IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 December 2023, accepted 13 January 2024, date of publication 30 January 2024, date of current version 5 February 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3359910*

RESEARCH ARTICLE

# Machine Learning-Based Cardiovascular Disease Detection Using Optimal Feature Selection

**TAHSEEN ULLAH[1], SYED IRFAN ULLAH[1], KHALIL ULLAH[2], MUHAMMAD ISHAQ[3], AHMAD KHAN[4], YAZEED YASIN GHADI[5], AND ABDULMOHSEN ALGARNI[6]**

[1]Department of Computing and Technology, Abasyn University, Peshawar 25000, Pakistan
[2]Department Software Engineering, University of Malakand, Chakdara 18800, Pakistan
[3]Department of Computer Science, Helping Hand Institute of Rehabilitation Sciences, Mansehra 22800, Pakistan
[4]Department of Software Engineering, Mirpur University of Science and Technology, Mirpur, Azad Jammu Kashmir 10250, Pakistan
[5]Department of Computer Science, Al Ain University, Al Ain, United Arab Emirates
[6]Department of Computer Science, King Khalid University, Abha 61421, Saudi Arabia

Corresponding author: Khalil Ullah (Khalil.ullah@uom.edu.pk)

**ABSTRACT** Cardiovascular disease (CVD) is a prevalent and serious condition causing a significant global mortality rate. According to the World Health Organization (WHO), in 2022, CVD claimed the lives of approximately 19.1 million people, accounting for 33% of global fatalities. ECG is widely used for automatic detection of CVD using traditional machine learning; however, it is usually difficult to select optimal features. Addressing this issue, a scalable machine learning-based architecture is proposed for early CVD detection based optimal feature selection. This architecture aims to revolutionize healthcare by enabling timely diagnosis and treatment, reducing CVD-related fatalities. Comprising data collection, storage, and processing components, the system employs machine learning classifiers to predict patients' heart conditions. Initially features are extracted from ECG signals then feature selection techniques like FCBF, MrMr, and relief, along with PSO-optimization are used to select optimal features. Extra Tree and Random Forest classifiers trained on the selected features have achieved notable performance rates with accuracy of 100% respectively. Furthermore, a comparison of the proposed method with state of the art on both small and large dataset is provided. The proposed system holds potential to revolutionize patient care and substantially lower CVD-related mortality, enhancing the quality of life for affected individuals. In summary, this architecture offers a promising solution to the pressing issue of CVD and paves the way for advanced healthcare systems.

**INDEX TERMS** Cardiovascular disease, feature selection, optimization, machine learning.

## I. INTRODUCTION

One of the most serious problems in the world today is public health. The World Health Organization (WHO) claims that the pursuit of health is a basic human right. Several epidemic diseases are currently posing a threat to the world's population, causing a fatal outcome. Consequently, chronic illnesses (CDs) impact a sizable percentage of the population and account for a considerable share of overall mortality. CDs are not only incurable but also have a far longer half-life in the body than diseases including cancer, diabetes, stroke, Parkinson's disease, and cardiovascular disease (CVD). Unhealthy eating habits, smoking cigarettes, drinking too much alcohol, and living in general are major contributors to these disorders. Half of all Americans suffer from at least one chronic health condition, and more than 80% of the population has financial concerns related to healthcare [1].

Improved lifestyle choices have a direct impact on reducing the prevalence of chronic diseases. Among nations, the United States (US) bears the greatest burden of chronic diseases. The US allocates approximately $2.70 trillion each year for the treatment of these diseases, which constitutes 18.0% of the nation's gross domestic product. In particular, cardiovascular disease (CVD) stands out as the primary cause of death in the Americas [2].

The associate editor coordinating the review of this manuscript and approving it for publication was Md. Kafiul Islam.

In a comparable way, other nations around the world are dealing with the issues associated with CVD. According to studies that were just recently made public, chronic diseases are responsible for 86.5% of deaths in China [3].

Cardiovascular diseases (CVD) have emerged as a major global cause of mortality, claiming a substantial number of lives annually. The underlying pathology of cardiac diseases is the inability of the heart to effectively circulate an adequate amount of blood to various organs. This condition poses a significant threat to life and is recognized as one of the most lethal and life-threatening chronic diseases worldwide. By affecting the heart or blood vessels, CVD disrupts the normal supply of blood, impairing the proper functioning of essential body organs [4].

Cardiovascular disease (CVD) is a leading cause of mortality worldwide, affecting both developed and developing nations. According to the World Health Organization (WHO), in 2022, CVD claimed the lives of approximately 19.1 million people, accounting for 33% of global fatalities. In the United States alone, CVD causes the death of around 647,000 individuals each year. Similarly, in Pakistan, CVD claims the lives of approximately 200,000 people annually, with mortality rates on the rise. The European Society of Cardiology (ESC) estimates that 26.5 million Europeans are currently living with CVD, and each year, 3.8 million new cases are diagnosed. Shockingly, 50% to 55% of CVD patients do not survive beyond a year, placing a significant burden on healthcare systems. Moreover, approximately 4% of healthcare spending is allocated to the treatment of CVD patients [5].

The symptoms associated with CVD encompass physical fragility, shortness of breath, inflammation in the feet, lethargy, and other related manifestations [6].

Cardiovascular disease (CVD) is a significant global health issue that can be attributed to various risk factors including hypertension, high cholesterol levels, smoking, sedentary lifestyle, and obesity. CVD encompasses a range of conditions such as congenital heart disease, congestive heart failure, and cardiac arrhythmias. Traditional approaches to predicting and diagnosing CVD were complex and often led to complications that impacted individuals' overall well-being. This disease remains the leading cause of mortality in both developed and developing countries, necessitating effective preventive and diagnostic measures [7].

In developing countries, clinicians face challenges in accurately diagnosing and treating cardiovascular disease (CVD) due to limited resources. Computer technology and machine learning have been introduced as aids in clinical decision-making, enabling early detection and assessment of CVD risk. Medical data mining technologies can extract useful information from the massive amounts of data in healthcare, which is vital due to the complexity of medical data. Our technology for CVD prediction could potentially save millions of lives by enabling more people to receive treatment faster [8].

## II. TRADITIONAL RISK ASSESSMENT SYSTEMS

Traditional risk assessment systems have certain limitations in terms of their accuracy and efficiency, despite the fact that early detection and prompt intervention can considerably improve patient outcomes. The application of machine learning algorithms has shown some encouraging results in predicting the risk of cardiovascular disease, and these algorithms have the potential to improve clinical decision-making and individualized care. However, there is a dearth of sufficient knowledge, empirical data, and competent research studies in this field, underscoring the need for additional exploration into the topic. As a result of addressing this research gap, the proposed project hopes to make a contribution to the creation of prediction models for cardiovascular disease that are more accurate and efficient, which would eventually improve the outcomes for patients and reduce the costs of healthcare.

In a previously carried out study, MrMr, FCBF, LASSO, and Relief were utilized to identify characteristics. However, the outcomes were neither satisfactory nor of the required quality. In addition to the characteristics listed above, we used ANOVA with join combinations. The authors predicted cardiovascular disease (CVD) with the help of a classifier, however, in the current research work, we optimized the findings of the model with the help of an optimization approach called particle swarm optimization (PSO), as described in the methodological section of this research study. This will save a lot of effort and produce empirical data to predict cardiovascular disease (CVD), and we utilized four different machine learning techniques to do this.

## III. RELATED WORK

Accurate prediction of cardiac issues is crucial for providing optimal care to patients. Machine learning (ML) techniques offer a promising approach to gaining a deeper understanding of heart disease symptoms and improving treatment strategies. In our study, we evaluated six ML models using a dataset comprising 74 features. The results demonstrated high accuracy rates of 98.7%, 99.0%, and 99.4% for the Cleveland, Hungarian, and Cleveland-Hungarian (CH) datasets, respectively, using the combined approach of chi-square and principal component analysis (CHI-PCA) with random forests (RF). Through the analysis, we identified significant features such as cholesterol levels, maximum heart rate, chest discomfort, ST depression characteristics, and coronary artery structure using the Chi-Square Selector. Our experiments revealed the powerful combination of chi-square and PCA in predicting cardiac issues [1]. Although this work achieves high performance however it can't be generalized as the model is applied to a limited dataset.

Deep learning-based ECG signal classification. Stacked de-noising autoencoders (SDAEs) with sparsity constraints extracted meaningful features from raw ECG data. A DNN was created using these features and a soft-max regression layer. Experts labeled the most relevant and uncertain ECG beats during interaction to update network weights. DNN

posterior probabilities and confidence measures classified ECG beats. Multiple databases showed improved accuracy, reduced expert interaction, and faster online retraining. The proposed method may improve ECG classification and cardiac disease diagnosis [2]. The performance of this model is good however their main focus is on use of Python language in heart disease detection.

A recent study demonstrated the benefits of automated classification methods in aiding physicians' treatment decisions for cardiac arrhythmias. The study focused on utilizing probabilistic n-grams to classify these arrhythmias and compared the performance of five unsupervised dimensionality reduction (DR) methods: principal component analysis (PCA), fast independent component analysis (fastICA) with tangential, kurtosis, and Gaussian contrast functions, kernel PCA (KPCA) with polynomial kernel, hierarchical nonlinear PCA (hNLPCA), and principal polynomial analysis (PPA). Notably, employing the fastICA DR algorithm with a tangential contrast function on at least 10 dimensions, along with a PNN classifier at a spread parameter of 0.4, resulted in a significant improvement in F score (99.83%). However, it is important to note that the calculation of low-dimensional mapping using hNLPCA or KPCA is more time-consuming. Additionally, PPA demonstrated a 10% higher effectiveness compared to PCA. These findings highlight the potential of utilizing specific DR methods in conjunction with a PNN classifier for accurate cardiac arrhythmia classification [3]. The study was conducted on a relatively small dataset of 100 ECG recordings. This limits the generalizability of the findings. The study only evaluated five dimensionality reduction methods and one classifier. It is possible that other methods and classifiers could achieve even better performance.

Cardiac Resynchronization Therapy (CRT) is a rhythm treatment for people with heart failure that has been used for a long time. People often use the New York Heart Association (NYHA) rating to figure out how well a patient is responding to CRT. Finding a patient's NYHA class regularly over time in an electronic health record (EHR) can help doctors learn more about how heart failure gets worse and how well CRT works. But NYHA is rarely kept as organized data in an EHR. Instead, this kind of information is often written down in unstructured clinical notes. They investigated the feasibility of using NLP to categorize NYHA from clinical notes. The authors analyzed 6,174 hospital-specific clinical records that had been linked to NYHA class diagnostic numbers. The results of machine learning-based methods were comparable to those of rule-based ones. Support vector machines using n-gram features performed best (93.0% F-measure). The study does not provide details of the feature selection method also it is trained and validated on a small dataset [4].

The accurate prediction and diagnosis of heart disease has become a critical challenge for healthcare systems worldwide. To reduce the mortality rate associated with heart diseases, it is crucial to develop a fast and efficient method of detection. Data mining techniques and machine learning methods play a vital role in this domain. Researchers are actively engaged in accelerating their efforts to develop software using machine learning algorithms that can assist doctors in making accurate predictions and diagnoses of heart disease. The primary objective of this study was to leverage machine learning techniques to determine the presence of heart disease in patients. Graphical representations of data were employed to evaluate the performance of various machine learning algorithms [5]. The article presents a significant overview of the articles in ML models for CVD however, it is limited to the work done until 2021.

To identify coronary artery disease (CAD) patients the HRV data from a standard library and self-recorded data from healthy people is used. The HRV time series was broken up into four levels, and 62 features were taken out of it using nonlinear methods. Numerical tests were done, and the suggested method of using the ten most important entropy features got close to a 100% detection accuracy. HRV signals can be used to find and study CAD patients by breaking them up into subspaces level 4 and level 3 [6]. This method can't be generalized to all the cases of heart diseases as it is applied to a very small dataset.

The study in [7] aimed to improve fuzzy clustering methods by automatically calculating feature weights and eliminating irrelevant feature components. They proposed a feature-reduction FCM (FRFCM) approach, which utilized a learning schema to optimize parameters and reduce irrelevant features. FRFCM outperformed existing feature-weighted FCM algorithms, as demonstrated through experimental results and comparisons on numerical and real datasets. The findings highlight the efficacy and practicality of FRFCM in enhancing fuzzy clustering techniques. Overall, the study provides some promising results on the use of machine learning for early and accurate heart disease prediction. However, it also has the overfitting problem and is trained on features which are not ranked for optimality.

A pre-processing technique is used to improve ECG signal classification accuracy by removing noise from raw data. Classifiers (KNN, Naive Bayes, and Decision Tree) are tested for detecting normal and abnormal heart rhythms, with the Decision Tree performing best. This technique proves effective in accurately diagnosing heart-related diseases. Machine learning is applied to predict and understand heart disease symptoms, using feature selection to reduce dimensionality. The CHI-PCA with RF approach achieves high accuracies on different datasets, identifying relevant features. Chi-square with PCA outperforms other classifiers, while PCA with raw data yields poorer results The study's drawbacks include its limited dataset size, a lack of investigation into feature selection, and a lack of comparison with state-of-the-art approaches. Ethical considerations in healthcare research are also not thoroughly explored in the work [9].

Heart disease has been a subject of significant research due to its detrimental impact on human health. It is a leading cause of death in the United States. Data mining has emerged as a crucial technique for analyzing healthcare data, enabling

easier interpretation of medical records. The paper focuses on predicting heart disease using supervised machine learning algorithms like support vector machines, k-nearest neighbors, and naive Bayes. The implementation of these algorithms is carried out using the R programming language. The accuracy metric is utilized to assess the performance of the algorithms, and the findings of the analysis are discussed however insufficient discussion on the model's generalizability and ethical considerations in heart disease prediction [10]. The method is validated on relatively small dataset thus can't be generalized more over use of different of optimization techniques are not analyzed. Accurate prediction of heart disease is crucial for saving lives, as misdiagnosis can have severe consequences. This research focuses on analyzing the UCI Machine Learning Heart Disease dataset using a range of machine learning and deep learning techniques, comparing the obtained results and analysis. The dataset consists of 14 key features, which are thoroughly examined. Evaluation is performed using accuracy metrics and confusion matrix, confirming promising outcomes. To enhance accuracy, irrelevant features are eliminated using Isolation Forest, and data standardization techniques are applied. The integration of this research with multimedia technology, particularly mobile devices, is also explored. By leveraging deep learning, an impressive accuracy of 94.2% is achieved [11]. The researchers aimed to predict the likelihood of heart disease based on medical criteria. They employed machine learning techniques such as logistic regression and K-Nearest Neighbors to classify individuals with cardiovascular disease. The proposed model demonstrated high flexibility and outperformed previous classifiers. It effectively alleviated concerns by accurately predicting the likelihood of identifying heart disease however Ambitious plans for the future don't have conversations about whether they are possible, and MICE's claim to be the best algorithm doesn't have any comparisons. It is imperative to consider these characteristics in order to improve the system's dependability in practical medical situations [12].

Using machine learning algorithms and Python programming to identify heart disease. Heart disease has become a common and dangerous disease in the last few decades. It is caused by fat. People get this disease when their bodies have too much pressure. The authors looked at a dataset with 13 attributes and 270 individual data points to study how well patients did. The main goal of the paper is to get better at detecting heart disease using algorithms whose aim output is a count of whether or not a person has heart disease but the investigation of alternative data mining methods was constrained, and there was a lack of comprehensive examination about the computing efficiency and interpretability of the selected algorithms [13].

In another study, the researchers developed an intelligent medical system utilizing machine learning to detect heart issues and aid in accurate diagnoses. They addressed information gaps between the Framingham dataset and the publicly available UCI Heart Disease dataset. By applying machine learning techniques, they sought to identify the most effective approach for detecting cardiovascular diseases. The system's performance was evaluated using accuracy, sensitivity, F-measure, and precision, highlighting the superiority of the proposed strategy compared to similar models but One problem with this survey paper is that it doesn't go into enough detail about the techniques and algorithms it looks at. It also only gives a general outline of automation in cardiovascular disease prediction without looking into each method in more detail [14]. A dataset from India was utilized to diagnose heart disease, and the performance of an automatic diagnosis system was evaluated based on classification accuracy, sensitivity, and specificity. The findings indicated that the Sequential Minimization Optimization (SMO) learning method in Support Vector Machines (SVM) outperformed other approaches for medical disease diagnosis applications [15].

Classification, data mining, machine learning, and deep learning algorithms for predicting cardiovascular diseases are compared and reported on. The survey is broken down into three sections: cardiovascular disease classification and data mining techniques, cardiovascular disease prediction using machine learning and deep learning models. In addition to collecting and reporting the accuracy metrics, dataset, and instruments used for prediction and classification, the survey also compiles and publishes the performance metrics utilized for reporting the accuracy however ignoring overfitting and high-dimensional data issues, the evaluation recommends wide research without specific solutions to cardiovascular disease prediction problems [16].

An effective machine learning system for heart disease diagnosis uses Support Vector Machine, Logistic Regression, Artificial Neural Network, K-nearest neighbors, Naïve Bayes, and Decision Tree. The study introduces Relief, Minimal Redundancy Maximal Relevance, Least Absolute Shrinkage Selection Operator, Local Learning, and a novel Fast Conditional Mutual Information algorithm to improve classification accuracy and execution time. The suggested approach (FCMIM-SVM) shows potential accuracy and feasibility for healthcare deployment using leave one subject out cross-validation. The study's weaknesses include a lack of discussion on dataset biases, system generalizability, and feature interpretability and clinical relevance [17].

Given the rising prevalence of cardiovascular disease, particularly among the younger population, it is imperative to adopt a proactive approach in the early detection of symptoms in order to mitigate future complications such as strokes. The feasibility of conducting costly ECG tests on a daily basis for the general population may be questionable. Therefore, it is imperative to establish a consensus on a reliable and readily available approach to predict the risk of heart disease. The objective of this work is to create a nursing framework Assistant that can identify the risk of heart disease by utilizing essential parameters such as age, gender, and heart rate.

The incorporation of neural codes into the learning process boosts the dependability and resilience of the predictive model, providing a viable method for timely evaluation of potential risks. A potential limitation of this technique is the possibility of oversimplifying the prediction of heart disease risk by relying on a restricted range of indications, which may result in the omission of other pertinent aspects [18].

Many sources, including wearable sensor devices in Internet of Things health monitoring and streaming systems, create an unprecedented amount of continuous data. In healthcare, streaming big data analytics and machine learning can detect early cardiac problems cost-effectively. A sophisticated large-scale distributed computing platform, Apache Spark, is used to predict cardiac illness in real time. Spark's in-memory computations match streaming data events, optimizing machine learning. Data storage/visualization and streaming processing are included. The former applies a classification model for real-time heart disease prediction using Spark MLlib with Spark streaming, while the latter saves much created data in Apache Cassandra however One of the potential obstacles in guaranteeing data privacy and security is the absence of a comparative analysis with other real-time prediction systems [19].

An improved machine learning method for heart disease risk prediction used mean-based splitting to randomly partition the dataset. A homogeneous ensemble is formed by building classification and regression tree (CART) models for each subgroup using an accuracy-based weighted ageing classifier ensemble. This weighted ageing classifier ensemble (WAE) adjustment optimizes performance. Classification accuracy on the Cleveland and Framingham datasets is 93% and 91%, respectively, outperforming previous machine learning techniques and similar scientific publications. The ensemble learning method's higher effectiveness in predicting heart disease risk is supported by receiver operating characteristic curves. The study lacks additional dataset validation despite excellent classification accuracy, limiting the implications for different demographics and healthcare settings [20]. Early diagnosis is essential for cardiovascular disease, the leading cause of death worldwide. The article introduced a Machine Intelligence Framework for Heart Disease Diagnosis (MIFH) using Factor Analysis of Mixed Data (FAMD) to extract features and train machine learning models using the UCI heart disease Cleveland dataset. Holdout-validated MIFH exceeds recent approaches in accuracy, helping healthcare professionals and radiologists diagnose heart patients however The research does not investigate multi-class classification for heart disease, which is important for medical institution conditions [21].

The study presented a novel ensemble method applying majority voting to predict the occurrence of heart disease using cost-effective medical tests conducted at community healthcare facilities. The objective is to enhance the level of confidence and precision in physicians' diagnostic abilities by using authentic patient data. The proposed model utilizes a hard voting ensemble technique, which combines multiple machine learning models to make predictions. This ensemble approach has been found to achieve a notable accuracy rate of 90% however The lack of a thorough examination of the interpretability and clinical significance of the ensemble model [22].

Over the past two decades, artificial intelligence (AI) has grown rapidly in computer engineering and has many applications in computer vision, medicine, philosophy, psychology, and robotics. Machine Learning, a subtype of AI, has shaped manufacturing automation, biometric recognition, medical diagnosis, and data science. Even though machine learning is used daily, cardiovascular diseases (CVDs) are a global health issue. The trustworthy Boolean Machine Learning Algorithm (RBMLA) is a unique heart disease prediction algorithm that emphasizes the requirement for a trustworthy and accurate system for timely identification and diagnosis. The proposed RBMLA has 86% accuracy, indicating its potential for real-time and new test data prediction. Due to limits and changes in the supervised machine learning algorithms it relies on, the proposed Reliable Boolean Machine Learning Algorithm (RBMLA) cannot attain 100% accuracy and ideal performance [23].

The study used multi-layer perceptron (MLP) and K-nearest neighbor (K-NN) machine learning algorithms to diagnose cardiovascular disease (CVD) early and automatically. Performance optimization through outlier removal and null value handling gave the MLP model 82.47% detection accuracy and 86.41% area-under-the-curve value over the K-NN model. The MLP model proposed for automatic CVD detection may also work for other diseases. One of the drawbacks of this study is the absence of an in-depth investigation of the probable factors contributing to the observed performance disparity between the Multilayer Perceptron (MLP) and K-Nearest Neighbors (K-NN) models [24].

The research aimed to examine different computational intelligence techniques, including Logistic Regression, Support Vector Machine, Deep Neural Network, Decision Tree, Naïve Bayes, Random Forest, and K-Nearest Neighbor, in their ability to predict coronary artery heart disease. A thorough analysis of performance measures was conducted. The deep neural network demonstrated a remarkable accuracy rate of 98.15%, along with sensitivity and precision values of 98.67% and 98.01% correspondingly. The proposed methodologies shown superior performance in comparison to state-of-the-art studies in cardiac disease prediction, as evidenced by comparative assessments. However, the work does not address imbalanced dataset issues or biases [25].

Analyzing features to create an effective system using enormous data is the proposed work. The study highlights the need to evaluate medical and pathological data from healthcare providers. The proposed approach is tested on whole and reduced feature sets for classifier precision and implementation time. Machine learning, notably the Decision Tree and Ada-Boost algorithms, helps medical professionals diagnose

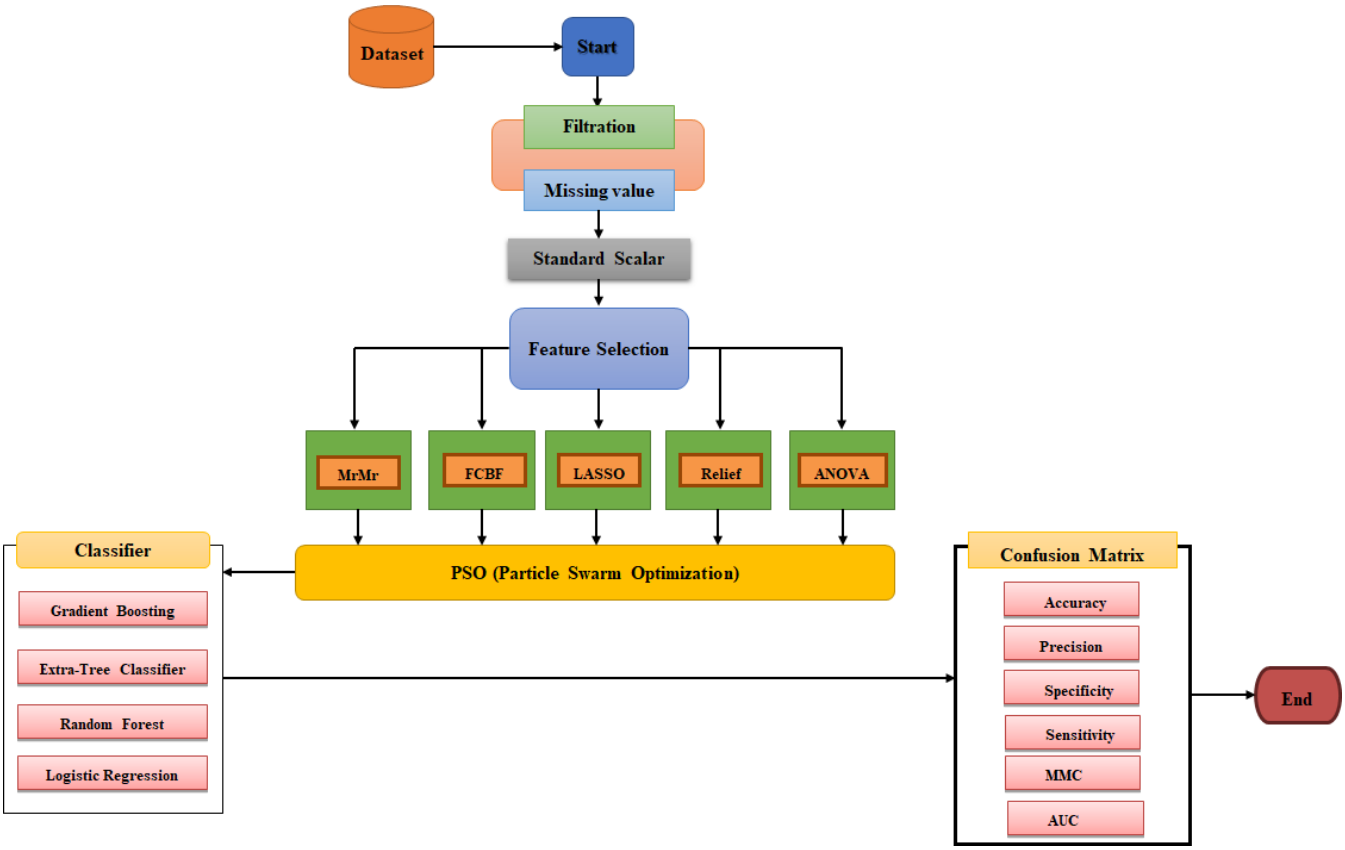



**FIGURE 1.** Flow chart of the proposed system for cardiovascular disease detection.

cardiac patients of all ages. This study incorporates machine learning techniques and critical features to understand cardiac disease however the Decision Tree algorithm overfits, and while the Ada-Boost algorithm optimizes output, the study is limited by its use of simulated data, suggesting the need for future validation on real-world datasets and exploration of a wider range of machine learning techniques for heart disease prediction [26].

Machine learning optimization issues were originally described. Then, they presented the basic principles and developments of well-known optimization techniques. Following this, they provided a brief overview of how optimization techniques have been used and developed in a variety of well-known ML applications. In conclusion, the authors outlined certain difficulties and unanswered questions regarding machine learning optimization [27].

Despite rich activity of research in the field, this research suggests that an analysis of the different feature selection methods combined with machine learning algorithms for predicting heart diseases have been understudied in the previous two decades, and the current body of literature lacks sufficient information and research studies to adequately address this gap in knowledge. Therefore, this study presents a comprehensive work of analyzing different feature selection and optimization methods. This research aims to bridge a gap in the existing body of knowledge by investigating various feature selection and machine learning algorithms for the prediction of cardiovascular diseases.

## IV. METHODOLOGY

A detailed schematic representation of the suggested research framework's design is depicted as flow chart in Figure 1. This diagram provides a thorough overview of the structure and components of the proposed framework.

### A. DATASET COLLECTION

The accuracy of classification metrics is heavily dependent on the quality of the dataset used for statistical predictions. For our research, we have picked the following datasets to both highlight the significance of the dataset and to assess its generalizability.

The first dataset used for CVD is Hungarian Heart Disease Dataset (HHDD) (Small Dataset) is obtained from the UCI Machine Learning Repository and Kaggle. It is an older and standard dataset developed in 1988. It comprises multiple databases, including those from Cleveland, Hungary, Switzerland, and Long Beach V. The dataset consists of 14 attributes and a total of 1025 instances. The target field in the dataset represents the patient's heart condition, with a numerical scale ranging from 0 (indicating no disease) to 1

(indicating severe disease). The 2nd dataset used in this study is the Kaggle (Large Dataset). In this dataset, the Behavioral Risk Factor Surveillance System (BRFSS), conducted by the Centers for Disease Control (CDC), involves annual phone surveys of over 400,000 Americans. The surveys gather information on health-related behaviors, chronic conditions, and the use of preventive services. This dataset specifically focuses on the 2015 BRFSS, containing 253,680 responses that have been cleaned and categorized into two groups based on the presence or absence of heart disease. It should be noted that there is a significant imbalance in the classes, with 229,787 individuals categorized as not having heart disease and 23,893 individuals having a history of heart disease.

### B. DATA PRE-PROCESSING

Preprocessing data transforms raw data into meaningful combinations. It is essential for accurate data representation and for proper training and testing of the classification algorithms.

Pre-processing refers to the techniques used to prepare data for analysis. The goal of pre-processing is to transform raw data into a format that is suitable for analysis, modeling, and interpretation.

### C. MISSING VALUES REMOVAL

The presence of missing values is a typical issue in data analysis and can be caused by a number of factors, including mistakes in data collection or entry, incomplete surveys, or omissions in the original data [28]. In this study the dataset was preprocessed for the removal of all the missing values.

### D. STANDARD SCALAR

Machine learning's standard scaler is a preprocessing tool for converting non-normally distributed (mean = 0, standard deviation = 1) continuous variables into a normal distribution. In many algorithms, the performance and convergence speed of the model are both affected by the scale of the features, making this a crucial step.

### E. FEATURE SELECTION LAYER

Five feature selection strategies are used: Minimum Redundancy Maximum Relevance (MrMr), Least Absolute Shrinkage and Selection Operator (LASSO), Fast Correlation-Based Filter Solution (FCBF), Relief, and Analysis of Variance (ANOVA).

MrMr represents a method for feature selection relying on filters. Its primary goal is to pinpoint a set of pertinent features while reducing redundancy. This is accomplished through a step-by-step process of eliminating features that exhibit the most redundancy with the remaining ones, while retaining a strong correlation with the target [29]. LASSO are regularization methods in regression that aid feature selection by finding the absolute values of regression coefficients. This way they shrink certain coefficients and eliminate the associated features from the model [30]. The FCBF evaluate the relevance and redundancy of features using mutual information of redundancy. It then selects features with high significance and lower redundancy to the target variables [31]. Relief assigns values to feature weights on the basis of their differentiating ability of different classes. Using weights, it selects the optimal features with no redundancy and more informative [32]. ANOVA on the other hand is a statistics based method which ensure the differences among different classes. It selects optimal features which have statistical significance on the target variable [33].

These are standard feature selection methods used in machine learning. Alongside these features selection methods, we proposed a novel Particle Swarm Optimization (PSO) for optimal feature selection.

### F. PARTICLE SWARM OPTIMIZATION (PSO)

Particle Swarm Optimization, or PSO, is an automated approach for finding the best solution for a given problem. It is based on how birds flock or fish school. PSO is a heuristic optimization method that is often used in machine learning, engineering, and operational research to solve optimization problems [34].

Equation of PSO

$$v_i^{t+1} = v_i^t c_1 r_1 \left(pbest_i^t - p_i^t\right) + c_2 r_2 \left(gbest^t - p_i^t\right) \quad (1)$$

In the above equation

- $v_i^{t+1}$ : The new value for something (like a variable or position) at location i and time t+1.
- $v_i^t$ : The current value for the same thing at location i and time t.
- $c_1$ and $c_2$: Coefficients that control the influence of the two terms that follow.
- $r_1$ and $r_2$: Random values or coefficients that introduce some randomness or scaling.
- $pbest_i^t$: The best-known position for something at location i at time t.
- $p_i^t$: The current position of something at location i at time t.
- $gbest^t$: The globally best-known position for something at time t.

In this study it is used to select optimal features for CVD.

### G. MACHINE LEARNING ALGORITHMS

Different widely used machine learning algorithms are used in this study for the classification of CVD including Gradient Boosting, Logistic regression, extra tree etc. These classifiers have shown high performance in the detection of heart diseases [35], [36], [37].

#### 1) GRADIENT BOOSTING

It's a type of ensemble learning in which several ineffective learners are generated and their predictions are averaged into one. Gradient Boosting also has good interpretability, as it can provide feature importance that indicate the relative importance of each feature in the final prediction. Equation

for Gradient Boosting is as follows [38].

$$F_o(x) = \text{argmin} \sum_{i=1}^{n} L(y_i, \gamma) \quad (2)$$

In the above equation

- $F_o(x)$ : Represents a function $F_o$ that takes input (x).
- argmin: Denotes the argument (input) that minimizes the following expression.
- $\sum i = 1n$: This is the summation symbol, indicating that we are summing up the terms that follow for each i from 1 to n
- $L(y_i, \gamma)$ : A loss function that measures the difference between the true value $y_i$ and a predicted or estimated value $\gamma$

In this study it is used to automatically detect CVD from the optimal features.

#### 2) EXTRA TREE CLASSIFIER

An ensemble machine learning algorithm, Extra Trees Classifier has been put to use in classification challenges. A random subset of features is used as the split points at each node to create each decision tree in this variant of Random Forest. The main advantage of Extra Trees Classifier is its fast training time, as the decision trees are grown using random features and split points at each node [39]. Equation for Extra Tree Classifier as follow.

$$\text{Entropy}(S) = \sum_{i=1}^{e} -p_i \log_2(p_i) \quad (3)$$

In the above equation

- Entropy (S): This represents a measure of uncertainty or randomness in a system, often denoted as S.
- $\sum i = 1e$: This is the summation symbol, indicating that we are summing up the terms that follow for each possible outcome i from 1 to e.
- $-pilog2(pi)$: This is the contribution of each possible outcome to the overall entropy. It consists of two parts:
- pi: The probability of the i-th outcome.
- log2(pi): The logarithm base 2 of the probability

In this study it is used to automatically detect CVD from the optimal features. Its performance is also compared with the state of the art ML models.

#### 3) RANDOM FOREST

It's a type of ensemble learning in which numerous decision trees are built and their predictions are averaged out. Random Forest also has good interpretability, as it can provide feature importance that indicate the relative importance of each feature in the final prediction.

Equation for Random Forest as follow [40].

$$RFfi_i = \sum_{j \in \text{alltrees}} \text{normfi}_{ij} \quad (4)$$

In the equation

- RFfii: This represents a quantity or value associated with index i.
- $\sum j \in$ all trees: This is the summation symbol, indicating that we are summing up the terms that follow for each j in the set of all trees.
- normfiij: This term represents the normalized value of fi associated with index i in the j-th tree

It is utilized in this research to detect CVD automatically based on the optimal features chosen by the proposed models. It is also evaluated in comparison to the most recent machine learning models.

#### 4) LOGISTIC REGRESSION

Logistic Regression is a statistical method used for binary classification problems, where the goal is to predict the probability of an instance belonging to one of two classes (e.g. yes/no, true/false). It is a type of generalized linear model that uses a logistic function to model the relationship between the input features and the output class. The logistic regression model is expressed as a mathematical equation of the form [41].

$$P = \frac{e^{a+bx}}{1 + e^{a+bx}} \quad (5)$$

In the above equation

- P: This represents the probability of an event occurring. In logistic regression, it's often the probability of a binary outcome being 1.
- e: This is the mathematical constant approximately equal to 2.71828.
- a and b: These are coefficients that are determined during the training of the logistic regression model. They influence the shape and position of the logistic curve.
- x: This is the input variable, and the logistic function is modeling how it influences the probability P.

### H. PERFORMANCE EVALUATION

The performance evaluation of each algorithm in this study is done using various widely used metrics such as the confusion matrix, accuracy, precision, sensitivity, specificity, area under the curve (AUC), F1-score, and Matthews correlation coefficient (MCC). These metrics provide a comprehensive assessment of the algorithms' performance and allow for a thorough analysis of their effectiveness in predicting and diagnosing heart disease [28].

#### 1) ACCURACY

The model's overall performance can be measured by calculating its accuracy using the formula in equation 6 [28]. In these equations (6)-(10), TP is number of true-positives, TN is number of true-negatives, FP is false positives and FN is false-negatives.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

#### 2) PRECISION

Precision is the ratio of the classifier's real positive scores, and it can be computed with the following formula [28].

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100 \tag{7}$$

#### 3) SENSITIVITY

Sensitivity (SN) is a metric that measures the accuracy of positive predictions by dividing the number of correctly predicted positives by the total number of actual positives. It is also known as recall (REC) or true positive rate (TPR). The calculation for sensitivity is as follows [28].

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100 \tag{8}$$

#### 4) SPECIFICITY

Specificity (SP) is a measure of the accuracy of negative predictions. It is determined by dividing the number of correctly predicted negatives by the total number of actual negatives. Specificity is also referred to as the true negative rate (TNR). The formula for calculating specificity is as follows [28].

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100 \tag{9}$$

#### 5) AREA UNDER THE CURVE (AUC)

The Area under the ROC curve (AUC) is a widely used metric for assessing the performance of a classifier in distinguishing between classes. It measures the classifier's ability to correctly classify instances from different classes. The AUC value always ranges between 0 and 1, representing the area under a unit square. It provides a quantitative measure of how well the classifier separates positive and negative instances, with higher values indicating better performance [28].

#### 6) MATHEW CORRELATION COEFFICIENT (MMC)

The Matthews correlation coefficient (MCC) is a robust statistical measure that provides a comprehensive evaluation of a classifier's performance by considering all four categories of the confusion matrix (true positives, false negatives, true negatives, and false positives). Unlike other metrics, the MCC yields a high score only when the classifier performs well across all categories. It is calculated using a specific formula, allowing for a more accurate assessment of the classifier's overall effectiveness [28].

$$\text{MMC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})\,(\text{TP} + \text{FN})\,(\text{TN} + \text{FP})\,(\text{TN} + \text{FN})}} \tag{10}$$

## V. EXPERIMENT RESULTS AND DISCUSSION

Feature selection approaches are indispensable tools in data analysis and machine learning, as they enable the identification of the most informative and influential features for building predictive models. These approaches play a critical role in enhancing the performance and interpretability of machine learning models, particularly in high-dimensional datasets with a multitude of features. The effectiveness of feature selection techniques hinges on several factors, including the size and characteristics of the dataset, the nature of the features, and the chosen feature selection algorithm.

In this study, we delve into the performance of various feature selection algorithms on datasets of varying sizes. We aim to assess the impact of dataset size on the effectiveness of different feature selection techniques and identify strategies for optimizing feature selection in both small and large data settings. The findings of this study will contribute to a deeper understanding of feature selection methodologies and their applicability in real-world data analysis scenarios for the detection of cardiovascular diseases.

First of all, we checked the statistical properties of small dataset of each feature that are given below in the following Table 1 and plotted in figure 2.

**TABLE 1. Statistical property of each feature of small data.**

| INDEX | COUNT | MEAN | STD | MIN | 25 % | 50 % | 75 % | MAX |
|---|---|---|---|---|---|---|---|---|
| AGE | 1025.0 | 54.434 | 9.072 | 29.0 | 48.0 | 56.0 | 61.0 | 77.0 |
| SEX | 1025.0 | 0.696 | 0.46 | 0.0 | 0.0 | 1.0 | 1.0 | 1.0 |
| CP | 1025.0 | 0.942 | 1.03 | 0.0 | 0.0 | 1.0 | 2.0 | 3.0 |
| TRESTBPS | 1025.0 | 131.612 | 17.517 | 94.0 | 120.0 | 130.0 | 140.0 | 200.0 |
| CHOL | 1025.0 | 246.0 | 51.593 | 126.0 | 211.0 | 240.0 | 275.0 | 564.0 |
| FBS | 1025.0 | 0.149 | 0.357 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 |
| RESTECG | 1025.0 | 0.53 | 0.528 | 0.0 | 0.0 | 1.0 | 1.0 | 2.0 |
| THALACH | 1025.0 | 149.114 | 23.006 | 71.0 | 132.0 | 152.0 | 166.0 | 202.0 |
| EXANG | 1025.0 | 0.337 | 0.473 | 0.0 | 0.0 | 0.0 | 1.0 | 1.0 |
| OLDPEAK | 1025.0 | 1.072 | 1.175 | 0.0 | 0.0 | 0.8 | 1.8 | 6.2 |
| SLOPE | 1025.0 | 1.385 | 0.618 | 0.0 | 1.0 | 1.0 | 2.0 | 2.0 |
| CA | 1025.0 | 0.754 | 1.031 | 0.0 | 0.0 | 0.0 | 1.0 | 4.0 |
| THAL | 1025.0 | 2.324 | 0.621 | 0.0 | 2.0 | 2.0 | 3.0 | 3.0 |
| TARGET | 1025.0 | 0.513 | 0.5 | 0.0 | 0.0 | 1.0 | 1.0 | 1.0 |

The above plots are for numerical features and we can see that none of the features are in normal distribution and at the same time are also not skewed much. So, a simple scaling technique can help us to reduce the skewness.

### A. TECHNIQUES FOR SMALL DATA

We performed different feature selection techniques on a small dataset the details are given below with Graph representation and Tables.

The accuracy of various models using different techniques is provided in Table 2. Notably, our approach achieved higher

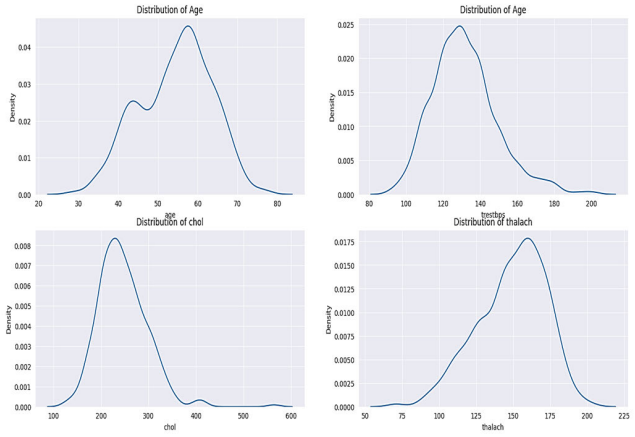



**FIGURE 2.** Distribution of numerical features.

**TABLE 2.** Accuracy of all model on small dataset.

| Feature Selection | Logistic Regression Accuracy | Extra Tree Accuracy | Random Forest Accuracy | Gradient Boosting Accuracy |
|---|---|---|---|---|
| MrMr | 0.854 | **1** | **1** | 0.966 |
| FCBF | 0.81 | **1** | **1** | 0.912 |
| Lasso | 0.761 | 0.888 | 0.893 | 0.859 |
| Relief | 0.776 | **1** | **1** | 0.912 |
| ANOVA | 0.722 | **0.941** | **0.937** | 0.805 |

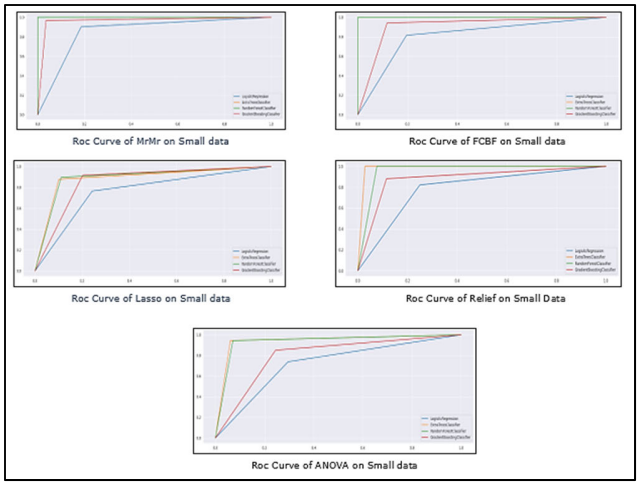



**FIGURE 3.** ROC curves for MrMr, FCBF, Lasso, relief and ANOV on small.

accuracy compared to the state of the art as we used a novel feature selection method based on PSO which finds the best features. Specifically, we attained 100% accuracy using MrMr, FCBF, and Relief along with PSO on Extra Tree Classifier and Random Forest, a significant improvement from the as compared to state of the art which is less than 96%. Additionally, we introduced the ANOVA selection technique, resulting in improved accuracy in our research compared to the state of the art.

The comparison highlighted in Figure 4 shows that the strong performance of the Extra Tree and Random Forest models is due to the optimal features provided by our proposed methods. Notably the MrMr, FCBF and Relief selection techniques achieved the highest accuracy of 100%. On the contrary, the Lasso technique showed the poorest performance among the methods evaluated because Lasso's L1 regularization penalty tends to favor features with large absolute values, potentially overlooking features that have smaller values. Also it has mostly overfitting issue which leads to low validation accuracy.

### B. TECHNIQUES FOR LARGE DATA

The performance of the proposed models was also evaluated using a large dataset related to heart diseases. The dataset has 253680 records and 22 columns. Four columns are of numerical type but remaining all features were of categorical type which are encode into integers values. There was no null value in whole dataset thus in pre-processing only the Pearson correlation of all features is computed for feature selection. As can be seen from figure 5 the correlation between heart disease and other medical features like cholesterol, BMI, stroke, High PB etc is positive and have higher values which means that these parameters have high effect on heart condition and are significant for disease detection on the other hand the non-medical features like sex, education and income etc. have negative correlation with the heart disease thus these are not significant.

The same feature selection and optimization techniques are applied on Large Dataset. The machine learning classifiers are then trained on the selected features for CVD. The results are summarized below in Table 4. In the analysis, we found that two models, Extra Tree and Random Forest, performed the best. Among the feature selection techniques, FCBF and Relief gave the highest accuracies of 78% and 77% respectively. On the other hand, MrMr and ANOVA had the lowest performance, achieving less than 70% accuracy for large datasets. Possible reasons for low performance of these models is that the classifier makes predictions using simple rules which is not useful when dealing with complex opitimized feature set.

The tabulated information in TABLE 5 offers an extensive insight into the collective outcomes derived from a diverse range of classifiers. Accompanied by their respective confusion matrices, these results summarie the performanceof the various models utilized within the scope of our study. Specifically focusing on a smaller dataset, this detailed table provides a comprehensive representation of the

We assessed the performance of each classification model by using only the chosen features as inputs for the large

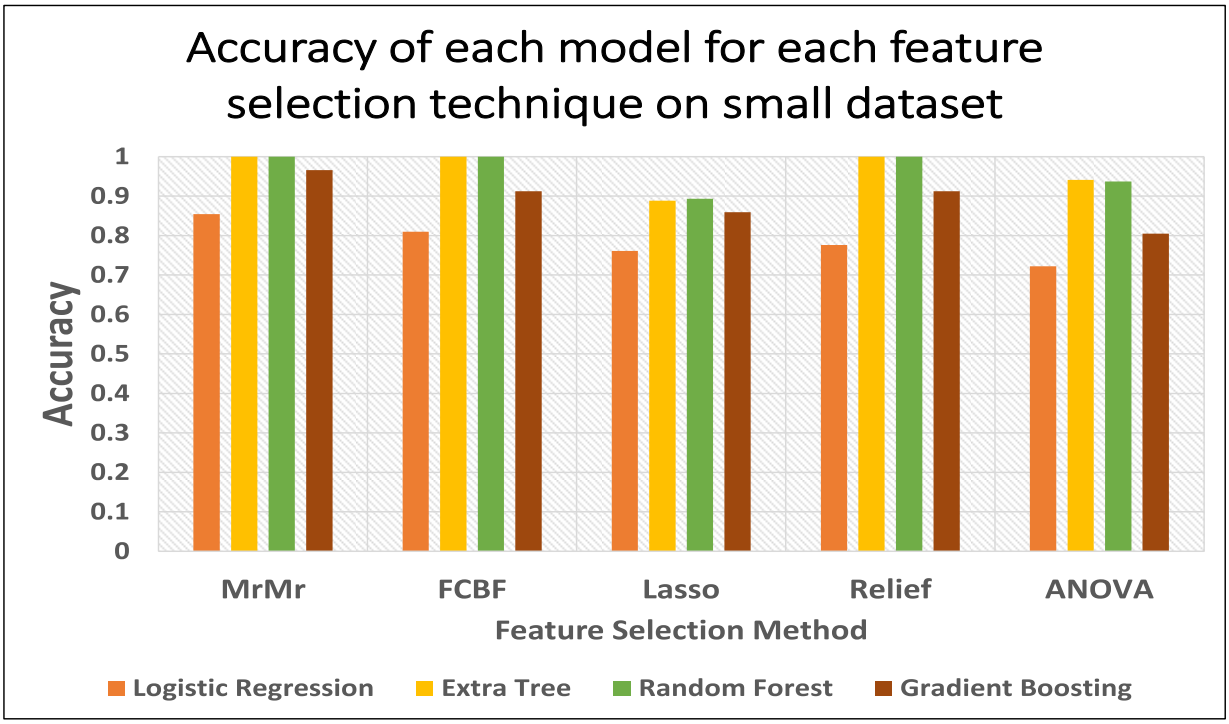



**FIGURE 4.** Accuracy of each model on each selection technique on small data.

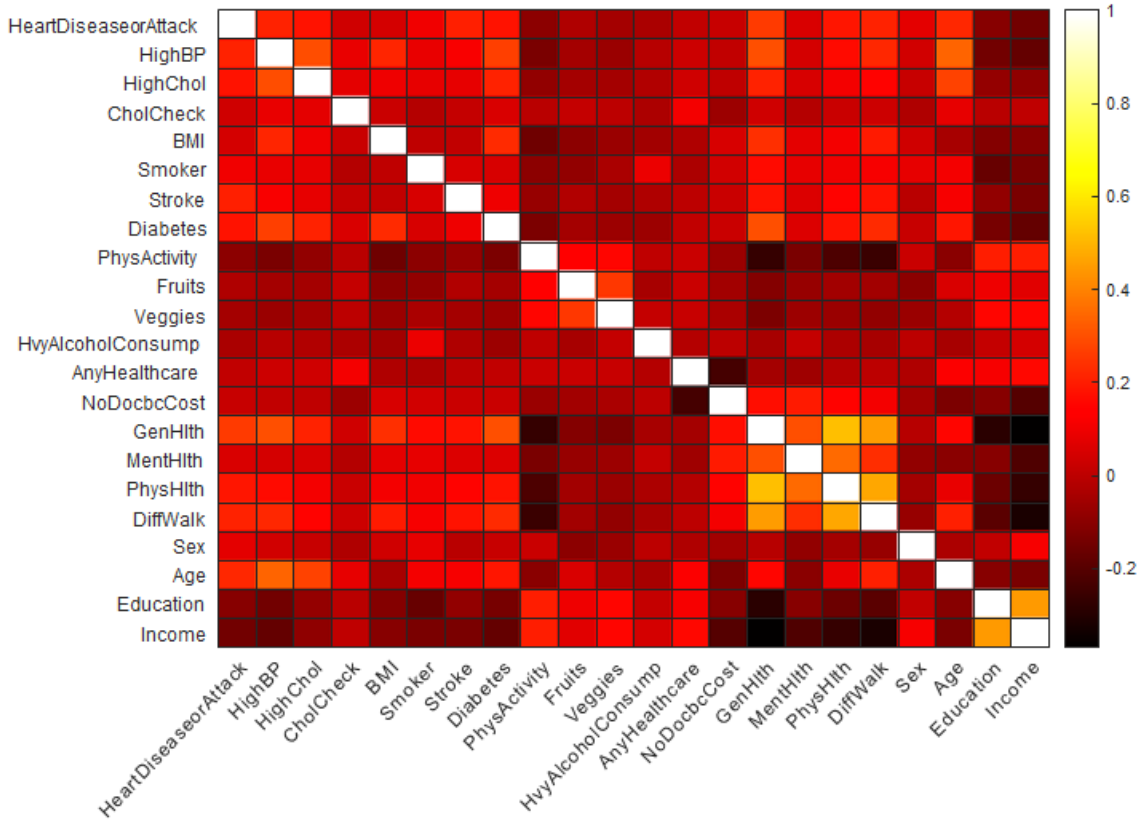



**FIGURE 5.** Pearson correlation between all the features for Large Data.

**TABLE 3.** Overall results of all classifiers with confusion matrix on small dataset.

| | Model | Selection Technique | Accuracy | Precision | Recall | Sensitivity | Specificity | AUC Score | MMC SCORE | F1_Score | Confusion Matrix |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | LRC | MrMr | 0.854 | 0.798 | 0.902 | 0.902 | 0.814 | 0.858 | 0.713 | 0.847 | [[92 2 1] [ 9 83 ]] |
| 1 | ETC | MrMr | **1** | **1** | **1** | **1** | **1** | **1** | **1** | **1** | [[113 0] [ 0 92]] |
| 2 | RFC | MrMr | **1** | **1** | **1** | **1** | **1** | **1** | **1** | **1** | [[113 0] [ 0 92]] |
| 3 | GBC | MrMr | 0.966 | 0.957 | 0.967 | 0.967 | 0.965 | 0.966 | 0.931 | 0.962 | [[109 4] [ 3 89]] |
| 4 | LRC | FCBF | 0.81 | 0.808 | 0.816 | 0.816 | 0.804 | 0.81 | 0.62 | 0.812 | [[82 20 ] [19 84]] |
| 5 | ETC | FCBF | **1** | **1** | **1** | **1** | **1** | **1** | **1** | **1** | [[102 0] [ 0 103]] |
| 6 | RFC | FCBF | **1** | **1** | **1** | **1** | **1** | **1** | **1** | **1** | [[102 0] [ 0 103]] |
| 7 | GBC | FCBF | 0.912 | 0.89 | 0.942 | 0.942 | 0.882 | 0.912 | 0.826 | 0.915 | [[90 12] [ 6 97]] |
| 8 | LRC | LASSO | 0.761 | 0.771 | 0.764 | 0.764 | 0.758 | 0.761 | 0.522 | 0.768 | [[75 24] [25 81]] |
| 9 | ETC | LASSO | 0.888 | 0.903 | 0.877 | 0.877 | 0.899 | 0.888 | 0.776 | 0.89 | [[89 10] [13 93]] |
| 10 | RFC | LASSO | 0.893 | 0.896 | 0.896 | 0.896 | 0.889 | 0.893 | 0.785 | 0.896 | [[88 11] [11 95]] |
| 11 | GBC | LASSO | 0.859 | 0.829 | 0.915 | 0.915 | 0.798 | 0.857 | 0.72 | 0.87 | [[79 20] [ 9 97]] |
| 12 | LRC | RELIEF | 0.776 | 0.713 | 0.809 | 0.809 | 0.75 | 0.779 | 0.554 | 0.758 | [[87 29] [17 72]] |
| 13 | ETC | RELIEF | **1** | **1** | **1** | **1** | **1** | **1** | **1** | **1** | [[116 0] [ 0 89]] |
| 14 | RFC | RELIEF | **1** | **1** | **1** | **1** | **1** | **1** | **1** | **1** | [[116 0] [ 0 89]] |
| 15 | GBC | RELIEF | 0.912 | 0.851 | **0.966** | **0.966** | 0.871 | 0.918 | 0.83 | 0.905 | [[101 15] [ 3 86]] |
| 16 | LRC | ANOVA | 0.722 | 0.731 | 0.738 | 0.738 | 0.704 | 0.721 | 0.443 | 0.735 | [[69 29] [28 79]] |
| 17 | ETC | ANOVA | **0.941** | **0.944** | **0.944** | **0.944** | **0.939** | **0.941** | 0.883 | **0.944** | [[ 92 6] [ 6 101]] |
| 18 | RFC | ANOVA | 0.937 | 0.935 | 0.944 | 0.944 | 0.929 | 0.936 | 0.873 | 0.94 | [[ 91 7] [ 6 101]] |
| 19 | GBC | ANOVA | 0.805 | 0.791 | 0.85 | 0.85 | 0.755 | 0.803 | 0.61 | 0.82 | [[74 24] [16 91]] |

**TABLE 4.** Results for large data using the selected features.

| Feature Selection | Logistic Regression Accuracy | Extra Tree Accuracy | Random Forest Accuracy | Gradient Boosting Accuracy |
|---|---|---|---|---|
| MrMr | 0.685 | 0.688 | 0.688 | 0.689 |
| FCBF | 0.729 | 0.779 | 0.782 | 0.736 |
| Lasso | 0.717 | 0.721 | 0.721 | 0.722 |
| Relief | 0.736 | 0.769 | 0.772 | 0.745 |
| ANOVA | 0.692 | 0.697 | 0.697 | 0.698 |

dataset. As shown in Tables 4 and 5, this reduced feature subset from the CVD dataset resulted in the classification performance for each model also plotted in Figure 6. The analyses conducted, highlighting the effectiveness and intricacies of each classifier in our research work using the reduced features set. In the above analysis we found that two models' extra tree and random forest have achieved highest performance with the limited number of features. Among the feature selection techniques, FCBF and relief achieved accuracies of 78% and 77% respectively. On the other hand, MrMr and ANOVA had lowest performance with accuracy of 70% each for large datasets. Incorporated within TABLE 5 are the comprehensive and overarching outcomes yielded by a diverse set of models, coupled with their corresponding confusion matrices. This table serves as an encompassing depiction of the cumulative performance of these varied models, meticulously evaluated within the framework of a larger dataset. It provides an extensive and detailed overview of the analyses conducted, shedding light on the intricate interplay between each classifier and its corresponding results in the context of our research endeavors. Moreover, employing

**TABLE 5. Over all results of all classifier with confusion matrix on large dataset using the selected features.**

| | Model | Selection Technique | Accuracy | Precision | Recall | Sensitivity | Specificity | AUC Score | MMC SCORE | F1_Score | Confusion Matrix |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | LRC | MrMr | 0.685383 | 0.677285 | 0.706246 | 0.706246 | 0.664588 | 0.685417 | 0.371147 | 0.691462 | [30593 15440] [13478 32404] |
| 1 | ETC | MrMr | 0.688952 | 0.653249 | 0.803256 | 0.803256 | 0.575022 | 0.689139 | 0.388475 | 0.720528 | [26470 19563] [ 9027 36855 ] |
| 2 | RFC | MrMr | 0.688952 | 0.653249 | 0.803256 | 0.803256 | 0.575022 | 0.689139 | 0.388475 | 0.720528 | [26470 19563] [ 9027 36855 ] |
| 3 | GBC | MrMr | 0.689605 | 0.65633 | 0.793884 | 0.793884 | 0.585667 | 0.689776 | 0.388003 | 0.718584 | [26960 19073] [ 9457 36425 ] |
| 4 | LRC | FCBF | 0.729652 | 0.714636 | 0.763153 | 0.763153 | 0.696261 | 0.729707 | 0.460422 | 0.738098 | [32051 13982] [10867 35015] |
| 5 | ETC | FCBF | 0.779329 | 0.755985 | 0.823852 | 0.823852 | 0.734951 | 0.779402 | 0.560978 | 0.788461 | [33832 12201] [ 8082 37800 ] |
| 6 | RFC | FCBF | 0.782756 | 0.752327 | 0.841986 | 0.841986 | 0.72372 | 0.782853 | 0.56964 | 0.794636 | [33315 12718] [ 7250 38632 ] |
| 7 | GBC | FCBF | 0.736093 | 0.704732 | 0.81119 | 0.81119 | 0.661243 | 0.736216 | 0.477778 | 0.754223 | [30439 15594] [ 8663 37219 ] |
| 8 | LRC | LASSO | 0.717021 | 0.703248 | 0.749292 | 0.749292 | 0.684857 | 0.717074 | 0.435032 | 0.72554 | [31526 14507] [11503 34379] |
| 9 | ETC | LASSO | 0.721808 | 0.694903 | 0.789198 | 0.789198 | 0.654639 | 0.721919 | 0.447866 | 0.739055 | [30135 15898] [ 9672 36210]] |
| 10 | RFC | LASSO | 0.721841 | 0.694801 | 0.789612 | 0.789612 | 0.654291 | 0.721952 | 0.447979 | 0.739179 | [30119 15914] [ 9653 36229 ] |
| 11 | GBC | LASSO | 0.721928 | 0.694244 | 0.791552 | 0.791552 | 0.652532 | 0.722042 | 0.448392 | 0.739712 | [30038 15995] [ 9564 36318 ] |
| 12 | LRC | ANOVA | 0.692781 | 0.698998 | 0.675385 | 0.675385 | 0.710121 | 0.692753 | 0.385747 | 0.686989 | [32689 13344] [14894 30988] |
| 13 | ETC | ANOVA | 0.697851 | 0.678073 | 0.751493 | 0.751493 | 0.644386 | 0.697939 | 0.39814 | 0.712897 | [29663 16370] [11402 34480] |
| 14 | RFC | ANOVA | 0.697721 | 0.678284 | 0.750338 | 0.750338 | 0.645276 | 0.697807 | 0.397788 | 0.712494 | [29704 16329] [11455 34427] |
| 15 | GBC | ANOVA | 0.698308 | 0.682212 | 0.740617 | 0.740617 | 0.656138 | 0.698378 | 0.398156 | 0.710216 | [30204 15829] [11901 33981] |
| 16 | LRC | RELIEF | 0.736213 | 0.72241 | 0.765834 | 0.765834 | 0.706689 | 0.736261 | 0.473329 | 0.743488 | [32531 13502] [10744 35138] |
| 17 | ETC | RELIEF | 0.769733 | 0.749098 | 0.810013 | 0.810013 | 0.729585 | 0.769799 | 0.541312 | 0.778365 | [33585 12448] [ 8717 37165] |
| 18 | RFC | RELIEF | 0.772562 | 0.745609 | 0.826294 | 0.826294 | 0.719006 | 0.77265 | 0.548411 | 0.783881 | [33098 12935] [7970 37912] |
| 19 | GBC | RELIEF | 0.745602 | 0.716632 | 0.811081 | 0.811081 | 0.680338 | 0.745709 | 0.49562 | 0.760937 | [31318 14715] [ 8668 37214] |

the reduced feature set during the training of classification models led to a decrease in computational iterations per second (it/s). These findings underscore the impact of feature selection techniques, demonstrating that they not only reduce the dimensionality of the feature space but also enhance the performance of ML models in various aspects.

This proposed framework holds promise for improving the early detection and diagnosis of cardiovascular disease, a significant global public health concern. The results obtained in this study also presents a comparative analysis of feature selection techniques on a small dataset and evaluated their impact on the performance of machine learning algorithms. To assess the performance, we considered various evaluation metrics such as accuracy, precision, sensitivity, specificity, AUC, F1-score, and MCC. The findings indicated that MRMR, FCBF, and Relief demonstrated superior performance compared to the other techniques in terms of accuracy. Notably, the Extra Tree Classifier and Random Forest achieved a remarkable accuracy of 100% when using features selected by MRMR, FCBF, and Relief on. This study underscored the significance of feature selection in enhancing the performance of machine learning algorithms and emphasized the importance of selecting an appropriate technique based on the dataset's size and characteristics. The large dataset used in this research work comprised 253,680 records and 22 columns, predominantly consisting of categorical features. Imbalanced data was observed, and to address this issue, oversampling techniques were employed to balance the dataset. The report compared the performance of various feature selection techniques, including MRMR, FCBF, LASSO, ANOVA, and RELIEF, in combination with different machine learning models.

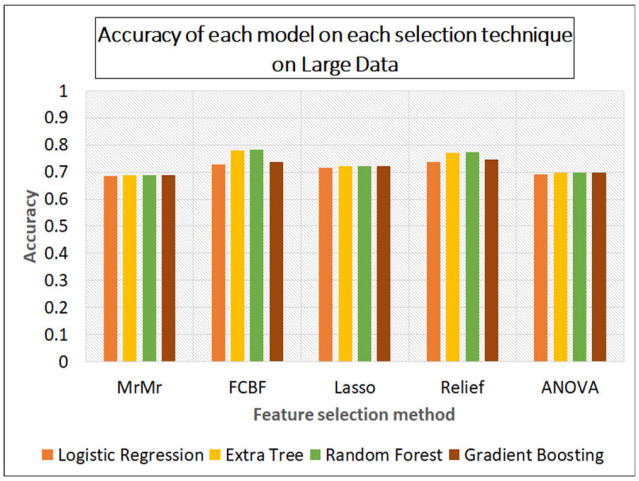



**FIGURE 6.** Accuracy of models on each technique for large data.

## VI. CONCLUSION

This study proposed a novel framework for detecting and classifying cardiovascular disease (CVD) using machine learning algorithms and optimal feature selection techniques. The proposed framework demonstrated the significant impact of feature selection on enhancing the performance of machine learning algorithms for CVD prediction. The study evaluated five different feature selection techniques: MRMR, FCBF, LASSO, Relief, and ANOVA. Among these techniques, FCBF exhibited superior performance, achieving an accuracy of 78% when combined with the Extra Tree and Random Forest models. This finding highlights the effectiveness of FCBF in selecting relevant features from large-scale CVD datasets. The study also highlights the importance of selecting an appropriate feature selection and optimization technique based on the characteristics of the dataset. For large datasets with predominantly categorical features, like the one used in this study, FCBF emerged as a promising technique for identifying relevant features and improving the performance of machine learning algorithms in CVD prediction.

## VII. FUTURE WORK

In addition to the future works mentioned earlier, there are several other areas that could be explored to improve the analysis and results of this study. Some of these potential future works include:

Consider using deep learning models, specifically neural networks, to enhance model accuracy on different datasets. These models capture complex feature relationships and have the potential to improve overall performance. Further exploration of deep learning could lead to valuable insights and advancements in the field Alternative feature selection techniques: While several feature selection techniques were used in this study, there are many other methods that could be explored. For example, genetic algorithms, decision trees, and mutual information-based methods could be used for feature selection. By testing various methods and comparing the results, the most effective feature selection method for this dataset could be identified. Additional data sources: While the dataset used in this study contained a large amount of information about heart disease risk factors, there may be additional data sources that could be used to further improve the accuracy of the models. For example, data on the environmental factors such as air quality and access to healthcare could be included and analyzed to see if they have an impact on heart disease risk. Improved data balancing techniques: In this study, oversampling was used to balance the imbalanced data. However, there are other techniques such as under sampling and SMOTE (Synthetic Minority Over-Sampling Technique) that could be explored to improve the balance of the data. Ensemble learning: Ensemble learning is a technique that combines multiple machine learning models to improve the overall accuracy of the predictions. By using ensemble learning techniques such as bagging or boosting, the accuracy of the models could potentially be improved by exploring these and other potential future works, it may be possible to further improve the accuracy of the heart disease risk prediction models and develop more effective strategies for preventing and managing heart disease.

## REFERENCES

[1] A. K. Gárate-Escamila, A. H. El Hassani, and E. Andrès, "Classification models for heart disease prediction using feature selection and PCA," *Informat. Med. Unlocked*, vol. 19, Jan. 2020, Art. no. 100330, doi: 10.1016/j.imu.2020.100330.

[2] V. Chang, V. R. Bhavani, A. Q. Xu, and M. Hossain, "An artificial intelligence model for heart disease detection using machine learning algorithms," *Healthcare Anal.*, vol. 2, Nov. 2022, Art. no. 100016, doi: 10.1016/j.health.2022.100016.

[3] M. Ganesan and N. Sivakumar, "IoT based heart disease prediction and diagnosis model for healthcare using machine learning models," in *Proc. IEEE Int. Conf. Syst., Comput., Autom. Netw. (ICSCAN)*, Mar. 2019, pp. 1–5, doi: 10.1109/ICSCAN.2019.8878850.

[4] D. P. Isravel, S. V. P. Darcini, and S. Silas, "Improved heart disease diagnostic IoT model using machine learning techniques," *Int. J. Sci. Technol. Res.*, vol. 9, no. 2, pp. 4442–4446, 2020.

[5] I. S. G. Brites, L. M. da Silva, J. L. V. Barbosa, S. J. Rigo, S. D. Correia, and V. R. Q. Leithardt, "Machine learning and IoT applied to cardiovascular diseases identification through heart sounds: A literature review," *Informatics*, vol. 8, no. 4, p. 73, Oct. 2021, doi: 10.3390/informatics8040073.

[6] D. T. Thai, Q. T. Minh, and P. H. Phung, "Toward an IoT-based expert system for heart disease diagnosis," in *Proc. 28th Mod. Artif. Intell. Cogn. Sci. Conf. (MAICS)*, 2017, pp. 157–164.

[7] B. Padmaja, C. Srinidhi, K. Sindhu, K. Vanaja, N. M. Deepika, and E. K. R. Patro, "Early and accurate prediction of heart disease using machine learning model," *Turkish J. Comput. Math. Educ.*, vol. 12, no. 6, pp. 4516–4528, 2021.

[8] S. Anitha and N. Sridevi, *Heart Disease Prediction Using Data Mining Techniques S Anitha, N Sridevi to Cite This Version*, document HAL Id Hal-02196156, 2019. [Online]. Available: https://hal.archives-ouvertes.fr/hal-02196156/document

[9] R. Bharti, A. Khamparia, M. Shabaz, G. Dhiman, S. Pande, and P. Singh, "Prediction of heart disease using a combination of machine learning and deep learning," *Comput. Intell. Neurosci.*, vol. 2021, pp. 1–11, Jul. 2021, doi: 10.1155/2021/8387680.

[10] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "Heart disease prediction using machine learning algorithms," *IOP Conf., Mater. Sci. Eng.*, vol. 1022, no. 1, Jan. 2021, Art. no. 012072, doi: 10.1088/1757-899x/1022/1/012072.

[11] B. Pavithra and V. Rajalakshmi, "Heart disease detection using machine learning algorithms," in *Proc. Int. Conf. Emerg. Current Trends Comput. Expert Technol.*, vol. 35, 2020, pp. 1131–1137, doi: 10.1007/978-3-030-32150-5_115.

[12] N. Louridi, S. Douzi, and B. El Ouahidi, "Machine learning-based identification of patients with a cardiovascular defect," *J. Big Data*, vol. 8, no. 1, pp. 1–5, Dec. 2021, doi: 10.1186/s40537-021-00524-9.

[13] P. Singh, G. K. Pal, and S. Gangwar, "Prediction of cardiovascular disease using feature selection techniques," *Int. J. Comput. Theory Eng.*, vol. 14, no. 3, pp. 97–103, 2022, doi: 10.7763/ijcte.2022.v14.1316.

[14] M. Swathy and K. Saruladha, "A comparative study of classification and prediction of cardio-vascular diseases (CVD) using machine learning and deep learning techniques," *ICT Exp.*, vol. 8, no. 1, pp. 109–116, Mar. 2022, doi: 10.1016/j.icte.2021.08.021.

[15] D. Vaddella, C. Sruthi, B. K. Chowdary, and S.-R. Subbareddy, "Prediction of heart disease using machine learning techniques," *Restaur. Bus.*, vol. 118, no. 1, pp. 125–129, 2019, doi: 10.26643/rb.v118i1.7621.

[16] V. V. Ramalingam, A. Dandapath, and M. Karthik Raja, "Heart disease prediction using machine learning techniques: A survey," *Int. J. Eng. Technol.*, vol. 7, no. 2, p. 684, Mar. 2018, doi: 10.14419/ijet.v7i2.8.10557.

[17] 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, doi: 10.1109/ACCESS.2020.3001149.

[18] P. Kalpana, S. S. Vignesh, L. M. P. Surya, and V. V. Prasad, "Prediction of heart disease using machine learning," *J. Phys., Conf. Ser.*, vol. 1916, no. 1, May 2021, Art. no. 012022, doi: 10.1088/1742-6596/1916/1/012022.

[19] A. Ed-Daoudy and K. Maalmi, "Real-time machine learning for early detection of heart disease using big data approach," in *Proc. Int. Conf. Wireless Technol., Embedded Intell. Syst. (WITS)*, Apr. 2019, pp. 1–5, doi: 10.1109/WITS.2019.8723839.

[20] I. D. Mienye, Y. Sun, and Z. Wang, "An improved ensemble learning approach for the prediction of heart disease risk," *Informat. Med. Unlocked*, vol. 20, Jan. 2020, Art. no. 100402, doi: 10.1016/j.imu.2020.100402.

[21] 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, doi: 10.1109/ACCESS.2019.2962755.

[22] R. Atallah and A. Al-Mousa, "Heart disease detection using machine learning majority voting ensemble method," in *Proc. 2nd Int. Conf. New Trends Comput. Sci. (ICTCS)*, Oct. 2019, pp. 1–6, doi: 10.1109/ICTCS.2019.8923053.

[23] M. Bheemalingaiah, G. R. Swamy, P. Vishvapathi, P. V. Babu, E. N. Rao, and P. N. Rao, "Detection of heart disease by using reliable Boolean machine learning algorithm," *J. Theor. Appl. Inf. Technol.*, vol. 99, no. 15, pp. 3856–3880, 2021, doi: 10.5281/zenodo.5353586.

[24] M. Pal, S. Parija, G. Panda, K. Dhama, and R. K. Mohapatra, "Risk prediction of cardiovascular disease using machine learning classifiers," *Open Med.*, vol. 17, no. 1, pp. 1100–1113, Jun. 2022.

[25] S. I. Ayon, M. M. Islam, and M. R. Hossain, "Coronary artery heart disease prediction: A comparative study of computational intelligence techniques," *IETE J. Res.*, vol. 68, no. 4, pp. 2488–2507, Jul. 2022, doi: 10.1080/03772063.2020.1713916.

[26] G. Choudhary and S. N. Singh, "Prediction of heart disease using machine learning algorithms," in *Proc. Int. Conf. Smart Technol. Comput., Electr. Electron. (ICSTCEE)*, Oct. 2020, pp. 197–202, doi: 10.1109/ICSTCEE49637.2020.9276802.

[27] Y. Khourdifi and M. Bahaj, "Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization," *Int. J. Intell. Eng. Syst.*, vol. 12, no. 1, pp. 242–252, Feb. 2019, doi: 10.22266/ijies2019.0228.24.

[28] Y. Muhammad, M. Tahir, M. Hayat, and K. T. Chong, "Early and accurate detection and diagnosis of heart disease using intelligent computational model," *Sci. Rep.*, vol. 10, no. 1, pp. 1–18, Nov. 2020, doi: 10.1038/s41598-020-76635-9.

[29] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, Univ. Waikato, Hamilton, New Zealand, 1999.

[30] L. Yu and H. Liu, "Feature selection for regression," *Data Mining Knowl. Discovery*, vol. 15, no. 3, pp. 259–285, 2003.

[31] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Stat. Soc., Ser. B, Methodolog.*, vol. 58, no. 1, pp. 267–288, Jan. 1996.

[32] K. Kira and L. Cooper, "A compressed sensing approach to feature extraction from high-dimensional data," School Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep., 2000.

[33] D. C. Montgomery, *Introduction to Linear Regression Analysis*. Hoboken, NJ, USA: Wiley, 2012.

[34] A. H. Shahid and M. P. Singh, "A novel approach for coronary artery disease diagnosis using hybrid particle swarm optimization based emotional neural network," *Biocybernetics Biomed. Eng.*, vol. 40, no. 4, pp. 1568–1585, Oct. 2020, doi: 10.1016/j.bbe.2020.09.005.

[35] R. Tr, U. K. Lilhore, P. M, S. Simaiya, A. Kaur, and M. Hamdi, "Predictive analysis of heart diseases with machine learning approaches," *Malaysian J. Comput. Sci.*, pp. 132–148, Mar. 2022.

[36] N. A. Baghdadi, S. M. Farghaly Abdelaliem, A. Malki, I. Gad, A. Ewis, and E. Atlam, "Advanced machine learning techniques for cardiovascular disease early detection and diagnosis," *J. Big Data*, vol. 10, no. 1, p. 144, Sep. 2023.

[37] K. M. Mohi Uddin, R. Ripa, N. Yeasmin, N. Biswas, and S. K. Dey, "Machine learning-based approach to the diagnosis of cardiovascular vascular disease using a combined dataset," *Intell.-Based Med.*, vol. 7, Jan. 2023, Art. no. 100100.

[38] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, Apr. 2006, doi: 10.1007/s10994-006-6226-1.

[39] M. M. Hameed, M. K. AlOmar, F. Khaleel, and N. Al-Ansari, "An extra tree regression model for discharge coefficient prediction: Novel, practical applications in the hydraulic sector and future research directions," *Math. Problems Eng.*, vol. 2021, pp. 1–19, Sep. 2021, doi: 10.1155/2021/7001710.

[40] M. Pal and S. Parija, "Prediction of heart diseases using random forest," *J. Phys., Conf. Ser.*, vol. 1817, no. 1, Mar. 2021, Art. no. 012009, doi: 10.1088/1742-6596/1817/1/012009.

[41] A. G, B. Ganesh, A. Ganesh, C. Srinivas, and K. Mensinkal, "Logistic regression technique for prediction of cardiovascular disease," *Global Transitions Proc.*, vol. 3, no. 1, pp. 127–130, Jun. 2022, doi: 10.1016/j.gltp.2022.04.008.

**TAHSEEN ULLAH** received the Master of Computer Science (M.C.S.) degree from Abdul Wali Khan University Mardan, Pakistan. He is currently pursuing the M.S. degree in computer science with Abasyn University, Peshawar, Pakistan. His research interests include artificial intelligence, machine learning, deep learning, computer vision, and the IoT.

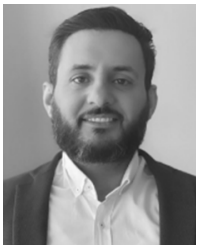
**SYED IRFAN ULLAH** received the Ph.D. degree from the Department of Computer Science, International Islamic University Peshawar. He is currently an Associate Professor with the Department of Computing, Abasyn University. He is a well known Researcher in the area of intelligent cryptosystems and he has a number of research articles in various fields of computer science.

**KHALIL ULLAH** received the Graduate degree in computer systems engineering from the University of Engineering and Technology, Peshawar, Pakistan, in 2006, the Master of Science (M.S.) degree in electronics and communications engineering from Myongji University, South Korea, in 2009, and the Ph.D. degree in biomedical engineering from LISiN, Politecnico di Torino, in 2016, under the Erasmus Mundus Expert II Fellowship. He is currently an Assistant Professor and the Head of the Software Engineering Department, University of Malakand. His research interests include extracting muscle anatomical and physiological information from high-density electromyography, computer vision, digital signal and image processing, and deep learning with applications to medical healthcare.

**MUHAMMAD ISHAQ** received the B.S. degree in computer science from The University of Haripur, Pakistan. He is currently pursuing the M.S. degree in telecommunication and networks with Abasyn University, Peshawar, Pakistan.

He is currently a Senior IT Officer with the Helping Hand Institute of Rehabilitation Sciences, Mansehra, Pakistan. His research interests include artificial intelligence, machine learning, deep learning, and the IoT.

**AHMAD KHAN** received the B.Sc. degree in computer systems engineering from the University of Engineering and Technology (UET) Peshawar, Pakistan, in 2006, the M.Sc. degree in computer software engineering from the University of Engineering and Technology, Peshawar, Pakistan, in 2014, and the Ph.D. degree in computer systems engineering from UET Peshawar, in October 2020. He is currently an Assistant Professor in software engineering with the Mirpur University of Science and Technology, Mirpur, Azad Jammu and Kashmir, Pakistan. His research interests include machine to machine communication, the Internet of Things (IoT), and computer vision.



**YAZEED YASIN GHADI** received the Ph.D. degree in electrical and computer engineering from Queensland University. His dissertation on developing novel hybrid plasmonicphotonic onchip biochemical sensors received the Sigma Xi Best Ph.D. Thesis Award. He is currently an Assistant Professor in software engineering with Al Ain University. He was a Postdoctoral Researcher with Queensland University, before joining Al Ain. His current research is on developing novel electro-acousto-optic neural interfaces for largescale high-resolution electrophysiology and distributed optogenetic stimulation. He has published more than 80 peer-reviewed journals and conference papers and he holds three pending patents. He is a recipient of several awards.

**ABDULMOHSEN ALGARNI** received the Ph.D. degree from the Queensland University of Technology, Australia, in 2012. He was a Research Associate with the School of Electrical Engineering and Computer Science, Queensland University of Technology, in 2012. He is currently an Associate Professor with the College of Computer Science, King Khalid University. His research interests include artificial intelligence, data mining, text mining, machine learning, information retrieval, and information filtering.