and AUC of 98.2%. In addition to individual investigations, the literature includes research investigating the precision and capabilities of several machine-learning algorithms for predicting heart disease. Vishnupriya et al. [66] propose a novel method to predict heart disease using hybrid ML and DL techniques. Achieving a 92% accuracy rate, they combine a hybrid random forest and deep learning approach. Integration of a tkinter program enhances prediction accuracy by comparing current healthcare data with reference distribution information. Chang et al. [67] utilize techniques such as KNN, DT, SVM, CN2, and SGD that explore modern information management practices, emphasizing the importance of global standards and advanced technologies like AI, ML, DL, and cloud/edge computing. It proposes new research directions and provides a comprehensive overview of the field's evolution and trends, making a significant contribution to information management. Noroozi et al. [68] explore the role of feature selection in optimizing ML algorithms for heart disease prediction. Using the Cleveland Heart disease dataset and sixteen techniques, seven algorithms were evaluated. SVM-based filtering achieved the highest accuracy at 85.5%, while filtering techniques with more features outperformed others in ACC, Precision, and F-measures.

Yang et al. [69] developed HY_OptGBM, a CHD prediction model with an optimized LightGBM classifier. Achieving an AUC of 97.8% on Framingham Heart Institute CHD data suggests potential cost savings in medical treatment through early CHD detection. Dutt et al. [70] developed a two-layer CNN model for classifying clinical data on CHD, addressing class imbalances. The model uses LASSO for feature weighting and majority voting for feature selection, and a unique training routine per epoch enhances accuracy. Despite the NHANES dataset's class imbalance, the model achieves 77% classification accuracy. Nandy et al. [71] developed a healthcare framework for predicting cardiovascular disease using a Swarm-ANN approach, generating and evaluating neural networks with a novel heuristic for weight adjustment. This strategy reached a 95.78% prediction accuracy on a benchmark dataset, surpassing traditional learning methods. Abdulrakeeb et al. [72] have developed a technique to improve early CHD detection by using a bagged decision tree algorithm and duplicated misclassified instances, achieving a 97.1% accuracy in 10-fold CV tests.

Table 3 provides a comprehensive overview of significant studies in CVD prediction, offering insights into their primary objectives, employed datasets, ML techniques, and notable findings. In a study published in [53], the accuracy of CVD predictions was improved using a quantum neural network, with a remarkable accuracy rate of 98.57%. In contrast, a research work conducted in [54] emphasizes the critical role of model selection, with the KKN model yielding the highest accuracy of 90.8% on the Cleveland heart disease dataset. In [73], ML was employed to identify significant CVD risk factors in patients with metabolic-associated fatty liver disease, demonstrating robust risk assessment capabilities. In [59], heart failure prediction using the Cleveland Clinic Foundation dataset is explored with the DT algorithm, achieving an impressive accuracy rate of 93.19%. In [63], extensive research on feature selection is conducted, culminating in the XGBoost classifier's accuracy of 73.74%, underscoring the critical role of feature selection in CVD prediction. In [66], RF, NB, DT, and Logistic Regression algorithms are evaluated for heart status prediction, with

each algorithm demonstrating varying degrees of accuracy. A study conducted in [74] recommends several methods, including KNN, DT, SVM, Combined Nomenclature (CN2) rule, and Stochastic Gradient Descent (SGD), for CVD prediction, with their respective accuracy rates providing valuable comparative data. In [67], a Python-based application for health monitoring systems is developed, contributing to the development of comprehensive healthcare solutions. Reference [75] addresses the class imbalance in CVD prediction, employing the Synthetic Minority Oversampling Technique (SMOTE) and various ML classifiers, enhancing predictive accuracy. In [76], the Korean National Health Sample Cohort dataset is utilized to develop multiple ML-based prediction models for 2-year and 10-year CVD risk, further expanding the array of prediction methodologies. Collectively, Table 3 highlights the diverse approaches and substantial contributions of these studies to the field of CVD prediction, showcasing their potential to improve predictive accuracy and patient outcomes.

III. PROPOSED METHODOLOGY

This study uses machine learning technology to predict early-stage CHD risk and identify risk factors. The results help doctors make accurate diagnoses and help residents manage their daily health. Early detection and treatment can prevent severe CHD progression and reduce the economic burden on families. The study explores the use of digital technology to predict CHD risk in residents, integrating digital healthcare with disease prediction and daily health management. The research aims to refine the predictive model for early CHD detection by introducing an enhanced dataset. Data from health records, physical examinations, medical data, and chronic disease screening are used to monitor each resident's health status. High-risk residents receive basic health interventions and recommended physical examinations, while low-risk residents can reduce their risk by adjusting their diet and increasing physical activity. The study improves research models by incorporating imbalanced datasets and improving interpretability, creating a new predictive model for predicting CHD. The research methodology provides a comprehensive insight into the research methodology and the proposed layered framework. By incorporating digital healthcare with disease prediction and daily health management, the study aims to enhance health awareness and promote healthy lifestyle habits, ultimately improving public health.

A. PROPOSED ARCHITECTURE

Figure 1 illustrates the architecture and the steps involved in the proposed architecture, and the flow is as follows:

1) Initially, data is gathered from different sources.
2) Data preprocessing, including feature extraction and feature engineering, is conducted.
3) Once data preprocessing is completed, model training begins using the preprocessed data.

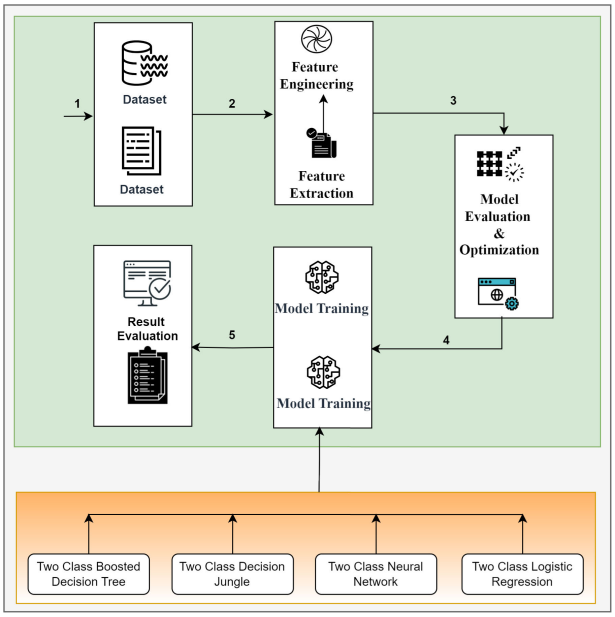

FIGURE 1. Workflow of the proposed model for early coronary artery disease detection.

4) Model evaluation and optimization are performed after model training to assess the performance and fine-tune the model.
5) Finally, the optimized model results are evaluated.

Figure 1 illustrates that data acquisition and feature engineering involve obtaining a dataset pertinent to the intended result. The data is subjected to preprocessing and altered using feature engineering approaches to ensure its appropriateness for model training. The preprocessed data is then inputted into a selected ML model. The model, which starts with randomly assigned weights, goes through an iterative training process to acquire the ability to recognize patterns and connections in the data. This training entails fine-tuning the model's weights to minimize discrepancies between its predictions and the actual results. After training, a distinct validation dataset assesses the model's performance. This evaluation aids in measuring the model's capacity to apply its knowledge to new situations and in pinpointing areas that need enhancement. According to the evaluation results, the model can be improved by adjusting its hyperparameters or selecting an alternative model architecture. The refined model is then utilized on fresh data to provide forecasts or categorizations. The veracity and dependability of these predictions are subsequently assessed to ascertain the model's efficacy in practical situations.

Figure 1 also emphasizes the periodicity of the model training process. As additional data becomes accessible, the model can be subjected to retraining and optimization processes to consistently enhance its performance. This iterative refinement process enables machine learning models to gradually improve and adjust to changing conditions, rendering them highly valuable for various activities and applications.

TABLE 3. Comparison of ML techniques in coronary heart disease risk prediction.

Reference	Study Focus	Dataset Used	ML Methods	Key Findings
[53]	Enhancing CVD Prediction	Framingham Research Dataset	Quantum Neural Network	Impressive 98.57% accuracy in CVD risk prediction. Significant potential for improving CVD risk assessment and early diagnosis.
[54]	CVD Prediction	Cleveland Heart Disease Dataset	NB, DTs, RF, KNN	KKN model achieved the highest accuracy at 90.8%, highlighting the importance of model selection in CVD prediction.
[59]	Heart Failure Prediction	Cleveland Clinic Foundation Dataset	Various ML Algorithms (DTs, etc.)	DT algorithm achieved 93.19% accuracy, emphasizing the potential of ML in heart failure prediction.
[63]	Feature Selection for CVD Prediction	N/A	XGBoost Classifier with Wrapper Technique	XGBoost with wrapper technique achieved 73.74% accuracy, highlighting the importance of feature selection in CVD prediction.
[76]	2-Year and 10-Year CVD Risk Prediction	Korean National Health Sample Cohort Dataset	Development of ML-based prediction models for 2-year and 10-year CVD risk.	—
[74]	CVD Prediction	N/A	KNN, DT, SVM, CN2 Rule, SGD	Recommendation of methods for CVD prediction.
[66]	Heart Status Prediction	UCI ML Repository Dataset	Random Forest, Naive Bayes, DT, Logistic Regression	Evaluation of various ML algorithms for predicting heart status using UCI ML repository dataset.
[67]	Health Monitoring Systems	N/A	Python-based Application for Health Monitoring Systems	Development of a Python-based application for health monitoring systems.
[75]	Addressing Imbalance in CVD Prediction	Synthetic Minority Oversampling Technique (SMOTE), ML Classifiers	Use of SMOTE and ML classifiers for addressing the imbalance issue in CVD prediction.	—
[73]	CVD Risk Identification	MAFLD Dataset	Multiple Logistic Regression, Univariate Feature Ranking, PCA	Identified key risk factors (hypercholesterolemia, plaque scores, diabetes duration) and achieved an accuracy of 85.11% for high-risk patients.

B. PROPOSED LAYERED FRAMEWORK

The suggested framework consists of five distinct levels: a data source layer, a data preprocessing layer, a data analysis and classification layer, a model evaluation layer, and a model training layer. The subsequent part provides a comprehensive assessment of the features and components of each layer.

1) DATA SOURCE

The research was carried out by utilizing the all-encompassing dataset on heart disease. This dataset is the most comprehensive one on heart disease that is currently accessible for use in research and available for free to use and download.[2] It consists of 1190 different cases and 14 different features, which include features such as age, gender, chest pain, cholesterol, resting blood pressure, cholesterol, fasting blood sugar, and maximum heart rate, as illustrated in Table 4. The output variables comprised category data. The values were used to determine the presence or absence of CHD, with a value of 1 indicating the diagnosis of CHD and a value of 0 showing the absence of CHD. On the other hand,

[2]: Downloaded from https:// ieee- dataport. org/ open- access/ heart-disease dataset- comprehensive

TABLE 4. Original study data.

age	sex	chest pain type	resting bp	cholesterol	max heart rate	target
40	1	2	140	289	172	0
49	0	3	160	180	156	1
37	1	2	130	283	98	0
48	0	4	138	214	108	1
54	1	3	150	195	122	0
39	1	3	120	339	170	0
45	0	2	130	237	170	0
54	1	2	110	208	142	0
37	1	4	140	207	130	1
48	0	2	120	284	120	0
37	0	3	130	211	142	0
58	1	2	136	164	99	1
39	1	2	120	204	145	0
49	1	4	140	234	140	1
42	0	3	115	211	137	0
54	0	2	120	273	150	0

TABLE 5. Features and their corresponding values, source of the document is available on the source website.

S.No.	Attribute	Description	Unit	Data type
1	age	Age	in years	Numeric
2	sex	Sex	1, 0	Binary
3	cp	Chest pain type	1,2,3,4	Nominal
4	rbp	Resting blood pressure	in mm Hg	Numeric
5	chl	Cholesterol	in mg/dl	Numeric
6	fbs	Fasting blood sugar	1,0 > 120 mg/dl	Binary
7	rstecg	Resting electrocardiogram	0,1,2	Nominal
8	maxhrate	Maximum heart rate	71–202	Numeric
9	exangina	Exercise-induced angina	0,1	Binary
10	oldpeak	Exercise produced ST depression	1,2,3	Numeric
11	slp	Peak performance ST segment slop	1,2,3	Numeric
12	fcvc	Main vessel count colored Fluoroscopically	0,1,2,3	Numeric
13	thal	Normal, Fixed, Reversible	0,1,2	Numeric
14	target	Disease is present or not	0,1	Binary

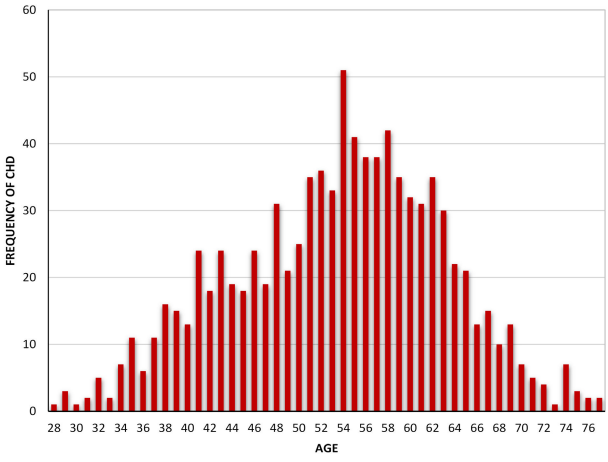

FIGURE 2. CHD age distribution.

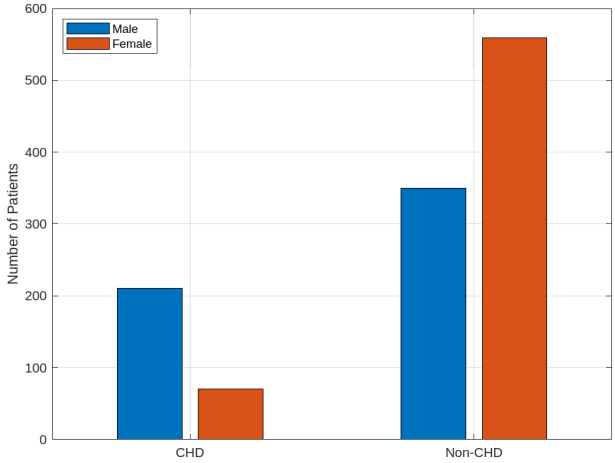

FIGURE 3. CHD gender distribution.

Table 5 presents the attributes together with the values that correspond to them.

The data intended for feature analysis was imported and represented in a visualized format. Following this,

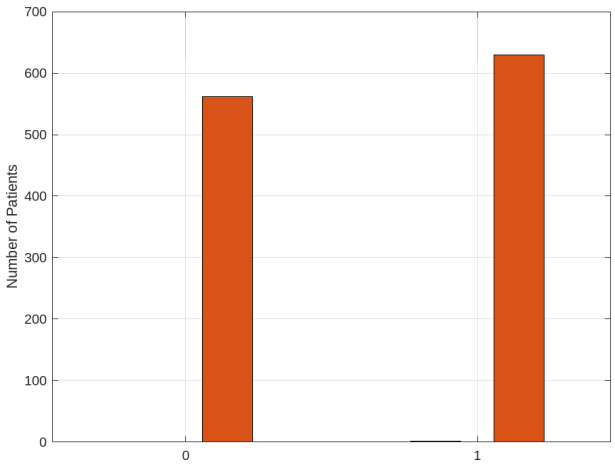

FIGURE 4. CHD all dataset distribution.

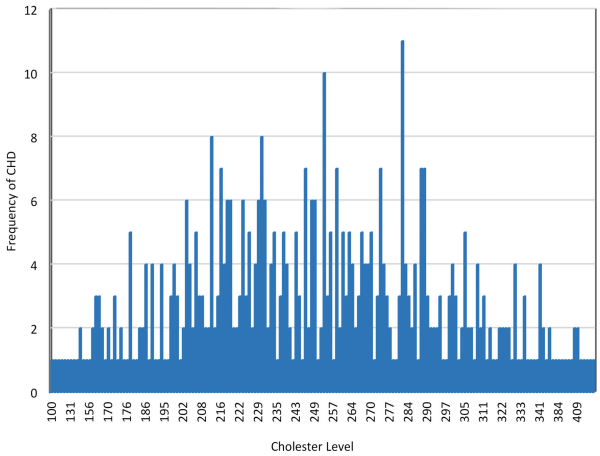

FIGURE 5. CHD cholesterol distribution.

the distribution of CHD cases was evaluated. The data distribution within individual fields was examined using visualization methods for confirmed CHD cases. Moreover, the correlations between CHD and all variables were explored using these visualization techniques.

Figure 2 illustrates the distribution of samples within different age groups for the Cleveland dataset that depicts the distribution pattern among those with CHD and those without. The vertical axis represents the cumulative count of samples belonging to each age group. In the present context, the X-axis is utilized for binary classification, where the value 0 denotes persons who do not have CHD. In contrast, the value 1 signifies those diagnosed with CHD. Figure 2 successfully demonstrates the segmentation of age groups within the dataset according to their CHD condition. The provided analysis offers valuable insights into CHD prevalence among individuals of various age groups. Moreover, it presents a visually intuitive picture of the correlation between age and the occurrence of this condition within the dataset under consideration.

Figure 3, labeled as "Gender Wise Distribution", presents a comprehensive analysis of the gender distribution in the Cleveland dataset, emphasizing the incidence of CHD for each gender category. Figure 3 represents the relationship between two types, "CHD" and "Non-CHD," while gender is divided into two groups, "Female" and "Male." The dataset indicates that among the 280 female participants, 211 individuals (75%) have been diagnosed with CHD, while the remaining 70 individuals (25%) do not exhibit this condition. The dataset comprises 909 people in the case of males. In the sample, 350 males (38.50%) had CHD, while the remaining 559 males (61.50%) did not exhibit this condition. The provided figure depicts the distribution of CHD among different genders within the dataset, offering valuable insights into the comparative prevalence of this disease among females and males; comparatively, males are more affected than females.

Figure 4 illustrates the dataset consisting of 1190" succinctly outlines the composition of the Cleveland dataset. The dataset consists of a total of 1190 people. Out of the overall sample size of 561 participants, constituting approximately 47% of the entire set, it was observed that none of these individuals had been diagnosed with CHD. On the other hand, the remaining 629 participants, accounting for approximately 53% of the overall sample, had received a diagnosis of CHD. The figure presented in this analysis clearly illustrates the distribution of CHD throughout the dataset, emphasizing the high occurrence of this condition among the patients involved in the study.

Figure 5 concisely depicts the dataset's composition and the inherent distribution of CHD within it. The dataset comprises 473 data points, representing approximately 39% of the population. The horizontal axis of the figure displays cholesterol levels, which span from 100 to 603. Meanwhile, the vertical axis represents the number of persons impacted by CHD. The notable observation in this graph pertains to the clustering of CHD instances within the cholesterol level interval of 200 to 290, suggesting an elevated vulnerability to CHD within this specific range. The visual depiction clearly explains the natural distribution of CHD within the dataset, providing insight into the prevalence of CHD across various cholesterol levels.

2) DATA CLEANING

Data cleansing is a crucial phase in data preprocessing, as many available data sets contain impurities and require extensive procedures. Addressing missing variables and discrepancies is essential for reliable outcomes in machine learning models. Data preprocessing is essential for improving data quality and optimizing the performance of subsequent models. Accurate data acquisition processes can lead to noisy data or missing values. To enhance the dataset's quality, various preprocessing approaches have been applied, including:

- **Filling the missing values:** Managing missing values is a significant challenge in the process of data preparation. Different techniques, such as imputation with global

values, mean imputation, or standard deviation imputation, are commonly utilized. However, it is important to highlight that upon evaluation of our original dataset using the 'pandas' library, all values were present.

- **Removing duplicate records:** The identification and removal of duplicate records is an additional crucial stage in the process. By utilizing the 'pandas' library, we successfully detected and eliminated a total of 151 duplicate entries from our dataset, thereby ensuring the integrity of our data.
- **Encoding categorical data:** Categorical data encoding refers to the process of transforming categorical variables into numerical representations to facilitate analysis and modeling in various domains. It is very important to encode categorical data to turn categorical values into numerical representations that can be used with machine learning algorithms. It is worth mentioning that we utilized one-hot encoding for properties such as 'cp,' 'fbs,' 'restecg,' 'exang,' and 'slope.' At the same time, the attributes 'sex' and 'target' were not modified, as they already yielded binary outputs (0 and 1).
- **Feature scaling:** is used to ensure that independent characteristics have consistent ranges. In our particular scenario, we decided to employ Z-score normalization where each data unit is modified as follows:

$$Z = \frac{(X - M)}{SD} \tag{1}$$

The calculation uses Eq. 1 to adjust each data point, with 'SD' representing the standard deviation and 'M' representing the mean of all values about a specific attribute.

- **Feature Selection:** It is important to note that not all attributes possess similar significance in the prediction of CHD. Consequently, a feature selection process was performed to keep only the most relevant properties. The wrapper-style Recursive Feature Elimination (RFE) method was utilized to rank the characteristics according to their significance. Through an iterative process, the least relevant attributes were deleted, resulting in a reduction of the attribute count from 14 to 12. To prevent overfitting, the attributes 'fcvc' and 'thal' were chosen to be excluded from our model.
- **Outlier detection:** The presence of outliers, which have the potential to generate noise, was discovered and mitigated using the Inter Quartile Range (IQR) approach. When values were found to be more than 1.5 times the IQR, they were changed to the mean value for continuous features or the mode value for categorical features. Fortunately, the datasets analyzed in our study did not exhibit any outliers, confirming the integrity of our data analysis.

The proposed model is based on a systematic data preprocessing approach, which includes handling missing values, removing duplicates, encoding categorical variables, normalizing features, feature selection, and outliers. The focus then shifts to implementing the model, which uses Python programming and data science libraries like Pandas and NumPy for label encoding. The model's details are explored in Section III-E.

C. THE PROPOSED FEATURE EXTRACTION APPROACHES

The process of feature engineering encompasses several fundamental processes. First and foremost, it is imperative to address the issue of missing values, especially in datasets of considerable size. This can be achieved by employing techniques such as mean imputation. Furthermore, it is essential to detect and remove outliers that can potentially affect the interpretation of data. Finally, the data transformation process is conducted to convert variables into suitable formats for analysis. This may involve translating categorical variables into numerical representations. The aforementioned strategies collectively aim to improve data quality, minimize errors, and address biases, guaranteeing precise and dependable outputs in data analysis.

D. PROPOSED ML-BASED ALGORITHMS

ML is a branch of artificial intelligence that uses computational models and automation to detect patterns and make intelligent decisions. It focuses on developing algorithms and technologies to learn from data, divided into supervised and unsupervised learning. ML models are used in the context of cardiovascular disease to identify patterns and predict disease presence. Decision trees are an effective data mining technique for disease prediction, offering advantages such as easy classification based on relevance, transparent and understandable tree structures, efficient data classification with numerous attributes, and the ability to use data without separate preprocessing steps, simplifying the process.

1) TWO-CLASS LOGISTIC REGRESSION

Logistic regression is an ML-based technique that estimates regression on proportional, proportional, or categorical data. It allows for probabilistic analysis of categorical data and can predict the occurrence of a specific event using an independent variable directly influencing the dependent variable. It is used in industries like medicine, telecommunications, and finance to predict event probability. Logistic regression is a version of linear regression, often applied in mining data, automatic diagnostics, and economic projections. It is particularly useful in predicting and classifying CHD. The model transforms input values into predicted values, assigned to a Sigmoid function.

The logistic regression function is expressed as in Eq. 2.

$$\hat{Y} = \sigma(Z) \tag{2}$$

$$Z = X^T \cdot W + b \tag{3}$$

In Eq. 3 Z is the linear transformation. This linear transformation is converted through the Sigmoid function to approach the true value Y of the predicted value $\hat{Y}$. The

Sigmoid function is defined as in Eq. 4.

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{4}$$

In this context, $\hat{Y}$ is the predicted output, X is the vectorized matrix of the sample set, nx represents the number of features per sample, making X an $nx \times m$ matrix, W is an $(nx \times 1)$ matrix, b is a constant, and A is the predicted result. After calculating the predicted output, the loss function is computed as in Eq. 5.

$$J(w, b) = \frac{1}{m} \sum_{i=1}^{m} L(a_i, y_i) \tag{5}$$

Subsequently, the gradients are calculated through the differentiation of the composite function, as in Eqs. 6, 7 and 8.

$$\frac{\partial J(w, b)}{\partial Z} = A - Y \tag{6}$$

$$\frac{\partial J(w, b)}{\partial W} = \frac{1}{m} X^T \cdot \frac{\partial Z}{\partial W} \tag{7}$$

$$\frac{\partial J(w, b)}{\partial b} = \frac{1}{m} \sum \frac{\partial Z}{\partial b} \tag{8}$$

After updating w and b and iterating several times, these steps are repeated until the derivatives minimize the cost function $J(w, b)$. LR is a classic classifier with the advantages of simple implementation, low computational cost, fast speed, and low storage requirements. It also allows convenient observation of sample probability scores. However, logistic regression faces the issue of overfitting, thus requiring regularization. Common regularization methods include $L1$ and $L2$ regularization. The loss function of logistic regression with L1 regularization includes an $L1$ norm as a penalty, with a hyperparameter α adjusting the penalty term. The loss function for binary logistic regression with $L1$ regularization is expressed as in Eq. 9.

$$J(\theta) = -[Y \log(h(X)) + (1 - Y) \log(1 - h(X))] + \alpha \|\theta\|_1 \tag{9}$$

where $\|\theta\|_1$ is the L1 norm of θ. Optimization methods for $L1$ regularized logistic regression commonly include coordinate descent and least angle regression. The loss function for binary logistic regression with L2 regularization as in Eq. 10.

$$J(\theta) = -[Y \log(h(X)) + (1 - Y) \log(1 - h(X))] + \alpha \|\theta\|_2^2 \tag{10}$$

where $\|\theta\|_2$ is the $L2$ norm of θ. The optimization methods for $L2$ regularized logistic regression are similar to those for ordinary logistic regression. Generally, $L1$ regularization can produce a sparse weight matrix, creating a sparse model useful for feature selection. $L2$ regularization can prevent model overfitting, and to some extent, $L1$ can also prevent overfitting. Since PCA has already been used for dimensionality reduction, feature selection is unnecessary. This study aims to address the multicollinearity problem, thus adopting $L2$ regularization. If the learning rate is too low, convergence is too slow; if too high, the cost function oscillates. Therefore, this study initially sets the regularization coefficient λ to 0 to determine a good learning rate. After fixing the learning rate, λ is gradually adjusted based on the accuracy of the test set. In this experiment, the regularization coefficient is set to 1.0.

2) DECISION TREE (DT)

Decision trees are a machine learning technique used to predict and diagnose CHD. They sort patient instances down a tree structure, starting at the root node and progressing to the leaf nodes. Each branch is created by selecting a feature to split the data, and intermediate nodes further divide the data based on specific testing conditions related to CHD risk factors. This structured approach helps isolate and classify patterns within the data, aiding in accurate prediction and diagnosis. This study constructs a decision tree model using entropy from information theory, which measures the complexity of raw data. The advantages of the decision tree classification algorithm include its simple structure, clear classification rules, and visual display of decision-making processes. It is highly interpretable and does not require prior knowledge beyond training data. Decision tree predictive models have been effectively applied to various clinical diseases, including CHD. The multi-valued attribute multi-label decision tree algorithm is a novel approach to the classic decision tree framework, focusing on selecting the splitting attribute and deciding when to stop nodes.

3) TWO-CLASS BOOSTED DECISION TREE

The study aims to develop an ML-based model for predicting cardiovascular disease using the Two-Class Boosted Decision Tree module. Boosted decision trees are an ensemble learning method that corrects errors of previous trees, making predictions based on the entire ensemble. However, they are memory-intensive and may not handle large datasets as efficiently as linear learners. To address this challenge, greedy heuristics have been proposed, which are recursive and employ a top-down approach. Decision trees combine mathematical and computational techniques to describe, categorize, and generalize data. The study focuses on cardiovascular disease by leveraging clinical data to identify patterns and predictors of disease presence. By incorporating patient attributes like age, cholesterol levels, and blood pressure, the model aims to accurately classify individuals into different risk categories for cardiovascular disease. The effectiveness of the model is evaluated through performance metrics.

4) TWO-CLASS NEURAL NETWORK

The Two-Class Neural Network algorithm is a popular machine learning method for binary classification tasks in cardiovascular disease. It uses ANNs modeled after the human brain's structure and function. ANNs are effective for predicting and classifying cardiovascular disease through

advanced computer simulations, and analyzing complex patterns using patient data like age, cholesterol levels, blood pressure, and lifestyle factors. This study developed an ANN model to predict cardiovascular disease risk, focusing on training functions, transfer functions, layers, and neurons per layer. The model can be clinically applied to diagnose and screen high-risk individuals early. However, further research is needed to systematize ANN applications and integrate them with established statistical theories for enhanced clinical use.

E. MODEL TRAINING

This study aims to improve the prediction model for timely CHD detection by providing an enhanced dataset for automated disease detection. It will allow us to achieve higher accuracy in our work. This part of the article explains the methodology utilized in modeling to predict CHD. The training process of the model is illustrated in Figure 6. The initial step involves acquiring a dataset, which is a compilation of data encompassing both the input and output values necessary for the model. The dataset is subsequently split into a training set and a test set. Subsequently, an ML algorithm is chosen to train the data. Figure 6 depicts two distinct ML models: an LR and NN models. The two ML models operate independently, but they are visually represented in a unified diagram to conserve space. These ML models can be executed in any order without a predetermined sequence. The ML algorithm refers to the specific algorithm employed by the model to acquire knowledge and insights from the provided training data. After selecting ML methods, the model undergoes training using the training set. This process entails providing the training data to the algorithm and enabling it to acquire knowledge from the data. After completing the training process, the model is evaluated using the test set. This process entails providing the test data to the model and evaluating its level of correctness. Once the model has achieved sufficient accuracy, it can be implemented in a production environment. This implies that the model can provide forecasts based on novel data.

The same process is repeated for training the next two ML models, as shown in Figure 7. The first step is obtaining a dataset containing the model's input and output values. The dataset is split into training and testing datasets. Next, an ML method is selected for data training. Two ML models run separately but are visually displayed in a single figure for space efficiency. These ML models can be run in any order without a predetermined sequence. Following training, the model is tested using the test set. This procedure involves feeding the model test data and assessing its accuracy. After achieving adequate precision, the model can be used in production. This means that the model can foresee novel data.

Following that, cross-validation and comparisons were carried out, as is demonstrated in Figure 8 and 9. In the end, the test results were collected, including the true positive (TP), true negative (TN), false positive (FP), and false negative (FN), accuracy, precision, recall, F1 score, and AUC.

TABLE 6. Confusion matrix.

Classifier prediction	Actual judgment	
	TRUE	FALSE
Postive	TP	FP
Negative	FN	TN

F. MODEL EVALUATION METRICS

The process of creating classifiers based on datasets involves evaluating their performance using metrics like precision, recall, F-score, and accuracy, using the confusion matrix.

The confusion matrix is a visual representation of a classification algorithm's performance, displaying four distinct results: TP, TN, FP, and FN. The table's dimensions are 2×2, and it was evaluated using various indicators, as suggested in previous research. studies [77], [78], [79]. When formulating predictions on occurrences, it is feasible to identify four unique potential outcomes.

1) TP: Individuals diagnosed with CHD who are accurately identified as having CHD. For instance, if a diagnostic test accurately detects 80 out of 100 individuals with CHD, the true positives would be 80.
2) FP: Individuals who do not have CHD are erroneously classified as having CHD. For instance, if a diagnostic test erroneously classifies 10 out of 100 patients who do not have CHD as having CHD, then the count of False Positives would be 10.
3) FN: refers to cases where individuals with CHD are inaccurately classified as not having CHD. If a diagnostic test cannot correctly detect 15 out of 100 patients with CHD, then the number of false negatives is 15.
4) TN: Individuals who do not have CHD are accurately identified as not having CHD. For instance, if a diagnostic test accurately detects 90 out of 100 people who do not have CHD, then the count of true negatives would be 90.

An FP is a Type I error, also known as an alpha error, while an FN is a Type II error, also known as a beta error. In hypothesis testing, a false null hypothesis (H0) implies a false alternative hypothesis (H1). These concepts are essential for assessing the effectiveness of diagnostic tests. They are frequently used to compute sensitivity, specificity, accuracy, and positive and negative predictive values. These parameters aid in evaluating the diagnostic test's ability to differentiate between individuals with the specific condition being studied and those who do not possess it.

The accuracy metric holds significance in evaluating models, as it measures the proportion of accurate predictions generated by the classifier compared to the true label values during the testing phase. The confusion metric shown in Table 6 denotes the ratio of accurately classified occurrences to the overall number of instances. The accuracy of a

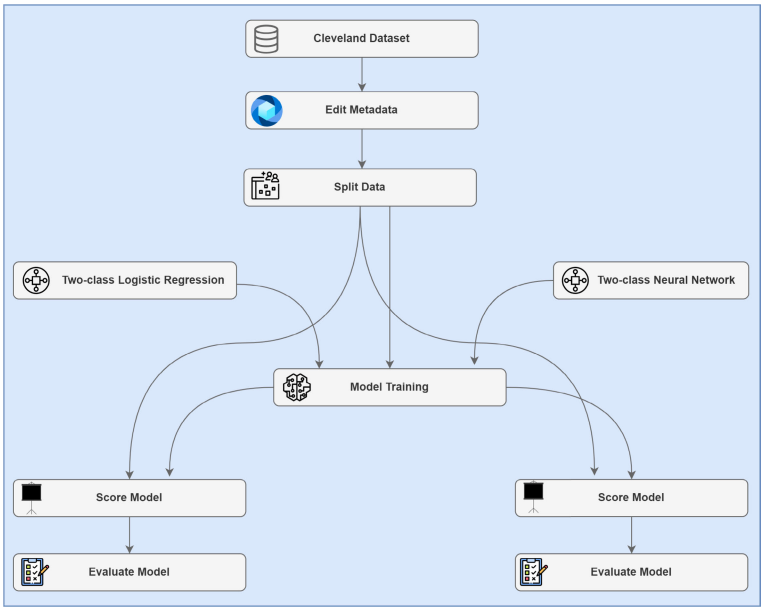

FIGURE 6. Two-Class LR and NN model training.

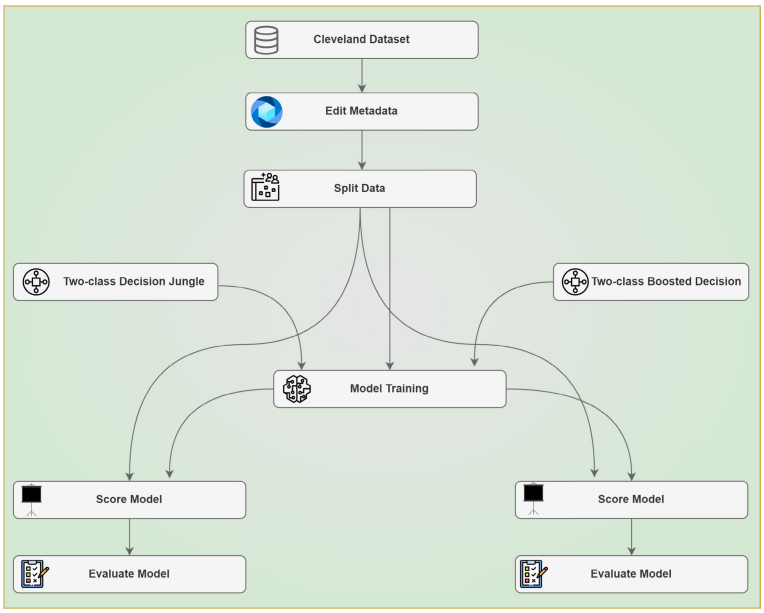

FIGURE 7. Two-Class decision jungle and boosted decision model training.

measurement can be determined by utilizing Eq. 11 as presented in reference [80].

$$Accuracy = \frac{(TN + TP)}{(TN + TP + FN + FP)} \tag{11}$$

Precision is an essential measurement for evaluating the degree of accuracy. Eq. 12 demonstrates the proportion of instances the classifier identifies as positive to the overall number of predicted positive occurrences.

$$Precision = \frac{TP}{(TP + FP)} \tag{12}$$

The concept of recall is related to the proportion of events belonging to a specific class that we correctly classified as belonging to that class. It shows how successfully your model identifies all positive cases (e.g., sick persons with medical conditions). Measuring the completeness of your model's positive predictions indicates the percentage of Tp correctly identified by the model. This means how many sick people your diagnostic test accurately diagnoses in medicine. Eq. 13 demonstrates the calculation of the Tp rate, which is the ratio of Tp to all positive instances [81].

$$Recall = \frac{TP}{(TP + FN)} \tag{13}$$

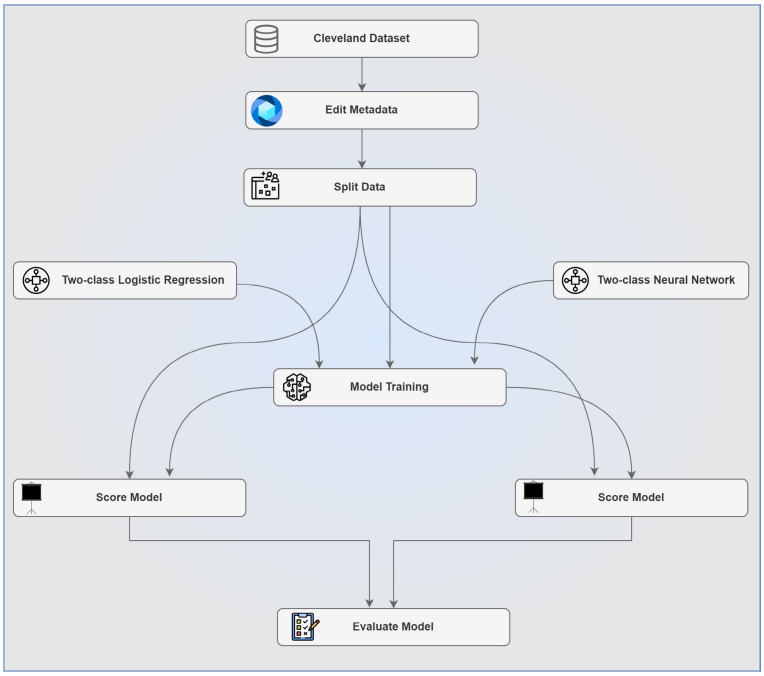

FIGURE 8. Cross validation of LR and NN.

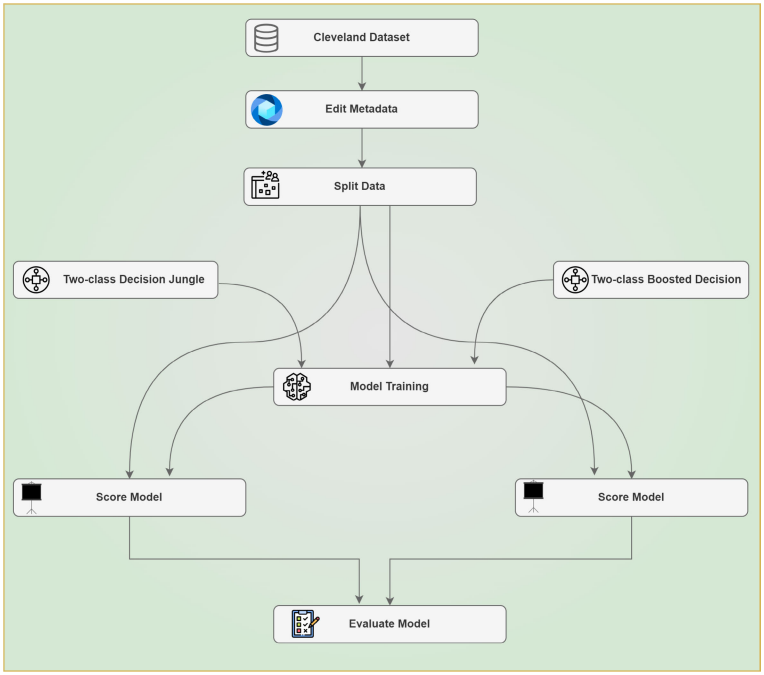

FIGURE 9. Cross validation of decision jungle and boosted decision.

Accuracy is not useful for classification issues involving a skewed distribution since accuracy measures just the center of the distribution. Instead, precision and recall are significantly more indicative of the whole. It is possible to obtain the F1 score by combining the two metrics of accuracy and recall. The F1 score is the weighted average (harmonic mean) of the precision and recall scores. The score can be anywhere from 0 to 1, with 1 being the highest possible F1 score (the harmonic mean is used when discussing ratios), as demonstrated in Eq. 14:

$$F1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} \tag{14}$$

The ROC curve is a tool used to analyze True Positive Rate (TPR) and False Positive Rate (FPR) differences across decision thresholds in academic discourse. The AUC is

used to measure a model's performance, making comparison easier. AUC equals 1 for accuracy, 0.5 for classification performance, and lower than 0.5 for classification performance. The ROC curve is a line connecting all sample points, with a higher TPR indicating a larger ratio of correct judgment. Reverse prediction would improve the model's performance. The ROC curve is a useful tool for comparing models in academic discourse. Put another way, the performance is greater when the area covered by the ROC curve (AUC) has a larger value [82].

IV. RESULTS ANALYSIS AND DISCUSSION

The findings of the proposed study show encouraging developments in machine learning-based CHD prediction. Several machine learning models were trained and assessed using a precise experimental setup and meticulous data preprocessing, including feature scaling and encoding. The Two-Class Decision Jungle and Two-Class Boosted Decision models, which achieved outstanding AUC ratings of 0.976 and 0.991, respectively, stood out as top performers. These high AUC values are higher than those predicted by earlier studies, which shows a significant improvement in the CHD prediction methods. The emphasis on AUC as a primary evaluation parameter made choosing models with greater discrimination abilities possible, thus boosting their clinical value. The models performed admirably on accuracy, Precision, Recall, and F1 Score parameters, highlighting their potential for precise CHD risk estimation. Visualizations of training and validation accuracy, as well as loss, provide insights into the dynamics of model training and help spot possible overfitting or underfitting problems. Additionally, sensitivity analysis using ROC curves and confusion matrices provided a thorough understanding of model behavior and its capacity to distinguish between positive and negative CHD instances. These findings represent a significant advance in the early diagnosis and treatment of CHD and provide insightful paths for further study using unsupervised machine learning and deep learning methods.

A. EXPERIMENTAL SETUP

Python was the main programming language used in the suggested study, and some fundamental ML libraries, including Scikit-Learn, Pandas, NumPy, and Matplotlib/Seaborn, enabled in-depth experimentation. Preprocessing of the obtained dataset was rigorous and included management of missing data, feature scaling, and categorical variable encoding. Different machine learning models were implemented and tested, including boosted decision trees, decision jungles, neural networks, and logistic regression. Performance parameters like Accuracy, Precision, Recall, F1 Score, and AUC were used for detailed model evaluation. The emphasis on AUC ratings made it easier to choose a model, while visualizations helped with result interpretation.

B. PERFORMANCE ANALYSIS OF THE PROPOSED METHODS

This study demonstrated that the participants had an increased risk of CHD due to the type of chest pain they experienced, high cholesterol levels, or higher blood pressure [83], [84]. This was determined by the data input and feature classification described above. After model training, outcome models were stored, tested, cross-validated, and compared after completion of storage and testing. The results are shown in Figures 10a-b, 11a-b and 12a-b. These metrics include the true positive, the false positive, the false negative, the true negative, and the true positive. Table 7, illustrates the evaluation results of different classification models, specifically Two-Class Logistic Regression, Two-Class Neural Network, Two-Class Decision Jungle, and Two-Class Boosted Decision. The table contains essential performance indicators: TP, FP, FN, TN, Accuracy, Precision, Recall, F1 Score, and AUC. These metrics provide a thorough summary of the efficacy of each model in differentiating between two groups. The Two-Class Boosted Decision model has outstanding performance, achieving the highest values in Accuracy (0.96), Precision (0.93), Recall (0.93), F1 Score (0.93), and AUC (0.991). The table is helpful for readers who want to compare and analyze different models. It helps them understand the strengths and weaknesses of each model relative to the evaluation criteria provided.

To validate the data presented in Table 6 and evaluate the accuracy of the models, the previously discussed Eqs. 11, 12, 13 and 14 were utilized to fill in the values. The Eq 11 represents the computation of accuracy, an essential measure in assessing classification models. Accuracy is a metric that quantifies how much a model's predictions are correct. It is computed by dividing the number of accurately classified examples (true negatives and true positives) by the total number of occurrences. The numerator (TN+TP) indicates the count of accurately classified cases, while the denominator (TN+TP+FN+FP) reflects the total count of instances. The accuracy is computed by dividing the correct predictions by the cases. The resulting value falls between 0 and 1, representing flawless categorization.

C. RESULTS ANALYSIS COMPARISON FOR VARIOUS PROPOSED ARCHITECTURES

This research assessed the effectiveness of multiple ML models in predicting the likelihood of CHD using a dataset containing various patient variables. These models' accuracy results are impressive. The accuracy of the Two-Class Logistic Regression model shown in Figure 10a, which measures its capacity to categorize cases accurately, is 0.84%. With an amazing 0.92%, the Two-Class Neural Network model outperformed this accuracy, demonstrating better prediction skills. The Two-Class Boosted Decision model outperformed all others with an accuracy of 0.96%, demonstrating its durability in accurate case classification. In addition, the Two-Class Decision Jungle model significantly increased the accuracy, reaching 0.94%.

TABLE 7. Evaluation and comparative assessment of outcomes utilizing diverse metrics in the context of training (Train) and validation (Val) phases.

Evaluation Criteria and Metrics for the Model														
Model	TP	FP	FN	TN	Accuracy		Precision		Recall		FScore		AUC	
					Train	Val	Train	Val	Train	Val	Train	Val	Train	Val
Two-Class Logistic Regression	285	90	144	905	0.84	0.87	0.76	0.79	0.66	0.67	0.71	0.73	0.87	0.88
Two-Class Neural Network	319	67	43	871	0.92	0.93	0.83	0.84	0.88	0.90	0.85	0.86	0.96	0.97
Two-Class Decision Jungle	345	37	44	845	0.94	0.95	0.90	0.91	0.89	0.90	0.89	0.91	0.97	0.98
Two-Class Boosted Decision	357	27	27	833	0.96	0.97	0.93	0.94	0.93	0.95	0.93	0.96	0.99	0.99

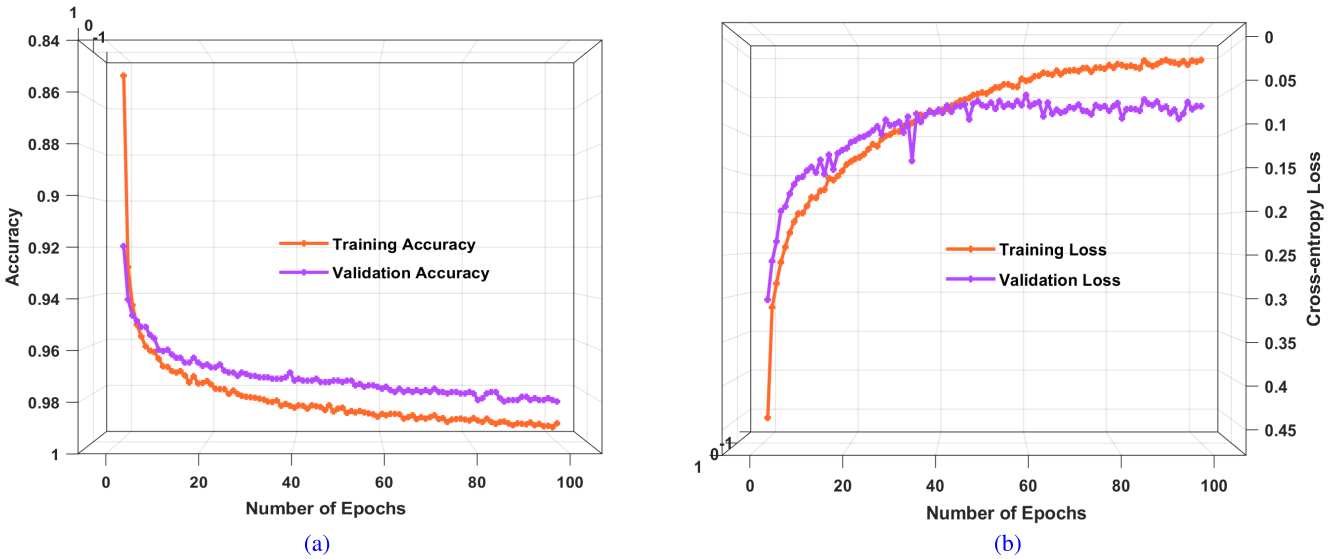

FIGURE 10. (a) Shows the training and the validation accuracy and (b) shows loss of the proposed fine-tuned VGGNet framework.

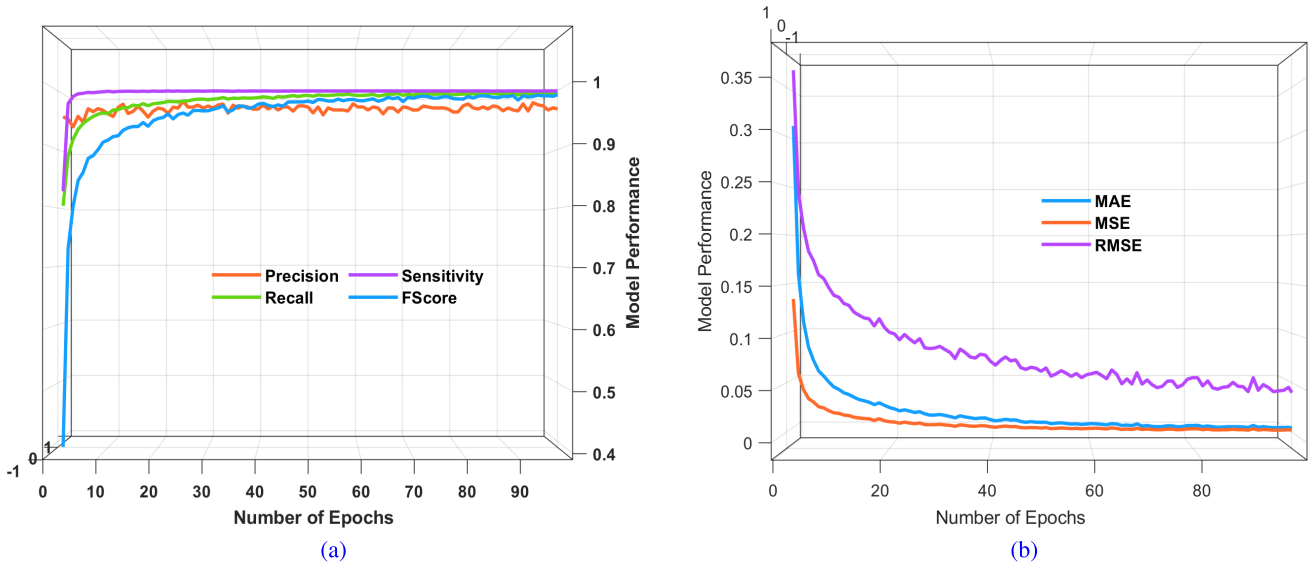

FIGURE 11. (The figure(a-b) represents a comprehensive performance comparison, featuring precision, sensitivity, recall, and F1-Score metrics, as well as MAE, MSE, and RMSE for the proposed model.

The "Two-Class Boosted Decision" technique shown in Figure 11a performs better than the other models assessed in predicting CHD within the provided dataset. The model attained a maximum accuracy score of 0.96%, suggesting

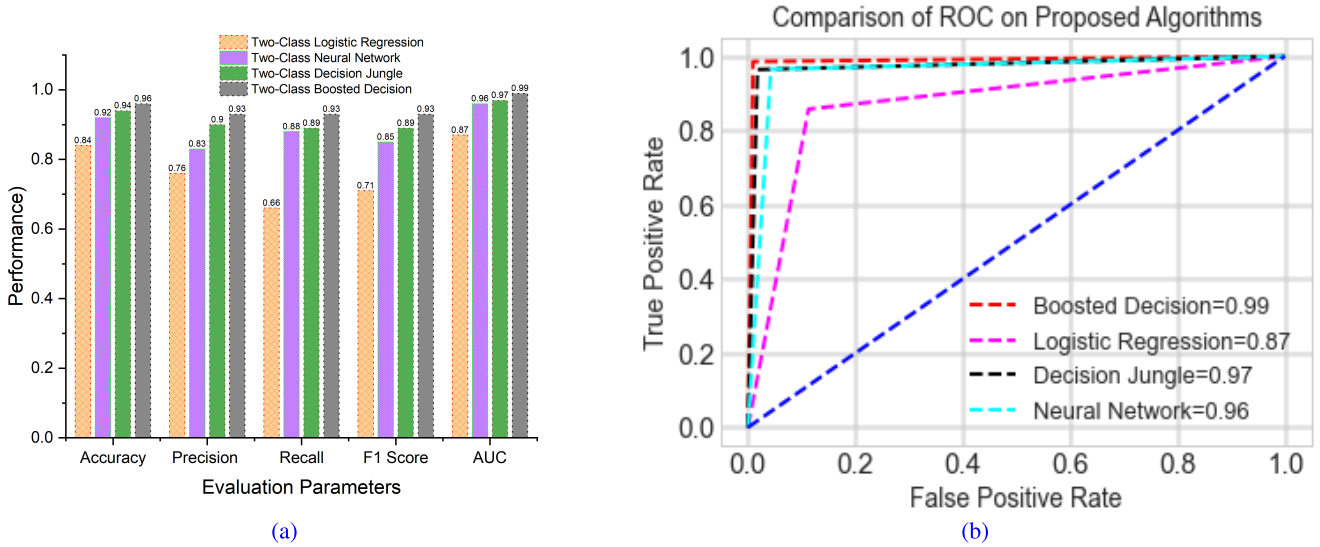

FIGURE 12. Shows the overall comparison of the proposed model and area under the curve of all the proposed models.

its superior performance in accurately classifying instances of CHD. Precision is a key factor for assessing models' performance, particularly in predicting CHD. Precision is a metric that accurately evaluates the model's capacity to classify positive instances (specifically, cases of CHD) among all the instances it predicts as positive. A greater level of precision indicates a reduced occurrence of false positives. The ''Two-Class Decision Jungle'' and ''Two-Class Neural Network'' approaches have precision scores of 0.9% and 0.83%, respectively. The ''Two-Class Logistic Regression'' model exhibits a precision value of 0.76%, suggesting a relatively elevated occurrence of false positives compared to the remaining models. Based on the precise findings in Figure 11a, it is evident that the technique known as ''Two-Class Boosted Decision'' has the highest precision score of 0.93%. This implies that the model's accuracy in correctly predicting positive cases of CHD is 93%, a noteworthy achievement. The result below shown in Figure 11b indicates that the model effectively identifies 93% of patients with CHD and demonstrates a low occurrence of false negatives, which is critical in predicting CHD, as overlooking positive cases can lead to significant repercussions. Furthermore, the ''Two-Class Decision Jungle'' and ''Two-Class Neural Network'' methodologies exhibit notable recall scores of 0.89% and 0.88%, correspondingly suggesting their efficacy in identifying a substantial fraction of positive CHD instances. On the other hand, the ''Two-Class Logistic Regression'' model exhibits a comparatively lower recall score of 0.66%, indicating a heightened likelihood of false negatives and the potential for overlooking certain cases of CHD. Upon examination of the F1 scores of the various models shown in Figure 12a, it becomes apparent that the ''Two-Class Decision Jungle'' model has attained the highest F1 score of 0.98%. This outcome indicates the model's strong performance, effectively balancing precision and recall. The F1 score is a statistic that holds particular utility in scenarios characterized by an imbalanced distribution of classes, as it considers both false positives and false negatives. Hence, the ''Two-Class Decision Jungle'' model demonstrates exceptional proficiency in attaining a harmonious equilibrium between accurately detecting CHD cases and minimizing misclassifications. The ''Two-Class Boosted Decision'' model demonstrates a noteworthy F1 score of 0.93%, suggesting its capacity to offer a favorable balance between precision and recall. The approach achieves a harmonious equilibrium by effectively discerning CHD cases while mitigating false positive results. In contrast, the ''Two-Class Neural Network'' model exhibits a commendable F1 score of 0.85%, and the ''Two-Class Logistic Regression'' F1 score is 0.71%, showcasing its efficacy in terms of the model's overall performance. Upon analysis of the Area Under the Curve (AUC) values related to the various models shown in Figure 12b, it becomes apparent that the ''Two-Class Boosted Decision'' model has attained the highest AUC score of 0.991%. This outcome signifies a remarkable ability to differentiate between positive and negative cases of CHD, thus highlighting its excellent discriminatory capacity. A large AUC indicates that the model has great discriminatory power in prioritizing positive events over negative cases, rendering it a promising choice for predicting CHD. The ''Two-Class Decision Jungle'' model also has a notable AUC score of 0.976%, indicating its strong ability to classify data. The method demonstrates proficiency in distinguishing between positive and negative instances, providing additional evidence to support its effectiveness in detecting CHD. The ''Two-Class Neural Network'' model demonstrated a notable AUC score of 0.966%, which suggests a robust discriminatory capacity. Although its predictive capabilities for CHD are not the highest compared to other models, it demonstrates significant predictive capabilities. The ''Two-Class Logistic

Regression'' model exhibits a commendable AUC score of 0.87%, yet it lags below the other models regarding its discriminative capacity. The ''Two-Class Boosted Decision'' model demonstrates superior performance in terms of AUC, with the ''Two-Class Decision Jungle'' closely trailing behind. These models are especially suitable for predicting CHD in situations where establishing a high level of differentiation between positive and negative instances is of utmost importance.

D. SENSITIVITY ANALYSIS OF CUTTING-EDGE ARCHITECTURES

A sensitivity assessment is a statistical approach used in coronary heart risk analysis to evaluate the impact of uncertainties in input data on the accuracy of lesion prediction models, as shown in Figure 13a-d. This analysis helps identify the critical features that affect the detection model's performance, and researchers can optimize the model's parameters and configuration accordingly. Statistical metrics like mean absolute error (MAE), mean squared error (MSE), root means squared error (RMSE), sensitivity, and AUC, are employed to assess the prediction error between the predicted and target values. Quantifying the influence of each input variable on the model's output can improve the reliability and effectiveness of the coronary disease diagnosis process, and more accurate detection and classification systems can be developed. The sensitivity analysis results are represented in the Table 8.

E. COMPARING WITH THE CUTTING-EDGE APPROACHES

Table 9 presents a comprehensive analysis of the current methodologies employed in the classification of cardiac disease, as described in the literature [54], [85], [86], [87], [88], [89]. The main purpose of this study is to investigate the implementation of different models within the specific domain of heart disease diagnosis. This is achieved by using datasets such as the UCI Cleveland Dataset, Cleveland Heart Disease Dataset, and the User Sensor Dataset. Various ML models were utilized in this study, including kNN, ANN, a hybrid model that combines DT and RF, Federated Learning (FL), SVM, and RFFS. The proposed approach involves integrating a multi-model strategy, which includes the utilization of Two-Class LR, Two-Class NN, Two-Class DJ, and Two-Class BD models. The novel methodology continually shows higher performance, with accuracy rates ranging from 84% to 96%, while maintaining the same set of 14 features as a reference point. The above result shows a big step forward in classifying heart disease. It shows an organized and systematic method that makes understanding the characteristics and their corresponding sizes easier.

The Analysis of Variance (ANOVA) test is a crucial statistical tool for comparing model performances in predicting Coronary Artery Disease using DL and ML algorithms. It helps determine if there is a significant difference in predictive capabilities between models like logistic regression, neural network, decision jungle and boosted DT. ANOVA supports the claim that the two-class Boosted DT outperforms its counterparts. It also assesses the significance of various features in CAD prediction, ensuring predictions are reliable and equitable for all demographic subgroups. ANOVA can also evaluate model stability across different periods or incorporate updates, highlighting the importance of statistical validation in implementing effective diagnostic approaches to reduce mortality rates associated with CAD.

Table 10 showcases an ANOVA test comparison of model performances in predicting Coronary Artery Disease using various algorithms, including LR, NN, DJ, and BDT, against existing methods like KNN, ANN, Hybrid model, Federated Learning, and SVM. Based on ANOVA test results and corresponding p-values, this comparison aims to discern the statistical significance of differences in predictive capabilities among the models. Significantly, the results indicate a pronounced superiority of the two-class Boosted DT model, particularly evident in its comparison with Federated Learning, where the p-value is remarkably low (0.0032). This marked performance enhancement underscores the efficacy of the Boosted DT model in accurately predicting CAD, offering promising avenues for early diagnosis and intervention. The emphasis on statistical validation underscores the importance of meticulous evaluation in assessing model performances, highlighting the pivotal role of the two-class Boosted DT model in advancing diagnostic approaches to mitigate CAD-related mortality rates. These insights contribute valuable knowledge to the realm of cardiovascular health, facilitating the development of more effective and timely interventions.

V. EXPERIMENTAL DISCUSSIONS

The significant implications of disease prediction on human life are evident. The study holds significant practical implications that stand to revolutionize clinical practices, patient care strategies, and healthcare system efficiencies. Through the application of sophisticated ML models for early detection of CHD, this research not only promises to elevate diagnostic precision but also facilitates the shift towards preemptive healthcare measures. Such advancements enable the customization of patient care, tailoring interventions to individual risk profiles well before the onset of critical symptoms. The rapid identification of diseases can result in prompt intervention and treatment, potentially mitigating death rates and diminishing the overall impact of diverse health conditions. Healthcare practitioners can optimize patient outcomes by implementing preventative measures and offering individualized treatment strategies by identifying health concerns at their early stages. Utilizing ML and DL algorithms assumes a pivotal significance in this particular context. The utilization of extensive datasets in analyzing patterns and indicators that may not be readily discernible to healthcare professionals enables the provision of a robust automated framework for disease prediction. The algorithms can effectively handle a diverse set of input variables, encompassing genetic information and lifestyle aspects,

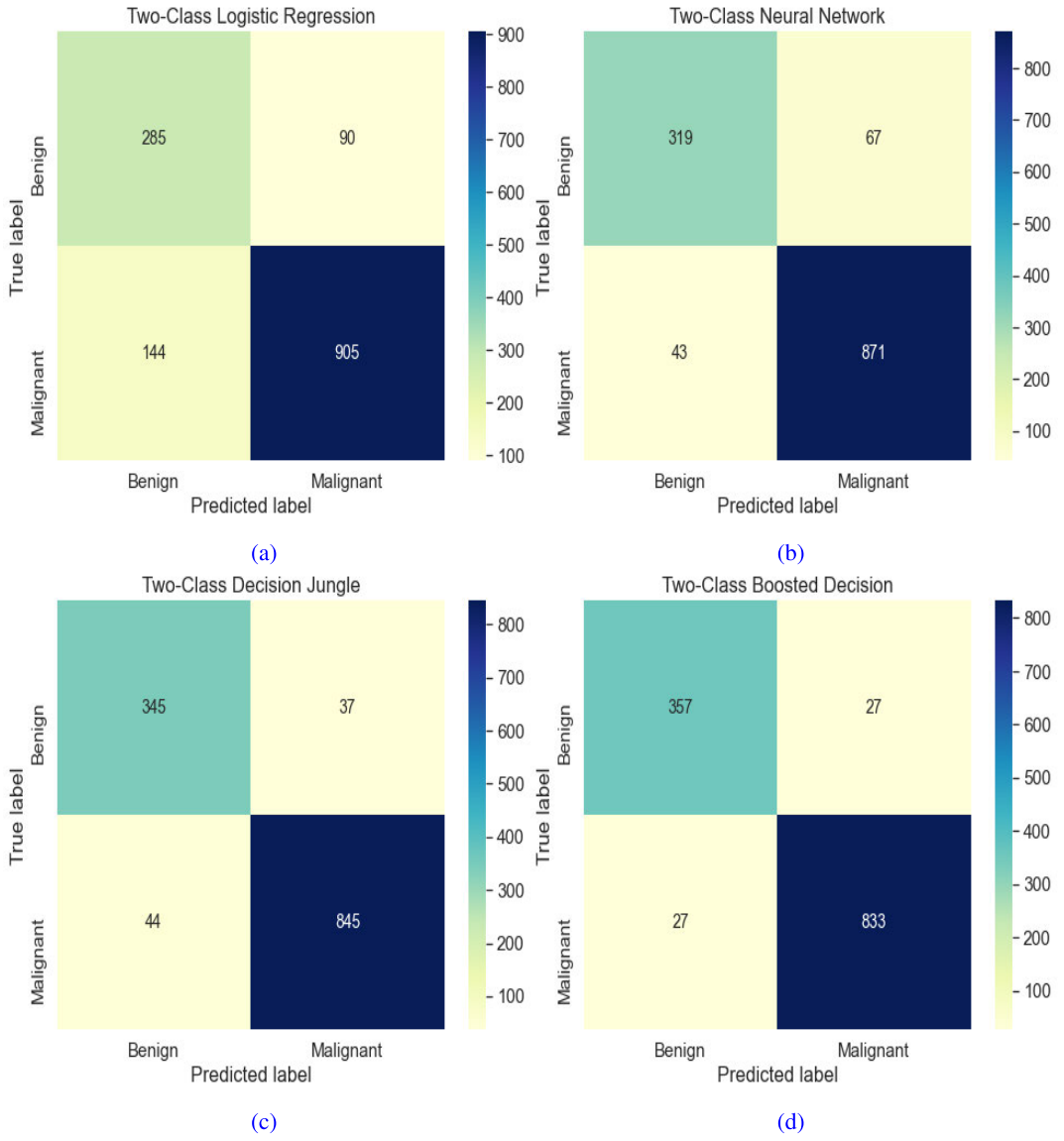

FIGURE 13. Accuracy and precision of the proposed models (a) Two-Class logistic regression, (b) Two-Class neural network, (c) Two-Class decision jungle and (d) Two-Class boosted decision.

TABLE 8. The proposed deep learning-based algorithms have certain environmental requirements to be met for optimal performance.

Proposed Models	MAE	MSE	RMSE	Sensitivity
Two-Class Logistic Regression	0.017	0.008	0.966	0.968
Two-Class Neural Network	0.012	0.009	0.091	0.979
Two-Class Decision Jungle	0.013	0.008	0.082	0.987
Two-Class Boosted Decision	0.010	0.005	0.069	0.997

facilitating a thorough examination of an individual's state of health. By utilizing this technique, it becomes feasible to generate forecasts regarding the susceptibility to diseases with a significant level of precision. Nevertheless, it is essential to acknowledge the existence of certain restrictions. A significant obstacle is the requirement for datasets of superior quality and diversity to train these algorithms adequately. Moreover, it is crucial to consider the ramifications of data privacy and ethical considerations in managing confidential medical information. Moreover, it is essential to note that although ML and DL algorithms have the potential to offer vital insights, their role should be seen as supplementary to, rather than a substitute for, the skill and judgment of medical professionals. Potential prospects in this subject include refining prediction models, augmenting data privacy and security measures, and developing user-friendly interfaces tailored specifically for healthcare providers. In conclusion, the significance of machine learning and deep algorithms in

TABLE 9. Comparison with state-of-the-art.

Reference	Year	Dataset	Model	Accuracy	Features
[54]	2020	UCI Cleveland Dataset	KNN	90%	14
[85]	2020	Cleveland Heart Disease Dataset	ANN	90%	14
[86]	2021	Cleveland Heart Disease Dataset	Hybrid model (DT and RF)	88%	14
[87]	2021	User Sensor Dataset	Federated Learning	81%	14
[18]	2022	Cleveland Heart Disease Datase	Hybrid model (RF and SVM)	88%	14
[88]	2023	Cleveland Heart Disease Dataset	SVM	90%	14
[89]	2023	Cleveland Heart Disease Dataset	Random forest-feature	86%	14
Proposed	2023	Cleveland Heart Disease Dataset	Two-Class LR Two-Class NN Two-Class DJ Two-Class BD	84% 92% 94% 96%	14

TABLE 10. Statistical ANOVA test comparison of the proposed model using accuracy with state-of-the-art models.

Proposed Model		Existing Methods	ANOVA Test	P-Value
Two-Class LR text	**VS**	KNN	4.0980	0.1802
		ANN	4.0980	0.1802
		Hybrid mode	1.0499	0.4132
		Federated Learning	6.1217	0.1318
		SVM	4.0980	0.1802
Two-Class NN text	**VS**	KNN	3.1671	0.2170
		ANN	3.1671	0.2170
		Hybrid mode	6.6537	0.1231
		Federated Learning	28.9328	0.0328
		SVM	3.1671	0.2170
Two-Class DJ text	**VS**	KNN	8.8450	0.0969
		ANN	8.8450	0.0969
		Hybrid mode	15.7907	0.0578
		Federated Learning	55.8407	0.0174
		SVM	8.8450	0.0969
Two-Class BD text	**VS**	KNN	24.6403	0.0382
		ANN	24.6403	0.0382
		Hybrid mode	39.6459	0.0243
		Federated Learning	46.1342	0.0032
		SVM	24.6403	0.0382

offering automated frameworks for disease prediction resides in their capacity to transform healthcare, potentially diminish mortality rates, and enhance the overall quality of life. Nevertheless, exercising caution and accountability when implementing these technologies while acknowledging their inherent limitations and ethical implications is imperative. This study investigates using different machine learning models to forecast using a dataset comprising multiple patient characteristics such as age, gender, cholesterol levels, chest pain type, resting blood pressure, resting ECG, maximal heart rate, and fasting blood sugar. The performance of the models is assessed by employing various metrics, including accuracy, precision, recall, F1 score, and AUC, as shown in Figure 12b. This proactive approach will likely enhance patient outcomes, reduce the incidence of emergency interventions, and substantially lower healthcare costs by minimizing the need for extensive treatments. Furthermore, the study acts as a beacon for integrating artificial intelligence in healthcare, demonstrating the potential of data analytics to transform patient data into actionable insights for risk prediction and management. It encourages a more informed clinical decision-making process, leveraging technology to support the early identification and management of health risks. Beyond immediate healthcare applications, the research sets a precedent for future innovation, encouraging the exploration of machine learning across various medical disciplines. Ultimately, this work not only contributes to the scientific understanding of CHD risk factors but also advocates for a more informed, efficient, and personalized healthcare ecosystem, showcasing the transformative power

of machine learning in advancing public health objectives. The findings of this study indicate favorable performance across the various models employed. However, it is essential to acknowledge certain areas that warrant more investigation and limits that should be considered.

A. STUDY LIMITATIONS

1) **Dataset Size:** The dataset's size is one limitation of the current study. A larger and more diverse dataset would improve the generalizability of the models and potentially lead to more accurate predictions.
2) **Data Imbalance:** Imbalanced class distribution, where positive CHD cases are relatively fewer than negative cases, can impact model performance. Future research should address this issue through techniques like oversampling or undersampling.
3) **External Validation:** The models in this study were trained and evaluated on the same dataset. External validation on an independent dataset would provide stronger evidence of the models' effectiveness.
4) **Clinical Interpretability:** While machine learning models can provide accurate predictions, their complex nature can make it challenging to interpret the underlying reasons for a prediction. Future research should focus on creating models that are not only accurate but also clinically interpretable.

B. FUTURE RESEARCH

Future studies will focus on enhancing diagnostic accuracy for CHD by incorporating advanced machine-learning techniques. This includes feature engineering and ensemble learning strategies, which enhance predictive models by analyzing complex healthcare data. Longitudinal patient datasets will provide insights into the disease's progression over time, enabling the development of predictive models. Unsupervised learning methodologies will be explored to discover latent risk factors and patterns associated with CHD, potentially uncovering novel predictors that have eluded conventional detection. This approach will enhance the precision of CHD forecasts, facilitating early intervention and tailored patient management strategies. Multimodal data analysis, including electronic health records, biomedical imaging, and genomic profiles, will provide a comprehensive lens for assessing CHD risk, aiming to refine predictions and develop personalized healthcare interventions. By navigating these advanced ML pathways, future research will make significant strides in the early detection and management of CHD, paving the way for a new era of precision medicine.

VI. CONCLUSION

The primary objective of the proposed study was to create and assess ML algorithms designed to predict the occurrence of CHD at an early stage. This was accomplished by utilizing a comprehensive dataset encompassing a range of patient variables. The primary focus of this research was to increase the accuracy of predicting CHD, a critical factor in ensuring prompt intervention and better patient outcomes. Multiple ML models were trained and tested, including Two-Class Logistic Regression, Two-Class Neural Network, Two-Class Decision Jungle, and Two-Class Boosted Decision. The outcomes exhibited great promise, as the Two-Class Boosted Decision model demonstrated superior performance compared to other models, obtaining a notable AUC value of 0.991. The AUC values are a crucial metric for assessing the model's capacity to distinguish between positive and negative cases of CHD. Furthermore, the models exhibited high accuracy, precision, recall, and F1 scores, suggesting their potential for accurately estimating the risk of CHD. Utilizing ML and DL algorithms is paramount in automating disease prediction, hence facilitating early detection and intervention. However, it is crucial to acknowledge the need for diverse and high-quality datasets, the significance of addressing data privacy issues, and the additional role that healthcare professionals play. Notwithstanding the constraints of the study, the results imply that ML has the potential to exert a substantial influence on the early detection of diseases, potentially mitigating death rates and enhancing the overall quality of healthcare.

CONFLICT OF INTEREST

The authors declare that they have no Conflict of interest.

ACKNOWLEDGMENT

This research is supported by Princess Nourah Bint Abdulrahman University Researchers Supporting Project number (PNURSP2024R513), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia. Authors are also thankful to Prince Sultan University, Riyadh, Saudi Arabia for supporting this research.

REFERENCES

[1] T. G. Richardson, E. Sanderson, T. M. Palmer, M. Ala-Korpela, B. A. Ference, G. D. Smith, and M. V. Holmes, "Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable Mendelian randomisation analysis," *PLOS Med.*, vol. 17, no. 3, Mar. 2020, Art. no. e1003062.
[2] C.-Y. Chou, D.-Y. Hsu, and C.-H. Chou, "Predicting the onset of diabetes with machine learning methods," *J. Personalized Med.*, vol. 13, no. 3, p. 406, Feb. 2023.
[3] E. Wilkins, L. Wilson, K. Wickramasinghe, P. Bhatnagar, J. Leal, R. Luengo-Fernandez, R. Burns, M. Rayner, and N. Townsend, *European Cardiovascular Disease Statistics 2017*. Brussels, Belgium: European Heart Network, 2017.
[4] J. Mackay and G. A. Mensah, *The Atlas of Heart Disease and Stroke*. Geneva, Switzerland: World Health Organization, 2004.
[5] S. S. Virani et al., "Heart disease and stroke statistics—2021 update: A report from the American Heart Association," *Circulation*, vol. 143, no. 8, pp. 535–552, 2021.
[6] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. van den Driessche, T. Graepel, and D. Hassabis, "Mastering the game of go without human knowledge," *Nature*, vol. 550, no. 7676, pp. 354–359, Oct. 2017.
[7] R. A. Ramadan, A. Y. Khedr, K. Yadav, E. J. Alreshidi, M. H. Sharif, A. T. Azar, and H. Kamberaj, "Convolution neural network based automatic localization of landmarks on lateral X-ray images," *Multimedia Tools Appl.*, vol. 81, no. 26, pp. 37403–37415, Nov. 2022.
[8] H. Yan, Q. Ye, T. Zhang, D.-J. Yu, X. Yuan, Y. Xu, and L. Fu, "Least squares twin bounded support vector machines based on L1-norm distance metric for classification," *Pattern Recognit.*, vol. 74, pp. 434–447, Feb. 2018.

[9] M. M. Hassan, S. Zaman, M. M. Rahman, A. K. Bairagi, W. El-Shafai, R. S. Rathore, and D. Gupta, "Efficient prediction of coronary artery disease using machine learning algorithms with feature selection techniques," *Comput. Electr. Eng.*, vol. 115, Apr. 2024, Art. no. 109130.

[10] M. Jaworski, P. Duda, and L. Rutkowski, "New splitting criteria for decision trees in stationary data streams," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 6, pp. 2516–2529, Jun. 2018.

[11] C.-Z. Gao, Q. Cheng, P. He, W. Susilo, and J. Li, "Privacy-preserving Naive Bayes classifiers secure against the substitution-then-comparison attack," *Inf. Sci.*, vol. 444, pp. 72–88, May 2018.

[12] S. Zhang, D. Cheng, Z. Deng, M. Zong, and X. Deng, "A novel kNN algorithm with data-driven k parameter computation," *Pattern Recognit. Lett.*, vol. 109, pp. 44–54, Jul. 2018.

[13] S. M. Javidan, A. Banakar, K. A. Vakilian, and Y. Ampatzidis, "Diagnosis of grape leaf diseases using automatic K-means clustering and machine learning," *Smart Agricult. Technol.*, vol. 3, Feb. 2023, Art. no. 100081.

[14] M. Abdar, N. Y. Yen, and J. C.-S. Hung, "Improving the diagnosis of liver disease using multilayer perceptron neural network and boosted decision trees," *J. Med. Biol. Eng.*, vol. 38, no. 6, pp. 953–965, Dec. 2018.

[15] P. Pławiak, "Novel genetic ensembles of classifiers applied to myocardium dysfunction recognition based on ECG signals," *Swarm Evol. Comput.*, vol. 39, pp. 192–208, Apr. 2018.

[16] E. Abdulhay, N. Arunkumar, K. Narasimhan, E. Vellaiappan, and V. Venkatraman, "Gait and tremor investigation using machine learning techniques for the diagnosis of Parkinson disease," *Future Gener. Comput. Syst.*, vol. 83, pp. 366–373, Jun. 2018.

[17] M. S. A. Reshan, S. Amin, M. A. Zeb, A. Sulaiman, H. Alshahrani, A. T. Azar, and A. Shaikh, "Enhancing breast cancer detection and classification using advanced multi-model features and ensemble machine learning techniques," *Life*, vol. 13, no. 10, p. 2093, Oct. 2023.

[18] K. N. Devi, S. Suruthi, and S. Shanthi, "Coronary artery disease prediction using machine learning techniques," in *Proc. 8th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS)*, vol. 1, Mar. 2022, pp. 1029–1034.

[19] World Health Org. (2009). *Cardiovascular Diseases (CVDs)*. [Online]. Available: http://www.who.int/mediacentre/factsheets/fs317/en/index.html

[20] R. O. Bonow, D. L. Mann, D. P. Zipes, and P. Libby, *Braunwald's Heart Disease E-Book: A Textbook of Cardiovascular Medicine*. Amsterdam, The Netherlands: Elsevier, 2020.

[21] P. Kierkegaard, "Electronic health record: Wiring Europe's healthcare," *Comput. Law Secur. Rev.*, vol. 27, no. 5, pp. 503–515, Sep. 2011.

[22] A. H. Gonsalves, F. Thabtah, R. M. A. Mohammad, and G. Singh, "Prediction of coronary heart disease using machine learning: An experimental analysis," in *Proc. 3rd Int. Conf. Deep Learn. Technol.*, Jul. 2019, pp. 51–56.

[23] S. R. N. Kalhori and X.-J. Zeng, "Evaluation and comparison of different machine learning methods to predict outcome of tuberculosis treatment course," *J. Intell. Learn. Syst. Appl.*, vol. 5, no. 3, pp. 184–193, 2013.

[24] F. Thabtah, "An accessible and efficient autism screening method for behavioural data and predictive analyses," *Health Informat. J.*, vol. 25, no. 4, pp. 1739–1755, Dec. 2019, doi: 10.1177/1460458218796636.

[25] P. Suryachandra and P. V. S. Reddy, "Comparison of machine learning algorithms for breast cancer," in *Proc. Int. Conf. Inventive Comput. Technol. (ICICT)*, vol. 3, Aug. 2016, pp. 1–6.

[26] K. M. M. Beyer and S. Namin, "Chronic environmental diseases: Burdens, causes, and response," in *Biological and Environmental Hazards, Risks, and Disasters* (Hazards and Disasters Series), 2nd ed., Boston, MA, USA: Elsevier, 2023, pp. 223–249.

[27] Puneet, Deepika, P. Singh, R. Bansal, and S. Sharma, "Coronary heart disease prediction using voting classifier ensemble learning," in *Proc. 3rd Int. Conf. Adv. Comput., Commun. Control Netw. (ICAC3N)*, Dec. 2021, pp. 181–185.

[28] P. C. Bizimana, Z. Zhang, A. H. Hounye, M. Asim, M. Hammad, and A. A. A. El-Latif, "Automated heart disease prediction using improved explainable learning-based technique," *Neural Comput. Appl.*, vol. 1, no. 1, pp. 1–30, 2024.

[29] C. J. Haug and J. M. Drazen, "Artificial intelligence and machine learning in clinical medicine, 2023," *New England J. Med.*, vol. 388, no. 13, pp. 1201–1208, 2023.

[30] H. Iftikhar, M. Khan, Z. Khan, F. Khan, H. M. Alshanbari, and Z. Ahmad, "A comparative analysis of machine learning models: A case study in predicting chronic kidney disease," *Sustainability*, vol. 15, no. 3, p. 2754, Feb. 2023.

[31] S. K. Mishra, A. Patel, A. R. Sahoo, N. K. Jena, M. Padhan, K. P. Kumar, R. K. Padhi, L. Dora, S. Agrawal, and R. Panda, "Heart failure detection using deep neural network," in *Proc. Int. Conf. Commun., Circuits, Syst. (IC3S)*, May 2023, pp. 1–6.

[32] 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.

[33] S. Khan and S. T. Rasool, "Current use of cardiac biomarkers in various heart conditions," *Endocrine, Metabolic Immune Disorders Drug Targets*, vol. 21, no. 6, pp. 980–993, Jun. 2021.

[34] A. T. Azar, "A bio-inspired method for segmenting the optic disc and macula in retinal images," *Int. J. Comput. Appl. Technol.*, vol. 72, no. 4, pp. 262–277, 2023.

[35] *The Enterprisers Project*. Accessed: Aug. 8, 2023. [Online]. Available: https://enterprisersproject.com

[36] World Health Org. (2020). *Health Topics: Cardiovascular Diseases*. Accessed: Dec. 11, 2020. [Online]. Available: http://www.who.int/cardiovascular_diseases/en/

[37] S. S. Virani et al., "Heart disease and stroke statistics—2020 update: A report from the American Heart Association," *Circulation*, vol. 141, no. 9, pp. e139–e596, 2020.

[38] R. Kavya, A. Kala, J. Christopher, S. Panda, and B. Lazarus, "DAAR: Drift adaption and alternatives ranking approach for interpretable clinical decision support systems," *Biomed. Signal Process. Control*, vol. 84, Jul. 2023, Art. no. 104793.

[39] A. Dubey, N. Baishwar, and H. Varshney, "Machine learning for heart disease prediction: Recent trends and major challenges," *AIP Conf. Proc.*, vol. 2782, Jun. 2023, Art. no. 020139.

[40] A. A. Almazroi, E. A. Aldhahri, S. Bashir, and S. Ashfaq, "A clinical decision support system for heart disease prediction using deep learning," *IEEE Access*, vol. 11, pp. 61646–61659, 2023.

[41] J. S. Rudman, A. Farcas, G. A. Salazar, J. Hoff, R. P. Crowe, K. Whitten-Chung, G. Torres, C. Pereira, E. Hill, S. Jafri, D. I. Page, M. von Isenburg, A. Haamid, and A. P. Joiner, "Diversity, equity, and inclusion in the United States Emergency medical services workforce: A scoping review," *Prehospital Emergency Care*, vol. 27, no. 4, pp. 385–397, May 2023.

[42] N. Noorbakhsh-Sabet, R. Zand, Y. Zhang, and V. Abedi, "Artificial intelligence transforms the future of health care," *Amer. J. Med.*, vol. 132, no. 7, pp. 795–801, Jul. 2019.

[43] E. Parimbelli, T. M. Buonocore, G. Nicora, W. Michalowski, S. Wilk, and R. Bellazzi, "Why did AI get this one wrong? Tree-based explanations of machine learning model predictions," *Artif. Intell. Med.*, vol. 135, Jan. 2023, Art. no. 102471.

[44] I. A. Scott, A. Abdel-Hafez, M. Barras, and S. Canaris, "What is needed to mainstream artificial intelligence in health care?" *Austral. Health Rev.*, vol. 45, no. 5, pp. 591–596, Jun. 2021.

[45] H. Naz, R. Nijhawan, N. J. Ahuja, T. Saba, F. S. Alamri, and A. Rehman, "Micro-segmentation of retinal image lesions in diabetic retinopathy using energy-based fuzzy C-means clustering (EFM-FCM)," *Microsc. Res. Technique*, vol. 87, no. 1, pp. 78–94, Jan. 2024.

[46] P. Melillo, N. De Luca, M. Bracale, and L. Pecchia, "Classification tree for risk assessment in patients suffering from congestive heart failure via long-term heart rate variability," *IEEE J. Biomed. Health Informat.*, vol. 17, no. 3, pp. 727–733, May 2013.

[47] G. Guidi, M. C. Pettenati, P. Melillo, and E. Iadanza, "A machine learning system to improve heart failure patient assistance," *IEEE J. Biomed. Health Informat.*, vol. 18, no. 6, pp. 1750–1756, Nov. 2014.

[48] G. Parthiban and S. K. Srivatsa, "Applying machine learning methods in diagnosing heart disease for diabetic patients," *Int. J. Appl. Inf. Syst.*, vol. 3, no. 7, pp. 25–30, Aug. 2012.

[49] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019.

[50] D. Waigi, S. Choudhary, P. Fulzele, and D. Mishra, "Predicting the risk of heart disease using advanced machine learning approach," *Eur. J. Mol. Clin. Med.*, vol. 7, no. 7, pp. 1638–1645, 2020.

[51] C. Zhou, A. Hou, P. Dai, A. Li, Z. Zhang, Y. Mu, and L. Liu, "Risk factor refinement and ensemble deep learning methods on prediction of heart failure using real healthcare records," *Inf. Sci.*, vol. 637, Aug. 2023, Art. no. 118932.

[52] C. C. Zwack, M. Haghani, M. Hollings, L. Zhang, S. Gauci, R. Gallagher, and J. Redfern, "The evolution of digital health technologies in cardiovascular disease research," *NPJ Digit. Med.*, vol. 6, no. 1, p. 1, Jan. 2023.

[53] A. Narin, Y. Isler, and M. Özer, "Early prediction of paroxysmal atrial fibrillation using frequency domain measures of heart rate variability," in *Proc. Med. Technol. Nat. Congr. (TIPTEKNO)*, Oct. 2016, pp. 1–4.
[54] D. Shah, S. Patel, and S. K. Bharti, "Heart disease prediction using machine learning techniques," *Social Netw. Comput. Sci.*, vol. 1, pp. 1–6, Oct. 2020.
[55] K. M. Gokhale, J. S. Chandan, C. Sainsbury, P. Tino, A. Tahrani, K. Toulis, and K. Nirantharakumar, "Using repeated measurements to predict cardiovascular risk in patients with type 2 diabetes mellitus," *Amer. J. Cardiol.*, vol. 210, pp. 133–142, Jan. 2024.
[56] P. Adekkanattu, L. V. Rasmussen, J. A. Pacheco, J. Kabariti, D. J. Stone, Y. Yu, G. Jiang, Y. Luo, P. S. Brandt, Z. Xu, V. Vekaria, J. Xu, F. Wang, N. C. Benda, Y. Peng, P. Goyal, F. S. Ahmad, and J. Pathak, "Prediction of left ventricular ejection fraction changes in heart failure patients using machine learning and electronic health records: A multi-site study," *Sci. Rep.*, vol. 13, no. 1, p. 294, Jan. 2023.
[57] L. Rasmy, Y. Wu, N. Wang, X. Geng, W. J. Zheng, F. Wang, H. Wu, H. Xu, and D. Zhi, "A study of generalizability of recurrent neural network-based predictive models for heart failure onset risk using a large and heterogeneous EHR data set," *J. Biomed. Informat.*, vol. 84, pp. 11–16, Aug. 2018.
[58] M. S. Parveen and S. Hiremath, "Cardiovascular disease prediction in retinal fundus images using ERNN technique," in *Proc. Frontiers ICT Healthcare (EAIT)*. Singapore: Springer, 2022, pp. 579–588.
[59] F. S. Alotaibi, "Implementation of machine learning model to predict heart failure disease," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 6, pp. 261–268, 2019.
[60] A. Chunduru, A. R. Kishore, B. K. Sasapu, and K. Seepana, "Multi chronic disease prediction system using CNN and random forest," *Social Netw. Comput. Sci.*, vol. 5, no. 1, p. 157, Jan. 2024.
[61] I. Mahmud, M. M. Kabir, M. F. Mridha, S. Alfarhood, M. Safran, and D. Che, "Cardiac failure forecasting based on clinical data using a lightweight machine learning metamodel," *Diagnostics*, vol. 13, no. 15, p. 2540, Jul. 2023.
[62] Ş. Ay, E. Ekinci, and Z. Garip, "A comparative analysis of meta-heuristic optimization algorithms for feature selection on ML-based classification of heart-related diseases," *J. Supercomput.*, vol. 79, no. 11, pp. 11797–11826, Jul. 2023.
[63] N. Hasan and Y. Bao, "Comparing different feature selection algorithms for cardiovascular disease prediction," *Health Technol.*, vol. 11, no. 1, pp. 49–62, Jan. 2021.
[64] S. Balasubramaniam, C. V. Joe, C. Manthiramoorthy, and K. S. Kumar, "ReliefF based feature selection and gradient squirrel search algorithm enabled deep maxout network for detection of heart disease," *Biomed. Signal Process. Control*, vol. 87, Jan. 2024, Art. no. 105446.
[65] E. Dritsas and M. Trigka, "Efficient data-driven machine learning models for cardiovascular diseases risk prediction," *Sensors*, vol. 23, no. 3, p. 1161, Jan. 2023.
[66] G. Vishnupriya, T. Pradeep, M. Mehanethra, and V. J. A. Mitra, "Prediction of pre-cardiac disease using ML & DL techniques in T-kinter," *J. Surv. Fisheries Sci.*, vol. 10, no. 4S, pp. 1874–1883, 2023.
[67] V. Chang, V. R. Bhavani, A. Q. Xu, and M. Hossain, "An artificial intelligence model for heart disease detection using machine learning algorithms," *Healthcare Analytics*, vol. 2, Nov. 2022, Art. no. 100016.
[68] Z. Noroozi, A. Orooji, and L. Erfannia, "Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction," *Sci. Rep.*, vol. 13, no. 1, p. 22588, Dec. 2023.
[69] H. Yang, Z. Chen, H. Yang, and M. Tian, "Predicting coronary heart disease using an improved LightGBM model: Performance analysis and comparison," *IEEE Access*, vol. 11, pp. 23366–23380, 2023.
[70] A. Dutta, T. Batabyal, M. Basu, and S. T. Acton, "An efficient convolutional neural network for coronary heart disease prediction," *Expert Syst. Appl.*, vol. 159, Nov. 2020, Art. no. 113408.
[71] S. Nandy, M. Adhikari, V. Balasubramanian, V. G. Menon, X. Li, and M. Zakarya, "An intelligent heart disease prediction system based on swarm-artificial neural network," *Neural Comput. Appl.*, vol. 35, no. 20, pp. 14723–14737, Jul. 2023.
[72] A. M. Al-Ssulami, R. S. Alsorori, A. M. Azmi, and H. Aboalsamh, "Improving coronary heart disease prediction through machine learning and an innovative data augmentation technique," *Cognit. Comput.*, vol. 15, no. 5, pp. 1687–1702, Sep. 2023.
[73] K. Drożdż, K. Nabrdalik, H. Kwiendacz, M. Hendel, A. Olejarz, A. Tomasik, W. Bartman, J. Nalepa, J. Gumprecht, and G. Y. H. Lip, "Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: A machine learning approach," *Cardiovascular Diabetol.*, vol. 21, no. 1, p. 240, Nov. 2022.
[74] I. M. Pires, G. Marques, N. M. Garcia, and V. Ponciano, "Machine learning for the evaluation of the presence of heart disease," *Proc. Comput. Sci.*, vol. 177, pp. 432–437, Jan. 2020.
[75] A. Abdellatif, H. Abdellatef, J. Kanesan, C.-O. Chow, J. H. Chuah, and H. M. Gheni, "An effective heart disease detection and severity level classification model using machine learning and hyperparameter optimization methods," *IEEE Access*, vol. 10, pp. 79974–79985, 2022.
[76] G. Joo, Y. Song, H. Im, and J. Park, "Clinical implication of machine learning in predicting the occurrence of cardiovascular disease using big data (nationwide cohort data in Korea)," *IEEE Access*, vol. 8, pp. 157643–157653, 2020.
[77] A. Mujumdar and V. Vaidehi, "Diabetes prediction using machine learning algorithms," *Proc. Comput. Sci.*, vol. 165, pp. 292–299, Jan. 2019.
[78] R. Birjais, A. K. Mourya, R. Chauhan, and H. Kaur, "Prediction and diagnosis of future diabetes risk: A machine learning approach," *Social Netw. Appl. Sci.*, vol. 1, no. 9, pp. 1–8, Sep. 2019.
[79] R. Katarya and S. Polipireddy, "Identifying risks in cardiovascular disease using supervised machine learning algorithms," *SSRN J.*, vol. 1, pp. 767–775, Jan. 2020.
[80] T. R. Gadekallu, N. Khare, S. Bhattacharya, S. Singh, P. K. R. Maddikunta, I.-H. Ra, and M. Alazab, "Early detection of diabetic retinopathy using PCA-firefly based deep learning model," *Electronics*, vol. 9, no. 2, p. 274, Feb. 2020.
[81] M. W. Nadeem, H. G. Goh, V. Ponnusamy, I. Andonovic, M. A. Khan, and M. Hussain, "A fusion-based machine learning approach for the prediction of the onset of diabetes," *Healthcare*, vol. 9, no. 10, p. 1393, Oct. 2021.
[82] K. S. Ryu, S. W. Lee, E. Batbaatar, J. W. Lee, K. S. Choi, and H. S. Cha, "A deep learning model for estimation of patients with undiagnosed diabetes," *Appl. Sci.*, vol. 10, no. 1, p. 421, Jan. 2020.
[83] G. Battineni, G. G. Sagaro, C. Nalini, F. Amenta, and S. K. Tayebati, "Comparative machine-learning approach: A follow-up study on type 2 diabetes predictions by cross-validation methods," *Machines*, vol. 7, no. 4, p. 74, Dec. 2019.
[84] N. G. Forouhi and N. J. Wareham, "Epidemiology of diabetes," *Medicine*, vol. 38, no. 11, pp. 602–606, 2010.
[85] M. Grama, M. Musat, L. Muñoz-González, J. Passerat-Palmbach, D. Rueckert, and A. Alansary, "Robust aggregation for adaptive privacy preserving federated learning in healthcare," 2020, *arXiv:2009.08294*.
[86] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai, and R. S. Suraj, "Heart disease prediction using hybrid machine learning model," in *Proc. 6th Int. Conf. Inventive Comput. Technol. (ICICT)*, Jan. 2021, pp. 1329–1333.
[87] J. Li, Y. Meng, L. Ma, S. Du, H. Zhu, Q. Pei, and X. Shen, "A federated learning based privacy-preserving smart healthcare system," *IEEE Trans. Ind. Informat.*, vol. 18, no. 3, pp. 2021–2031, Mar. 2022.
[88] B. Mali, S. Saha, D. Brahma, R. Pinninti, and P. K. Singh, "Towards building a global robust model for heart disease detection," *Social Netw. Comput. Sci.*, vol. 4, no. 5, pp. 1–12, Aug. 2023.
[89] G. Saranya and A. Pravin, "A novel feature selection approach with integrated feature sensitivity and feature correlation for improved prediction of heart disease," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 9, pp. 12005–12019, Sep. 2023.

TARIQ MAHMOOD received the master's degree in computer science from the University of Lahore, Pakistan, and the Ph.D. degree in software engineering from Beijing University of Technology, China. He is currently an Assistant Professor with the Faculty of Information Sciences, University of Education, Vehari Campus, Vehari, Pakistan. He is a renowned expert in image processing, healthcare informatics and social media analysis, ad-hoc networks, and WSN. He has contributed research articles in well-reputed international journals and conferences. His research interests include image processing, social media analysis, medical image diagnosis, machine learning, and data mining. He aims to contribute to interdisciplinary research of computer science and human-related disciplines. He is an Editorial Member and a Reviewer of various journals, including *PloS One*, *The Journal of Supercomputer*, *Journal of digital Imaging*, and *International Journal of Sensors, Wireless Communications and Control*.

AMJAD REHMAN (Senior Member, IEEE) received the Ph.D. and Postdoctoral degrees (Hons.) from the Faculty of Computing, Universiti Teknologi Malaysia, with a specialization in forensic documents analysis and security, in 2010 and 2011, respectively. He is currently a Senior Researcher with the Artificial Intelligence and Data Analytics Laboratory, College of Computer and Information Sciences (CCIS), Prince Sultan University, Riyadh, Saudi Arabia. He is the author of more than 200 ISI journal articles and conferences. He is also a PI in several funded projects and also completed projects funded by MOHE Malaysia and Saudi Arabia. His research interests include data mining, health informatics, and pattern recognition. He received the Rector Award for the 2010 Best Student from Universiti Teknologi Malaysia.

TANZILA SABA (Senior Member, IEEE) received the Ph.D. degree in document information security and management from the Faculty of Computing, Universiti Teknologi Malaysia (UTM), Malaysia, in 2012. She is currently an Associate Chair with the Information Systems Department, College of Computer and Information Sciences, Prince Sultan University, Riyadh, Saudi Arabia, where she is also the Leader of the Artificial Intelligence and Data Analytics Research Laboratory. Her research interests include medical imaging, pattern recognition, data mining, MRI analysis, and soft computing. She is an active professional member of ACM, AIS, and IAENG organizations. She is the PSU Women in Data Science (WiDS) Ambassador at Stanford University and the Global Women Tech Conference. She was a recipient of the Best Student Award from the Faculty of Computing, UTM, in 2012.

TAHANI JASER ALAHMADI received the B.S. degree in computer science and the M.S. degree in information technology (data management), and the Ph.D. degree from the Faculty of Information Technology, Griffith University, Australia, in 2019. She is currently an Assistant Professor with the IS Department, Faculty of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Saudi Arabia. Her research interests include innovative research methods in data analysis and mining, machine learning, pattern recognition, image processing, accessibility, usability, and sentiment analysis. She is a member of the Golden Key Society and Media Access Australia. She received multiple awards, such as the Google Doctoral Consortium Award and the Institute for Integrated and Intelligent Systems (IIIS) Award for Quality and Impact Research.

MUHAMMAD TUFAIL received the B.Sc. and M.Sc. degrees in computer science from the University of Peshawar, Khyber Pakhtunkhwa, Pakistan, in 1999 and 2007, respectively, and the Ph.D. degree in computer science from Liverpool University, in U.K., in 2017. In 2006, he was a Lecturer with the Higher Education Department, KP, until 2017, during the Ph.D. study, and was promoted to an Assistant Professor. His research interests include cybersecurity, artificial intelligence, simulation, and video pattern mining.

SAEED ALI OMER BAHAJ received the Ph.D. degree from Pune University, India, in 2006. He is currently an Associate Professor with the Department of Management Information Systems, College of Business Administration (COBA), Prince Sattam Bin Abdulaziz University, and the Department of Computer Engineering, Hadramout University, Yemen. His main research interests include artificial intelligence, information management, forecasting, information engineering, big data, and information security.

ZOHAIB AHMAD received the Ph.D. degree from the School of Electronics and Information Engineering, Beijing University of Technology, Beijing. He is currently with the Department of Criminology and Forensic Sciences, Lahore Garrison University, Lahore, Pakistan. His research interests include artificial intelligence, machine learning, algorithms, intelligent computation, and intelligent optimization.