Heart Disease Prediction Using Novel Quine McCluskey Binary Classifier (QMBC)

KAPILA, Ramdas, RAGUNATHAN, Thirumalaisamy, SALETI, Sumalatha, TANGIRALA, Jaya Lakshmi <http://orcid.org/0000-0003-0183-4093> and AHMAD, Mohd Wazih

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 10 June 2023, accepted 19 June 2023, date of publication 26 June 2023, date of current version 30 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3289584

RESEARCH ARTICLE

Heart Disease Prediction Using Novel Quine McCluskey Binary Classifier (QMBC)

RAMDAS KAPILA[1], THIRUMALAISAMY RAGUNATHAN[2], (Member, IEEE), SUMALATHA SALETI[1], T. JAYA LAKSHMI[1], (Member, IEEE), AND MOHD WAZIH AHMAD[3]

[1]Data Science Research Laboratory, Department of Computer Science and Engineering, SRM University AP, Amaravati 522502, India
[2]Department of Computer Science and Engineering, Sri Ramachandra Institute of Higher Education and Research, Chennai 600116, India
[3]Adama Science and Technology University, Adama, Ethiopia

Corresponding author: Mohd Wazih Ahmad (wazih.ahmad@astu.edu.et)

ABSTRACT Cardiovascular disease is the primary reason for mortality worldwide, responsible for around a third of all deaths. To assist medical professionals in quickly identifying and diagnosing patients, numerous machine learning and data mining techniques are utilized to predict the disease. Many researchers have developed various models to boost the efficiency of these predictions. Feature selection and extraction techniques are utilized to remove unnecessary features from the dataset, thereby reducing computation time and increasing the efficiency of the models. In this study, we introduce a new ensemble Quine McCluskey Binary Classifier (QMBC) technique for identifying patients diagnosed with some form of heart disease and those who are not diagnosed. The QMBC model utilizes an ensemble of seven models, including logistic regression, decision tree, random forest, K-nearest neighbour, naive bayes, support vector machine, and multilayer perceptron, and performs exceptionally well on binary class datasets. We employ feature selection and feature extraction techniques to accelerate the prediction process. We utilize Chi-Square and ANOVA approaches to identify the top 10 features and create a subset of the dataset. We then apply Principal Component Analysis to the subset to identify 9 prime components. We utilize an ensemble of all seven models and the Quine McCluskey technique to obtain the Minimum Boolean expression for the target feature. The results of the seven models ($x_0, x_1, x_2, \ldots, x_6$) are considered independent features, while the target attribute is dependent. We combine the projected outcomes of the seven ML models and the target feature to form a foaming dataset. We apply the ensemble model to the dataset, utilizing the Quine McCluskey minimum Boolean equation built with an 80:20 train-to-test ratio. Our proposed QMBC model surpasses all current state-of-the-art models and previously suggested methods put forward by various researchers.

INDEX TERMS Machine learning, chi-square, ANOVA, principal component analysis, Quine McCluskey technique, ensemble approach.

I. INTRODUCTION

The term ''Heart Disease'' (HD) is used to refer to a variety of pathological disorders that have an impact on the heart and blood vessels. It encompasses a variety of heart-related conditions, including but not limited to vascular diseases and disturbances in heart rhythm [1]. As per the World Health Organization (WHO), it is the deadliest and most devastating disease, taking over 18 million in lives a year [2]. To diagnose it, healthcare professionals rely on a patient's medical history and various tests, such as blood pressure, blood sugar, and cholesterol tests. Additionally, modern medical procedures like electrocardiograms, exercise stress tests, X-rays, echocardiography, coronary angiography, radionuclide tests, MRI scans, and CT scans can aid in the identification of cardiac conditions [3]. Heart failure is the result of chronic issues that damage or weaken the heart muscles, leading to reduced ejection fraction. It is a condition that can affect both adults and children and cause severe damage to other vital organs in the body. The primary risk factors associated with heart failure are age, ethnicity, family history, hereditary factors, lifestyle choices, and pre-existing Cardiovascular

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

disease (CVD) or genetics. While it affects both men and women equally, women are more likely to develop heart failure later in life [4]. To diagnose diseases at an early stage, ML is becoming an increasingly important tool. It aims to identify patterns hidden in observations and draw conclusions that are consistent with new information. Researchers have investigated the grouping of various techniques to create hybrid models that can outperform standalone models. Typically, these models have two phases. A subset of characteristics is chosen in phase-1 using Feature Selection (FS) and Feature Extraction (FE) techniques. The classifiers used in the phase-2 are then applied to this subset as input [5], [6], [7], [8], [9].

Heart disease datasets often contain various attributes, including both relevant and irrelevant as well as duplicate attributes. Relevant attributes are those that have an impact on how the target class is defined, whereas irrelevant do not contribute to the output class's description. Redundant attributes, on the other hand, introduce noise rather than adding any new information to the target class's definition [10]. Eliminating some traits that not only have an impact on the classification outcomes but also decrease system performance is crucial for improving the classification models. So, the HD diagnosis system requires the use of dimensionality reduction or FS techniques, as the datasets contain irrelevant and redundant features that contribute to noise rather than providing any information about the target class. The chance of overfitting is decreased, the model's capacity to generalize is increased, predictability is improved, and less computation is needed, which results in fewer features [10], [11].

To enhance the performance of a model, ensemble techniques have been proven effective. A considerable increase in performance improvement has also arisen from the inclusion of FS [12], [13]. To improve the ML Models, researchers are continuously exploring new approaches. Ensemble learning is a strategy that has been shown to enhance ML issues [14]. Ensemble learning involves combining predictions from multiple classifiers using a process such as a majority voting. According to research, ensemble classifiers frequently outperform classical classifiers [15].

An ensemble is a type of ML model that produces a final prediction by integrating the predictions from multiple individual models. These models can be of similar or diverse types, and there are numerous techniques available to combine them. Bagging and boosting are the two primary kinds of ensemble models [5], [16]. This research proposes an ML model to predict patients diagnosed with some form of HD and not diagnosed, using LR, DT, RF, KNN, NB, SVC, and MLP models. To enhance the model's performance, we use feature selection and feature extraction methods, including Chi-Square, ANOVA, and PCA techniques, to select and extract critical attributes from the dataset. An integrated strategy is developed by combining FS and FE techniques, reducing the dimensionality and accelerating the model's computation while retaining its best performance. This approach improves the model's efficiency while maintaining its effectiveness, allowing for more accurate predictions. We introduce a new ensemble technique, named QMBC, which aggregates the outputs of multiple individual models to generate a single prediction. Moreover, a variety of assessment measures have been used to rate classifier performance. The effectiveness of the proposed approach has been evaluated using three different datasets, namely the Cleveland HD dataset, the comprehensive HD dataset, and the CVD dataset. The effectiveness of the proposed method has also been compared to techniques that have been stated in the literature; among them are MIFH [17], RSA and RF model [18], weighted-average voting (WAVEn) [19], XGBoost with Bayesian optimization [16], stacking classifier [20]. Below is a list of the research study's achievements.

- The first step of this work involves preprocessing the dataset, followed by applying FS and FE techniques to optimize computation time.
- Specifically, we utilized the Chi-Square and ANOVA techniques to eliminate irrelevant and redundant attributes and create subsets of features.
- We then applied the PCA FE technique to extract prime components from these subsets, further improving the efficiency of our model.
- We introduced a novel ensemble technique called Quine McCluskey Binary Classifier (QMBC) that combines predictions from multiple models to predict patients diagnosed and not diagnosed with some form of HD. In particular, we believe the Quine McCluskey method for predicting diagnosed and not diagnosed patients with HD is a new addition to this field, as no prior research has been conducted in this specific area.
- The effectiveness of the proposed QMBC is evaluated using 3 benchmark datasets, including the Cleveland HD dataset [21], the HD dataset (comprehensive) [22], and the CVD dataset [23].
- A detailed comparison of QMBC with current research shows that, in terms of accuracy, precision, recall, specificity, and f1-score, the proposed approach is more efficient than the state-of-the-art models.

Below are the remaining sections included in this research article. Section II provides a review of related literature on predicting heart disease. Section III outlines the proposed methodology, algorithm, and architecture. Section IV presents the results and corresponding discussions. Finally, Section V provides the conclusion and recommendations for future scope.

II. RELATED WORK

The focus of this section is on discussing various ML models utilized by scholars earlier to successfully anticipate HD. ML algorithms that perform classification are widely used in disease prediction and other disciplines. To improve model performance and reduce time complexity, FS/FE methodologies are commonly employed.

A diagnostic system that employed rough sets and an Interval Type-2 Fuzzy Logic System (IT2FLS) to anticipate HD has been introduced in [11]. The goal is to manage high-dimensional datasets, reduce computational time and enhance model performance. The accuracy, sensitivity, and specificity for models with the dataset, namely BPSORS-AR and CFARS-AR, are 86%, 87.1%, and 90%. By utilizing Binary Particle Swarm Optimization and Rough Sets-based Attribute Reduction (BPSORS-AR), the accuracy, sensitivity, and specificity improved to 87.0%, 93.3%, and 79.2%, respectively. Furthermore, by applying the Chaos Firefly Algorithm and Rough Sets-based Attribute Reduction (CFARS-AR), the accuracy, sensitivity, and specificity increased to 88.3%, 84.9%, and 93.3%, respectively. A model for predicting HD using 9 key features has been proposed by Amin et al. [24] utilizing a voting classifier with an accuracy of 87.41%. By combining the Random Search Algorithm (RSA) and RF model to forecast the HD based on chosen subset characteristics, Javeed et al. [18] suggested a model to solve the overfitting issues. With an accuracy of 93.33%, 95.12% sensitivity, and 89.79% specificity, the suggested RSA-RF model with 7 features outperforms all existing mechanisms and reduces the execution time.

HD forecasting in advance can save a person's life. In this regard, Factor Analysis of Mixed Data (FAMD) + RF ML intelligent framework developed by Gupta et al. [17] surpassed all other models with an accuracy of 93.44%, a sensitivity of 89.28%, and specificity values of 96.06%, respectively. Latha et al. [13] look at the usefulness of ensemble classification, which combines many classifiers to increase the accuracy of weaker algorithms to enhance the performance of unreliable approaches and demonstrate the algorithm's worth in identifying illnesses at an early stage using a medical dataset. The research tested this technique on a heart disease dataset through various experiments. The findings indicate that ensemble methods enhance the predictive accuracy of weaker models and are successful to predict the risk of HD. Mehta et al. [25] hypothesized that human deaths might be avoided with early HD identification. Five DM methods are employed to find the disease. SVM outperformed with an accuracy of 97.91%. A model for HD prediction called HRFLM by Mohan et al. [26], integrates a hybrid RF with a linear model. The proposed model scored an accuracy of 88.7% more than the existing models. To effectively forecast, Raza and Khalid [27] suggested an ML model using LR, MLP, NB, and ensemble majority voting classifier. The ensemble technique with an accuracy of 88.88% suppresses all models.

An ensemble technique demonstrated by Mienye et al. [12] utilizes several Classification and Regression Trees (CART) models with a Weighted Aging classifier Ensemble (WAE) and significantly improves the performance of predicting illnesses. The proposed approach achieved 93% for Cleveland and 91% for the Framingham datasets. HD Prediction model (HDPM) introduced by Fitriyani et al. [28], firstly balances the dataset and removes the noise data and outliers, and then applies the XGBoost model to predict the disease. FCMIM FS technique to enhance efficiency and reduce computation time for the Support Vector Classifier (SVC) model proposed by Li et al. [8], which achieved an accuracy of 92.37% surpassing the performance of existing models. Using 5 distinct ML models, Pasha et al. [29] suggested a Novel Feature Reduction (NFR) method to effectively forecast the illness. With 92.53% accuracy, NFR + LR for the Cleveland dataset received the highest scores. The benefits of FS and FE are achieved by integrating the essential and vital features that are selected and extracted by Shah et al. [30]. An improved sparse autoencoder-based Artificial Neural Network (ANN) has been proposed by Mienye et al. [12] for predicting HD.

To identify and forecast HD at an early stage and prevent deaths, Rahim et al. [31] introduced the Machine Learning based Cardiovascular Disease Diagnosis (MaLCaDD) system. With an accuracy of 99.1%, 98.0%, and 95.5% the model scored accuracy for Framingham, Heart Disease, and Cleveland datasets. In this study, missing values are substituted with mean values, an imbalanced dataset is handled using the Synthetic Minority Over-sampling Technique (SMOTE) technique, the crucial features are chosen using the feature importance FS approach, and then LR and KNN models are integrated to improve the performance. Valarmathi et al. [32] proposed a Hyper Parameter Optimization (HPO) to increase the efficiency of the RF model to predict the HD with the highest accuracy of 97.52%. The CART model is utilized to anticipate the early state of HD with an accuracy of 88.33% and recall of 84.62% to save human lives as proposed by Miranda et al. [33]. In order to diagnose CVD effectively, Velusamy et al. [19] suggested an ensemble approach by incorporating KNN, RF, and SVM classifiers. Boruta FS in [34] approach is applied to pick the top attributes and feature importance of SVM utilized. The WAVEn approach improves classification accuracy, sensitivity, specificity, and precision for the original dataset by 98.97%, 100%, 96.3%, and 98.3% with the use of the top five features. The balanced dataset allows the WAVEn algorithm to diagnose CAD with 100% accuracy, sensitivity, specificity, and precision. A novel hybrid ensemble approach with a majority voting classifier using a genetic algorithm method is proposed by Ashri et al. [35] with an accuracy of 98.18% to predict the HD. A hybrid model that ensembles the RF & DT models developed by Kavitha et al. [36] bagged an accuracy of 88.7% to predict HD. An improved method utilizing a gradient decent optimizer is utilized to forecast HD with an accuracy of 98.54%, recall of 99.43%, and precision of 97.76% by Nawaz et al. [37]. A new MLP for Enhanced Brownian Motion based on Dragonfly Algorithm (MLP-EBMDA) proposed by Deepika et al. [38] predicts HD and suppressed all the existing models. SMOTE and edited nearest neighbour is a balancing approach with RF hyper tuning proposed by Muntasir Nishat et al. [4] scores high accuracy of 90%, f1-score of 92.3%, and recall of 97.3%.

A fused prediction model that combines six ML models using a weighted score fusion to differentiate between patients diagnosed with some form of HD and without diagnosed is proposed by Kibria et al. [39]. The suggested models are used for binary and multiclass classification, with oversampling used for balancing the multiclass dataset. The highest accuracy scores are 95% for binary classification and 75% for multiclass classification. Bayesian optimization model and an XGBoost with hyper tuning are suggested by Budholiya et al. [16] and are able to accurately forecast the HD with 91.8% effectiveness. A fusion of NB & RF models is proposed by Archana et al. [40] to diagnose with an accuracy of 92%. Using Bidirectional long short-term memory (Bi-LSTM), Nancy et al. [41] proposed a healthcare monitoring system to predict HD with an accuracy of 98.86%, precision of 98.9%, recall of 98.8%, and f1-score of 98.89%, respectively. The proposed framework that combines multiple ML techniques, and the stacked ensemble classifier achieved an accuracy of 92.34% by Tiwari et al. [42] surpassing the existing models reported in the literature.

Essential 11 features are selected by the Gradient Boosting-based Sequential FS (GBSFS) technique proposed by Chaurasia et al. [43] and a stacking methodology is to forecast the HD with an accuracy of 98.78% surpassing state-of-the-art models. The CART model is proposed by Ozcan et al. [3] to extract the rules to predict HD and achieved accuracy, sensitivity, specificity, and precision scores of 87.25%, 84.51%, 89.74%, and 88.24%, respectively. Fusion of PCA and Correlation approach with ensemble model by hyper-tuning proposed by Reddy et al. [44] to enhance the performance of the model with an accuracy of 97.91%. Hawks Optimizer (HO) with the stacked classification technique developed by Kumar et al. [45] outperformed all other models with accuracy and f1-score of 97% each, the precision of 98%, and recall of 96%.

III. PROPOSED METHODOLOGY

Heart disease is a significant cause of mortality, affecting people of all ages and genders. According to recent data, it's among the main causes of mortality. It is critical to identify the disease as swiftly as possible in order to recognize it early and maybe save lives. To accomplish this, we have developed an innovative ensemble classifier known as QMBC, which is discussed in this section. The proposed architecture and Algorithm are illustrated in Fig. 1 and Algorithm 1.

In this study, we have chosen to utilize ensemble learning techniques to address the problem at hand, despite the availability of standalone ML models and deep learning architectures. This decision is rooted in the recognition that ensemble learning offers distinct advantages over individual models, leading to improved performance and robustness in predictive tasks. By combining multiple models, ensemble learning harnesses the diversity and collective wisdom of the constituent models, resulting in enhanced accuracy and generalization capabilities. Ensemble learning excels in scenarios where individual models may suffer from limitations such as overfitting, biased predictions, or incomplete representation of the underlying patterns in the data. By aggregating the predictions of multiple models, ensemble methods are able to mitigate these shortcomings, leading to more reliable and accurate predictions. Furthermore, ensemble learning facilitates the exploration of complementary model architectures, learning algorithms, or feature representations, enabling a comprehensive exploration of the solution space. This flexibility allows us to leverage the strengths of different models and exploit their complementary nature, ultimately improving the overall performance. By adopting ensemble learning, we aim to maximize the predictive power of our models and provide more robust and reliable predictions for the problem under investigation. This choice is supported by previous studies and empirical evidence that demonstrate the efficiency of ensemble learning in a variety of domains and tasks. Overall, the decision to employ ensemble learning is driven by our pursuit of improved performance, enhanced generalization, and the desire to extract the full potential from the available data. Through this approach, we expect to achieve more accurate and reliable predictions, thereby contributing to advancements in the field of the health section.

A. EXPERIMENTAL SETUP

The experiments are conducted on a system featuring an Intel(R) Core(TM) i5-1135G7 processor from the 11^{th} generation, with a base clock of 2.40GHz, four cores, and eight logical processors. The system is equipped with 8 GB of RAM and ran on the Windows 10 operating system. The programming environment used for this study was Jupyter Notebook (Anaconda3) version 6.3.0. To implement the algorithms, several essential libraries were employed, including Pandas, Numpy, Matplotlib, Seaborn, Warnings, and Scikit-Learn. Pandas is utilized for efficient data preprocessing, while Scikit-Learn provided essential functionalities for FS, FE scaling, and classification. The models' performance is evaluated using various metrics, such as the confusion matrix, f1-score, accuracy, precision, and recall. Matplotlib and Seaborn have been used to create clear and insightful visualizations, while the Standard Scaler played a crucial role in data scaling procedures.

B. DATA COLLECTION

This study utilizes three benchmark datasets that are publicly available in open repositories. Several ongoing research studies utilize the Cleveland dataset from the University of California (UCI) as a standard to predict coronary heart disease. It is the first dataset, which is openly accessible [21]. Usually, only up to 14 features out of the 76 total features in the Cleveland dataset are utilized for the analysis. The second dataset employed in this study is the HD dataset (comprehensive) [22], which merges patient records from Cleveland (303), Hungarian (294), Switzerland (123), Long Beach VA (200), and Statlog (Heart) (270) datasets, resulting in a total of 1190 patient records. This dataset contains 11 independent features and one target feature. The CVD dataset [23], which

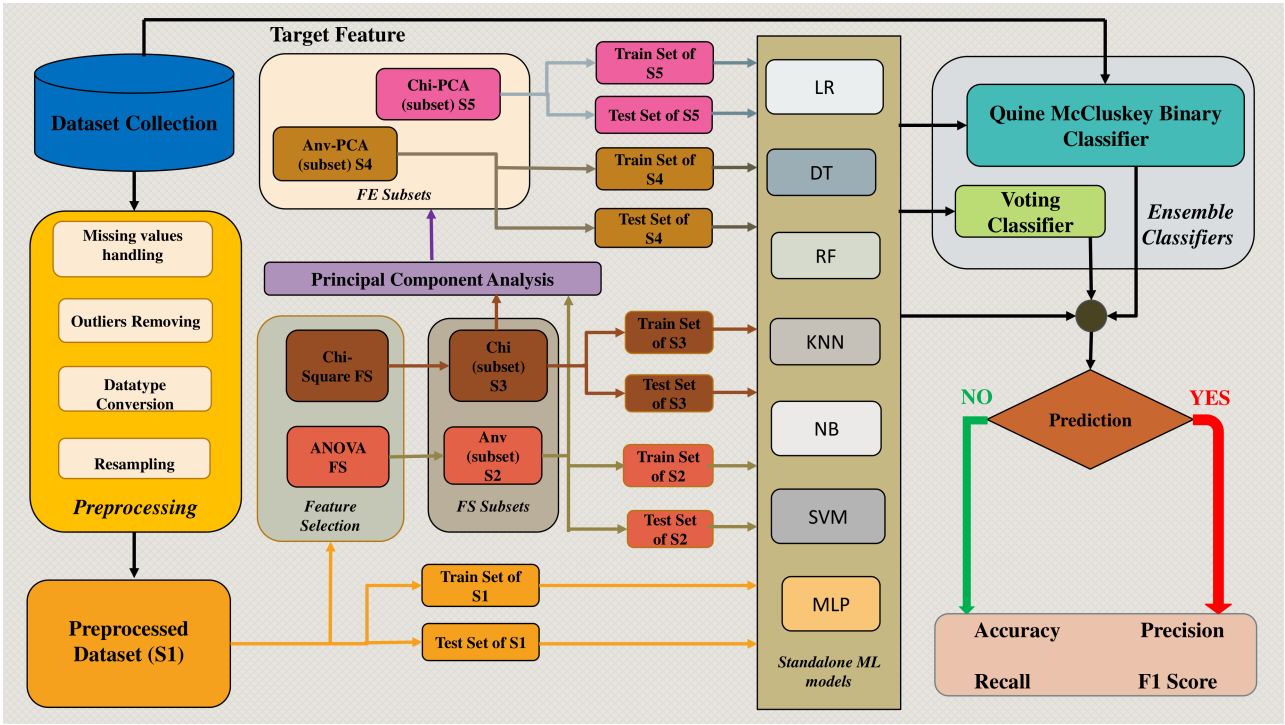


FIGURE 1. Proposed architecture.

Algorithm 1 Proposed Algorithm

1: Read the dataset from the Kaggle repository.
2: Preprocess the dataset by removing null values, outliers, datatype conversions, and resampling if any.
3: Split the dataset into train and test datasets as 80:20 ratios and build all MLT models. MLT = LR, DT, RF, KNN, SVC, and MLP.
4: Using MLT results build the Voting Classifier and QMBC Algorithm 7 with an 80% train set. VC = MLT (results). QMBC = [MLT (results), CHDD (target)].
5: Test the MLT, VC, and QMBC using a 20% test set and find the accuracy, precision, recall, and f1-score of the models.
6: Apply the Chi-2 test using Algorithm 2 to select the best features from CHDD and create a subset X_{selected}.
7: Using the X_{selected} subset repeat steps 3 to 5.
8: Apply PCA to extract the best features from X_{selected} and create a super subset X_{Chi}selected using Algorithm 4.
9: Use the super subset X_{Chi}selected and repeat steps 3 to 5.
10: Apply Anova test using Algorithm 3 on the preprocessed dataset and select the best features and create a subset Anv_{features}.
11: Using the subset Anv_{features} repeat steps 3 to 5.
12: Use Anv_{features} to extract the best features by applying PCA Algorithm 4 and create a super subset Anv_{PCA}selected.
13: Use Anv_{PCA}selected and repeat steps 3 to 5.
14: Compare the models and identify the best model.

is the third dataset used in this study, is available for free in the Kaggle repository. It contains 70, 000 patient records, 11 attributes, and a target variable. Out of the three datasets used in the study, the CVD dataset stands out as the largest. The Cleveland HD, HD (Comprehensive), and CVD dataset's attributes have been shown in Tables 1, 2, and 3 along with their relevance to heart disease prediction. The features in the datasets on heart disease offer important new information on the factors affecting cardiovascular health and heart disease. Age, gender, type of chest pain, blood pressure, cholesterol levels, and exercise-induced symptoms are only a few of the features that are important in understanding and predicting the development of heart disease. Researchers and medical practitioners can better understand risk factors, diagnostic signs, and potential therapies for the prevention and management of heart disease by examining these characteristics and their connection to the disease.

C. DATASET PREPROCESSING

To ensure optimal performance and accuracy in ML, preprocessing is a crucial step that involves cleaning and preparing data before it is fed into a model for training. The role

of preprocessing is pivotal in enhancing an ML model's performance and improving the accuracy of its predictions. In order for ML to be effective, handling missing values is an essential step. This is because most algorithms depend on complete data. As a result, missing data entries are eliminated from the dataset in this study to make it complete and effective. Data points known as outliers, which frequently result from measurement, data collecting, or input mistakes, differ dramatically from the rest of the data. These outliers may introduce bias and lower accuracy in ML models, which would be hazardous. Therefore, handling outliers is a critical preprocessing step to obtain reliable and accurate ML outcomes. If any outliers are discovered in this study, they are dealt with by being eliminated to maintain the integrity of the data. In ML, converting data from one data type to another is a crucial step in preprocessing known as data type conversion. This step is necessary because the input data may not always be in a format that ML algorithms can process effectively. This process involves converting text data to numerical data, transforming numerical data to categorical data, or scaling numerical data to standardize each feature. Properly converting data types is vital for ML algorithms to handle input data accurately and efficiently, which is necessary to generate accurate predictions. Resampling is a useful method for ML when dealing with unbalanced datasets. Unbalanced datasets are those that have an uneven distribution of cases across classes, which can lead to biased models. Resampling entails changing a dataset's class distribution to produce a balanced dataset. The minority class can be over-sampled or the dominant class can be under-sampled to achieve this. An ML model's performance can be enhanced by using resampling to make sure it has a representative dataset. Many techniques, including SMOTE, random under-sampling, random over-sampling, and ADASYN (Adaptive Synthetic Sampling), can be used for re-sampling.

As all three datasets used in this research are balanced, resampling techniques are not utilized. However, it is important to assess whether a dataset is balanced or not and implement resampling methods when deemed necessary. In this study, the three datasets exhibit an even distribution, as depicted in Fig. 2.

D. SPLITTING DATASET

Data is essential to ML models, and splitting the data into training and testing sets is an important phase in model development. The model may experience problems like underfitting or overfitting, which could produce biased results if the data are not split appropriately. The importance of considering the train and test sets without developing a separate validation set is emphasized in this section as we explore our strategy for splitting the dataset. We used a common procedure to divide the data into train and test datasets using an 80:20 ratio, where 80% of the data has been allocated for training and 20% for testing, to ensure a trustworthy evaluation of our models. The training set is the starting point for training the models, which helped them discover patterns

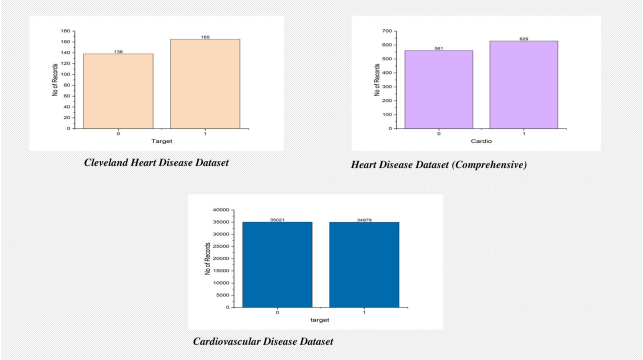


FIGURE 2. Balanced HD dataset.

and relationships in the information being analyzed. On the other hand, the test set served as an unbiased measure of the model's performance on unseen data, providing insights into its ability to generalize. Table 4, which displays the split of records for the Cleveland, HD (Comprehensive), and CVD datasets into train and test sets, provides a summary of the dataset partitioning details. The table shows how many records have been allocated to each set, emphasizing the percentage that has been employed for model training and testing.

Even though it is typical for ML research to include a separate validation set, we decided to only use the train and test sets for the following reasons:

- Our primary goal in doing this study is to concentrate on using ensemble approaches rather than standalone models to increase the efficiency of the models. The Voting Classifier (VC) and the novel QMBC are two ensemble methods that we wanted to investigate for their potential to improve prediction performance. Therefore, rather than adjusting individual models through hyperparameter tweaking, our focus is on the integration of many models.
- Existence of enough data: The datasets used in our analysis, namely the Cleveland dataset, the HD (Comprehensive) dataset, and the CVD dataset, consisted of 303, 1,190, and 70,000 records, respectively. Due to the substantial amount of available data, we did not require an additional validation set. Sufficient data is available for training and testing our models.
- Furthermore, given the size and goals of our study, we decided against engaging in a thorough hyperparameter tweaking process. Even while hyperparameter tuning might boost model performance, it frequently necessitates using a validation set to compare various configurations. We chose a straightforward strategy without considerable hyperparameter adjustment because ensemble techniques are our main focus.

Considering these factors, we made a conscious decision to only utilize the train and test sets for our evaluation. The train set allowed us to train the models and optimize their internal parameters, while the test set provided an independent assessment of their generalization performance on

TABLE 1. Attributes and their Relevance to Heart Disease Prediction.

Feature	Relevance to Heart Disease Prediction
Age	Age is an important risk factor for heart disease
Sex	Gender can play a role in heart disease risk and symptoms
Chest Pain Type (cp)	Different types of chest pain may indicate different heart conditions
Resting Blood Pressure (trestbps)	High blood pressure is a significant risk factor for heart disease
Cholesterol (chol)	Elevated cholesterol levels can contribute to heart disease
Fasting Blood Sugar (fbs)	Elevated fasting blood sugar levels can be a risk factor for heart disease
Resting Electrocardiographic Results (restecg)	Abnormal electrocardiographic findings can indicate heart conditions
Maximum Heart Rate Achieved (thalach)	Abnormal heart rate responses can be relevant for diagnosing heart disease
Exercise-Induced Angina (exang)	Angina during exercise can indicate underlying heart disease
ST Depression Induced by Exercise (oldpeak)	ST depression can indicate reduced blood flow and heart disease
Slope of the Peak Exercise ST Segment (slope)	The slope can provide additional information for diagnosing heart disease
Number of Major Vessels Colored by Fluoroscopy (ca)	The presence of vessel blockages can indicate the severity of heart disease
Thallium Stress Test Result (thal)	Abnormal results can suggest the presence of coronary artery disease
Presence or Absence of Heart Disease (num)	This attribute serves as the target variable for heart disease prediction

TABLE 2. Attributes and their relevance to Heart Disease Prediction (Comprehensive).

Feature	Relevance to Heart Disease Prediction
Age	Age is an important risk factor for heart disease
Sex	Gender can play a role in heart disease risk and symptoms
Chest Pain Type (cp)	Different types of chest pain may indicate different heart conditions
Resting Blood Pressure (trestbps)	High blood pressure is a significant risk factor for heart disease
Cholesterol (chol)	Elevated cholesterol levels can contribute to heart disease
Fasting Blood Sugar (fbs)	Elevated fasting blood sugar levels can be a risk factor for heart disease
Resting Electrocardiographic Results (restecg)	Abnormal electrocardiographic findings can indicate heart conditions
Maximum Heart Rate Achieved (thalach)	Abnormal heart rate responses can be relevant for diagnosing heart disease
Exercise-Induced Angina (exang)	Angina during exercise can indicate underlying heart disease
ST Depression Induced by Exercise (oldpeak)	ST depression can indicate reduced blood flow and heart disease
Slope of the Peak Exercise ST Segment (slope)	The slope can provide additional information for diagnosing heart disease
Presence or Absence of Heart Disease (num)	This attribute serves as the target variable for heart disease prediction

unseen data. By adopting this approach, we aimed to ensure a robust evaluation framework while focusing on the core objectives of our study. This strategy has been employed to reduce the possibility of leakage and overfitting. To ensure a fair evaluation of their performance, the models underwent scrutiny using data that they had not encountered on the test set. The test sets hadn't been exposed to any algorithms or preprocessing procedures, including feature selection. This ensured a reliable assessment of the model's performance on unseen data, free from any potential influence or bias in order to ensure their independence and unbiased nature. In future research, the inclusion of a separate validation set and exploring hyperparameter tuning could be considered to further enhance the performance of the models.

E. MACHINE LEARNING TECHNIQUES

ML has revolutionized disease prediction by allowing medical professionals to create predictive models using a huge quantity of patient data. By using ML models to identify risk factors and trends that may not be readily apparent, doctors can develop individualized treatment plans for patients based on their unique risk factors.

Recent innovations in ML have generated interest in utilizing its potential benefits alongside established statistical analysis methods used in cardiology. While survival curves and statistical modeling have historically been used by cardiologists to predict cardiovascular outcomes, the incorporation of ML has the opportunity to identify complex patterns and relationships that may be missed by traditional methods. Healthcare professionals aspire to improve their ability and reliability in predicting a patient's vulnerability to heart disease or the likelihood of experiencing a heart attack by leveraging the capabilities of ML algorithms. This developing relationship between ML and cardiology has the potential to improve modern techniques for prediction and move the field closer to more accurate and reliable cardiovascular risk

TABLE 3. Attributes and their relevance to Cardiovascular Disease Prediction.

Feature	Relevance to Cardiovascular Disease Prediction
Age	Age of the patient is an important risk factor for cardiovascular disease
Height	Height of the patient may be related to cardiovascular health
Weight	Weight of the patient may be related to cardiovascular health
Gender	Gender can play a role in cardiovascular disease risk and symptoms
Systolic Blood Pressure	High blood pressure is a significant risk factor for cardiovascular disease
Diastolic Blood Pressure	High blood pressure is a significant risk factor for cardiovascular disease
Cholesterol	Elevated cholesterol levels can contribute to cardiovascular disease
Glucose	Elevated glucose levels can be a risk factor for cardiovascular disease
Smoking	Smoking is a known risk factor for cardiovascular disease
Alcohol Consumption	Excessive alcohol consumption can increase the risk of cardiovascular disease
Physical Activity	Regular physical activity can help prevent cardiovascular disease
Target (Cardiovascular Disease)	Presence or absence of cardiovascular disease; serves as the target variable for prediction

TABLE 4. Datasets partitioning.

Dataset	Train set	Test set
Cleveland	242	61
HD (Comprehensive)	952	238
CVD	56,000	14,000

assessment. These models can use a variety of variables, including age, gender, family history, lifestyle factors, and medical history, to calculate a risk score. To predict the disease accurately, this study employs seven ML techniques, including LR, DT, RF, KNN, NB, SVC, and MLP, as well as an ensemble model such as VC and QMBC.

1) LOGISTIC REGRESSION (LR)

A logistic function is used in a LR model, which is a statistical method for modeling binary dependent variables. Classification problems commonly make use of LR.

$$P(A = 1|B) = \frac{1}{1 + c^{-\beta_0 - \beta_1 B_1 - \beta_2 B_2 - \ldots - \beta_q B_q}} \quad (1)$$

where

P(A=1|B) is the conditional likelihood that event A will occur given event B, event A is a binary attribute with a value of either 1 or 0 and event B is a collection of predictor variables. $\beta_0, \beta_1, \beta_2, \ldots, \beta_q$ is the LR model's predictor variable correlation coefficients. During the model-fitting process, these coefficients are calculated from the data. $B_1, B_2, \ldots, B_q$ is the LR model's predictor variables. These variables, which might be categorical or continuous, are thought to be independent of one another. c is the odds ratio for a unit variation among the predictor variables. For ease of use, this value is frequently set to 1, however, it may also be approximated using the data. e is the natural logarithm's base, which is roughly equivalent to 2.71828. This is incorporated into the equation for the LR model.

We may calculate the likelihood that event A will occur using LR on the basis of the values of predictors $B_1, B_2, \ldots, B_q$. The LR model calculates the coefficients $\beta_0, \beta_1, \beta_2, \ldots, \beta_q$ that show the correlation between the predictors and the likelihood that event A will occur. For ease of usage, the parameter c is frequently set to 1, and the LR equation is exponentiation using e. With respect to the values of predictors $B_1, B_2, \ldots, B_q$ the resultant probability, which is limited among 0 and 1, may be understood as the likelihood that event A will occur.

2) DECISION TREE (DT)

A versatile technique is the DT model, which employs a set of decision rules defined by branches and attributes represented by nodes, each of which is a leaf node expressing a class label. It may be used for problems involving classification and regression.

$$A(B) = -\sum_{i=1}^{|C|} a_i \log_2(a_i), \quad (2)$$

$$\text{Gain}(B, X) = A(B) - \sum_{b \in \text{Values}(X)} \frac{|B_b|}{|B|} A(B_b) \quad (3)$$

where

A(B) is the entropy of the target variable, a_i is the probability of class 'i' in the target variable, y is the number of classes in the target variable, B is the current dataset, X is a feature in the dataset, B_b is the subset of B for which feature X has value c, $|B_b|$ is the number of instances in B_b, Values(X) are the set of all possible values of feature X.

3) RANDOM FOREST (RF)

It is an ensemble approach that combines different decision trees to enhance performance and reduce overfitting. The model functions by building several decision trees on various

random subsets of data and then combining all of the trees predictions. It may be applied to both classification and regression problems.

$$\hat{y}_i = \frac{1}{M}\sum_{j=1}^{M} f_j(x_i) \tag{4}$$

where

$\hat{y}_i$ is the predicted output for observation i, M is the number of trees, f_j is j^{th} decision tree, and x_i is the input vector for observation i.

4) K NEAREST NEIGHBOURS (KNN)

An approach for non-parametric ML that forecasts the result based on the training set's KNN. It is a versatile algorithm that can be used for both regression and classification tasks. The number of nearest neighbors in a KNN used to create predictions for a new observation is represented by the value of k. KNN measures the distance between both the new entity and every other entity in the training set and figures out the classification of the new entity. A forecast for the new observation is then made using the class labels of the k nearest neighbours.

Several distance metrics, including the Euclidean, Manhattan, and Minkowski distances, can be used by KNN to determine the separation between instances.

$$\hat{y}(x) = \text{mode}\{y_i : x_i \in N_k(x)\} \tag{5}$$

where

$\hat{y}(x)$ is the predicted value of the target variable for the input x, y_i is the target value for the i^{th} training instance, $N_k(x)$ is the set of KNN of the input x in the training set, mode{ } denotes the most common value in a set.

5) NAIVE BAYES (NB)

A probabilistic algorithm that calculates the probability of each class based on the input features is known as Naive Bayes. It is commonly used for text classification and other high-dimensional datasets.

$$P(g|d_1, d_2, \ldots, d_n) = \frac{P(g)\prod_{i=1}^{n} P(d_i|g)}{P(d_1, d_2, \ldots, d_n)} \tag{6}$$

where

g is the class attribute, $d_1, d_2, \ldots, d_n$ are the features, $P(g|d_1, d_2, \ldots, d_n)$ is the posterior probability of g given $d_1, d_2, \ldots, d_n$, $P(g)$ is the prior probability of g, $P(d_i|g)$ is the probability of d_i given g, $P(d_1, d_2, \ldots, d_n)$ is the marginal probability of $d_1, d_2, \ldots, d_n$.

6) SUPPORT VECTOR MACHINE (SVM)

SVM is a technique that divides data into groups using a hyperplane in a high-dimensional environment. Given that it works well for both linear and non-linear classification tasks, SVM is a versatile ML technique.

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2} wv^T wv + E\sum_{j=1}^{n} \zeta_j \\ \text{subject to} \quad & h_j(wv^T \phi(z_j) + f) \geq 1 - \zeta_j, \\ & \zeta_j \geq 0, \quad j = 1, \ldots, n \end{aligned} \tag{7}$$

In this formula, wv is vector weight, f is the bias, E is the penalty parameter, h_j is the label of the j^{th} training example, z_j is the j^{th} training example, and $\phi(z_j)$ is feature map of x_j. The ζ_j variables are slack variables that allow for some misclassification of the training examples.

7) MULTILAYER PERCEPTRON (MLP)

A particular kind of neural network called an MLP is made up of several layers of linked nodes. The MLP method is a flexible tool that may be applied to a variety of applications, such as classifiers, regression, and outlier detection.

$$h_1 = af(Wm_1 v + bv_1), \tag{8}$$
$$h_2 = af(Wm_2 h_1 + bv_2), \tag{9}$$
$$\vdots \tag{10}$$
$$h_L = af(Wm_L h_{L-1} + bv_L), \tag{11}$$
$$q = ao(Wm_{L+1} h_L + bv_{L+1}). \tag{12}$$

where

The i^{th} hidden layer produces the output h_i, for the i^{th} layer, Wm_i represents the weight matrix, and bv_i represents the bias vector, the input vector is represented by bv, the activation function used for the hidden layers is denoted by af, while ao represents the activation function used for the output layer, the quantity L represents the number of hidden layers in the MLP, while q denotes its output.

8) VOTING CLASSIFIER (VC)

It is a method of ensemble learning that combines the results of various independent models to produce a single final prediction. It is a straightforward but efficient strategy that can raise the model's overall effectiveness. Every model in a VC generates its own prediction after being trained using the same dataset. A final prediction is then made by the VC. In a hard vote, the result is determined by a majority of votes. When three separate models are used, for instance, and two of them predict class A and the third predicts class B, class A will be the final prediction, representing the majority opinion.

$$\hat{j} = \text{model}[f_1(l), f_2(l), \ldots, f_{mi}(l)] \tag{13}$$

where

$\hat{j}$ represents the final forecast made by the voting classifier. The model function takes the model (most common) value among the predictions of each individual model $f_1(l), f_2(l), \ldots, f_{mi}(l)$, mi is the number of models used in the ensemble.

After preprocessing, the dataset is fed to both individual ML models and an ensemble model to anticipate whether the patient is diagnosed with HD or not. The findings and analyses of the prediction are presented in the results section IV.

F. FEATURE SELECTION (FS) STRATEGY

FS is a critical process in ML that involves selecting the most relevant and informative features from a dataset to enhance a model's performance. This process aims to eliminate irrelevant or redundant data, which ultimately minimizes the dimensionality. By doing so, the model's accuracy and efficiency can be improved, as it will have a smaller number of input variables to work with.

In this study, two FS techniques are utilized to enhance the disease prediction model. The two FS techniques are ANOVA and Chi-Square.

1) CHI-SQUARE

Chi-Square is a statistical method used in ML to pick the most relevant attributes. The Chi-Square is calculated for each attribute, and the highest scores are chosen. This is the basic functioning of the Chi-Square technique. If a feature is determined to be independent of the target attribute, it is excluded. On the other hand, if a feature has a relatively high Chi-Square (Chi2) score, it is deemed more relevant to the target attribute. By selecting the most informative attribute, performance can be significantly boosted while also reducing the complexity of the data.

$$\chi^2(X, y) = \sum_{j=1}^{2} \sum_{k=1}^{2} \frac{(T_{ij} - E_{ij})^2}{E_{ij}} \tag{14}$$

In this equation, X stands for the independent attributes and y for the dependent. The target variable's y stands for the number of categories, while T and E stand for the actual and anticipated frequency of each attribute pair. The top attributes with the greatest Chi-Square values are chosen for the final model after the Chi-Square value is computed for each feature.

Algorithm 2 computes the Chi-Square score for each independent and dependent attribute pair. *Step-1* of Algorithm 2, a contingency table is generated to show the frequency of each pair of independent and dependent variable values. Then, the anticipated frequency for each cell in the contingency table is calculated under the assumption that the feature and the dependent variable are independent. *Step-2*, the Chi-Square statistic is computed for each feature, which indicates the degree to which the observed frequency in the contingency table differs from the expected frequency. Finally, *Step-4*, the algorithm generates a new input data matrix by selecting the top k attributes with high Chi-Square scores.

Step-6 & *7*, the process picks the relevant attributes in the dataset for the target variable. The model's speed and high computational cost can be enhanced by minimizing the amount of pointless or redundant elements.

The attributes with the best *chi*2 scores are chosen, and a new input dataset called X_{selected} is produced. X_{selected} is then applied to ML models and ensemble approaches, and the model's performance is discussed in the result section IV.

Algorithm 2 Chi-Square Feature Selection

Require: X: input data matrix with n samples and m features, y: target variable

Ensure: X_{selected}: input data matrix with selected features

1: Calculate the contingency table T for each feature i and target variable y.
2: **for** $i = 1$ to m **do**
3: Calculate the expected frequency for each cell in $T_{i,\cdot}$ as $\frac{\sum_{j=1}^{n} T_{ij} \sum_{j=1}^{n} T_{\cdot j}}{n^2}$
4: Calculate the chi-square statistic for feature i as $\chi_i^2 = \sum_{j=1}^{2} \sum_{k=1}^{2} \frac{(T_{ij} - E_{ij})^2}{E_{ij}}$
5: **end for**
6: Choose the best k attributes with the best *chi*2 scores.
7: Create a new input data matrix X_{selected} with the selected features
8: **return** X_{selected}

These three tables represent the top features selected by Chi-Square on three different datasets. Table 5 displays the top 10 attributes selected on the Cleveland dataset, which is a dataset of heart disease patients. Table 6 shows the top 10 attributes selected on the HD dataset, which is a comprehensive dataset of heart disease patients. Table 7 displays the top 9 attributes selected on the CVD dataset, which is a dataset related to cardiovascular diseases. The scores in each table represent the Chi-Square values for each feature, with higher scores indicating a stronger association with the target variable. These tables can be used as a starting point for FS in ML models and can assist in selecting the essential features for predicting HD or CVD.

2) ANOVA

ANOVA is a statistical technique that measures the significance of variations among categories or groups of data. The F-test score is used to determine the degree of variance in the target variable that can be attributed to the variance in a particular feature. The following describes how the ANOVA Algorithm 3 works:

1) By contrasting the variance of the feature's values among several classes of the target variable, find the F-test score for each characteristic in the dataset.
2) Arrange the characteristics in descending order according to their F-test results.
3) To generate the ultimate feature subset, pick the top k characteristics with the greatest F-test outcomes.
4) In order to exclude attributes that do not strongly correlate to the target variable, a threshold may need to be set on the F-test result.
5) A fresh dataset created by the ANOVA technique, which only includes the chosen features, can be utilized to create an ML model. ANOVA can improve precision and effectiveness by focusing on the most crucial features.

TABLE 5. Top 10 attributes ranked by Chi-Square feature selection on the Cleveland dataset.

Feature	thalach	oldpeak	ca	cp	exang	chol	age	trestbps	slope	sex
Score	188.32	72.64	66.44	62.59	38.91	23.93	23.28	14.82	9.80	7.57

TABLE 6. Top 10 attributes ranked by Chi-Square feature selection on the HD dataset (Comprehensive.)

Feature	exercise angina	fasting blood sugar	chest pain type	sex	ST slope	max heart rate	oldpeak	age	cholesterol	resting ecg
Score	168.98	43.95	32.88	27.22	23.24	11.67	7.18	5.67	3.79	3.44

TABLE 7. Top 9 attributes ranked by Chi-Square feature selection on the CVD Dataset.

Feature	cholesterol	gluc	age	weight	active	smoke	ap_lo	alco	gender
Score	2158.98	403.69	215.82	39.24	17.46	15.30	5.82	3.55	2.99

TABLE 8. Top 10 attributes ranked by Anova feature selection on the Cleveland dataset.

Feature	thalach	slope	oldpeak	thal	cp	ca	age	trestbps	chol	restecg
Score	65.12	55.33	53.77	45.77	29.79	20.95	16.11	6.45	2.20	1.41

Algorithm 3 ANOVA Feature Selection Algorithm

Input: Data matrix $X \in \mathbb{R}^{n \times d}$, labels $y \in \{0, 1\}^n$, number of top features k
Output: Top k features according to ANOVA F-test (Anv_{features})
Calculate the mean value for each feature:
for $j = 1$ to d **do**
    $\mu_j \leftarrow \frac{1}{n} \sum_{i=1}^{n} x_{ij}$
end for
Calculate the between-class variability SS_B and within-class variability SS_W:
$SS_B \leftarrow \sum_{j=1}^{d} n_j(\mu_j - \mu)^2$
$SS_W \leftarrow \sum_{j=1}^{d} \sum_{i=1}^{n_j} (x_{ij} - \mu_j)^2$
Calculate the ANOVA F-score for each feature:
for $j = 1$ to d **do**
    $F_j \leftarrow \frac{SS_B/d}{SS_W/(n-d)}$
end for
Select the top k features with the highest F-scores:
$idx \leftarrow \text{argsort}(F)[:: -1]$
$Anv_{\text{features}} \leftarrow X[:, idx[: k]]$
return Anv_{features}

The best k features are chosen based on their F-scores and a new dataset is created namely Anv_{features}. ML models and ensemble approaches are applied to the Anv_{features} dataset. In section IV, all of the models' performance is discussed.

These three tables represent the top attributes selected by Anova on three different datasets. Table 8 displays the top 10 attributes selected on the Cleveland dataset. Table 9 shows the top 10 attributes selected on the HD (comprehensive) dataset. Table 10 displays the top 9 attributes selected on the CVD dataset. The scores in each table represent the Chi-Square values for each feature, with higher scores indicating a stronger association with the target variable. These tables can be used as a starting point for FS in ML models and can assist in selecting the essential features for predicting HD or CVD.

G. FEATURE EXTRACTION (FE) TECHNIQUE

In this study, a grouping of FS and FE approaches is employed to increase the performance of the model. Specifically, the Chi-Square with PCA and Anova with PCA approaches are utilized to select relevant features and extract primary components from the dataset. By fusion of these techniques, the most important features are selected and the dimensionality of the dataset is reduced, resulting in improved model performance. To extract the prime components from the datasets, the PCA technique is applied to the two selected datasets, namely X_{selected} and Anv_{features}.

1) PRINCIPAL COMPONENT ANALYSIS (PCA)

PCA is a data analysis technique that aims to reduce dimensionality. It achieves this by compressing a high-dimensional into a lower-dimensional one while retaining as much of the variability of the original data as possible. PCA starts with selecting the most relevant features using FS techniques like ANOVA or Chi-Square. PCA is then applied to these chosen characteristics to extract principle components, linear variants of the feature set that effectively represent the most important variances in the data. The principal components are sorted by their corresponding eigenvalues, and the components with the highest eigenvalues are selected as they capture most of the variability in the data.

Here are the mathematical formulas for PCA:

Centering the data:

$$Z_{\text{centered}} = Z - \bar{Z} \tag{15}$$

where

Z_{centered} is the centered data, Z is the original data, and $\bar{Z}$ is the mean of the original data.

TABLE 9. Top 10 attributes ranked by Anova feature selection on the HD dataset (Comprehensive.)

Feature	ST slope	exercise angina	chest pain type	max heart rate	oldpeak	sex	age	fasting blood sugar	cholesterol	resting bp s
Score	408.00	358.49	319.07	244.70	224.11	127.45	87.58	58.53	48.66	17.77

TABLE 10. Top 9 attributes ranked by Anova feature selection on the CVD dataset.

Feature	age	cholesterol	weight	gluc	ap_lo	ap_hi	active	smoke	height
Score	4209.007	3599.36	2388.77	562.77	303.62	208.33	89.09	16.79	8.19

Computing the covariance matrix:

$$\text{Cov}(Z) = \frac{1}{n-1}(Z_{\text{centered}})^T(Z_{\text{centered}}) \tag{16}$$

where

Cov(Z) is the covariance matrix, n is the number of records, and T denotes the transpose operation.

Calculating the eigenvalues and eigenvectors:

$$\text{Cov}(Z)\text{ev} = \lambda\text{ev} \tag{17}$$

where

Cov(Z) is the covariance matrix, λ is the eigenvalue, and ev is the corresponding eigenvector.

Sorting the eigenvectors by their corresponding eigenvalues:

$$\text{ev}_{\text{sorted}} = [\text{ev}_1, \text{ev}_2, \ldots, \text{ev}_d] \tag{18}$$

where

$\text{ev}_{\text{sorted}}$ is the sorted eigenvector matrix, and d is the number of dimensions.

Choosing the number of principal components: k principal components are selected such that they account for a large percentage of the variance. This can be determined by examining the eigenvalues and selecting the top k eigenvectors that account for the majority of the variance.

Projecting the data onto the new feature space:

$$Z_{\text{new}} = Z_{\text{centered}}\text{ev}_{\text{sorted}}[:, 1:k] \tag{19}$$

where

Z_{new} is the data projected onto the new feature space, and $[:, 1:k]$ denotes the first k columns of the sorted eigenvector matrix.

The PCA technique is presented in Algorithm 4.

Algorithm 4 is employed to extract the principal components from the X_{selected} and Anv_{features} datasets. After applying PCA on these datasets, two subsets X_{Chi}selected and Anv_{PCA}selected are obtained. The performance of ML models and ensemble models is then evaluated using these subsets to predict the presence or absence of HD in patients. All the model results are discussed in the result section IV.

H. QUINE McCluskey BINARY CLASSIFIER (QMBC)

In this study, we introduce a novel ensemble technique, the Quine McCluskey Binary Classifier (QMBC), for predicting whether a patient has been diagnosed with HD or not. The QMBC is applied to all three datasets and their corresponding subsets after FS, FE, and the fusion of FS with FE approaches. The proposed methodology is an ensemble of seven standalone ML models, and the workflow of QMBC is shown in Fig. 3.

All the datasets are tested by the seven standalone ML models, and the predicted results along with the target feature are saved in CSV format and serve as the input for the proposed model. The QMBC architecture uses seven distinct ML models, with each patient record from the dataset labeled with 0s and 1s after training and testing all the models. When a patient is not diagnosed with HD, the model predicts 0, and when they do, it predicts 1. The Quine McCluskey method is used to generate an equation by classifying all patient data according to whether they have been diagnosed with HD or not. For each of the seven models, LR, DT, RF, KNN, NB, SVM, and MLP are represented as x_0, x_1, x_2, x_3, x_4, x_5, and x_6, respectively, and the dependent variable represented as 'target', the following findings are obtained. Each column in the dataset contained the predicted model results for each of the M patient records. The dataset is labeled as D. The D consists of 7 features and M patient-predicted outcomes from 7 different ML models. Similarly $\{\sim x_0, \sim x_1, \sim x_2, \sim x_3, \sim x_4, \sim x_5, \text{and} \sim x_6\}$ is represented as $\{\bar{x}0, \bar{x}1, \bar{x}2, \bar{x}3, \bar{x}4, \bar{x}5, \text{and } \bar{x}6\}$. Where, if $x_i = 0$ then $\bar{x}_i$ will be 1, similarly if $x_i = 1$ then $\bar{x}_i$ will be 0.

Algorithm 4 Principal Component Analysis Algorithm

1: Mean centering: Calculate the mean μ for each feature and subtract it from the corresponding feature in Z: $Z_{\text{centered}} = Z - \mu$
2: Covariance matrix: Calculate the covariance matrix Σ for Z_{centered}: $\Sigma = \frac{1}{n-1}X^T_{\text{centered}}Z_{\text{centered}}$
3: Eigen decomposition: Compute the eigenvalues λ_i and eigenvectors ev_i for Σ $ev = [, ev_1; ev_2; \cdots; ev_m,]$
4: Feature transformation: Project Z_{centered} onto the k-dimensional space spanned by the top k eigenvectors: $Z' = Z_{\text{centered}}ev_k$
5: Variance explained: Compute the variance explained by each principal component: $Var_i = \frac{\lambda_i}{\sum_{j=1}^{m}\lambda_j}$
6: Choose the number of components: Choose the number k of principal components to keep based on the variance explained and/or the desired dimensionality reduction

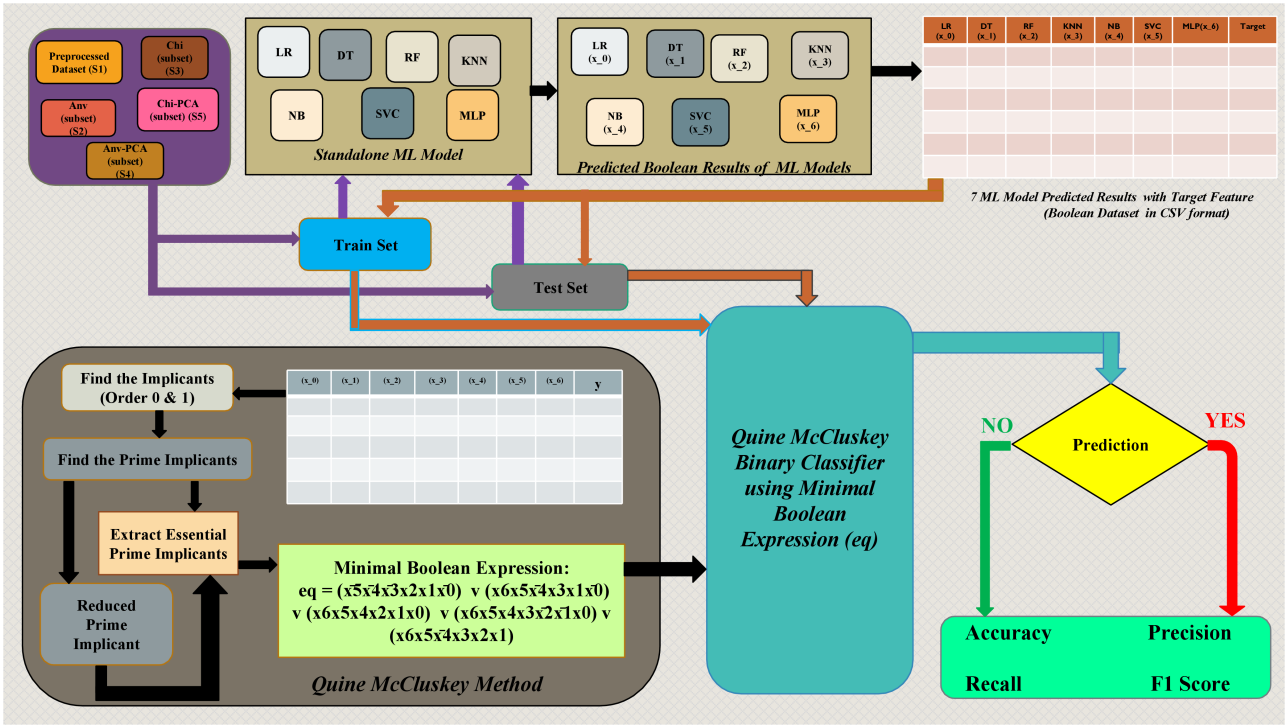


FIGURE 3. Proposed workflow.

The Quine McCluskey method is an algorithmic approach to minimize the number of terms in a boolean expression. It is a two-step process that involves finding all the prime implicants and then grouping them to obtain a minimal boolean expression shown in Algorithm 5. The QMBC equation is displayed in Algorithm 6. The following is the step-by-step procedure of the Quine McCluskey method:

Step-1 of Algorithm 5 is the Initialization: The algorithm initializes three sets: M_0, which represents the set of all possible minterm groupings, P_0, which represents the set of all prime implicants of f, and C_0. *Step-2* Iteration: The process iterates until no more prime implicants are created or can be ascribed to the undiscovered minterm groupings. *Step-3* Create M_{i+1}: The algorithm creates M_{i+1} by finding all of the minterms that are represented by each prime implicant in P_i for each iteration. When at least one prime implicant covers a given minterm, the two minterms are joined to form a new minterm. *Step-4* In order to create P_{i+1}, the method first identifies all prime implicants that cover at least one minimum term in M_{i+1}. After that, the algorithm eliminates any unnecessary prime implicants from this collection. *Step-5* Build C_{i+1}: The algorithm creates a fresh collection of potential minterm groupings by selecting all viable minterm groupings from M_{i+1}. *Step-7* Simplification: The algorithm conducts a simplification phase to assign prime implicants to exposed minterm groupings. The method selects a grouping from C_{i+1} that is uncovered in each iteration of this step, c, and determines if a prime implicant p exists that covers c. If there is, the algorithm eliminates c from C_{i+1} and assigns p to cover c. The procedure combines the minterms in c to create a new prime implicant q and adds q to P_{i+1} if there is no prime implicant that covers c. *Step-19* Update Iteration Counter: After each iteration, the algorithm increases the iteration counter i. *Step-20* Output: After the process finishes, its result is the Boolean function that has been simplified and is represented by the remaining prime implicants.

The Quine McCluskey method is an iterative process that may require multiple passes to obtain the minimal expression. It is a systematic and efficient approach to minimize boolean expressions with many variables.

The QM boolean minimized equation is generated using Algorithm 6. This equation is then used to create a novel ensemble classifier called QMBC, which is used to predict the HD datasets. The QMBC Algorithm is shown in Algorithm 7.

To train the proposed model, the features x_0, x_1, x_2, x_3, x_4, x_5, and x_6 are used as independent features, and the target variable is used as the dependent feature in Algorithm 6. The model is trained on 80% of the training dataset. The QMBC model is then built using algorithm 7. Afterward, the model is tested on 20% of the testing set, and its performance is evaluated using the accuracy, precision, recall, specificity, and f1-scores of all the models. A detailed discussion of all the model's results are available in the result section IV.

I. PERFORMANCE EVALUATION METRICS

To evaluate the performance of all models and proposed methodologies in this work, a confusion matrix is utilized. It provides an efficient and effective means of quantifying a model's accuracy, precision, specificity, recall, and f1-score. The confusion matrix is comprised of True Positive (TP),

Algorithm 5 Quine McCluskey Method

Initialization:
M_0 = set of minterms of f,
P_0 = set of prime implicants (PI) of f,
C_0 = set of all possible minterm groupings,
$i = 0$.
while P_i is not empty **do**
    Generate M_{i+1}:
Identify all minterms covered by each PI in P_i,
Combine the minterms that are covered by more than one PI.
    Generate P_{i+1}:
Identify all PI that covers at least one minterm in M_{i+1},
Remove redundant prime implicants.
    Generate C_{i+1}:
Identify all potential groupings of minterms from M_{i+1}.
end while
Simplification:
while there are groupings in C_{i+1} **do**
    Choose an uncovered grouping c from C_{i+1}.
    if there is a PI p that covers c **then**
        Assign p to cover c.
        Remove c from C_{i+1}.
    else
        Combine the minterms in c to form a new PI q.
        Add q to P_{i+1}.
        Remove c from C_{i+1}.
    end if
end while
$i = i + 1$.
eg is the simplified Boolean function represented by the remaining PI.

Algorithm 6 Generating Boolean Equation Using Quine McCluskey Method

1: Read the min-terms from dataset D which are in the form of decimal values.
2: Read $x_0, x_1, \ldots, x_6$ of min-terms from dataset D.
3: Using the min-terms from D find the Prime Implicates (PI) and Essential Prime Implicates (EPI).
4: By using PI & EPI generate the Boolean minimized equation.
5: The Generated Binary Boolean Minimized QM method equation using Algorithm 5 for given D is as follows: $eq = (\bar{x}_5 * \bar{x}_4 * \bar{x}_3 * x_2 * x_1 * \bar{x}_0) v (x_6 * x_5 * \bar{x}_4 * x_3 * x_1 * \bar{x}_0) v (x_6 * x_5 * x_4 * x_2 * x_1 * x_0) v (x_6 * x_5 * x_4 * x_3 * \bar{x}_2 * \bar{x}_1 * x_0) v (x_6 * x_5 * \bar{x}_4 * x_3 * x_2 * x_1)$

False Positive (FP), False Negative (FN), and True Negative (TN) values. Two types of errors are present in the confusion matrix, namely type-1 errors (FP) and type-2 errors (FN). The emphasis placed on each error type should be based on specific needs. Since the focus of this study is HD prediction, minimizing type-2 errors is crucial. The goal is to minimize

Algorithm 7 Quine McCluskey Binary Classifier

1: Read the dataset and apply it to a 7 standalone ML models.
2: Create a new dataset that consists of the predicted results of all 7 ML models along with the target feature in CSV format.
3: Using Algorithm 6 find the Boolean minimized equation.
4: Split the boolean dataset into train and test of 80 : 20 ratio.
5: Using generated minimized equation from Algorithm 6 build the QMBC model with the train set.
6: Test the model with the remaining test set and calculate the performance of the QMBC model.

TABLE 11. Confusion matrix.

	Actual : Yes	Actual: No
Predicted: Yes	TP	FP
Predicted: No	FN	TN

errors across all models. By using the confusion matrix all the model's efficiency is calculated shown in Table 11.

The percentage of instances that are correctly classified out of all instances is known as accuracy.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{20}$$

Precision is a metric that calculates the ratio of true positives to all predicted positives by the model.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{21}$$

Recall measures the ratio of true positives to all actual positives in the dataset.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{22}$$

F1-Score is a combined metric that uses the harmonic mean of both precision and recall to give a balanced score.

$$\text{F1-Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{23}$$

Specificity measures the ratio of true negatives to all actual negatives in the dataset. It indicates the model's ability to correctly identify negative instances.

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{24}$$

IV. RESULTS AND DISCUSSIONS

This section aims to present the outcomes of the conducted experiments and compare them with the findings of relevant prior studies. After extracting the datasets from an open repository, the data is preprocessed and used to evaluate seven standalone ML models, an ensemble VC, and the proposed QMBC methodology. In order to improve model performance, essential features are selected using Chi-Square and

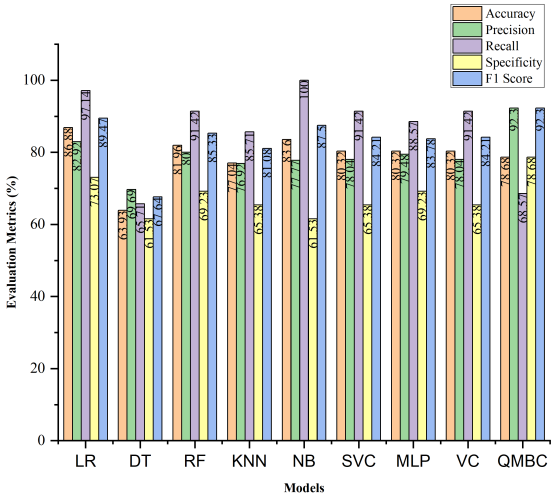


FIGURE 4. Model results of preprocessed (S1) Cleveland dataset.

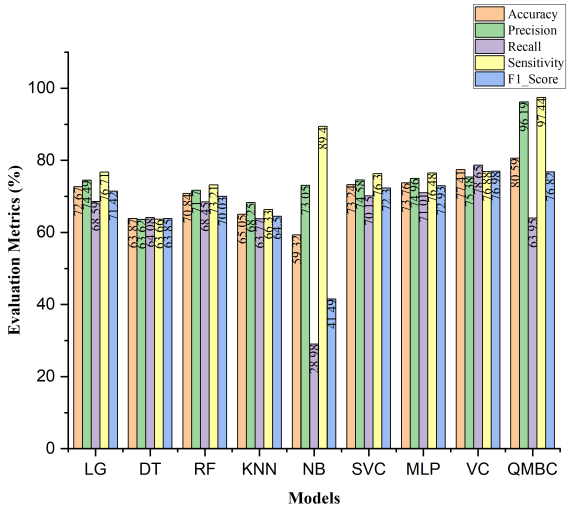


FIGURE 5. Model results of preprocessed (S1) CVD dataset.

Anova techniques. These subsets of features are then applied to all ML models and the QMBC methodology. To further enhance the models and increase their robustness, FS is combined with FE. PCA technique is applied to the X_{selected} and Anv_{features} subsets, resulting in new subsets named $X_{\text{Chi-selected}}$ and $Anv_{\text{PCA-selected}}$. The results of all the models and methodologies are thoroughly discussed in this section.

A. MODEL RESULTS OF PREPOSSESSED DATASETS

After performing data preprocessing, the preprocessed dataset is used to train and test all seven standalone ML models and the ensemble methods of the VC and the proposed QMBC, with an 80:20 ratio. The results for all models are presented in a table and a bar chart for the Cleveland dataset in Fig. 4, a bar chart for the CVD dataset in Fig. 5 and a bar chart for the HD dataset (Comprehensive) in Fig. 6. Among the models without FS & FE LR scored the highest accuracy and precision of 86.88%, and 82.92%, NB scored the highest

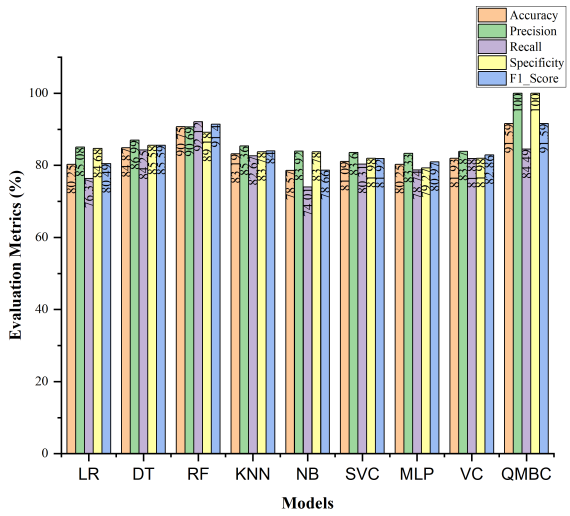


FIGURE 6. Model results of preprocessed (S1) HD dataset (Comprehensive.)

recall of 100%, QMBC outperformed all the models in terms of specificity and f1-score with 78.59%, and 78.68% for Cleveland HD dataset. For the CVD dataset, the proposed QMBC method achieved 80.59%, 96.19%, 63.95%, 97.44%, and 76.82%, respectively. Finally, for the HD dataset (comprehensive), the proposed QMBC method achieved 91.52%, 100%, 84.44%, 100%, and 91.59%, respectively. Specifically, for the Cleveland dataset, the proposed QMBC method achieves an accuracy, precision, recall, specificity, and f1-score of 78.68%, 92.3%, 68.57%, 78.68%, and 92.3%, respectively.

B. MODEL RESULTS OF PREPOSSESSED DATASETS AFTER FS TECHNIQUE

The implementation of the FS approach is crucial in ML for minimizing dataset dimensions, computational speed, and removing irrelevant and duplicate features. A detailed explanation of the FS approach used in this research is provided in the relevant section. In summary, ANOVA and Chi-Square techniques are employed to select the top features from the preprocessed dataset. The selected subsets are then utilized to train and test 7 standalone ML models as well as ensemble strategies, such as the VC and proposed QMBC. The results are presented in the bar charts depicted in Fig. 7 for the Cleveland dataset, CVD dataset as shown in Fig. 8, and HD dataset (Comprehensive) displayed in Fig. 9, respectively. Among the models with ANOVA and without FE, RF achieved the highest accuracy and specificity with values of 89.49% and 87.38% respectively. For precision, QMBC performed the best with 94.01%, while DT had the highest recall of 85.82%. In terms of the f1-score, RF obtained the highest value of 90.27%. In the case of models with Chi-Square and without FE, RF demonstrated the highest accuracy and f1-score with values of 90.75% and 91.47% respectively. For precision, NB achieved the highest score of 84.61%, while DT had

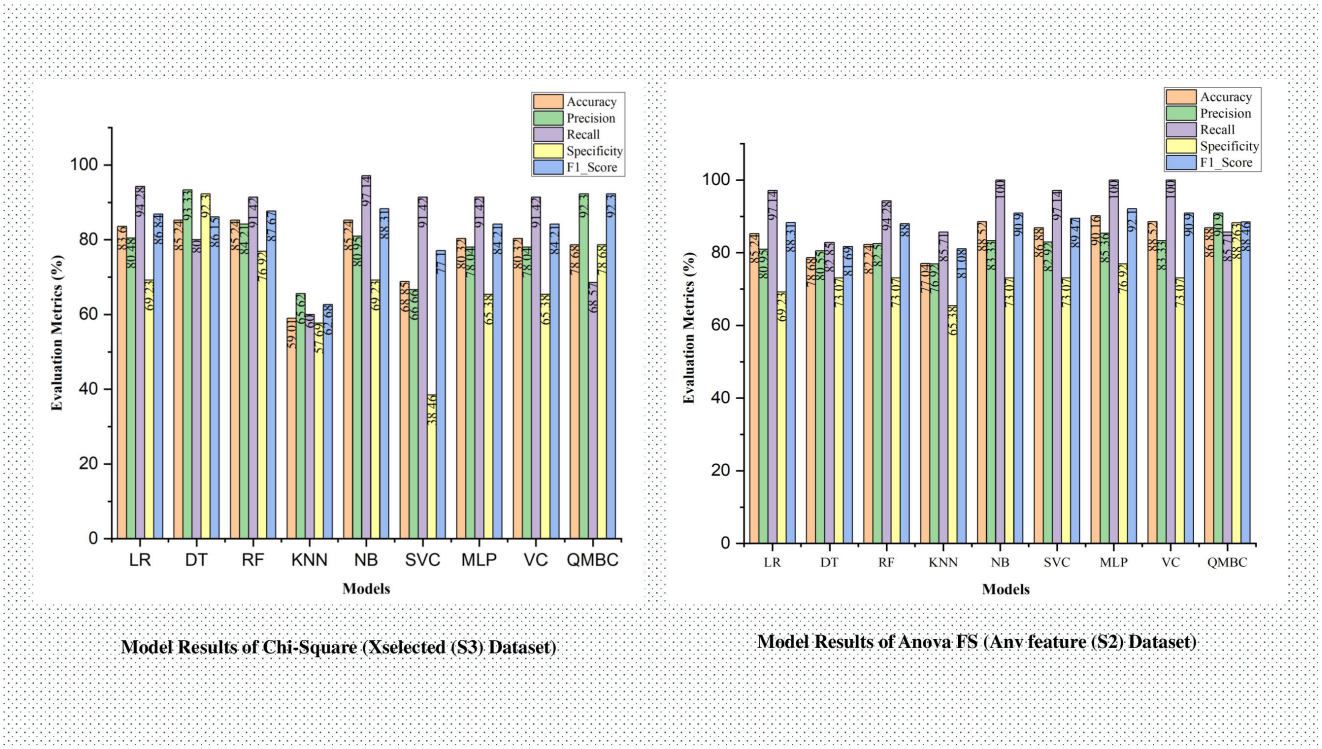


FIGURE 7. Model results of FS technique on Cleveland dataset.

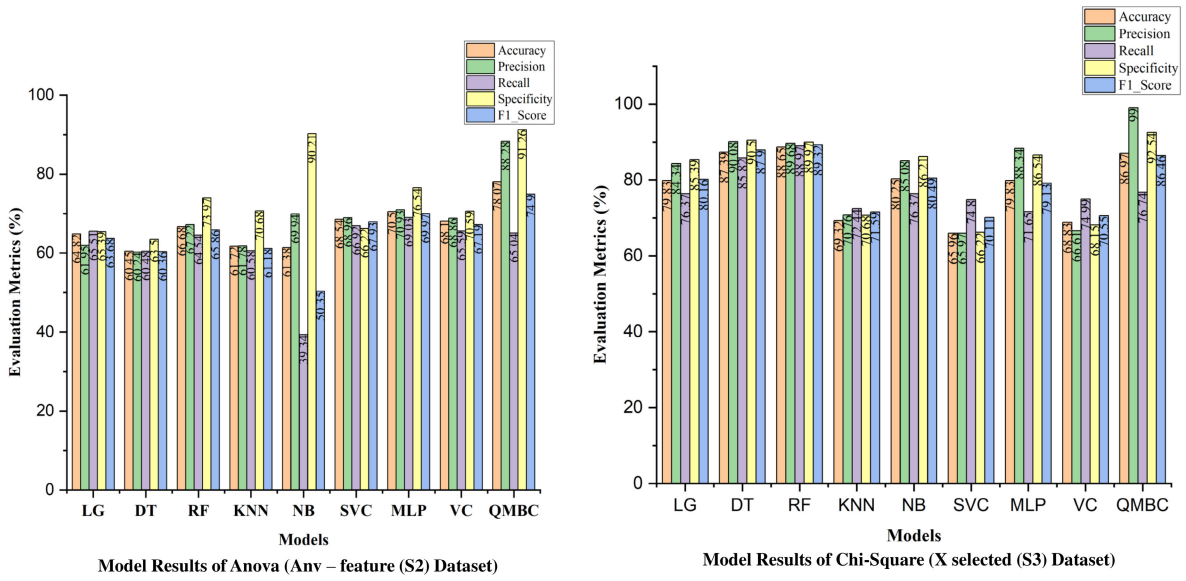


FIGURE 8. Model results of FS technique on CVD dataset.

the highest recall of 87.4%. When considering specificity, RF outperformed the other models with a value of 88.28%.

Overall, RF consistently performed well across both FS methods, showing strong accuracy, precision, recall, specificity, and f1-score. QMBC also demonstrated notable performance in terms of precision and recall in the ANOVA case (Anv_{features}), and in terms of accuracy and f1-score in the Chi-Square case (X_{selected}).

C. MODEL RESULTS OF PREPOSSESSED DATASETS BY FUSION OF FS AND FE TECHNIQUES

In this research, a fusion of FS and FE strategies has been implemented to make the model more efficient and effective. Specifically, PCA is used to extract the prime components from subsets of the preprocessed data, including X_{selected} and Anv_{features}. The outcomes of all the models are exhibited after the amalgamation of FS and FE techniques for the Cleveland

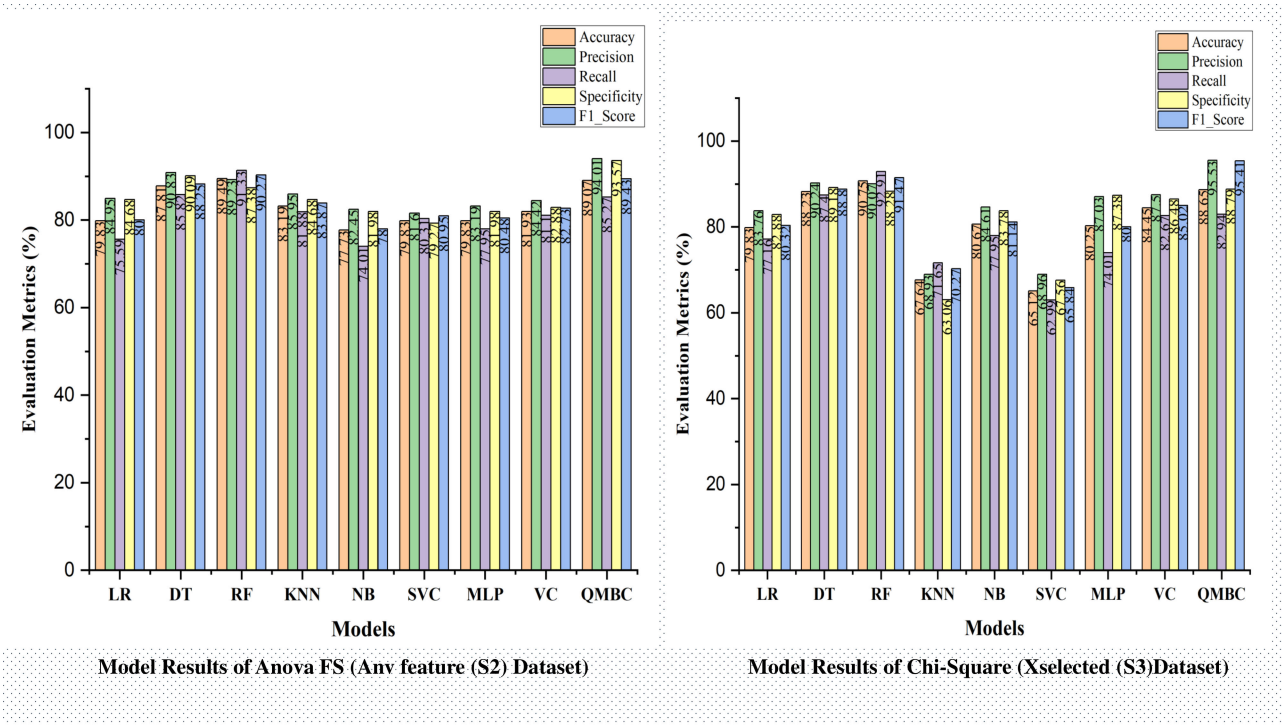


FIGURE 9. Model results of FS technique on HD dataset (Comprehensive.)

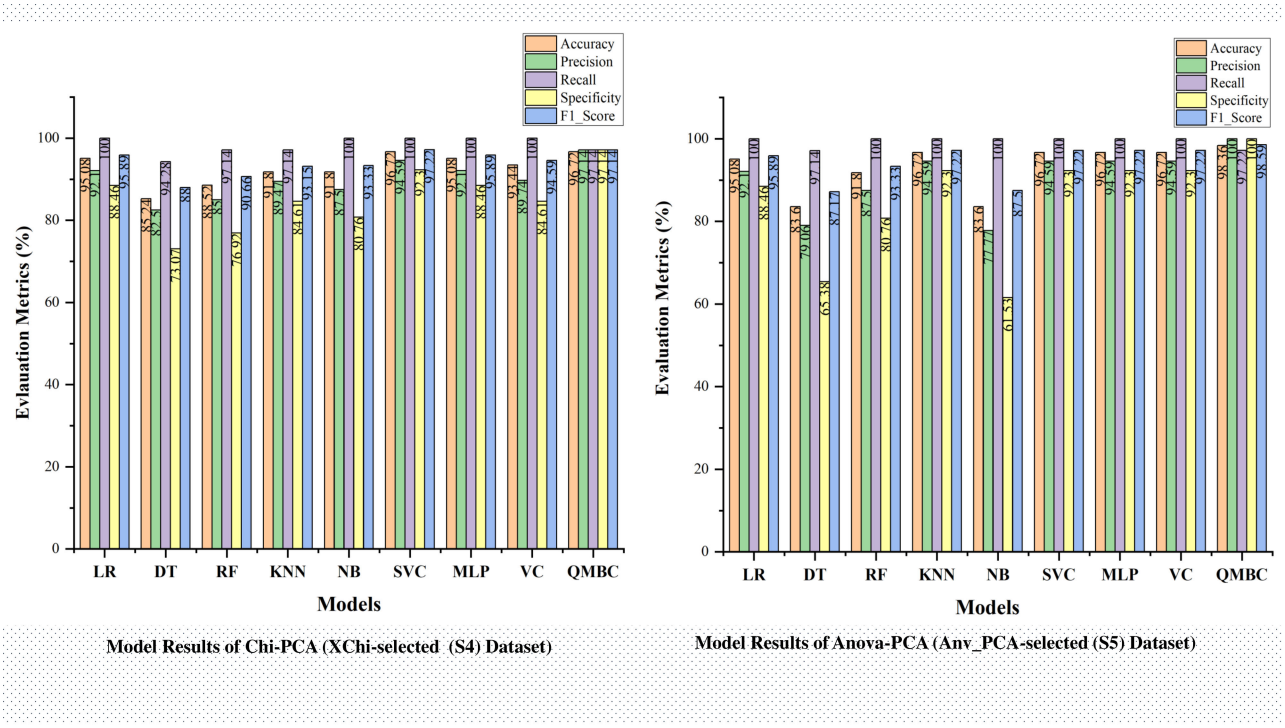


FIGURE 10. Model results of FS & FE technique on Cleveland dataset.

dataset in Fig. 10, for the CVD dataset in Fig. 11, and for the HD dataset (Comprehensive) in Fig. 12. The results on the Cleveland dataset indicate that the fusion of Anova with PCA technique by the QMBC model outperformed all state-of-the-art models, achieving an accuracy of 98.36%, precision of 100%, recall of 97.22%, specificity of 100%, and

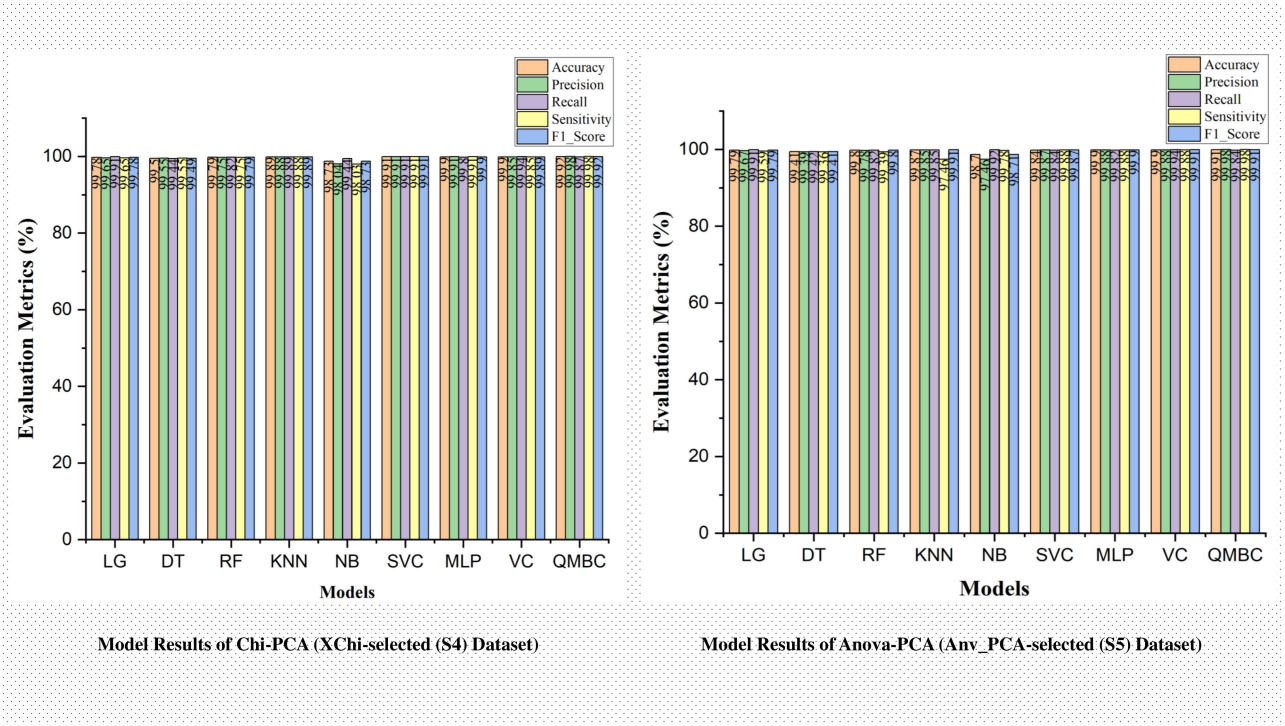


FIGURE 11. Model results of FS & FE technique on CVD dataset.

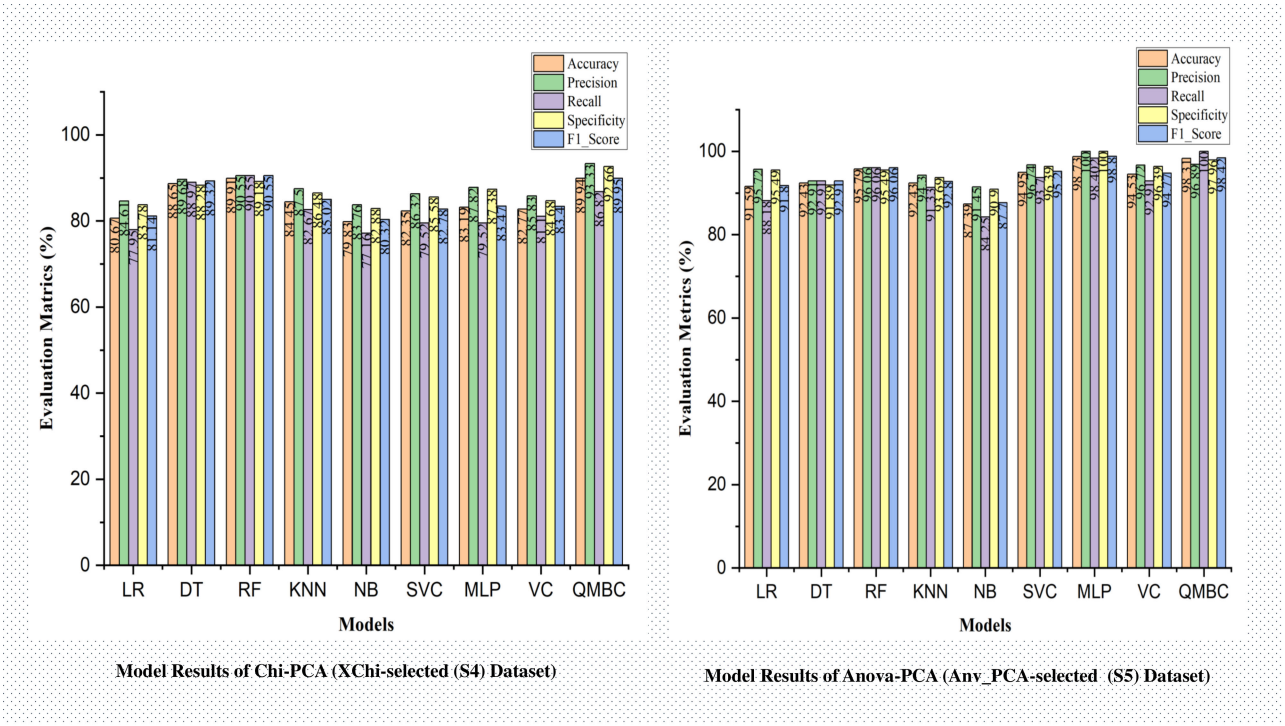
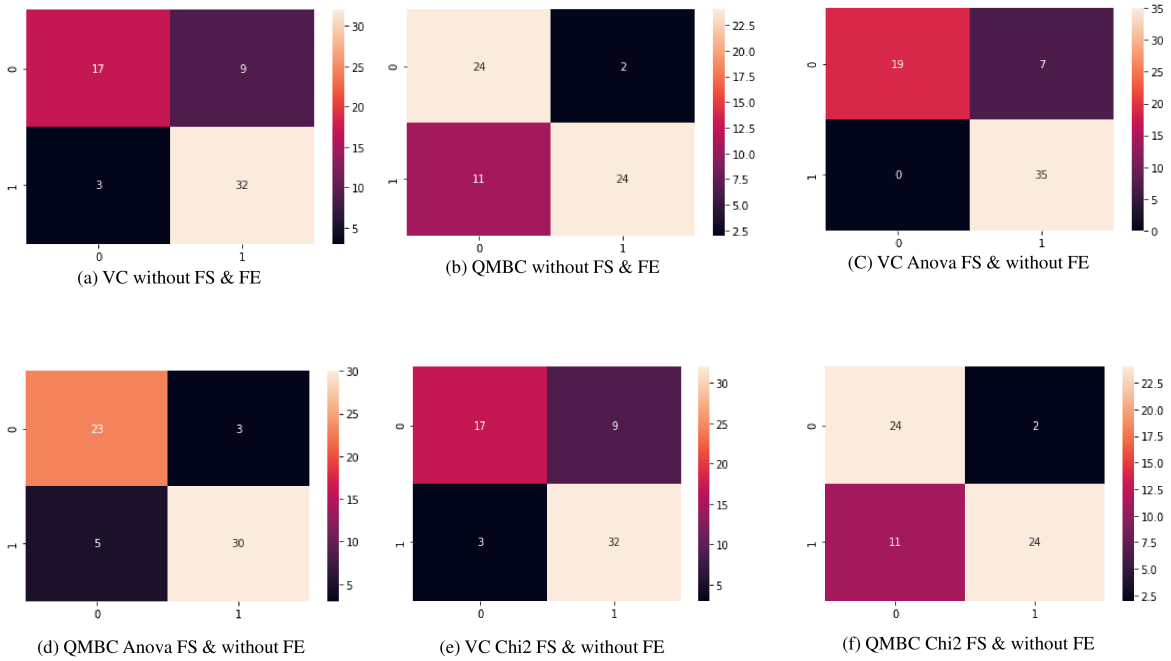


FIGURE 12. Model results of FS & FE technique on HD dataset (Comprehensive.)

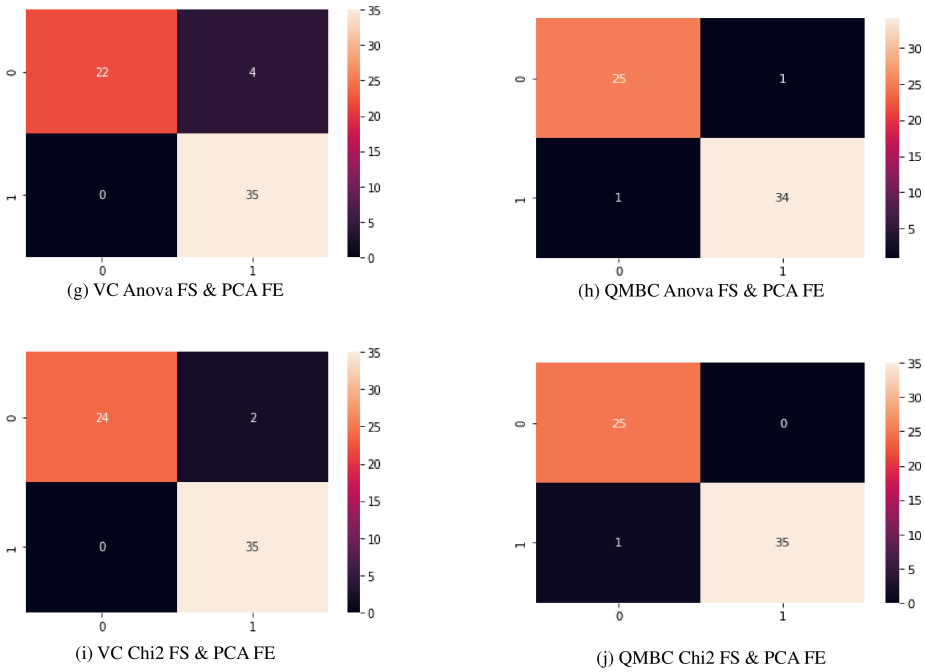
f1-score of 98.59%. Similarly, on the CVD dataset, the fusion of Chi-Square with PCA technique by the QMBC model outperformed all state-of-the-art models, with an accuracy of 99.92%, precision of 99.98%, recall of 99.87%, specificity of 99.98%, and f1-score of 99.92%. Finally, on the HD dataset (comprehensive), the fusion of Anova with PCA technique by



(a) VC without FS & FE (b) QMBC without FS & FE (C) VC Anova FS & without FE

(d) QMBC Anova FS & without FE (e) VC Chi2 FS & without FE (f) QMBC Chi2 FS & without FE

FIGURE 13. The figure presents a grid of subplots illustrating the performance of ensemble models (VC & QMBC) on the Cleveland Dataset: (a) VC model on the preprocessed (S1) dataset, (b) QMBC model on the preprocessed (S1) dataset, (c) VC model on Anv (S2) dataset without FE, (d) QMBC model on Anv (S2) dataset without FE, (e) VC model on Chi-Square (S3) dataset without FE, (f) QMBC model on Chi-Square (S3) dataset without FE.



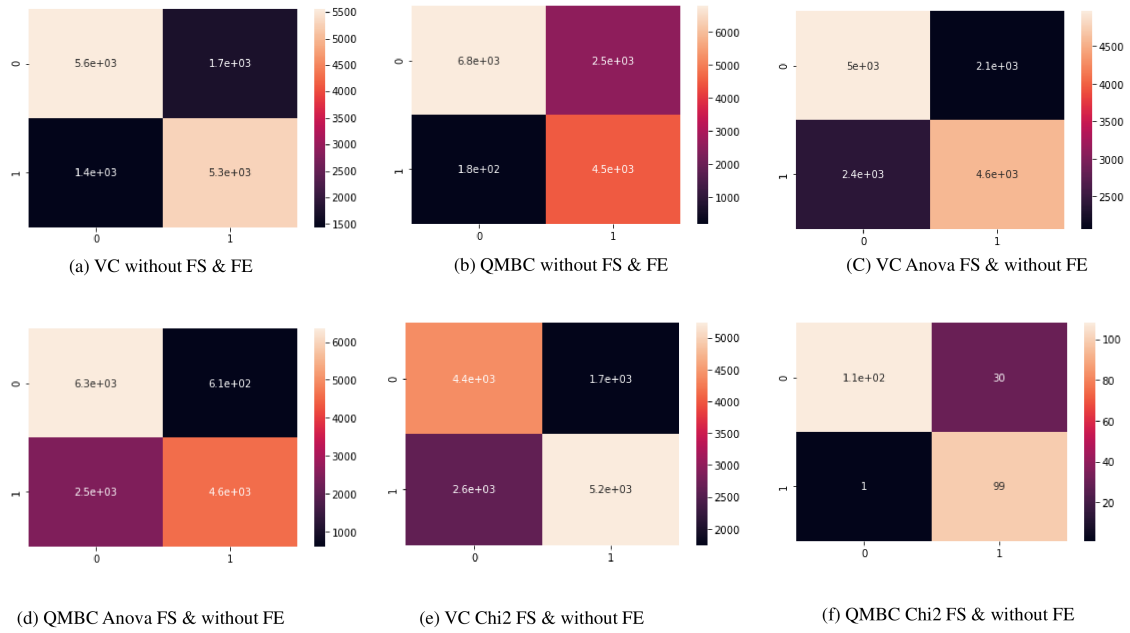
(g) VC Anova FS & PCA FE (h) QMBC Anova FS & PCA FE

(i) VC Chi2 FS & PCA FE (j) QMBC Chi2 FS & PCA FE

FIGURE 14. The figure presents a grid of subplots illustrating the performance of ensemble models (VC & QMBC) on the Cleveland Dataset: (g) VC model on Anv with PCA (S5) dataset, (h) QMBC model on Anv with PCA (S5) dataset, (i) VC model on Chi-Square with PCA (S4) dataset, and (j) QMBC model on Chi-Square with PCA (S5) dataset.

the QMBC model outperformed all state-of-the-art models, with an accuracy of 98.31%, precision of 96.89%, recall of 100%, specificity of 97.96%, and f1-score of 98.42%.

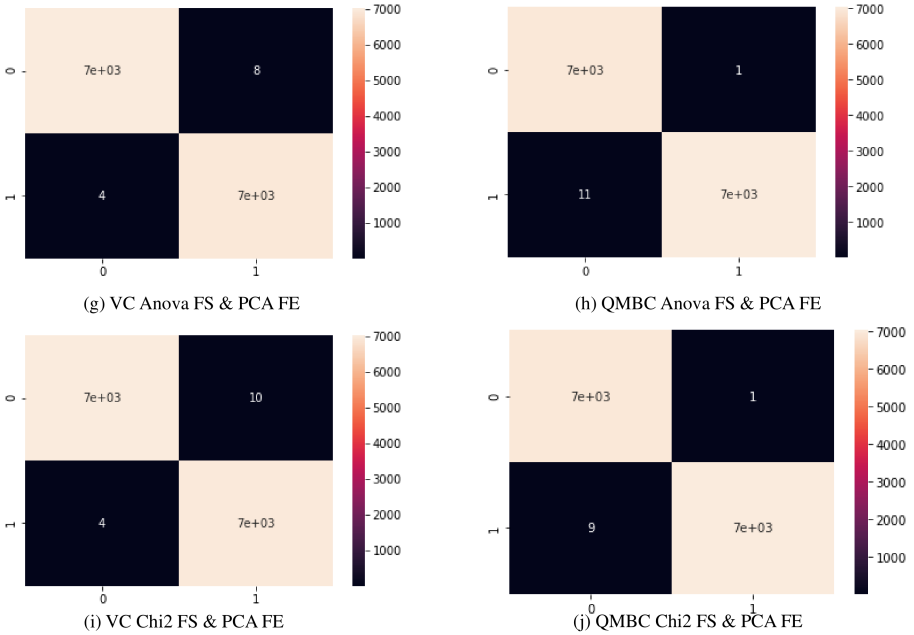
The presentation of confusion matrices for the ensemble approaches, specifically the VC and QMBC, across all the models applied to the Cleveland HD dataset, CVD HD



(a) VC without FS & FE (b) QMBC without FS & FE (C) VC Anova FS & without FE

(d) QMBC Anova FS & without FE (e) VC Chi2 FS & without FE (f) QMBC Chi2 FS & without FE

FIGURE 15. The figure presents a grid of subplots illustrating the performance of ensemble models (VC & QMBC) on the CVD Dataset: (a) VC model on the preprocessed (S1) dataset, (b) QMBC model on the preprocessed (S1) dataset, (c) VC model on Anv (S2) dataset without FE, (d) QMBC model on Anv (S2) dataset without FE, (e) VC model on Chi-Square (S3) dataset without FE, (f) QMBC model on Chi-Square (S3) dataset without FE.

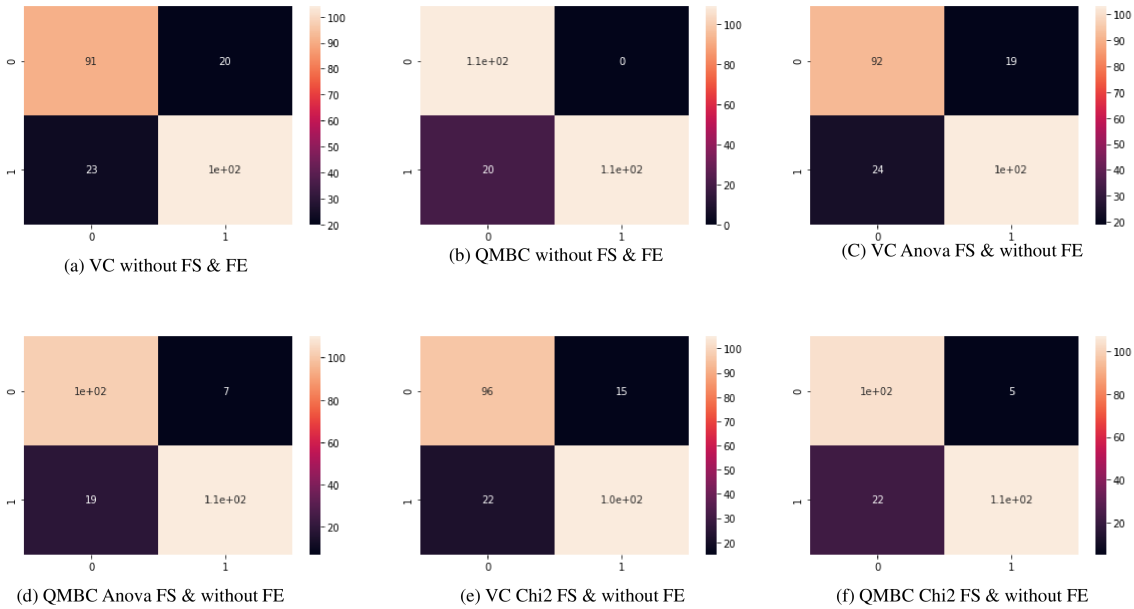


(g) VC Anova FS & PCA FE (h) QMBC Anova FS & PCA FE

(i) VC Chi2 FS & PCA FE (j) QMBC Chi2 FS & PCA FE

FIGURE 16. The figure presents a grid of subplots illustrating the performance of ensemble models (VC & QMBC) on the CVD Dataset: (g) VC model on Anv with PCA (S5) dataset, (h) QMBC model on Anv with PCA (S5) dataset, (i) VC model on Chi-Square with PCA (S4) dataset, and (j) QMBC model on Chi-Square with PCA (S5) dataset.

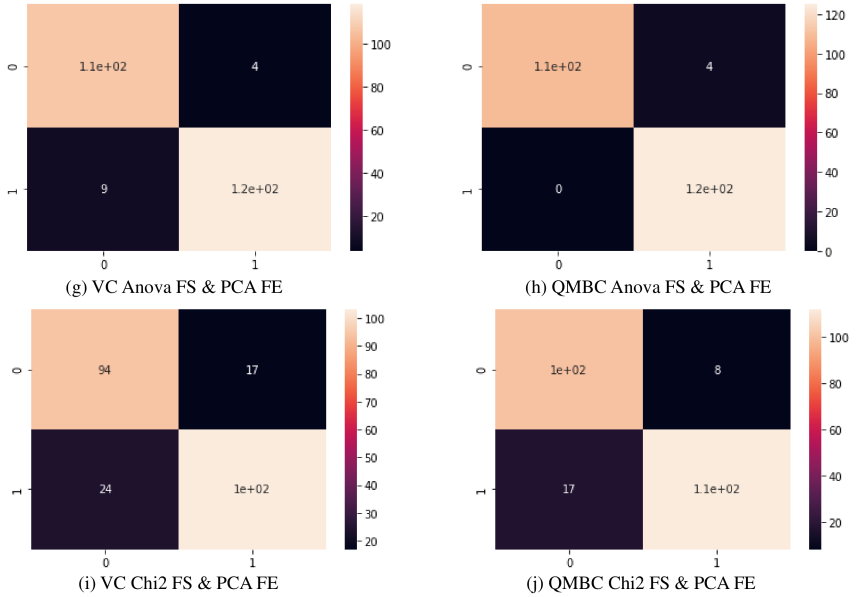
dataset, and HD (Comprehensive) dataset is included to provide a thorough evaluation of the classifier's performance and to offer a deeper insight into the results. These metrics offer insightful information on how well the classifier performs



(a) VC without FS & FE (b) QMBC without FS & FE (C) VC Anova FS & without FE

(d) QMBC Anova FS & without FE (e) VC Chi2 FS & without FE (f) QMBC Chi2 FS & without FE

FIGURE 17. The figure presents a grid of subplots illustrating the performance of ensemble models (VC & QMBC) on the HD Dataset (Comprehensive): (a) VC model on the preprocessed (S1) dataset, (b) QMBC model on the preprocessed (S1) dataset, (c) VC model on Anv (S2)dataset without FE, (d) QMBC model on Anv (S2) dataset without FE, (e) VC model on Chi-Square (S3) dataset without FE, (f) QMBC model on Chi-Square (S3) dataset without FE.



(g) VC Anova FS & PCA FE (h) QMBC Anova FS & PCA FE

(i) VC Chi2 FS & PCA FE (j) QMBC Chi2 FS & PCA FE

FIGURE 18. The figure presents a grid of subplots illustrating the performance of ensemble models (VC & QMBC) on HD Dataset (Comprehensive): (g) VC model on Anv with PCA FE (S5) dataset, (h) QMBC model on Anv with PCA (S5) dataset, (i) VC model on Chi-Square with PCA (S4) dataset, and (j) QMBC model on Chi-Square with PCA (S5) dataset.

in classifying instances correctly and detecting TP, FP, TN, and FN. The study produces reliable and thorough assessments of accuracy, precision, recall, specificity, and f1-score for each model by utilizing these confusion matrices. These performance measures are crucial indicators of the classifier's efficiency and play a key role in determining the validity and efficiency of the proposed method.

Figures. 13, 14, 15, 16, 17, and 18 present a grid of subplots illustrating the performance of ensemble models (VC & QMBC) on the Cleveland Dataset, CVD Dataset, and HD

TABLE 12. Comparison of existing approaches with proposed methodology.

Author (s)	Best Model	Year	Acc	Precision	Recall / Sensitivity	Specificity	F1-score
[11]	CFARS-AR	2015	88.3	93.3	84.9	93.3	-
[24]	Voting Classifier	2018	88.7	-	-	-	-
[18]	Random Search Algorithm (RSA) and RF model	2019	93.33	89.79	95.12	95.91	-
[17]	(FAMD) + RF	2019	93.44	96.06	89.28	96.96	-
[13]	Majority Vote with NB, BN, RF and MP	2019	85.48	-	-	-	-
[25]	SVM	2019	97.91	-	-	-	-
[26]	Hybrid RF with a Linear Model (HRFLM)	2019	88.4	90.1	92.8	82.6	90
[27]	Ensemble Model	2019	88.88	89	85	92	87
[12]	Weighted Aging Classifier Ensemble (WAE)	2020	93	96	91	-	93
[28]	Heart Disease Prediction Model (HDPM)	2020	98.4	97.14	94.67	-	95.35
[8]	FCMIM - SVM	2020	92.37	-	89	98	-
[29]	NFR + LR	2020	92.53	-	-	-	-
[30]	SVM Hyper-tuning Model using MFFSA & AFSA FS/FE	2020	82.9	90.57	88.48	90.56	-
[12]	ANN	2020	90	89	91	-	90
[31]	MaLCaDD (ML based Cardiovascular Disease Diagnosis) system	2021	95.5	-	-	-	-
[32]	RF with Hyper Parameter Optimizer	2021	97.52	97	97.29	97.7	97
[33]	CART Classification	2021	88.33	88	84.62	-	-
[46]	CNN	2021	97	97.06	96.35	-	96.7
[35]	Hybrid Ensemble Majority VC	2021	98.18	-	-	-	-
[36]	Hybrid model (RF + DT)	2021	88.7	-	-	-	-
[47]	GA - LDA with Bagging Technique	2021	93.65	89.25	96	-	-
[38]	MLP-EBMDA	2022	97.63	-	98.92	96.47	96.45
[4]	SMOTE-ENN	2022	90	-	97.3	-	92.3
[48]	SVC Hyper-tuning	2022	96.72	-	-	-	-
[39]	Weighted Score Fusion Approach	2022	95.08	95	95	-	95
[16]	XGBoost with Hyper tuning by Adopting Bayesian Optimization	2022	91.8	-	85.71	96.96	90.56
[40]	Hybrid ML model	2022	92	-	84	85	-
[49]	Ensemble Model (LR + NB)	2022	92.7	92.5	-	91.5	93.1
[50]	MLP - PSO Hybrid Algorithm	2022	84.61	80.08	88.3	-	84.4
[42]	SC	2022	92.34	92	93.49	91.07	92.74
[3]	CART Classification	2023	87.25	88.24	84.51	89.74	-
[44]	Rotation Forest Ensemble with RF	2023	97.91	97.92	97.91	97.66	97.91
[45]	Hawks Optimizer (HO) Optimizer	2023	97	98	96	-	97
[20]	SC	2023	87.3	88	88.3	-	-
Our Proposed Methodology	**Proposed QMBC (Anova with PCA) (Cleveland Dataset)**		**98.36**	**100**	**97.22**	**100**	**98.59**
	Proposed QMBC (Chi-Square with PCA) (CVD Dataset)		**99.92**	**99.98**	**99.87**	**99.98**	**99.92**
	Proposed QMBC (Anova with PCA) (HD (Comprehensive)Dataset)		**98.31**	**96.89**	**100**	**97.96**	**98.42**

(Comprehensive) Dataset. Each dataset consists of two subplots showcasing different model configurations. Subplots (a) and (b) represent the VC model and QMBC model on the preprocessed (S1) dataset without FS and FE. Subplots (c) and (d) display the performance of the VC model and QMBC model on the Anv (S2) dataset without FE. Subplots (e) and (f) depict the VC model and QMBC model on the Chi-Square (S3) dataset without FE. Subplots (g) and (h) showcase the VC model and QMBC model on the Anv with PCA (S5) dataset. Subplots (i) and (j) illustrate the VC model and QMBC model on the Chi-Square with PCA (S4) dataset. These figures provide a comprehensive evaluation of the ensemble models' performance across different configurations for each dataset, allowing for a deeper understanding of their effectiveness in classifying instances correctly and detecting TP, FP, TN, and FN.

The effectiveness of the proposed methodology is evaluated and compared with existing methodologies and state-of-the-art models in Table 12.

V. CONCLUSION

The present study investigates the performance of seven standalone ML models and a voting classifier using the Cleveland dataset, cardiovascular dataset, and HD dataset (Comprehensive). All datasets are preprocessed to ensure the suitability of ML models. To minimize dimensions, computational speed, and remove irrelevant and duplicate features from the dataset, the study utilizes Chi-Square and Anova techniques with

PCA FE strategy. Furthermore, a novel Quine McCluskey Binary Classifier (QMBC) is proposed, which ensembles seven standalone ML models to predict the presence of HD in patients.

The QMBC model with the fusion of Anova with PCA FE techniques outperformed all state-of-the-art models and existing methodologies, achieving remarkable accuracy, precision, recall, and f1-score for the Cleveland dataset, cardiovascular dataset, and HD dataset (Comprehensive) that are available in an open repository. Specifically, the QMBC model with the fusion of Anova with PCA FE techniques achieved an accuracy of 98.36%, precision of 100%, recall of 97.22%, specificity of 100%, and f1-score of 98.59% for the Cleveland dataset, an accuracy of 99.95%, precision of 100%, recall of 99.91%, specificity of 99.98%, and f1-score of 99.95% for the CVD dataset, and accuracy of 98.31%, the precision of 96.89%, recall of 100%, specificity of 97.96%, and f1-score of 98.42% for the HD dataset (Comprehensive).

In the future, the authors intend to work on imbalanced datasets and explore deep learning approaches to predict the presence of HD and ultimately save human lives.

ABBREVIATIONS

The following are the abbreviations used in this paper:

1) BC: Binary Classifier
2) CVD: Cardiovascular Disease dataset
3) DM: Data Mining
4) DT: Decision Tree
5) FS: Feature Selection
6) FE: Feature Extraction
7) KNN: K- Nearest Neighbours
8) LR: Logistic Regression
9) ML: Machine Learning
10) MLP: Multilayer Preceptor
11) NB: Naive Bayes
12) QM: Quine McCluskey Method
13) QMBC: Quine McCluskey Binary Classifier
14) RF: Random Forest
15) SVM: Support Vector Machine
16) VC: Voting Classifier

REFERENCES

[1] C. G. D. S. E. Silva, G. C. Buginga, E. A. D. S. E. Silva, R. Arena, C. R. Rouleau, S. Aggarwal, S. B. Wilton, L. Austford, T. Hauer, and J. Myers, "Prediction of mortality in coronary artery disease: Role of machine learning and maximal exercise capacity," *Mayo Clinic Proc.*, vol. 97, no. 8, pp. 1472–1482, Aug. 2022.

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

[3] M. Ozcan and S. Peker, "A classification and regression tree algorithm for heart disease modeling and prediction," *Healthcare Anal.*, vol. 3, Nov. 2023, Art. no. 100130.

[4] M. M. Nishat, F. Faisal, I. J. Ratul, A. Al-Monsur, A. M. Ar-Rafi, S. M. Nasrullah, M. T. Reza, and M. R. H. Khan, "A comprehensive investigation of the performances of different machine learning classifiers with SMOTE-ENN oversampling technique and hyperparameter optimization for imbalanced heart failure dataset," *Sci. Program.*, vol. 2022, pp. 1–17, Mar. 2022.

[5] P. Ghosh, S. Azam, M. Jonkman, A. Karim, F. M. J. M. Shamrat, E. Ignatious, S. Shultana, A. R. Beeravolu, and F. De Boer, "Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques," *IEEE Access*, vol. 9, pp. 19304–19326, 2021.

[6] A. K. Gárate-Escamila, A. Hajjam El Hassani, and E. Andrès, "Classification models for heart disease prediction using feature selection and PCA," *Informat. Med. Unlocked*, vol. 19, 2020, Art. no. 100330.

[7] S. Bashir, Z. S. Khan, F. H. Khan, A. Anjum, and K. Bashir, "Improving heart disease prediction using feature selection approaches," in *Proc. 16th Int. Bhurban Conf. Appl. Sci. Technol. (IBCAST)*, Jan. 2019, pp. 619–623.

[8] J. P. Li, A. U. Haq, S. U. Din, J. Khan, A. Khan, and A. Saboor, "Heart disease identification method using machine learning classification in e-healthcare," *IEEE Access*, vol. 8, pp. 107562–107582, 2020.

[9] M. Ayar, A. Isazadeh, F. S. Gharehchopogh, and M. Seyedi, "Chaotic-based divide-and-conquer feature selection method and its application in cardiac arrhythmia classification," *J. Supercomput.*, vol. 78, pp. 5856–5882, Mar. 2022.

[10] S. Shilaskar and A. Ghatol, "Feature selection for medical diagnosis: Evaluation for cardiovascular diseases," *Expert Syst. Appl.*, vol. 40, no. 10, pp. 4146–4153, Aug. 2013.

[11] N. C. Long, P. Meesad, and H. Unger, "A highly accurate firefly based algorithm for heart disease prediction," *Expert Syst. Appl.*, vol. 42, no. 21, pp. 8221–8231, Nov. 2015.

[12] I. D. Mienye, Y. Sun, and Z. Wang, "Improved sparse autoencoder based artificial neural network approach for prediction of heart disease," *Informat. Med. Unlocked*, vol. 18, Jan. 2020, Art. no. 100307.

[13] C. B. C. Latha and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," *Informat. Med. Unlocked*, vol. 16, Jan. 2019, Art. no. 100203.

[14] R. K. Sevakula and N. K. Verma, "Assessing generalization ability of majority vote point classifiers," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 12, pp. 2985–2997, Dec. 2017.

[15] H. Li, Y. Cui, Y. Liu, W. Li, Y. Shi, C. Fang, H. Li, T. Gao, L. Hu, and Y. Lu, "Ensemble learning for overall power conversion efficiency of the all-organic dye-sensitized solar cells," *IEEE Access*, vol. 6, pp. 34118–34126, 2018.

[16] K. Budholiya, S. K. Shrivastava, and V. Sharma, "An optimized XGBoost based diagnostic system for effective prediction of heart disease," *J. King Saud Univ., Comput. Inf. Sci.*, vol. 34, no. 7, pp. 4514–4523, Jul. 2022.

[17] A. Gupta, R. Kumar, H. S. Arora, and B. Raman, "MIFH: A machine intelligence framework for heart disease diagnosis," *IEEE Access*, vol. 8, pp. 14659–14674, 2020.

[18] A. Javeed, S. Zhou, L. Yongjian, I. Qasim, A. Noor, and R. Nour, "An intelligent learning system based on random search algorithm and optimized random forest model for improved heart disease detection," *IEEE Access*, vol. 7, pp. 180235–180243, 2019.

[19] D. Velusamy and K. Ramasamy, "Ensemble of heterogeneous classifiers for diagnosis and prediction of coronary artery disease with reduced feature subset," *Comput. Methods Programs Biomed.*, vol. 198, Jan. 2021, Art. no. 105770.

[20] 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.

[21] D. Dua and C. Graff. (2017). *UCI Machine Learning Repository*. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Heart+Disease

[22] UCI Machine Learning Repository. (2018). *Heart Disease Dataset (Comprehensive)*. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Heart+Disease

[23] Kaggle. (2019). *Cardiovascular Disease Dataset*. [Online]. Available: https://www.kaggle.com/sulianova/cardiovascular-disease-dataset

[24] M. S. Amin, Y. K. Chiam, and K. D. Varathan, "Identification of significant features and data mining techniques in predicting heart disease," *Telematics Informat.*, vol. 36, pp. 82–93, Mar. 2019.

[25] D. B. Mehta and N. C. Varnagar, "Newfangled approach for early detection and prevention of ischemic heart disease using data mining," in *Proc. 3rd Int. Conf. Trends Electron. Informat. (ICOEI)*, Apr. 2019, pp. 1158–1162.

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

[27] K. Raza, "Improving the prediction accuracy of heart disease with ensemble learning and majority voting rule," in *U-Healthcare Monitoring Systems*. Amsterdam, The Netherlands: Elsevier, 2019, pp. 179–196.

[28] N. L. Fitriyani, M. Syafrudin, G. Alfian, and J. Rhee, "HDPM: An effective heart disease prediction model for a clinical decision support system," *IEEE Access*, vol. 8, pp. 133034–133050, 2020.
[29] S. J. Pasha and E. S. Mohamed, "Novel feature reduction (NFR) model with machine learning and data mining algorithms for effective disease risk prediction," *IEEE Access*, vol. 8, pp. 184087–184108, 2020.
[30] S. M. S. Shah, F. A. Shah, S. A. Hussain, and S. Batool, "Support vector machines-based heart disease diagnosis using feature subset, wrapping selection and extraction methods," *Comput. Electr. Eng.*, vol. 84, Jun. 2020, Art. no. 106628.
[31] A. Rahim, Y. Rasheed, F. Azam, M. W. Anwar, M. A. Rahim, and A. W. Muzaffar, "An integrated machine learning framework for effective prediction of cardiovascular diseases," *IEEE Access*, vol. 9, pp. 106575–106588, 2021.
[32] R. Valarmathi and T. Sheela, "Heart disease prediction using hyper parameter optimization (HPO) tuning," *Biomed. Signal Process. Control*, vol. 70, Sep. 2021, Art. no. 103033.
[33] E. Miranda, M. Aryuni, C. Bernando, and A. Hartanto, "Application for early heart disease prediction based on data mining approach," in *Proc. 4th Int. Conf. Comput. Informat. Eng. (IC2IE)*, Sep. 2021, pp. 375–378.
[34] M. B. Kursa and W. R. Rudnicki, "Feature selection with the Boruta package," *J. Stat. Softw.*, vol. 36, no. 11, pp. 1–13, 2010.
[35] S. E. A. Ashri, M. M. El-Gayar, and E. M. El-Daydamony, "HDPF: Heart disease prediction framework based on hybrid classifiers and genetic algorithm," *IEEE Access*, vol. 9, pp. 146797–146809, 2021.
[36] 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.
[37] M. S. Nawaz, B. Shoaib, and M. A. Ashraf, "Intelligent cardiovascular disease prediction empowered with gradient descent optimization," *Heliyon*, vol. 7, no. 5, May 2021, Art. no. e06948.
[38] D. Deepika and N. Balaji, "Effective heart disease prediction using novel MLP-EBMDA approach," *Biomed. Signal Process. Control*, vol. 72, Feb. 2022, Art. no. 103318.
[39] H. B. Kibria and A. Matin, "The severity prediction of the binary and multi-class cardiovascular disease—A machine learning-based fusion approach," *Comput. Biol. Chem.*, vol. 98, Jun. 2022, Art. no. 107672.
[40] K. S. Archana, B. Sivakumar, R. Kuppusamy, Y. Teekaraman, and A. Radhakrishnan, "Automated cardioailment identification and prevention by hybrid machine learning models," *Comput. Math. Methods Med.*, vol. 2022, pp. 1–8, Feb. 2022.
[41] A. A. Nancy, D. Ravindran, P. M. D. R. Vincent, K. Srinivasan, and D. G. Reina, "IoT-cloud-based smart healthcare monitoring system for heart disease prediction via deep learning," *Electronics*, vol. 11, no. 15, p. 2292, Jul. 2022.
[42] A. Tiwari, A. Chugh, and A. Sharma, "Ensemble framework for cardiovascular disease prediction," *Comput. Biol. Med.*, vol. 146, Jul. 2022, Art. no. 105624.
[43] V. Chaurasia and A. Chaurasia, "Novel method of characterization of heart disease prediction using sequential feature selection-based ensemble technique," *Biomed. Mater. Devices*, pp. 1–10, Jan. 2023.
[44] K. V. V. Reddy, I. Elamvazuthi, A. A. Aziz, S. Paramasivam, H. N. Chua, and S. Pranavanand, "An efficient prediction system for coronary heart disease risk using selected principal components and hyperparameter optimization," *Appl. Sci.*, vol. 13, no. 1, p. 118, Dec. 2022.
[45] A. S. Kumar and R. Rekha, "An improved hawks optimizer based learning algorithms for cardiovascular disease prediction," *Biomed. Signal Process. Control*, vol. 81, Mar. 2023, Art. no. 104442.
[46] A. Mehmood, M. Iqbal, Z. Mehmood, A. Irtaza, M. Nawaz, T. Nazir, and M. Masood, "Prediction of heart disease using deep convolutional neural networks," *Arabian J. Sci. Eng.*, vol. 46, no. 4, pp. 3409–3422, Apr. 2021.
[47] V. Jothi Prakash and N. K. Karthikeyan, "Enhanced evolutionary feature selection and ensemble method for cardiovascular disease prediction," *Interdiscipl. Sci., Comput. Life Sci.*, vol. 13, no. 3, pp. 389–412, Sep. 2021.
[48] K. G. Dinesh, K. Arumugaraj, K. D. Santhosh, and V. Mareeswari, "Prediction of cardiovascular disease using machine learning algorithms," in *Proc. Int. Conf. Current Trends Towards Converging Technol. (ICCTCT)*, Mar. 2018, pp. 1–7.
[49] R. Rajendran and A. Karthi, "Heart disease prediction using entropy based feature engineering and ensembling of machine learning classifiers," *Expert Syst. Appl.*, vol. 207, Nov. 2022, Art. no. 117882.
[50] A. Al Bataineh and S. Manacek, "MLP-PSO hybrid algorithm for heart disease prediction," *J. Pers. Med.*, vol. 12, no. 8, p. 1208, Jul. 2022.

RAMDAS KAPILA received the B.Tech. degree in computer science and information technology from JNTUH, Hyderabad, India, in 2008, and the M.Tech. degree in computer science and engineering from JNTUK, Andhra Pradesh, India, in 2015. He is currently pursuing the Ph.D. degree in computer science and engineering with SRM University AP, Amaravati, India. His research interests include machine learning, data mining, and deep learning.

THIRUMALAISAMY RAGUNATHAN (Member, IEEE) received the Ph.D. degree in computer science and engineering from IIIT Hyderabad, Hyderabad. He has 27 years of teaching experience and 17 years of research experience. He is currently the Dean of the Faculty of Engineering and Technology, Sri Ramachandra Institute of Higher Education and Research. He conducts research activities with regard to the development of a fast distributed file system for the efficient storage and retrieval of large data. He also conducts research on the task scheduling and load balancing issues in cloud computing systems and developing a health information system for disease identification and treatment plan generation.

SUMALATHA SALETI received the Ph.D. degree from the National Institute of Technology, Warangal, India, in 2020. She is currently an Assistant Professor with the Department of Computer Science and Engineering, SRM University AP, Amaravathi, Andhra Pradesh, India. Her research interests include big data, data mining, and pattern discovery.

T. JAYA LAKSHMI (Member, IEEE) received the Ph.D. degree from the School of Computer and Information Sciences, University of Hyderabad, India, in 2019. She is currently an Assistant Professor with the Department of Computer Science and Engineering, SRM University AP, Amaravathi, Andhra Pradesh, India. Her Ph.D. work on "link prediction in heterogeneous social networks." She is a reviewer of reputed international journals. She has an overall teaching experience of 22 years. Her research interests include graph mining, recommender systems, natural language processing, and security analytics.

MOHD WAZIH AHMAD is currently an Assistant Professor with the Department of Computer Science and Engineering, Adama Science and Technology University, Adama, Ethiopia. His current research interests include machine learning, the Internet of Things, information retrieval and soft computing applications in agriculture, health, and other areas. He has supervised more than 20 post graduate thesis and he is the Leader of Intelligent Systems SIG, ASTU Campus.