Potential directions on coronary artery disease prediction using machine learning algorithms: A survey

Anu Ragavi Vijayaraj, Subbulakshmi Pasupathi
School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, India

Article Info

Article history:

Received Feb 9, 2023
Revised May 25, 2023
Accepted Jun 3, 2023

Keywords:

Classification
Clinical variables
Coronary artery disease
Machine learning

ABSTRACT

Coronary artery disease (CAD) is the most ubiquitous and protuberant cause of fatal death. The hit in mortality rate is because of certain lifestyle variables including unhealthy diet, usage of tobaccos and drugs, physical inactivity, and environmental pollution. Traditional screening tests including computed tomography, angiography, electrocardiography, and magnetic resonance imaging are employed for diagnosis and would necessitate more manpower. Machine learning (ML) has been utilized in healthcare to create early predictions from massive volumes of data. The Scopus, Web of Science databases were exhaustively searched utilizing a search strategy that comprised CAD prediction, cardiac illness detection, and heart disease categorization. After applying the inclusion and exclusion criteria to the 99 articles obtained, the population of the study was composed of 30 articles. This review study offers an organized look at the articles published in ML-based CAD detection and classification models that include clinical variables. The use of ML could produce amazing results in CAD detection, as evidenced by the classifiers random forest, decision tree, and k-nearest-neighbour with accuracy being >90%. The use of ML in CAD diagnosis lowers false-positive, and false-negative errors, and presents a special opportunity by providing patients quick, and affordable diagnostic services.

Corresponding Author:

Subbulakshmi Pasupathi
School of Computer Science and Engineering, Vellore Institute of Technology
Chennai, Tamil Nadu, India
Email: subbulakshmi.p@vit.ac.in

ABBREVIATIONS

LR	: Logistic regression	DNN	: Deep neural network
SVM	: Support vector machine	FNN	: Fuzzy neural network
KNN	: K nearest neighbor	ADB	: AdaBoost
DT	: Decision tree	XGB	: Extreme gradient boosting
RF	: Random forest	SMO	: Sequential minimal optimization
RT	: Random tree	MLR	: Multinomial logistic regression model
NB	: Naïve Bayes	SVC	: Support vector classifier
GB	: Gradient boosting	PSO	: Particle swarm optimization
GNB	: Gaussian Naive Bayes	CHAID	: Chi-squared automatic interaction detection
NN	: Neural network	NFC	: Neuro fuzzy classifier
GA	: Genetic algorithm	CART	: Classification and regression trees
MLP	: Multi layer perceptron	CNN	: Convoilutional neural networks
ANN	: Artificial neural network		

1. INTRODUCTION

People are worried about their hectic schedule in day to day lives and get addicted to drugs and tobaccos as stress relievers. Ultimately the individuals grow with obesity parallelly. This piles up with serious life threading disorders namely heart issues, cancer, tuberculosis and many more. The most challenging task is to predict them before it gets worse. Cardio vascular diseases (CVD) are one such life threading disease accounting 31.8% of all global deaths, according to the recent World Health Organisation (WHO) statistics [1] as depicted in Figure 1. A rise in mortality rate can be avoided if it is predicted in advance and proper lifestyle choices are made.

There are four main types of CVD's; a stroke, also known as a brain attack, happens when blood supply to a portion of the brain is cut off or when a blood artery in the brain bursts. Peripheral artery disease (PAD) is a common disorder characterized by constricted arteries that restrict blood flow to the arms or legs. Heart valve disorders include the aortic disease. Aortic valve dysfunction results in improper operation of the valve between the left ventricle, the lower left heart chamber, and the aorta, the main artery to the body. The last type is coronary heart disease, commonly referred to as coronary artery disease (CAD) and ischemic heart disease (IHD), which is caused by an accumulation of plaque deposits in the arteries that obstruct the heart's blood flow.

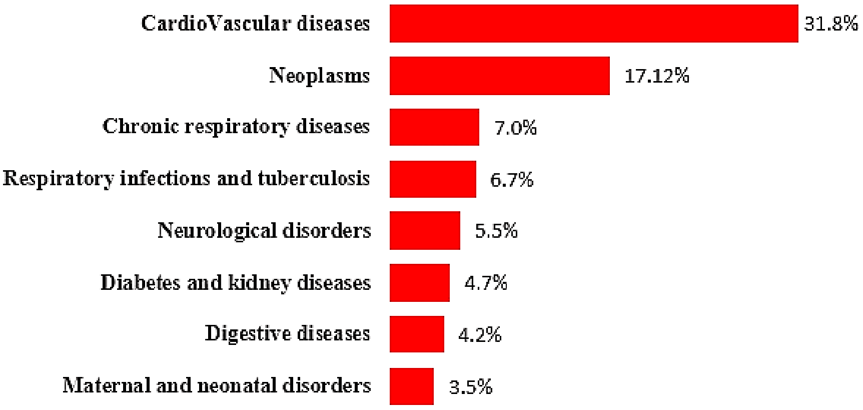


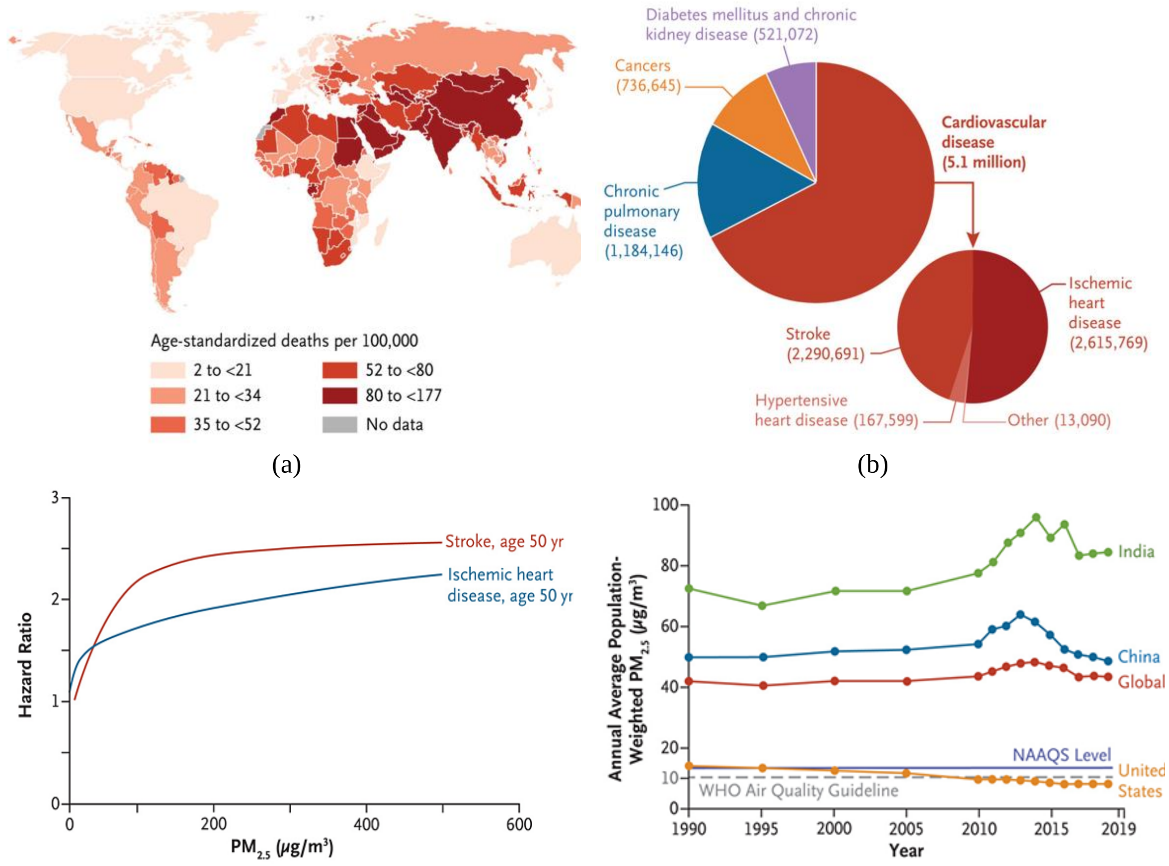
Figure 1. Global causes of death [1]

In accordance with American Heart Association, IHD was predicted to affect 244.1 million persons worldwide in 2020, and it was more common in men than in women (141.0 and 103.1 million people, respectively) [2]. The small arteries in the heart muscles are affected by coronary microvascular disease, another type of heart disease. In women, coronary microvascular disease is more prevalent.

Although doctors have not provided a definitive explanation for why CAD develops, risk factors are a significant contributing element. It includes obesity, usage of drugs and tobaccos, cholesterol, and also because of family history. In addition to all these risk variables, the environmental pollution stands the foremost for the development of any disease. The small particles from the polluted air can affect the heart and blood vessels.

According to the global burden of disease (GBD) report, pollution caused 9 million deaths worldwide in 2019, with cardiovascular disease, including IHD (31.7%), and stroke (27.7%), accounting for 61.9% of all fatalities. This data underscores the substantial impact of pollution on mortality worldwide, particularly its strong association with heart-related health issues. Figure 2 shows Cardiovacular disease impact on pollution, Figure 2(a) displays the age-standardized number of fine particulate matter (PM2.5) related deaths per 100,000 people in 2019, Figure 2(b) displays the total number of noncommunicable disease-related deaths caused by pollution worldwide in 2019, Figure 2(c) displays the annual mean population-weighted PM2.5 concentrations in China, India, and the United States from 1990 through 2019 and Figure 2(d) displays the model for the exposure-response relationships between CVD and PM2.5 air pollution in a 50-year-old person.

The three types of arteries are right coronary artery (RCA), left circumflex artery (LCX), and left anterior descending artery (LAD). CAD is caused when blood flow to the heart stops partially or completely because of the plaque which develops in the arteries. The plaque narrows the arteries blocking the oxygen-rich blood flow to the heart as shown in Figure 3. Because of the reduction in oxygen supply to the heart, chest pain, shortness of breath, and heart attack are caused.



(a) (b)

(c) (d)

Figure 2. Cardiovascular disease impact on pollution [3]: (a) worldwide mortality from cardiovascular disease associated with air pollution, (b) deaths caused by global pollution, (c) annual mean PM$_{2.5}$ Pollution levels (1990-2019), and (d) cardiovascular disease and PM$_{2.5}$ Pollution: an exposure-response connection

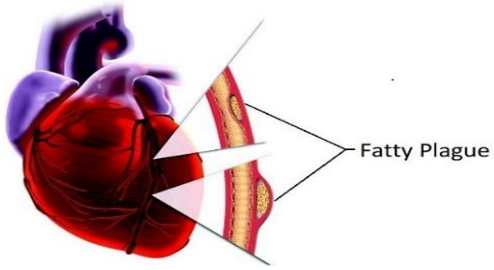


Figure 3. Coronary artery disease [4]

Medical experts recommended a variety of expensive and technically complex tests, such as electrocardiograph (ECG), angiography, computed tomography, and magnetic resonance imaging to diagnose individuals with positive signs of CAD. Currently, these examinations are high-priced and involves methodical professionals. In order to replace the aforementioned standard tests, researchers are striving to develop a less expensive yet equally effective test.

Machine learning (ML) and data mining methods are widely used in the analysis and extraction of information from medical data. ML-based techniques are effective in a variety of disciplines, including agriculture, credit card fraud detection, speech recognition, and is frequently recommended for predicting heart disease since it extracts more efficient and accurate data from large datasets, making predictions straightforward. It is the primary foundation of ML, assisting in the management of large volumes of data, having a fast-processing speed, and generating predictions in the early phases of development.

There exist two different types of studies that can be found in the literatures. Some research employed clinical indicators to categorize CAD patients, including age, blood pressure, smoking history, while other studies used signal recordings like electrocardiograph (ECG), photoplethysmography (PPG), and phonocardiography (PCG) to identify CAD symptoms. Consequently, in order to direct the evaluation of future works, our study concentrates on the workflow on the clinical factors. The remainder of the paper is structured: section 2 highlights ML for CAD diagnosis. Sections 3 and 4 discusses about data collection and classification. Finally, sections 5 and 6 discourses the results and conclusion.

In ML enables computers to learn and develop without being explicitly programmed with little or no human intervention. Both scientists and medical professionals are looking for affordable, precise, and quick CAD diagnosis and treatment options. In order to help researchers better address numerous difficulties in their future work, this review paper will highlight the strongest findings of earlier studies. This study is added.

2. ML FOR CAD DIAGNOSIS

Artificial intelligence (AI) has a substantial impact on the diagnosis of heart disease through the analysis of medical data. AI facilitates accurate forecasting of cardiovascular results and the non-invasive identification of CAD. A branch of AI called machine learning ressed:

- General flow diagram of the prediction model
- Data acquisition-CAD prediction
- Review of the articles on CAD prediction using machine learning algorithms (Back ground study)

2.1. General flow diagram of the prediction model

Figure 4 is the general heart disease prediction model. The patient details include age, cholesterol, sugar, blood pressure, are maintained in the hospital’s patient database. With the patient database, the data is collected and handled to the next phase for pre-processing. With the preprocessed data, the best classification method is identified. In this layer, we also train our model using the processed dataset, and prepare it for the next layer. We test the model we trained in the final phase, thus after dividing the dataset into training and testing data. The testing part of the dataset is used to assess the algorithm's classification accuracy.

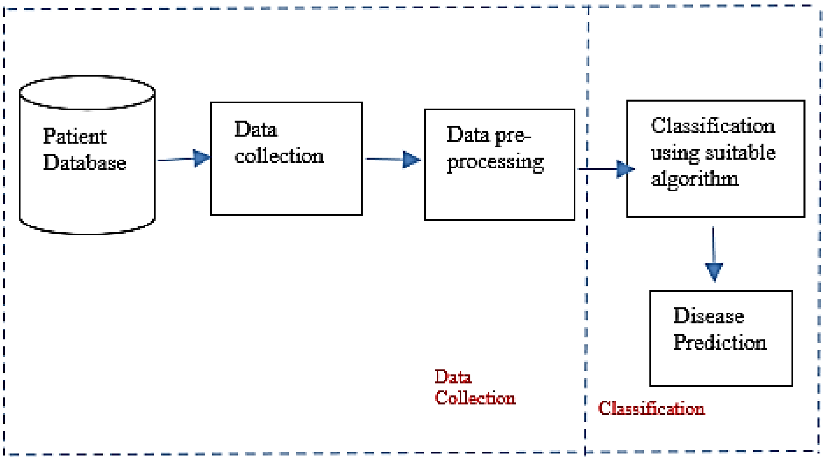


Figure 4. Heart disease prediction model

2.2. Dataset

2.2.1. Kaggle database

Four databases make up the 1988 created Kaggle heart disease dataset: Cleveland, Hungary, Switzerland, and Long Beach V. Despite having 76 features, including the anticipated attribute, only 14 of them are used in all reported studies. The "target" field is concerned with the patient's cardiac status. It takes the integer values 0 to indicate no disease and 1 to indicate disease [5].

2.2.2. UCI database

The UCI heart disease databases includes Cleveland with 303 instances, Hungarian with 294 instances, Switzerland with 123 instances, and Long Beach VA with 200 instances. There exist 76 features in the UCI database and all published experiments employ a section of 14 of them [6]. Cleveland database in particular is commonly employed by ML investigators. The "goal" attribute indicates the patient's possessing heart disease and it has integer values ranging from zero (No presence) to four. Experimentations carried out with Cleveland database focused mainly to distinguish between the disease's existence (Values 1, 2, 3, 4) and absence (Value 0).

2.2.3. Z-Alizadeh Sani database

The 303 patient records that makes up the Z-Alizadeh Sani dataset have 56 features each. The characteristics are divided into four categories: demographic, exam and symptom, ECG, laboratory and echo aspects. Every patient could be classified as either CAD or normal. If a patient's diameter narrowing exceeds 50%, they are classified with CAD; otherwise, they are considered to be normal [7].

2.2.4. Indira Gandhi Medical College database, Shimla, India

The dataset collected in real time from Indira Gandhi Medical College contains 335 instances of patients. All of the participants agreed to undergo an Angiography after developing CAD suspicions. Each patient had 27 characteristics, including historical, demographic, and laboratory information [8].

2.2.5. General hospital, Nigeria database

The dataset includes 506 instances with 12 features from two general hospitals in Kano, Nigeria. Eleven clinical risk factors for CAD are present in the dataset, along with one demographic component. The binary classification with integer values 0 for no disease and 1 for disease is the “goal” variable [9], [10].

2.2.6. Extension of Z-Alizadeh Sani dataset

This dataset is Z-Alizadeh Sani's extended version. The main difference between Z-Alizadeh Sani and extension Z-Alizadeh Sani dataset is that the extended dataset not only helps in the classification of CAD but also the stenosis of the arteries is predicted with the 3 added features LAD, LCX, and RCA to the existing dataset [7] The list of datasets utilized for CAD diagnosis by the researchers and authors is attached in Table 1.

Table 1. The list of datasets for CAD diagnosis

Dataset No.	Dataset Name	Sample No.	Input features Features No.	Stenosis
1	Heart disease dataset-Kaggle	1026	14	
2	UCI repository	920	14	
3	Z-Alizadeh Sani	303	56	
4	Extension of Z-Alizadeh Sani	303	59	✓
5	Indira Gandhi Medical College Database, Shimla, India	335	25	
6	General hospital, Nigeria	506	18	

2.3. Review of the articles on CAD prediction using machine learning algorithms (Back ground study)

Mostly CAD detection systems employ supervised ML techniques. There have also been few reports of clustering algorithms being used for CAD diagnosis. The list of articles published using ML for CAD diagnosis is depicted in Table 2 (see appendix).

Alizadehsani *et al.* [7] proposed a CAD diagnosis model which calculates the stenosis of each vessel. The effect of features on these three vessels was evaluated using the information gain. ML classifiers including C4.5, NB, and KNN were applied on the new features added to the Alizadeh dataset. C4.5 reaches the highest accuracy with 74.20%, 63.76%, and 68.33% for LAD, LCX, and RCA vessels, respectively. In a study by Garavand *et al.* [39] the efficiency of the various ML classifiers MLP, SVM, LR, J48, RF, KNN, and NB in predicting CAD was compared. The most effective algorithms for diagnosing CAD from patient examination data were SVM and RF.

With the survey carried out in Table 2, all the papers used public datasets including UCI repository, Z-Alizadeh Sani dataset. Only few researchers utilize real world datastore. The limitation of the Cleveland dataset is that few instances are found to be missing and suitable missing value imputation methods are incorporated to eliminate missing values from the dataset. Unlike other datasets, Cleveland dataset is used for multilevel classification problems where target variable indicates the level of disease ranging from 0 to 4.

Some studies [40]–[42] on CAD diagnosis achieved 90%, 70%, and 75% accuracy, respectively. However, they did not assess the stenosis of each vessel independently. Researchers can employ extension of Z-Alizadeh dataset for stenosis arteries. 37 features of the extension of Z-Alizadeh dataset were examined and the stenosis of these arteries was significantly influenced by the features age and typical chest pain.

Researchers generally employ performance estimators including accuracy, sensitivity, specificity, F1 score as shown in (2), (3), (4), and (5). Table 2 makes it clear that accuracy was the selection criterion used by all authors. As additional selection criteria, [11], [15], [21], [28], [31], [37], [38] used F1 score, sensitivity, and specificity.

3. DATA COLLECTION

Collecting data is the foremost step in ML pipeline. It is the process of gathering, measuring, and analysing information gleaned from a profusion of diverse sources. The information gathered is utilised to create ML and AI solutions. Data collection includes preprocessing, feature extraction, and selection.

3.1. Data preprocessing

Four stages make up preprocessing, which is done to ensure high-quality data. In order to produce an accurate result, cleaning involves removing noisy and missing values from the dataset. Once the data has been cleaned of noise and missing values, it is translated into a different format without changing the contents of the datasets by transformation. It involves aggregation, standardization, and smoothing. The process of merging data from numerous sources into a single database by integration. For the gathered data to give relevant findings, it must be structured, which is known as reduction.

3.2. Dimensionality reduction

Feature extraction is a dimensionality reduction procedure that reduces an initial collection of raw data to more manageable groups for processing. Algorithms for linear transformations that are often utilized include principal component analysis. It looks for mutually orthogonal directions in the feature space as well as directions that maximize variance.

Feature selection (FS) eliminates the redundant and irrelevant data by increasing the accuracy and provide a better understanding of the model. By choosing the most prominent features, a unique fast conditional mutual information feature selection technique (FCMIM) improves accuracy [25]. The algorithm is feasible with the classifier SVM in order to detect cardiac problems. For the features that less contributes for the improvement in the system, Ali *et al.* [17] An optimally configured and improved deep belief network (OCI-DBN) approach for heart disease prediction based on Ruzzo-Tompa and stacked genetic algorithm a novel feature selection algorithm Ruzzo-Tompa which eliminates the irrelevant features from the dataset. With the selected features, an optimally configured and Improved deep belief network is created and the accuracy is improved up to 94.61%.

With the feature selection techniques namely MLR, and sequential feature selection (SFS), age, slope, exang, fluoroscopy, and thalach are the features selected from [34] NFC with the feature selection method MLR (MLR+NFC), attains the accuracy of 84% than with SFS+NFC. With suitable feature selection through PSO [32], the attributes are selected from IGMC, Shimla. The selected attributes include smoking, diabetes mellitus, high density lipoprotein. These attributes are then used with the classification algorithms and MLR achieved an accuracy of 84.17% by properly identifying the wrong instances.

4. CLASSIFICATION

4.1. Machine learning algorithms

ML algorithms discover hidden patterns in data, anticipate outcomes, and enhance performance based on their own experiences. These algorithms and models are intended to learn from data and generate predictions or choices in the absence of explicit instructions. After selecting the attributes from suitable feature selection techniques, the selected feature subset along with classifier results are compared.

4.1.1. K-nearest neighbor

Hodges and Evelyn introduced the KNN rule, a nonparametric technique for classification and regression, in 1951. KNN is a straightforward but efficient classification method where little to no prior knowledge about the distribution of the data is available since it makes no assumptions on the data. The strategy entails locating the k data points in the training set that are most similar to the data point for which a target value is missing and assigning the average value of those data points to the missing data point.

A single diagnostic method for the prediction of 3 cardiac abnormalities namely CAD, myocardial infarction, and congestive heart failure are developed by Acharya *et al.* [15] using ECG. Additionally, it pinpoints the precise cardiac abnormalities seen in patients during an ECG test, obviating the need for other diagnostic techniques. This non-invasive, cost-effective method can be further extended in near future by detecting the cardiac abnormalities in the early phase by using a single ECG pulse.

4.1.2. Logistic regression

LR is a supervised method for binary problems either true/false, yes/no, pass/fail. The independent variables can be categorical/numerical whereas the dependent variable is always categorical. The application of LR includes credit scoring, predicting user behavior, and discrete choice analysis. By removing the insignificant features from the Cleveland dataset, the optimal attributes were chosen using MLR, and SFS. With the selected attributes, a novel NFC is proposed by Marateb and Goudarzi [34] performance is achieved when integrated NFC with the MLR with 84% than SFS+MLR.

4.1.3. Naïve Bayes

NB uses Bayes' theorem for classification problems. Application of NB includes medical diagnosis, spam filters, text analysis. Tarawneh and Embarak [23] created a hybrid model for CAD diagnosis with the selected 12 features, and different classification algorithms including SVM, KNN, GA, RF, NN, and J4.8 is implemented. NB, SVM shows good performance results with an accuracy of 89.2%. The following is the formula for Bayes' theorem.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{1}$$

$P(A|B)$ = Posterior probability of class given predictor
$P(B|A)$ = Likelihood is the probability of predictor given class
$P(A)$ = Class prior probability
$P(B)$ = Predictor Prior probability

4.1.4. Decision tree

It is a supervised approach that solves prediction and classification problems. DT is a tree-like structure with the nodes indicating the test on the attribute, branch nodes expressing the test's outcome, and leaf nodes providing the classification label. With the path created from root to the leaf node, DT can be easily transformed into a set of rules. Finally, appropriate conclusions are reached by following these rules. Applications include sentiment analysis, and products selection.

Fuzzy expert system for the prediction of CAD is developed by Muhammad and Algehyne [10] which includes knowledge base, inference engine, and defuzzification phases. 87 diagnostic rules are framed in the knowledge base. Instead of using traditional methods like interviews, questionnaires, the upgraded C4.5 is used to incorporate human knowledge into the system's knowledge base. C4.5 based fuzzy expert system attained an accuracy of 94.55%. With the real-world dataset collected from Indira Gandhi Medical College, Shimla, the data set is evaluated for missing values using a hybrid and new k-means cluster centroid-based method, and C4.5, NB Tree, and MLPs are utilized to predict CAD. When compared with the other predictive models, C4.5 constructed with 25 features yielded the highest accuracy, sensitivity and specificity [8].

With the Z-Alizadeh dataset, CART model is created by calculating the feature importance score of the features. A typical feature attains the highest feature score of 16.5%. Features which attained with 0% importance is neglected for classification [28]. Idris *et al.* [33] created an embedded method DT using RF with the features selected from Gini impurity. 20 features were used for the classification of CAD. The model of NN with Embedded DT features acquired the highest precision and accuracy 94.5%. University Of Malaya Medical Centre (UMMC) and subarachnoid hemorrhage (SAH) dataset is compared on the performance metrics.

4.1.5. Random forest

RF is a technique for solving regression and classification problems. A random forest model, as an ensemble approach, is built from a number of small DTs or estimators, each of which provides its own predictions. The estimators' estimates are combined by the random forest model to get a prediction that is more accurate. The 'forest' in RF is trained using bootstrap or bagging aggregation. Applications include credit card default, stock market prediction, and product recommendation.

According to Muhammad *et al.* [11], the DT created using the RF approach was the most effective model in terms of accuracy and receiver operating characteristic (ROC). With this combined DT from RF model, production rules are created, and this expert system diagnosed CAD victims very accurately at the rate of 92.04% in Nigeria. According to Jinny and Mate [21], it is evident that, of all the classifiers, by using the most features possible for the classification model, the RF classifier was successful in getting an accuracy of about 91%.

Hybrid RF with a linear model–HRFLM proposed by Mohan *et al.* [24] utilized with 11 attributes and achieved an accuracy of 88.7% when applied to the UCI heart dataset. Rajdhan *et al.* [36] developed a ML model that contrasts four different methods using the UCI Cleveland dataset. Examining the classification accuracy of RF, LR, NB, and DT algorithms, it was found that the RF approach had an accuracy rate of 90.16%. Proposed method by Joloudari *et al.* [16], data mining techniques SVM, CHAID, C5.0, and RTs were implemented on the Z-Alizadeh dataset. Random trees are the best amongst the other with 91.47% accuracy using the 10-fold validation. The following metrics were examined and evaluated: accuracy, AUC, Gini, return on investment (ROI), profit, confidence, response, and gain.

The Cleveland heart dataset was used to train classifiers NB, SVM, LR, RF, and Adaboost [35]. A comparison of the results without feature selection and with feature selection is made. 8 features selected from the hybrid model yielded the highest accuracy results.

4.1.6. Support vector machine

SVM is a type of supervised learning approach that categorizes data with the hyperplane. The new data is assigned to the appropriate category by the hyperplane, which then categorizes the n-dimensional space. The data points which is very close to the hyper plane is called support vectors. SVM maximizes the margin to reduce any chance of misclassification. The application includes face detection, image classification. The algorithm can classify both linear and non-linear data.

Li *et al.* [25] contrasted various feature selection methodologies, such as relief, maximum relevance minimum redundancy, least absolute shrinkage and selection operator, LLBFS FS algorithms, and presented a novel feature selection methodology, FCMIM. SVM is utilized as a classifier that combines with FCMIM for the prediction of cardiac illnesses among LR, KNN, ANN, NB, and DT. Abdar *et al.* [26] created a novel ML methodology for accurate CAD prediction. The Z-Alizadeh Sani dataset is preprocessed and normalized. The GA and PSO algorithms are employed in the selection of the features. The classification algorithms including NB, generalized linear models (GLM), LR, DL, DT, RF, GBT, and 3 types of SVM namely C-SVC, nu-SVC, and linear SVM were also tested on the dataset. The N2Genetic-NuSVM method has the highest accuracy (93.08%) and F1-score rate (0.9151).

4.1.7. Artificial neural network

A NN that mimics the human brain, also known as an artificial neural network. Interconnected neurons can be found in the human brain. Similar to this, the neurons in ANN are arranged in several layers and connected to one another. The various layers include input layers, hidden layers, and the output layers.

Kahramanli and Allahverdi [19] created a hybrid system for diabetes and heart disease prediction. Using the UCI dataset, ANN and FNN hybridization has been implemented. FNN1, FNN2, FNN3 are the three types of FNN. While the weights are fuzzy, the inputs to the FNN1 are crisp values and vice-versa with FNN2. Both the input and the weights for FNN3 are fuzzy values. A hybrid system is created by combining a FNN2 and an ANN trained with the backpropagation. The model shows 84.2% accuracy for Pima Indians database and 86.8% for Cleveland database.

A computer-aided system called heart disease prediction system (HDPS) [29] was created using the C and C# environments to forecast heart disorders. With the 13 clinical features, ANN classifies with a degree of precision of 80%. HDPS interface is created with data input panel, ROC curve display section and performance display section. A low compact sensor is designed by Dixit and Kala [18] that records the ECG signals and the preprocessed signal is then segmented. With each window, suitable features are extracted and the classification is made. Finally fusion of the windows is done by 1D CNN model with an accuracy of 93%.

4.1.8. Fuzzy logics

Fuzzy logics is a multi-valued logic in which variables takes the value between 0 and 1. A Fuzzy expert system-based prediction [37] focused on modules, meta-rules, and consistency checks in the rule base for improved rule organization. In the current expert system, a specific emphasis has been placed on effective rule organization methods. Because the patient may be unaware of the values of all clinical parameters, the various combinations of the criteria are created including age, blood pressure, cholesterol. In the rule base, two consistency checks namely i) contradictory rule checking and ii) redundant rule checking is defined for the improvement in results.

When diagnosing CAD using a fuzzy expert system based on PSO, the membership functions are optimised using PSO, and a fuzzy rule basis is produced using the optimised membership functions. With the fuzzy rule base, Mamdani inferenceing is implemented that yielded the highest accuracy of 93.27% with the Cleveland and Hungarian dataset [20]. The fuzzy system inferenced with Mamdani approach predicted the heart disease with the rules generated by C4.5. The highest accuracy of 94.55% is achieved by Muhammed and Algehyne [10].

4.1.9. Other learning methods

With the two-level stacking, Wang *et al.* [30] utilized the enumeration algorithm to identify the best classifiers with the Z-Alizadeh dataset. A novel hybrid dataset ‘Sathvi’ [38] by integrating public health dataset is created. The objective of this hybrid dataset is to make datasets free from noise. The attributes ‘ca’, ‘thal’ were eliminated from the dataset. With the comparison of the classifiers taken for this research, CatBoost classifier outperformed the other classifier with the accuracy from 88% to 98.11%.

4.2. Performance measurements

The classification of heart disease data requires the use of several supervised ML algorithms. The categorization models were assessed using the eight quality factors such as true positive (TP), false positive (FP), true negative (TN), false negative (FN), accuracy, specificity, sensitivity, and F1 score. These performance criteria for the categorization analysis were looked into.
TP – No. of. victims with presence of heart disease predicted as presence of heart disease.
FP – No. of. victims with absence of heart disease predicted as presence of heart disease.
TN – No. of. victims with absence of heart disease predicted as absence of heart disease.
FN – No. of. victims actually have presence of heart disease predicted as absence of heart disease.

An accuracy score indicates how effectively a model performs. It is calculated as the sum of TPs and TNs, divided by the sum of TPS, FPS, TNS, and FNS. The formula is:

$$Accuracy = \frac{(\mathrm{TP+TN})}{(\mathrm{TP+FP+TN+FN})} \tag{2}$$

following accuracy, specificity is a measure of negative cases recognized as negative by the classifier. The formula is:

$$Specificity = \frac{(TN)}{(TN+FP)} \tag{3}$$

the proportion of cases that were actually positive but were predicted to be positive is known as sensitivity. Another name for sensitivity is recall. To put it another way, an unhealthy person was predicted to be unhealthy. The formula is:

$$Sensitivity = \frac{(TP)}{(TP+FN)} \tag{4}$$

the harmonic mean of precision and recall is called F1 measure. The value is 1 for the finest performance and 0 for the worst. The formula is:

$$F1 = \frac{2(precision*recall)}{precision+recall} \tag{5}$$

5. RESULT ANALYSIS

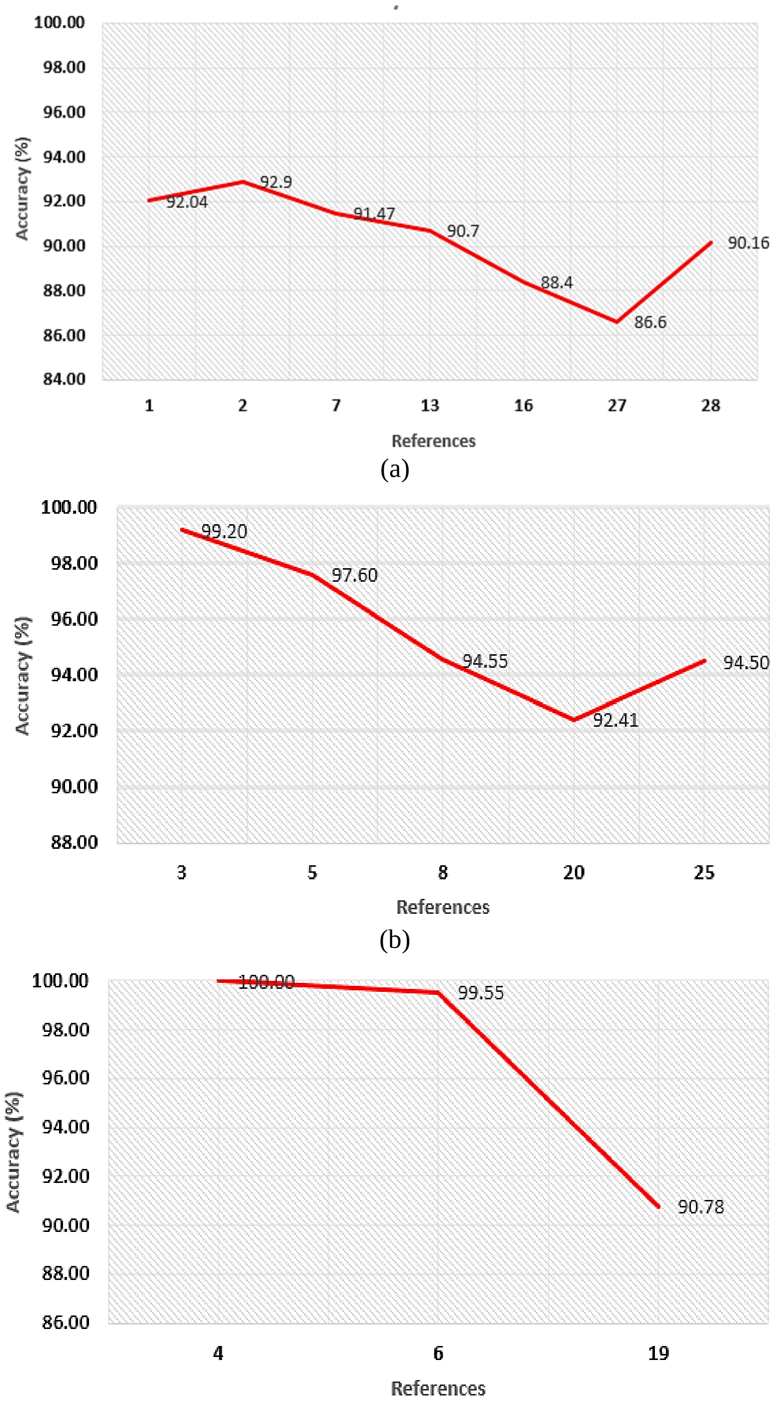
In this review, traditional classifiers and boost classifiers applied on the public datasets was studied. RF, DT, and KNN outperforms state-of-the-art techniques, and achieved accuracy of 100% using KNN algorithm. The performance evaluation results from the literature are shown in Figures 5(a)-5(c). The performance of all 3 classifiers resulted above 85%.

With regard to the articles examined in Table 2, 7 references with Sr.No [1], [2], [7], [13], [16], [27], [28] exhibit the best results for RF, and the authors used, correspondingly, 5, 6, 6, 12, 7, 4, and 5 studies for comparison. 5 references with Sr.No [3], [5], [8], [20], [25] showed best results for DT, and the authors used 4, 7, 5, 34, and 6 studies for comparison. 3 references with Sr.No [4], [6], [19] showed the best results for KNN where 2, 29, and 6 studies are compared respectively. Figure 5 shows the performance comparison of different CAD prediction models, Figure 5(a) shows the accuracy achieved by RF, Figure 5(b) shows the accuracy achieved by DT, and Figure 5(c) shows the accuracy achieved by KNN. With RF, the maximum attained accuracy is by Sr.No [2] with 92.90%. With DT, Sr.No [3] showed 99.2% accuracy. 100% accuracy is attained by Sr.No [4] with KNN and proved to be one of the best classifiers in CAD prediction.

5.1. Challenges and research directions

ML is frequently employed to solve categorization problems in the health care sector. Our study on ML algorithms opens up several research issues especially in healthcare. We observe that ML research is actively taking place in the field of cardiology, with some intriguing proofs-of-concept, and proprietary solutions being developed by the research community, healthcare industries respectively.

In general, the nature and quality of data, in addition to the quality of learning algorithms, will regulate a machine learning-based solution's success and efficiency. It is challenging to gather real world data in the pertinent fields, such as agriculture, IoT, healthcare. Therefore, a more thorough investigation of data collection is required. Also, it is a challenging task to accurately pre-process the data collected from different sources. Therefore, to effectively employ the learning algorithms, it is required to modify or expand existing pre-processing methods or to suggest new data preparation strategies. The hybrid learning model, for instance, the ensemble of methods, the modification or refinement of the current learning approaches, or the construction of new learning methods, could be a potential future work in classification and prediction of heart related disease.



(a)

(b)

(c)

Figure 5. Accuracy of CAD prediction models (a) accuracy of random forest classifier, (b) accuracy of decision tree classifier, and (c) accuracy of k-nearest neighbor classifier

6. CONCLUSION

ML algorithms have enormous potential in predicting heart-related ailments. Each of the aforementioned algorithms performed excellently in certain cases while failed terribly in others. In order to do this, we searched extensively across a number of search engines and databases. From 2006 to 2022, the most significant research on CAD diagnosis using ML algorithms was done, according to our review. The outcomes also shows that the most popular CAD detection methods are RFs, DTs, KNN, SVMs, ANNs, Fuzzy logics, and NB. Because of the inherent variety of datasets and ML algorithms, different performance metrics have been documented. According to the results, RF, DT, and KNN have the highest accuracy levels for the majority of the CAD datasets. Additionally, hybridizing classifiers and feature selection can enhance performance for precise CAD diagnosis. With the same methodology and dataset, the studies can be extended to other heart-related diseases. With the proper utilization of the real-time datasets, the studies can be improvised.

APPENDIX

Table 2. The list of articles published using ML for CAD diagnosis

Sr. No	Author	Year	Techniques	Conclusion	Dataset	Observations
1	Muhammad *et al.* [11]	2021	LR, SVM, KNN, RF, NB, GB	RF Accuracy = 92.04%	General Hospital, Nigeria	Results from various ML algorithms were obtained, and they were compared.
2	Yilmaz *et al.* [12]	2022	RF, SVM, LR	RF Accuracy = 92.9%	IEEE DataPort	Three traditional ML approaches were employed, and RF is used as a best classifier for the prediction of CHD.
3	Soni *et al.* [13]	2011	DT with GA, NB, Classification via clustering	DT with GA Accuracy = 99.2%	NR	With GA, the optimal subset of features is retrieved, and DT, Bayesian classification shows improvement in results.
4	Jabbar *et al.* [14]	2013	KNN, J48, NB	KNN Accuracy = 100%	UCI Repository	Optimal feature selection method, along with KNN classifier is compared with other datasets. Classifier hybridization with feature selection can best choose the features, enhancing accuracy.
5	Verma *et al.* [8]	2017	DT, MLP, NB tree	C4.5 Accuracy = 97.6% Sensitivity = 97.5% Specificity = 97.6%	Indira Gandhi Medical College, Shimla, India	Three predictive data mining approaches were compared.
6	Acharya *et al.* [15]	2018	DT, KNN	KNN Accuracy = 99.55% Sensitivity= 99.93% Specificity= 99.24%	St.-Petersburg Institute of Cardiological Technic (ECG)	With the selected 20 features from the ECG signal, KNN achieved the highest accuracy from contourlet coefficients.
7	Joloudari *et al.* [16]	2020	SVM, CHAID, C5.0, and RT	RT Accuracy = 91.47%	Z-Alizadeh Sani heart disease dataset	Different classifiers were experimented and RT reports best in Accuracy, (area under the curve) AUC, Gini. The most important rules were extracted using the random trees model for CAD diagnosis.
8	Muhammad and Algehyne [10]	2021	C4.5 Fuzzy Expert system	Fuzzy Logics Accuracy = 94.55% Sensitivity=95.35% Specificity = 95.0%	General Hospital, Nigeria	The C4.5 data mining algorithm is used to incorporate human expertise into the knowledge base of the system.
9	Ali *et al.* [17]	2020	OCI-DBN, ANN, DNN	OCI-DBN using SGA Accuracy = 94.61% Sensitivity= 96.03% Specificity= 93.15%	Cleveland heart disease dataset	Ruzzo-Tompa method eliminates features that less contributes to system performance. Stacked genetic algorithm is used to build the best-configured DBN.
10	Dixit and Kala [18]	2021	RF, GB, Deep Learning models (1D CNN)	1D CNN Accuracy = 93.0%	Swaroop Rani Nehru Hospital-Allahabad,	Oversampling with 1D CNN and voting strategy emerges as the suitable classification technique with 93% accuracy.

Table 2. The list of articles published using ML for CAD diagnosis *(Continue)*

Sr. No	Author	Year	Techniques	Conclusion	Dataset	Observations
11	Kahramanli and Allahverdi [19]	2008	ANN, FNN on 2 datasets	HNN (Pima dataset) Accuracy= 84.2% Sensitivity=80.3% Specificity 87.3% HNN (Cleveland dataset) Accuracy= 87.4% Sensitivity = 93.0% Specificity 78.5%	Pima Indians diabetes and Cleveland heart disease	The databases cover Cleveland heart disease and Pima Indian diabetes. The categorization accuracy of these datasets was assessed using k-fold cross-validation.
12	Muthukaruppan *et al.* [20]	2012	Fuzzy expert system	PSO based Fuzzy expert system Accuracy = 93.27%	UCI Repository	Cleveland and Hungarian Heart Disease dataset from UCI repository. DT was employed to identify the characteristics that influence the diagnosis. To fine-tune the fuzzy membership functions, PSO was used. The optimized MFs yielded the highest accuracy.
13	Jinny and Mate [21]	2021	DT, RF, AdaBoost, GNB, LR, KNN, GB, XGB	RF Accuracy = 90.7%	Framingham's Dataset	Performance comparison of both conventional ML methods, and cutting-edge Gradient Boosting techniques with feature selection and without feature selection is made.
14	Tiwari *et al.* [22]	2022	ET, RF, XG, RF, MLP, KNN, XGB, SVM, SGD, AdaBoost, CART, GB, NB	Stacked ensemble classifier Accuracy = 92.34%	IEEE Data Port	Hungarian, Cleveland, Long Beach VA, Switzerland, and Statlog datasets were combined into a one dataset. The stacked ensemble classifier implemented achieved highest accuracy of 92.34% than the other classifiers.
15	Tarawneh and Embarak [23]	2019	NB, SVM, KNN, NN, J4.8M, RF, GA	NB, SVM Accuracy = 89.2%	Cleveland heart disease dataset	A variety of ML classifiers were deployed on Cleveland datasets to predict cardiac disease, and NB, SVM shows improvement in accuracy when compared to other classifiers.
16	Mohan *et al.* [24]	2019	NB, LR, DL, DT, RF, GB Tree, SVM, VOTE, HRFLM	HRFLM Accuracy = 88.4% Sensitivity=92.8% Specificity=82.6%	UCI Repository	With the DT features, the dataset has been clustered. The classifiers performance is then estimated by applying them to each clustered dataset.
17	Li *et al.* [25]	2020	LR, KNN, ANN, SVM, NB, DT	FCMIM – SVM Accuracy = 92.37% Specificity = 98.0% Sensitivity = 89.0%	Cleveland heart disease dataset	The feature extraction process used Fast Conditional Mutual Information, which improved prediction accuracy with SVM.
18	Abdar *et al.* [26]	2019	SVC, nuSVM, LinSVM	N2Genetic-nuSVM Accuracy = 93.08%	Z-Alizadeh Sani heart disease dataset	Several ML techniques were explored, and three different forms of SVM were deployed, along with feature tuning for improved accuracy.
19	Shah *et al.* [27]	2020	NB, KNN, DT, RF	KNN Accuracy = 90.78%	UCI Repository	Four traditional ML approaches were employed, and KNN is used as a classifier for the prediction of CAD.
20	Ghiasi *et al.* [28]	2020	Bagging, SMO, Bagging SMO, NB, C4.5, J48, SVM, ANN, ANN-GA, CART.	CART Accuracy = 92.41%	Z-Alizadeh Sani heart disease dataset	The important and non-important features were categorized and the CART model for CAD diagnosis was created using those important features. CART SMO and Bagging SMO models, ANN-GA model gives the accurate results but CART reaches the highest accuracy.
21	Chen *et al.* [29]	2011	ANN	ANN Accuracy = 80.0%	UCI Repository	A user-friendly ANN based Heart Disease Prediction System (HDPS) is developed, and reliably predicts outcomes.

Table 2. The list of articles published using ML for CAD diagnosis *(Continue)*

Sr. No	Author	Year	Techniques	Conclusion	Dataset	Observations
22	Wang *et al.* [30]	2020	LR, RF, GNB, SVC, DT, KNN, ADB, GB, MLP, XGB, Stacking model	Stacking model Accuracy = 95.43% Sensitivity=95.84% Specificity=94.44%	Z-Alizadeh Sani heart disease dataset	The stacking model is tested on three distinct datasets and its performance is compared to that of other classifiers. The best combining classifiers are then discovered using the enumeration approach.
23	Sayadi *et al.* [31]	2022	DT, DL, LR, RF, SVM, XGB	LR, SVM Accuracy=95.08%	Z-Alizadeh Sani heart disease dataset	With the features selected from Pearson feature selection, six traditional ML approaches were compared. LR, SVM reached the highest accuracy level fo CAD detection.
24	Verma *et al.* [32]	2016	MLP, MLR FURIA, C4.5	MLR Accuracy = 88.4%	Indira Gandhi Medical College (IGMC), Shimla, India	Dimension reduction using correlation-based feature subset selection using PSO, and data clustering to spot cluster data points that were improperly assigned. Finally, C4.5, MLP, MLR, and FURIA were used to build the hybrid model.
25	Idris *et al.* [33]	2020	LR, NN, kNN, DT, NB, SVM, DL, Vote	NN with Embedded DT features Accuracy = 94.5%	Malaysian National Cardiovascular Disease Database (NCVD)–ACS registry	With the significant features selected from suitable feature selection methods, a prediction model was built and its performance metrics was compared.
26	Marateb and Goudarzi [34]	2015	Fuzzy expert system (MLR+NFC)	Fuzzy rule-based system (MLR+NFC) Accuracy = 84.0%	Cleveland heart disease dataset	The features are selected with MLR and integrated with the Neuro-Fuzzy classifier for the prediction of heart diseases.
27	Rani *et al.* [35]	2021	NB, SVM, LR, RF, AdaBoost	RF Accuracy = 86.6%	Cleveland heart disease dataset	A hybrid technique that included GA and RFE was applied for feature selection. Missing values in the dataset were treated using Multiple Imputation by Chained Equations algorithm. RF provided the highest performance in combination with MICE, GARFE, Scaling and SMOTE.
28	Rajdhan *et al.* [36]	2020	DT, LR, RF, NB	RF Accuracy = 90.16%	Cleveland heart disease dataset	When the accuracy of multiple ML algorithms was compared, RF performed well.
29	Pal *et al.* [37]	2012	ANN, ID3, CART, Fuzzy	Fuzzy Logics Accuracy = 84.20% Sensitivity= 95.85% Specificity=83.33%	NR	The risk factors for developing CAD, approaches for gathering and representing knowledge, a strategy for organizing rules, fuzzifying clinical parameters, and defuzzification to crisp value are all described in this study.
30	Kanagarathinam *et al.* [38]	2022	NB, XGB, KNN, SVM, MLP, CatBoost	CatBoost classifier Accuracy = 94.34%	Sathvi dataset	Sathvi dataset is created by combining the 4 CVD datasets. The dataset has 531 instances and 12 attributes with absence of missing data.

*NR = Not reported

REFERENCES

[1] S. Pouriyeh, S. Vahid, G. Sannino, G. De Pietro, H. Arabnia, and J. Gutierrez, "A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease," in *Proceedings - IEEE Symposium on Computers and Communications*, Jul. 2017, pp. 204–207, doi: 10.1109/ISCC.2017.8024530.

[2] C. W. Tsao *et al.*, "Heart disease and stroke statistics-2022 update: A report from the American Heart Association," *Circulation*, vol. 145, no. 8, pp. E153–E639, 2022, doi: 10.1161/CIR.0000000000001052.

[3] C. Abbafati *et al.*, "Global burden of 87 risk factors in 204 countries and territories, 1990–2019: a systematic analysis for the Global

Burden of Disease Study 2019," *The Lancet*, vol. 396, no. 10258, pp. 1223–1249, 2020, doi: 10.1016/S0140-6736(20)30752-2.

[4] P. Libby and P. Theroux, "Pathophysiology of coronary artery disease," *Circulation*, vol. 111, no. 25, pp. 3481–3488, Jun. 2005, doi: 10.1161/CIRCULATIONAHA.105.537878.

[5] D. Lapp, "Heart disease dataset," *Kaggle*, pp. 1–6, 2020, [Online]. Available: https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive.

[6] A. Frank and A. Asuncion, "UCI machine learning repository," *UCI*, 2010, [Online]. Available: http://archive.ics.uci.edu/ml.

[7] R. Alizadehsani *et al.*, "A data mining approach for diagnosis of coronary artery disease," *Computer Methods and Programs in Biomedicine*, vol. 111, no. 1, pp. 52–61, Jul. 2013, doi: 10.1016/j.cmpb.2013.03.004.

[8] L. Verma, S. Srivastava, and P. C. Negi, "An intelligent noninvasive model for coronary artery disease detection," *Complex & Intelligent Systems*, vol. 4, no. 1, pp. 11–18, Jul. 2018, doi: 10.1007/s40747-017-0048-6.

[9] A. A. Haruna, L. J. Muhammad, B. Z. Yahaya, E. J. Garba, N. D. Oye, and L. T. Jung, "An improved C4.5 data mining driven algorithm for the diagnosis of coronary artery disease," in *Proceeding of 2019 International Conference on Digitization: Landscaping Artificial Intelligence, ICD 2019*, Nov. 2019, pp. 48–52, doi: 10.1109/ICD47981.2019.9105844.

[10] L. J. Muhammad and E. A. Algehyne, "Fuzzy based expert system for diagnosis of coronary artery disease in nigeria," *Health and Technology*, vol. 11, no. 2, pp. 319–329, Feb. 2021, doi: 10.1007/s12553-021-00531-z.

[11] L. J. Muhammad, I. Al-Shourbaji, A. A. Haruna, I. A. Mohammed, A. Ahmad, and M. B. Jibrin, "Machine learning predictive models for coronary artery disease," *SN Computer Science*, vol. 2, no. 5, Jun. 2021, doi: 10.1007/s42979-021-00731-4.

[12] R. Yilmaz and F. H. Yagin, "Early detection of coronary heart disease based on machine learning methods," *Medical Records*, vol. 4, no. 1, pp. 1–6, Jan. 2022, doi: 10.37990/medr.1011924.

[13] J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive data mining for medical diagnosis: An overview of heart disease prediction," *International Journal of Computer Applications*, vol. 17, no. 8, pp. 43–48, Mar. 2011, doi: 10.5120/2237-2860.

[14] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Heart disease classification using nearest neighbor classifier with feature subset selection," *Annals Computer Science Series*, vol. XI, no. 1, pp. 47–54, 2013, [Online]. Available: http://www.anale-informatica.tibiscus.ro/download/lucrari/11-1-06-Jabbar.pdf.

[15] U. R. Acharya *et al.*, "Automated characterization of coronary artery disease, myocardial infarction, and congestive heart failure using contourlet and shearlet transforms of electrocardiogram signal," *Knowledge-Based Systems*, vol. 132, pp. 156–166, Sep. 2017, doi: 10.1016/j.knosys.2017.06.026.

[16] J. H. Joloudari *et al.*, "Coronary artery disease diagnosis; ranking the significant features using a random trees model," *International Journal of Environmental Research and Public Health*, vol. 17, no. 3, p. 731, Jan. 2020, doi: 10.3390/ijerph17030731.

[17] S. A. Ali *et al.*, "An optimally configured and improved deep belief network (OCI-DBN) approach for heart disease prediction based on ruzzo-tompa and stacked genetic algorithm," *IEEE Access*, vol. 8, pp. 65947–65958, 2020, doi: 10.1109/ACCESS.2020.2985646.

[18] S. Dixit and R. Kala, "Early detection of heart diseases using a low-cost compact ECG sensor," *Multimedia Tools and Applications*, vol. 80, no. 21–23, pp. 32615–32637, Aug. 2021, doi: 10.1007/s11042-021-11083-9.

[19] H. Kahramanli and N. Allahverdi, "Design of a hybrid system for the diabetes and heart diseases," *Expert Systems with Applications*, vol. 35, no. 1–2, pp. 82–89, Jul. 2008, doi: 10.1016/j.eswa.2007.06.004.

[20] S. Muthukaruppan and M. J. Er, "A hybrid particle swarm optimization based fuzzy expert system for the diagnosis of coronary artery disease," *Expert Systems with Applications*, vol. 39, no. 14, pp. 11657–11665, Oct. 2012, doi: 10.1016/j.eswa.2012.04.036.

[21] P. R. L, S. V. Jinny, and Y. V. Mate, "Early prediction model for coronary heart disease using genetic algorithms, hyper-parameter optimization and machine learning techniques," *Health and Technology*, vol. 11, no. 1, pp. 63–73, Nov. 2021, doi: 10.1007/s12553-020-00508-4.

[22] A. Tiwari, A. Chugh, and A. Sharma, "Ensemble framework for cardiovascular disease prediction," *Computers in Biology and Medicine*, vol. 146, p. 105624, Jul. 2022, doi: 10.1016/j.compbiomed.2022.105624.

[23] M. Tarawneh and O. Embarak, "Hybrid approach for heart disease prediction using data mining techniques," *Lecture Notes on Data Engineering and Communications Technologies*, vol. 29, pp. 447–454, 2019, doi: 10.1007/978-3-030-12839-5_41.

[24] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[25] 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.

[26] M. Abdar, W. Książek, U. R. Acharya, R. S. Tan, V. Makarenkov, and P. Pławiak, "A new machine learning technique for an accurate diagnosis of coronary artery disease," *Computer Methods and Programs in Biomedicine*, vol. 179, p. 104992, Oct. 2019, doi: 10.1016/j.cmpb.2019.104992.

[27] D. Shah, S. Patel, and S. K. Bharti, "Heart disease prediction using machine learning techniques," *SN Computer Science*, vol. 1, no. 6, Oct. 2020, doi: 10.1007/s42979-020-00365-y.

[28] M. M. Ghiasi, S. Zendehboudi, and A. A. Mohsenipour, "Decision tree-based diagnosis of coronary artery disease: CART model," *Computer Methods and Programs in Biomedicine*, vol. 192, 2020, doi: 10.1016/j.cmpb.2020.105400.

[29] A. H. Chen, S. Y. Huang, P. S. Hong, C. H. Cheng, and E. J. Lin, "HDPS: Heart disease prediction system," *Computing in Cardiology*, vol. 38, pp. 557–560, 2011.

[30] J. Wang *et al.*, "A stacking-based model for non-invasive detection of coronary heart disease," *IEEE Access*, vol. 8, pp. 37124–37133, 2020, doi: 10.1109/ACCESS.2020.2975377.

[31] M. Sayadi, V. Varadarajan, F. Sadoughi, S. Chopannejad, and M. Langarizadeh, "A machine learning model for detection of coronary artery disease using noninvasive clinical parameters," *Life*, vol. 12, no. 11, p. 1933, Nov. 2022, doi: 10.3390/life12111933.

[32] L. Verma, S. Srivastava, and P. C. Negi, "A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data," *Journal of Medical Systems*, vol. 40, no. 7, Jun. 2016, doi: 10.1007/s10916-016-0536-z.

[33] N. Md Idris, Y. K. Chiam, K. D. Varathan, W. A. Wan Ahmad, K. H. Chee, and Y. M. Liew, "Feature selection and risk prediction for patients with coronary artery disease using data mining," *Medical and Biological Engineering and Computing*, vol. 58, no. 12, pp. 3123–3140, Nov. 2020, doi: 10.1007/s11517-020-02268-9.

[34] H. R. Marateb and S. Goudarzi, "A noninvasive method for coronary artery diseases diagnosis using a clinically-interpretable fuzzy rule-based system," *Journal of Research in Medical Sciences*, vol. 20, no. 3, pp. 214–223, 2015.

[35] P. Rani, R. Kumar, N. M. O. S. Ahmed, and A. Jain, "A decision support system for heart disease prediction based upon machine learning," *Journal of Reliable Intelligent Environments*, vol. 7, no. 3, pp. 263–275, Jan. 2021, doi: 10.1007/s40860-021-00133-6.

[36] A. Rajdhan, A. Agarwal, M. Sai, D. Ravi, and P. Ghuli, "Heart Disease Prediction using Machine Learning," *International Journal of Engineering Research and*, vol. V9, no. 04, May 2020, doi: 10.17577/IJERTV9IS040614.

[37] D. Pal, K. M. Mandana, S. Pal, D. Sarkar, and C. Chakraborty, "Fuzzy expert system approach for coronary artery disease screening using clinical parameters," *Knowledge-Based Systems*, vol. 36, pp. 162–174, Dec. 2012, doi: 10.1016/j.knosys.2012.06.013.

[38] K. Kanagarathinam, D. Sankaran, and R. Manikandan, “Machine learning-based risk prediction model for cardiovascular disease using a hybrid dataset,” *Data and Knowledge Engineering*, vol. 140, p. 102042, Jul. 2022, doi: 10.1016/j.datak.2022.102042.

[39] A. Garavand, C. Salehnasab, A. Behmanesh, N. Aslani, A. H. Zadeh, and M. Ghaderzadeh, “Efficient model for coronary artery disease diagnosis: a comparative study of several machine learning algorithms,” *Journal of Healthcare Engineering*, vol. 2022, pp. 1–9, Oct. 2022, doi: 10.1155/2022/5359540.

[40] H. G. Lee, K. Y. Noh, and K. H. Ryu, “A data mining approach for coronary heart disease prediction using HRV features and carotid arterial wall thickness,” in *BioMedical Engineering and Informatics: New Development and the Future - Proceedings of the 1st International Conference on BioMedical Engineering and Informatics, BMEI 2008*, May 2008, vol. 1, pp. 200–206, doi: 10.1109/BMEI.2008.189.

[41] C. M. Chu *et al.*, “A Bayesian expert system for clinical detecting coronary artery disease,” *Journal of Medical Sciences*, vol. 29, no. 4, pp. 187–194, 2009.

[42] M. A. Karaolis, J. A. Moutiris, D. Hadjipanayi, and C. S. Pattichis, “Assessment of the risk factors of coronary heart events based on data mining with decision trees,” *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 3, pp. 559–566, May 2010, doi: 10.1109/TITB.2009.2038906.

BIOGRAPHIES OF AUTHORS

Anu Ragavi Vijayaraj is a full-time research scholar in the school of Computer Science and Engineering at VIT, Chennai, India. She received her B.Tech. in Information Technology from Kongu Engineering College, Perundurai (India) in 2013 and M.E. in Computer and Communication Engineering from Kongu Engineering College, Perundurai (India) in 2015. Her research interests include machine learning, deep learning, and artificial intelligence. She can be contacted at email: anuragaviraj@gmail.com.

Subbulakshmi Pasupathi is an Assistant Professor in the school of Computer Science and Engineering at VIT, Chennai, India. She received her B.Tech. in Information Technology in 2009, M.E. in Computer Science and Engineering in 2011 and Ph.D. from Anna University, Chennai (India) in 2019. She is the author of many scientific publications in international journals and conferences. Her current research interests include cognitive networks, artificial intelligence, and machine learning. She can be contacted at email: subbulakshmi.p@vit.ac.in.