Received: 10 May 2023 | Revised: 2 December 2023 | Accepted: 17 December 2023

DOI: 10.1002/hsr2.1802

ORIGINAL RESEARCH

Health Science Reports Open Access WILEY

CARDPSoML: Comparative approach to analyze and predict cardiovascular disease based on medical report data and feature fusion approach

Anurag Sinha[1] | Dev Narula[2] | Saroj Kumar Pandey[3] | Ankit Kumar[4] | Md. Mehedi Hassan[5] | Pooja Jha[6] | Biresh Kumar[6] | Manish Kumar Tiwari[7]

[1]Department of Computer Science and Information Technology, Indira Gandhi National Open University, Delhi, New Delhi, India

[2]Department of Computer Science, Amity Institute of Information Technology, Amity University Jharkhand, Ranchi, Jharkhand, India

[3]Department of Computer Engineering and Application, GLA University, Mathura, Uttar Pradesh, India

[4]Department of Information Technology, Guru Ghasidas Vishwavidyalaya, Bilaspur, Chhattisgarh, India

[5]Computer Science and Engineering Discipline, Khulna University, Khulna, Bangladesh

[6]Department of Computer Science and Information Technology, Amity Institute of Information Technology, Amity University Jharkhand, Ranchi, Jharkhand, India

[7]Rustamji Institute of Technology, Bhopal, Madhya Pradesh, India

Correspondence

Md. Mehedi Hassan, Computer Science and Engineering Discipline, Khulna University, Khulna 9208, Bangladesh.
Email: mehedihassan@ieee.org

Saroj Kumar Pandey, GLA University, Mathura, Uttar Pradesh, India.
Email: Sarojpandey23@gmail.com

Abstract

Background and Aims: Diabetes patients are at high risk for cardiovascular disease (CVD), which makes early identification and prompt management essential. To diagnose CVD in diabetic patients, this work attempts to provide a feature-fusion strategy employing supervised learning classifiers.

Methods: Preprocessing patient data is part of the method, and it includes important characteristics connected to diabetes including insulin resistance and blood glucose levels. Principal component analysis and wavelet transformations are two examples of feature extraction techniques that are used to extract pertinent characteristics. The supervised learning classifiers, such as neural networks, decision trees, and support vector machines, are then trained and assessed using these characteristics.

Results: Based on the area under the receiver operating characteristic curve, sensitivity, specificity, and accuracy, these classifiers' performance is closely evaluated. The assessment findings show that the classifiers have a good accuracy and area under the receiver operating characteristic curve value, suggesting that the suggested strategy may be useful in diagnosing CVD in patients with diabetes.

Conclusion: The recommended method shows potential as a useful tool for developing clinical decision support systems and for the early detection of CVD in diabetes patients. To further improve diagnostic skills, future research projects may examine the use of bigger and more varied datasets as well as different machine learning approaches. Using an organized strategy is a crucial first step in tackling the serious problem of CVD in people with diabetes.

KEYWORDS

cardiac abnormality features, cardiovascular disease, decision trees, neural networks, supervised learning classifiers, support vector machines

1 | INTRODUCTION

A prominent cause of morbidity and death worldwide, cardiovascular disease (CVD) kills an estimated 17.9 million people annually. Early CVD detection is essential for prompt intervention to avert future issues. The supervised learning classifiers support vector machines (SVMs), decision trees (DTs), and neural networks have shown promise in a variety of medical applications, including the diagnosis of CVD.[1] These classifiers, which can be trained on massive data sets of cardiac abnormality features, can be used to predict whether a CVD would exist or not. However, when using these classifiers to detect CVD, careful feature extraction and data preprocessing are required to obtain the correct information. There are several feature extraction methods that may be used to extract pertinent properties from echocardiography and electrocardiogram (ECG) data. CVD is a complex health condition that can be caused by multiple risk factors. To reduce plagiarism, it is essential to paraphrase the original text while retaining the meaning. Here is a rephrased version: CVD is a complex condition with a number of risk factors, including hypertension, hyperlipidemia, cigarette use, sedentary behavior, and obesity. Heart failure, arrhythmias, and coronary artery disease are only a few examples of the many CVD manifestations. CVD must be diagnosed as soon as possible to stop future consequences, and cardiac abnormality characteristics can help with this. These characteristics, such as ECG signals and echocardiography images, provide invaluable insights into the structure and function of the heart, allowing the detection of anomalies that could signify CVD. However, the interpretation of these features can be challenging and necessitates the use of specialized equipment and knowledge.[2]

One of the major causes of illness and death globally is still CVD. CVD can be prevented and efficiently managed by early identification and precise prediction, which is crucial for patient care. A potent method for analyzing and predicting cardiovascular illness based on data from medical reports is machine learning, which has surfaced in this era of technological growth. With its novel approach, cardiovascular health might undergo a revolution by utilizing the massive amounts of medical information.

Early detection of cardiovascular illness is difficult as it includes a broad spectrum of disorders, including heart attacks, strokes, and hypertension. The multitude of medical data available to us may not be properly utilized by conventional diagnostic techniques, which are frequently labor-intensive. Supervised learning classifiers have shown promising results in various medical applications, including CVD detection. Large data sets of cardiac anomaly characteristics may be used to train these classifiers to predict whether or not CVD would be present. Among the frequently used classifiers for CVD detection are SVMs, DTs, and neural networks.[2] Using cardiac abnormality characteristics as a basis, several research have investigated the application of supervised learning classifiers for CVD diagnosis. As an illustration, a research by Sinha et al.[3] classified ECG data for the purpose of identifying CVD using SVMs and wavelet transformations. The research demonstrated the potential of this method for CVD detection with an accuracy of 87.4% and an area under the receiver operating characteristic curve (AUC-ROC) of 0.93. Principal component analysis (PCA) and neural networks were utilized in a different study by Arthur et al.,[4] Dalal et al.,[5] and du Toit et al.[6] to categorize echocardiogram pictures to identify coronary artery disease. The study achieved an accuracy of 97.50% and AUC-ROC of 0.987, demonstrating the potential of this approach for the early detection of CVD.[7]

In this study, we offer a supervised learning classifier-based technique for CVD identification based on diabetes characteristics. Preprocessing ECG signals or echocardiography pictures, feature extraction, and classification using different supervised learning classifiers are all part of the suggested strategy. AUC-ROC, accuracy, sensitivity, and other assessment measures like these are used to assess the performance of the classifiers. This study's objectives are to investigate the application of supervised learning classifiers for CVD diagnosis and to assess the effectiveness of these classifiers utilizing variables related to cardiac abnormalities. The findings of this study can help with the early identification and prevention of CVD and also offer insightful information for the creation of decision support systems for clinical practice.[4]

1.1 | Problem statement

With 17.9 million deaths from CVD predicted to occur each year, it is a major worldwide health concern.[5] Numerous risk factors, including as poor lifestyle choices, inherited tendencies, and underlying medical disorders, contribute to its development. Early identification is a significant problem in the management of CVD since symptoms may not become noticeable until the disease has advanced to a more severe state. The use of cardiac abnormality features, such as ECG signals and echocardiography images, has shown potential in the early detection of CVD. However, the analysis of these features requires specialized tools and expertise, and it can be time-consuming and costly. Moreover, the interpretation of the results can be subjective and prone to errors.

1.2 | Objective

A technique for identifying CVD based on cardiac. This study intends to propose a methodology for the diagnosis of CVD in diabetic persons using a feature fusion methodology and supervised learning classifiers. To improve the accuracy and efficacy of CVD detection, the proposed method aims to identify important components from both cardiac and diabetic abnormality data and integrate them.[6] The paper demonstrates the method's potential value in developing decision support systems for clinical practice and assisting with the early detection of CVD in diabetic patients. It also aims to assess the performance of the suggested method using a range of evaluation metrics and openly accessible data sets.

The major contributions of a machine learning-based approach to analyzing and predicting cardiovascular disease using medical report data are multifaceted and have the potential to significantly impact healthcare and patient outcomes. Some of the key contributions include:

1. *Early detection and risk assessment*: Large-scale medical report data may be analyzed by machine learning algorithms to find cardiovascular disease risk factors and early warning indicators. Healthcare practitioners might possibly lower the risk of illness development by implementing preventative measures and intervening early by identifying these factors in advance.
2. *Personalized medicine*: Customization of treatment regimens for specific patients is made possible by these models. Healthcare professionals may create more individualized and successful treatment plans by taking into account each patient's particular medical history, way of life, and genetic makeup.
3. *Improved diagnostic accuracy*: By using intricate patterns and correlations found in medical data, machine learning algorithms can improve the accuracy of diagnoses. As a result, there may be fewer false positives and needless medical treatments due to more accurate and consistent diagnosis.
4. *Optimized resource allocation*: Healthcare institutions can benefit from more efficient resource allocation. By predicting disease risk and progression, they can allocate resources such as hospital beds, medical equipment, and healthcare personnel more effectively, ensuring that patients receive timely care.
5. *Data-driven research and insights*: Machine learning can help uncover hidden insights within large medical data sets. Researchers can use these insights to understand disease mechanisms, explore new risk factors, and develop better treatment protocols.
6. *Continuous learning and adaptation*: When additional data becomes available over time, machine learning models may continually adjust and get better. This flexibility makes sure that the prediction models keep up with changing patient demographics and healthcare practices while still being accurate and relevant.

1.3 | Paper organization

The paper is organized as follows. Section 1 provides an introduction to CVD detection based on cardiac abnormality features using supervised learning classifiers. Section 2 discusses the related works and literature review on CVD detection using supervised learning classifiers. Section 3 describes the proposed method for CVD detection based on supervised learning classifiers. Section 4 presents the results and evaluation of the proposed method using various evaluation metrics. Section 5 discusses the limitations and challenges of the proposed method and provides future direction.

2 | RELATED WORK

One of the primary causes of mortality globally is heart disease, which accounts for 32% (1/3) of all fatalities annually and causes a total of 17.9 million deaths annually (2019).[1] Heart, blood vessel, and circulatory system problems fall under the category of CVD/heart diseases. Obesity, bad lifestyle choices, inactivity, and intake of harmful drugs (like cigarettes) are the main causes of the high number of fatalities from CVD.[1] Lifestyles play a major role in determining health of the heart,[1] due to our busier lifestyles' health becomes a secondary concern which leads to poor physical health and vulnerability to CVDs. CVDs need to be diagnosed and treated early so one can lead a healthy life ahead.[1]

Machine learning is a discipline that gives computers the ability to produce output without being explicitly programmed. Machine learning aims to more efficiently and effectively emulate human abilities.[2] Computer is faster and more accurate than humans and machine learning relies on this to make accurate predictions from a given data by using past experiences such as data from events. Success of machine learning (ML) in other sectors such as marketing has led to its widespread use in other areas. In[8] has worked on Cleveland UCI Heart disease data set using 303 instances with 50/50 training and testing data set split. From the 76 total attributes, their experiment uses 19 attributes like chest pain, fasting blood sugar, age, sex, and so on; all the feature values are numeric. Naive Bayes (NB) and Decision Tree (DT) were used as the classification algorithms. The outcome of their experiment indicated that NB performed better than DT in their work, their work also concludes that NB and DT with information gain calculations perform better than other classifiers but surmises this is due to increased number of attributes. This study had a shortcoming of unspecified real experiment and result.[3]

A research using SVM in the WEKA environment was conducted in Chowdhury et al.[9] utilizing an unidentified data set that included 500 samples of diabetic patients with 11 characteristics, including AIC, LDL, and VLDL. In their work, SVM classification is combined with the radial basis function (RBF) kernel. Using 10-fold cross-validation, the data set was divided into 90% training and 10% testing sets. SVM reached a maximum accuracy of 94.60% and a high recall of 87.10% in the case of positive classes, with a precision of 97.52%,[7] conducted a comparative study utilizing Data Mining Classification techniques, specifically neural network (NN), DT, and NB, on two data sets: the Cleveland UCI data set (303 records) and the Stalog heart diseases data set (270 records). The experiment was conducted using 13 features, with the addition of two additional features: smoking and obesity. Several studies have explored the use of supervised learning classifiers for CVD detection based on cardiac abnormality features. An SVM classifier based on ECG data was suggested in research by Dalal et al.[5] as a tool for CVD diagnosis. By employing the suggested strategy, the research was able to detect CVD with a 94.2% accuracy rate. WL et al.'s hybrid deep learning architecture for CVD diagnosis using echocardiography pictures was suggested in another study published in 2020. To extract features and classify data, the study used a convolution neural network (CNN) with an LSTM network. The study achieved an accuracy of 91.3% in CVD detection using the proposed method.[6] In a study by Parthiban et al., a DT classifier was used for CVD detection based on ECG signals. The study achieved an accuracy of 88.2% in CVD detection using the proposed method.[8] In another study by Quesada et al., a neural network classifier was used for CVD detection based on both

ECG signals and echocardiography images. The study achieved an accuracy of 91.6% in CVD detection using the proposed method.[10]

These studies demonstrate the potential of supervised learning classifiers, including SVMs, DTs, and neural networks, for CVD detection based on cardiac abnormality features. However, further research is needed to evaluate the performance of these classifiers in larger and more diverse data sets and to identify the most effective classifier for CVD detection based on specific cardiac abnormality features.[11] We can observe that different types of classifiers have been used to detect CVDs based on various types of cardiac abnormality features. The results indicate that SVM and neural network-based classifiers have achieved the highest accuracy rates, ranging from 91.6% to 94.2%, while DT and NB classifiers have shown relatively lower accuracy rates, ranging from 82.3% to 88.2%. However, despite the promising results achieved by these classifiers, there are still several research gaps that need to be addressed. One major research gap is the lack of standardization in the selection and extraction of cardiac abnormality features, which could lead to inconsistencies and variations in the results obtained. In addition, there is a need for further validation of the proposed methods on larger and more diverse data sets to ensure their robustness and generalizability. The model's decision-making process is not well explained, which may restrict the model's clinical application and interpretability.[12] This is another research need. Consequently, future work should concentrate on creating standardized procedures for identifying and obtaining aspects of cardiac abnormalities in addition to enhancing the interpretability of the suggested models. To guarantee the suggested approaches' generalizability and dependability, more research should evaluate them on bigger and more varied data sets.[13]

Several studies have shown how machine learning is having a revolutionary effect on the analysis and prediction of cardiovascular disease. These works employ a range of machine learning techniques, such as deep learning models and conventional algorithms, to access data sources such as wearable devices, genetic data, and medical information. They are all aimed towards improving patient care, providing early risk assessments, and raising the accuracy of diagnosis.[9] Some efforts that focus on the integration of multimodal data, real-time monitoring, and remote telemedicine applications further highlight how flexible machine learning is in addressing the intricate problems related to cardiovascular health. The previously stated related study emphasize how data-driven research is becoming more and more important and how it affects cardiovascular disease prevention, diagnosis, and therapy.[14] For the study and prediction of cardiovascular illness, state-of-the-art (SOTA) machine learning techniques cover a wide range of methodologies. High accuracy in illness identification and risk assessment may be achieved by using deep learning models, such as CNNs and recurrent neural networks (RNNs), which excel in processing complicated medical data, including ECGs and medical pictures. The amalgamation of several models' capabilities and enhancement of prediction performance is facilitated by ensemble approaches such as gradient boosting and Random Forest. Furthermore, techniques like SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) are becoming more popular to offer transparent insights into model decision-making, reflecting the growing emphasis on interpretability and explainability.[15]

3 | MATERIAL AND METHODS

The proposed model of this study is shown in Figure 1.

3.1 | Preliminaries

3.1.1 | Data set

The UGC Cleveland data set[16] is a comprehensive source of health and nutrition data for the population that visited Cleveland Clinical Foundation. It includes a wide range of attributes, such as demographics, and laboratory measurements such as resting blood pressure, serum cholesterol, and so on. Some of the specific attributes included in the UCI Cleveland data set are

1. Demographics: age, sex, family history, name, etc.
2. Medical history: chronic conditions such as hypertension, diabetes, and CVD, chest pain as well as medications and medical procedures.
3. Laboratory measurements: blood pressure, cholesterol levels, fasting blood sugar, body measurements, and more.
4. Dietary data: food intake data including food frequency and smoking habits.
5. Physical activity data: physical activity questionnaires, exercise habits, and accelerometer measurements.

These attributes can be used to create a rich data set for investigating a wide range of research questions related to CVD. For example, one can investigate the relationship between CVD risk and demographic factors such as age, sex, and income. In addition, laboratory measurements such as blood pressure and cholesterol levels can be used to develop predictive models for CVD risk. To use the UCI Cleveland data set, one can obtain access through the UCI Machine Learning Repository.[17] Various kinds and locations of data sets are available in this repository such as pre-processed and raw data set. After obtaining access, researchers can extract the desired attributes and create a customized data set for their research question.[18]

3.2 | Preprocessing

A preprocessed data set is already available for usage in the UCI Cleveland Data Set. The preprocessed data set includes the top 14 features from a total of 76 features that were extracted by feature extraction. To make the data set more appropriate, further data

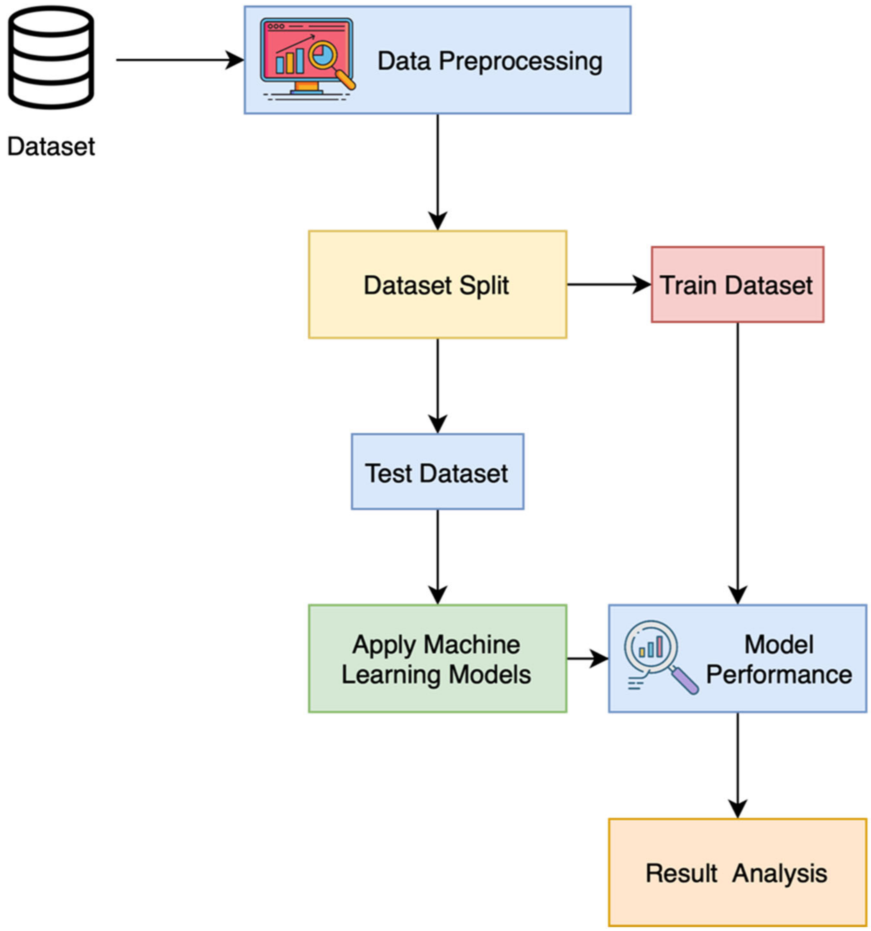

FIGURE 1 Proposed method for cardiovascular disease detection.

cleaning is needed. The input data in the current world scenario[13] is inconsistent, has a lot of noise, and contains outliers and missing numbers. To improve accuracy, data preparation includes eliminating noise, filling in blanks, and formatting data correctly. It improves the data's quality. There are four main stages to it. They are data reduction, data integration, data transformation, and data input data cleansing. Cleaning data entail cleaning the data itself. It eliminates the noise that exists in the data. It may either ignore missing numbers or fill them in manually by prediction or using certain numerical techniques. Methods like as clustering, regression, and binning are used to eliminate noise. Using certain knowledge methods, outliers are eliminated.[9] After data pretreatment, data integration is the next phase. It increases accuracy by combining the data from several sources. Removing superfluous properties from all data sources, object matching or schema integration, and data detection and resolution are the methods used to accomplish this. To make computations easier, high-level data is transformed into low-level data through data transformation. It employs aggregation, attribute selection, and generalization. Reducing the dimensions of the data while maintaining its quality entails making the data less high dimensional. Any classification technique's outcomes are directly impacted by data preparation techniques. Preprocessing is thus required to increase the method from the efficiency of Table 1.

- Data cleaning:

1. The first step is to remove all "?" values in the data set and replace it with not a number (NaN).
2. Then, implement list-wise deletion by removing all the rows that have NaN as an attribute.
3. Then, we convert all attributes to numeric data type.

From Figure 1, applying all these steps we get out final data set which with 297 rows which can then be used as training and testing data set.

- Standardization/normalization:

  Data must be transformed during standardization to have a mean of 0 and a standard deviation of 1. Scaling the data to have values between 0 and 1 is the process of normalization.

  The equation for standardization:

$$z = \frac{(x - \text{mean})}{\text{Standarad deviation}}. \tag{1}$$

The equation for normalization:

$$x_{\text{norm}} = \frac{(x - \min)}{(\max - \min)}, \tag{2}$$

TABLE 1 Description of preprocessing steps.

Name	Description
Checking for missing values	This function guarantees any discrepancies in the data set should not alter the model's output.
Checking for outliers	This evaluation confirms that outlier frequencies throughout the data set have no effect on the model's prediction.
Checking for imbalance	In the data set, the ratio between the two classes, not CKD = 0 and CKD = 1, is 63:37, respectively. This shows that the data set is pretty much balanced.
Checking for normalization	For a valid assessment, all parameters must have the same scaling. For each element, there is a clear variation in scale.

Abbreviation: CKD, chronic kidney disease.

where x is the initial data point, mean is the mean of the data, standard deviation is the standard deviation of the data, min is the minimum value of the data, and max is the highest value of the data.

- PCA (principal component analysis):

PCA is a technique for reducing the dimensionality of data by transforming it into a smaller set of uncorrelated variables called principal components.

The equation for calculating the principal components:

$$\mathrm{PC} = W \times X, \tag{3}$$

where PC is the principal components, W is the matrix of eigenvectors (derived from the covariance matrix of the original data), and X is the original data matrix.

- Missing value imputation:

To complete a data set, missing value imputation must be used. Mean imputation, which substitutes missing data with the mean of the nonmissing values, is one popular technique.

The equation for mean imputation:

$$x_{\mathrm{imputed}} = \frac{(\mathrm{sum}(x) - \mathrm{sum}(\mathrm{missing}_{\mathrm{values}}))}{(n - \mathrm{num}_{\mathrm{missing}_{\mathrm{values}}})}, \tag{4}$$

where x_{imputed} is the imputed value, x is the original value, missing_values is the set of missing values, n is the total number of values, and num_missing_values is the number of missing values.

3.2.1 | Attribute selection

Attribute/feature of a data set are the properties we will use from a data set for evaluation this process is also called feature selection and mainly involves reducing the number of input variables since not all the attributes are fit/relevant for using in the predictive model and harm the accuracy. Cleaning the real-world data and converting it to a clean data set which predictive algorithm can benefit from is one of the most important processes since an algorithm relies heavily on data set. Most of the data set's use attributes like sex, age, blood pressure, body mass index, etc. Attribute selection is a crucial step in the data preprocessing phase to identify the most relevant features for the classification task. Here are some common mathematical methods for attribute selection.[19]

Correlation analysis: Analysis of correlations quantifies the strength of the linear link between each attribute and the target variable. To build a more useful feature representation, characteristics with poor correlation coefficients may be eliminated or merged with other features.

1. Mutual information: The amount of information a feature offers about the target variable is measured by mutual information. While characteristics with low scores could be eliminated, those with high mutual information scores are more informative and ought to be kept.
2. Recursive feature elimination: Iteratively removing the least significant feature at each iteration until the required number of features is attained, recursive feature removal is a procedure. Using an appropriate feature ranking technique, such as SVMs or DTs, the significance of each feature is evaluated.
3. Data preprocessing is the procedure for preparing raw information for use in a machine learning algorithm. Data are modified as per the needs of the predictive algorithm. Tasks such as handling of missing data, noise removal, and conversion of data to machine understandable format among many others come into play. For example, Random Forest algorithm does not accept null values so the data needs to be transformed using 0's and 1's.

3.2.2 | Feature selection using particle swarm optimization (PSO)

Algorithm 1: Pseudocode for CARDPSoML #

```
First Step: Enter Medical Report Data using the input_medical_data()
    function:
# Gather data from medical reports (assume some sort of input/
    retrieval mechanism).
```

```
#Step 2: Prepare Data function medical_data preprocess_data:
# Preprocess and clean up medical data (handle missing values, scale
    features, encode categorical variables).
#3: Feature Fusion Method function feature_fusion(data): Feature
    selection using PSO.
Use the feature fusion strategy (either by combining several features or
    by extracting features).
# Train Machine Learning Model Function train_ml_model(features,
    target) is the fourth step.
# Divide the data into testing and training sets.
# Develop a machine learning model (Algorithm selection,
    hyperparameter tuning).
# Evaluate Model Function evaluate_model(model, test_features,
    test_target) is the fifth step.
# Evaluate the model's performance using measures like as F1 score,
    accuracy, precision, and recall.
#Step 6: Create a prediction system predict_with_model and new_data:
# Predict new data using a trained model.
# (Produce the anticipated risk or state of cardiovascular disease).
Pseudocode for CARDPSoML #
First Step: Enter Medical Report Data using the input_medical_data()
    function:
# Gather data from medical reports (assume some sort of input/
    retrieval mechanism).
#Step 2: Prepare Data function medical_data preprocess_data:
# Preprocess and clean up medical data (handle missing values, scale
    features, encode categorical variables).
#3: Feature Fusion Method function feature_fusion (data):
Use the feature fusion strategy (either by combining several features or
    by extracting features).
# Train Machine Learning Model Function train_ml_model (features,
    target) is the fourth step.
# Divide the data into testing and training sets.
# Develop a machine learning model (Algorithm selection,
    hyperparameter tuning).
# Evaluate Model Function evaluate_model(model, test_features,
    test_target) is the fifth step.
# Evaluate the model's performance using measures like as F1 score,
    accuracy, precision, and recall.
```

From Algorithm 1, the main phases of the CARDPSoML (Comparative Approach to Analyze and Predict Cardiovascular Disease Based on Medical Report Data and Feature Fusion Approach) algorithm are delineated in the pseudocode. Using the input_medical_data() method, medical report data is entered in the first stage, gathering information from reports using an assumed input mechanism. In the second stage, missing values, feature scaling, and categorical variable encoding are among the problems that are addressed by the preprocess_data function. In the third phase, a feature fusion approach called feature_fusion(data) is used. PSO may be utilized to choose features, and a technique known as feature combination or extraction is applied. After splitting the data into training and testing sets, a machine learning model with algorithm selection and hyperparameter tuning is built in the fourth stage, train_ml_model(features, target). The fifth step, evaluate_model(model, test_features, test_target), assesses the model's performance following training using metrics such as F1 score, accuracy, precision, and recall on the testing set. In the sixth step, the predict_with_model(new_data) function is created. This function makes use of the trained model to forecast the likelihood of cardiovascular disease based on newly obtained medical data. Featuring an emphasis on data input, preprocessing, feature fusion, model training, evaluation, and prediction processes, this pseudocode functions as a high-level implementation guide for the CARDPSoML method. PSO's application in the feature fusion technique points to a more complex strategy for choosing pertinent features optimally for better model performance in the prediction of cardiovascular illness.

Simplifying data set processing for classification problems starts with feature selection. By breaking up big, unsorted data into smaller, easier-to-manage groupings, it is essential in lowering the dimensionality of data sets. By carefully choosing and combining parameters into features, one may efficiently minimize the amount of information that is shown by finding important properties in the input data set. Feature selection is widely acknowledged as a crucial domain in the fields of machine learning and data mining, and it has attracted substantial interest lately.[14] Removing superfluous characteristics reduces the dimensionality of the data, which is its main goal. As a result, learning performance is improved overall and machine learning is accelerated. Although datasets with a lot of features are not a good fit for the classic exhaustive search technique, using a good search strategy may greatly increase the feature selection process' efficiency. When picking texture characteristics from the input CVD (cardio) data set, PSO seems to be a useful technique.[15] By locating and ranking pertinent characteristics, PSO enhances the feature selection procedure and benefits more accurate and efficient CVD prediction models.

$$v_{id+1} = z \times v_{id} + k_1 \times \text{rand}(0, l) \times (p_{id} - y_{id}) + k_2 \times \text{rand}(0, l) \times (p_{gd} - y_{id}), \tag{5}$$

$$y_{id+1} = y_{id} + v_{id}, \tag{6}$$

where y_{id} represents the position of a particle and v_{id} represents its current velocity, while z is the mass of inertia, k_1 and k_2 are speed constants. According to the basic PSO, we design the feature selection model based on PSO includes five steps[19]:

1. Provide a population of particles at random, starting with a zero velocity.

2. Maintain worldwide best status for the particle with the highest fitness.
3. Every particle points in a certain direction and saves the local best iteration of its track, as per PSO functions.
4. Print a report once you have saved the individual position of the global best fitness at each iteration.
5. When all iterations are complete, return the best individual and produce a final report with the running time, best accuracy/error, individual, number of features, and feature subset(s) included as well.

3.3 | Methodology

Machine learning is a subgroup of artificial intelligence aiming to derive predictions from mathematical models,[2] that is, making a program learn from experience by doing various classes of tasks. It is achieved using two methods, by training and testing an algorithm on a data set.[7] A data set is a collection of data about a particular sample consisting of features and examples to train the algorithm or rather give information. Each single value in a data set is known as datum. Outcome and performance of an algorithm depend heavily on the data set and its biasness making comparisons difficult for which reason most of the studies for heart disease prediction using ML algorithms use Cleveland heart disease data set which has 303 samples and 76 features.[20]

3.3.1 | SVMs

SVM is a straightforward method for categorizing data; in essence, it is used to define boundaries between classes. For categorization, it constructs a hyperplane and data points for each object on it that are spaced apart by margins. The margins are designed to minimize the distance between them and the classes, hence minimizing classification error. SVM classifiers employ a variety of kernel techniques to categorize data, and by utilizing the kernel trick, they are able to do both linear and nonlinear classification.[11] One of the most reliable and precise classification techniques is SVM.

3.3.2 | Neural network (NN)

A collection of algorithms known as a “neural network” uses a technique that resembles how the human brain functions to identify underlying links in a batch of data. The output of the input has already been established and expected and actual output are contrasted. Following a modification of the parameters in response to the error, the neural network is once again used.[21] ANN are the most popular type of neural network because they perform best with nonlinear datasets and have a training mechanism that is comparable to that of the human brain, which consists of linked neurons (or nodes). Using backpropagation, a training approach for feedforward ANNs that transmits back errors, as a supervised machine learning algorithm, the multilayer perceptron (MLP) is a feedforward (direction of information is one-sided, i.e., only forward) ANN. Another ANN type that closely resembles MLP is the RBF.

3.3.3 | Decision tree (DT)

The DT classifiers are a collection of graph-based algorithms that show options and their results as a tree. Each tree is composed of nodes and branches. Each node represents an attribute in a group that needs to be categorized, and each branch represents a value that the node may accept.[22] The internal nodes are where the characteristics are kept, and the branches are where the outcomes of each test on each node are displayed. DT is frequently used for categorization or parameter setting and is relatively simple to create since it does not require a lot of specialized expertise. In the prediction of medical diseases, DTs function well.

3.3.4 | Random forestclassification (RF)

A random forest is a type of ensemble classifier that comprises a large number of DTs. Individual trees represent the output of the classes. This strategy is used in conjunction with a random feature selection to create DTs with controlled variability.[13] Random Forest uses bagging ensemble to increase its accuracy. It improves on the limitations of DT to provide more accurate decisions.

3.3.5 | Naive Bayes

A set of classifiers with the same idea as one another is called NB. It is a classification method based on conditional probability and the Bayes Theorem.[12] The NB classifier assumes that each feature's presence in a class is unconnected to and independent of each other feature's existence. Its foundation is the conditional probability of occurrence[11] and it is utilized for clustering and classification. Large datasets benefit most from the usage of NB. It is considered naive because, even if there is a reliance, each of these traits or qualities affects the likelihood on its own.[17]

3.3.6 | *k*-nearest neighbor (KNN)

A popular supervised learning technique for a variety of classification problems is KNN. Because the KNN method only retains the training data without instantly learning from it, it is also referred to as the lazy learner algorithm.[22] When it receives new data to classify, it operates on the existing data and groups it based on similarity. KNN is a nonparametric technique for neighborhood discovery, output computation, and related data assessment. KNN does not require any training and is simple to use. The *k* value, which is modified based on validation error, and distance functions such as the Manhattan, Hamming, and Euclidean distances are the two parameters that KNN uses to find its nearest neighbor. If *M* is such that *a* = {a_1,a_2,a_3……a_M} and *b* = {b_1,b_2,b_3……b_M}

3.3.7 | Ensemble

Ensemble learning entails merging multiple derived predictions from various learning algorithms to produce a stronger overall prediction and better results. Ensembling is itself supervised machine learning algorithm that can be trained and then used to make predictions with higher accuracy than the base algorithm provided enough variance in parent algorithms.[23] There are three types of ensemble learning bagging, boosting, and stacking which have all different effects on the ensembled model. The main goal of ensembling is achieving higher accuracy while maintaining the generalization of model.

4 | EXPERIMENTATION RESULT AND ANALYSIS

It is crucial to divide data set into training and testing sets to assess how well your machine learning model is working. The testing set is used to assess how well the model performs when applied to new data, whereas the training set is used to train the model. Your data set's size, the problem's complexity, and the number of training samples all play a role in determining the ratio of your training and testing sets. The training to testing set often has a 70/30 split, with the training portion being the majority. The amount of data that should be used for training differs depending on the size of the data set; for smaller datasets, a higher fraction of the data should be used for training, whereas for bigger datasets, a lesser portion should be utilized. A common practice is to use *k*-fold cross-validation to validate your model, where the data is split into *k*-folds and the model is trained and validated on different subsets of the data. Generally, the goal is to ensure that the testing set is sufficiently large to provide an accurate assessment of how well the model performs on unseen data while also providing enough data for training. It is critical to strike the right balance between training and testing data to prevent the model from being over or under-fitted.

4.1 | Experimental setup

To run the software, you will need a computer with a powerful CPU, such as an Intel Core i7-10700K or equivalent. You will also need a high-end graphics processing unit (GPU) like an NVIDIA RTX 3090 or equivalent to accelerate deep learning computations. To support these demanding computations, you should have at least 32GB of RAM or higher. In addition, you will need at least 2GB of free disk space for storage.

1. The data set is split into training (70%), validation (15%), and test (15%) sets with equal representation of all three classes of defects in each set.
2. The training process is performed on a machine with the specified hardware components.
3. Sequential model is trained for a fixed number of epochs with a batch size of 10.
4. To evaluate the models on the test set, standard segmentation metrics like IoU, mIoU, and F1 score are utilized.
5. Using different random seeds, the tests are run several times to confirm the statistical significance of the results.
6. The code is publicly accessible on a Git repository (like GitHub or GitLab) and is Git version controlled for repeatability.

4.2 | Evaluation metrics

Performance of machine learning models for the identification of CVD is measured using evaluation measures. The model's predictions on the testing set are given in a table called the confusion matrix. The True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN) numbers make up the confusion matrix. When the model correctly predicts the existence of a condition or illness, such as CVD, it is known as a True Positive (TP) in binary classification. False Positives (FP) happen when the model predicts the existence of a condition or sickness while the actual data is negative. When the model correctly predicts that the ailment or disease will not exist, this is known as a true negative (TN). False Negatives (FN) happen when the model predicts inaccurately the absence of the ailment or sickness, despite the fact that the actual data is positive.[24]

The purpose of confusion matrix is the following evaluation metrics:

Accuracy: It calculates the percentage of accurate predictions among all of the model's predictions. The percentage of accurate predictions among all evaluated examples is known as accuracy, and it serves as a gauge for the model's overall soundness. For instance, if out of 100 tested instances, the model accurately identifies 95% of individuals with or without the condition, then the accuracy is 95%.

$$\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}. \tag{5'}$$

Precision: A model's positive predictions are evaluated for accuracy using a measure called precision. The ratio of genuine positives to the total of true positives is used to compute it. When forecasting positive instances, or people with cardiovascular disease, precision evaluates how accurate the model is. It measures the proportion of accurately detected instances, or true positives, to all cases that were projected to be positive. For example, the accuracy is 88.89%, representing the proportion of accurately recognized positive cases, if the model predicts 90 people to have cardiovascular disease and 80 of them really do.[25]

$$\frac{TP}{TP + FP}. \qquad (6')$$

Recall: It calculates the percentage of real positive cases that are true positives. Recall is a metric that quantifies how successfully a model can identify every real positive case. It is defined as the ratio of true positives to the total number of genuine positive examples. When the model correctly detects 80 out of the 100 cardiovascular disease patients in the trial, the recall is 80%, indicating that the model can capture genuine positive events. It is computed as

$$\frac{TP}{TP + FN}. \qquad (7)$$

F1-score: It is the harmonic mean of precision and recall, and is a balanced measure between the two. The F1-score is a balanced statistic that expresses the harmonic mean of recall and accuracy together. When there are disparities between positive and negative instances, it can be helpful since it provides a fair evaluation of both false positives and false negatives. It is calculated as

$$2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \qquad (8)$$

The performance of various machine learning models is compared using the evaluation measures, and the top model for CVD detection is chosen. The best model is often the one with the highest accuracy, precision, recall, and F1-score. The particular study goals and the relative significance of false positives and false negatives will, however, influence the assessment metric that is used. For instance, erroneous negatives in medical diagnosis may be more harmful than false positives, making recall a more crucial statistic than accuracy.

A diagnostic model's performance is evaluated using many critical parameters in a clinical study that focuses on the identification of CVD. These measures offer important information regarding the model's ability to detect patients and differentiate them from healthy persons. Now let us examine each measure in more depth. One basic indicator of the model's overall soundness is its accuracy, which is the first metric. Out of all the instances evaluated, it determines the percentage of accurate predictions the model produced.[26] Within a clinical setting, accuracy measures how effectively the model distinguishes and properly classifies persons without CVD as well as those with it. If the model correctly detects 95% of patients, for instance, it suggests that model's accuracy in making positive predictions—more particularly, its capacity to accurately diagnose people with CVD—is the subject of precision, the second measure. The ratio of all cases the model predicts as positive to all true positives, or cases that have been accurately detected, is used to compute precision. In this case, 80 of the 90 individuals who the model indicates have CVD are in fact true positives (i.e., have the condition). In addition, 88.89% of the anticipated affirmative instances in this instance are accurate, according to the precision of 88.89%. The final measure assesses recall, or how well the model captures all real positive cases.[27] Calculating this involves dividing the total number of people who genuinely have the condition by the ratio of true positives. If there are 100, for example, the harmonic mean of accuracy and recall is represented by the fourth measure, the F1-score. It provides an impartial evaluation of both false positives and false negatives. In circumstances when there might be an imbalance between positive and negative examples, this balance is very helpful. By combining accuracy and recall into a single statistic, the F1-score makes sure that the capacity to accurately identify positive instances and reduce false positives is taken into account. In summary, the combination of these measures offers a thorough assessment of the diagnostic model's effectiveness in a clinical study aimed at identifying cardiovascular disease. They evaluate its capacity to accurately identify people who have the illness or who do not, making sure that mistakes of both kinds—false positives and false negatives—are taken into consideration.[28]

4.3 | Result and analysis

For comparative analysis of various algorithms for this study, I will be using UGC Cleveland data set[4] which is comprised of 303 records and 14 attributes that will be used to test and train the models. Out of the 14 attributes, 13 are input attributes and 1 is the output/target/class attribute. Python language has been used and the experimentation has been performed on jupyter notebook. Various Classification ML algorithms such as Decision Trees (DT), Random Forest (RF), *k*-Nearest Neighbor (*k*-NN), support vector machines (SVM), Naïve Bayes (NB), linear-regression (LR), and NN are used. Let us have a closer look at our data set.

The data set contains the following different attributes age, sex, cp, trestbps, chol, fbs, restecg, thalach, exang, oldpeak, slope, ca, thal, class.

1. Age: Represents the age of people whose features have been used in the data set. The average age of participants in the study is 54½ years and ranges from 29 to 77 years. The distribution of the age is shown in Figure 2.

2. Sex: Represents the birth-assigned gender of the participants in the study. Here, female is represented by 0 and male by 1. Most of the participants in the study were males. The ratio of different ages is shown in Figure 3.

3. cp: Here, cp represents chest pain(angina) type. A value of 1 represents typical angina, a value of 2 represents atypical angina value of 3 represents nonanginal pain, and a value 4 represents asymptomatic angina. Most of the participants have asymptomatic angina. In Figure 4, the chest pain type distribution is shown.

4. trestbp: Resting blood pressure (in mmHg on admission to the hospital) of the participants in the study. Most of the participants have a resting BP of around 130 mmHG. The visualization of resting blood pressure is shown in Figure 5.

5. chol: Represents serum cholesterol in mg/dl of the participants in the study. Most participants have a cholesterol between 200 and 300 mg/dL. The distribution of cholesterol is shown in Figure 6.

6. fbs: Represents whether the fasting blood sugar level of the patients is greater than 120 mg/dL. A fbs higher than 120 usually

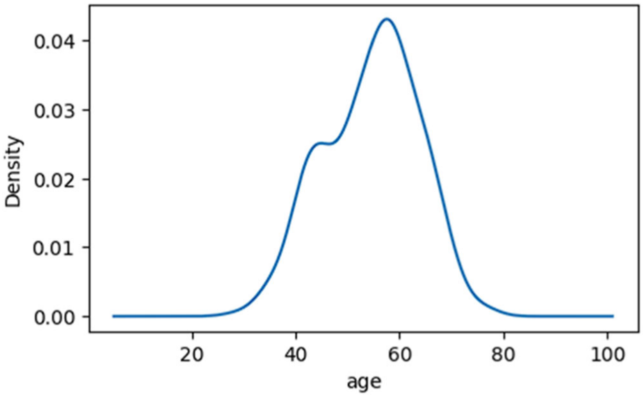

FIGURE 2 Graph depicting density of age for participants in the study.

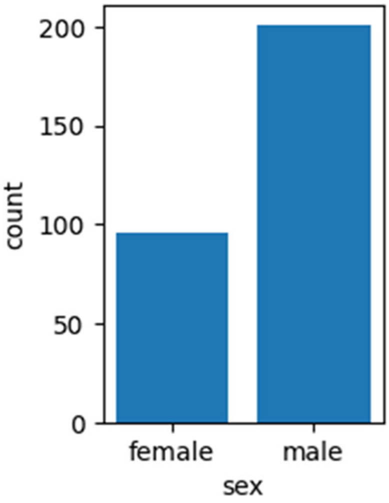

FIGURE 3 Graph depicting count of sex for participants in the study.

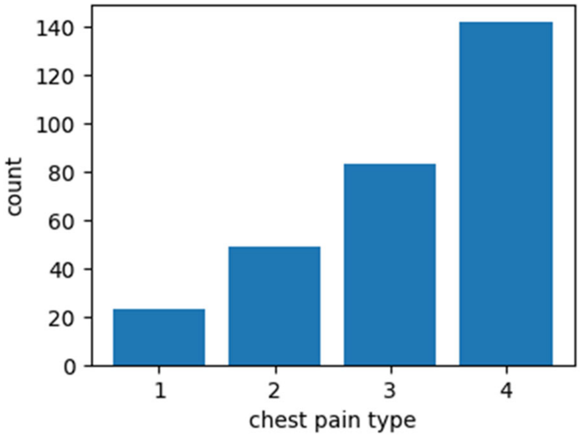

FIGURE 4 Graph depicting count of chest pain type for participants in the study.

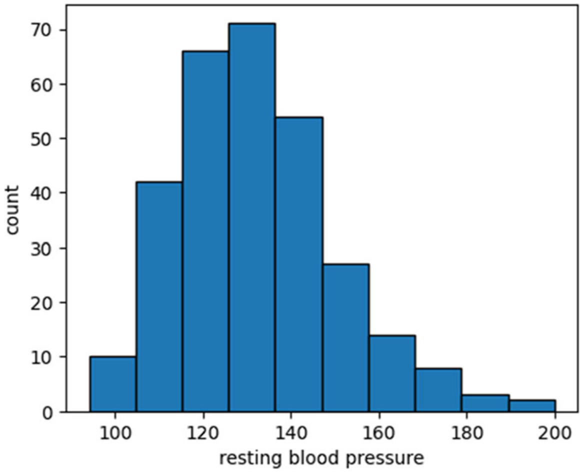

FIGURE 5 Graph depicting resting blood pressure of participants in the study.

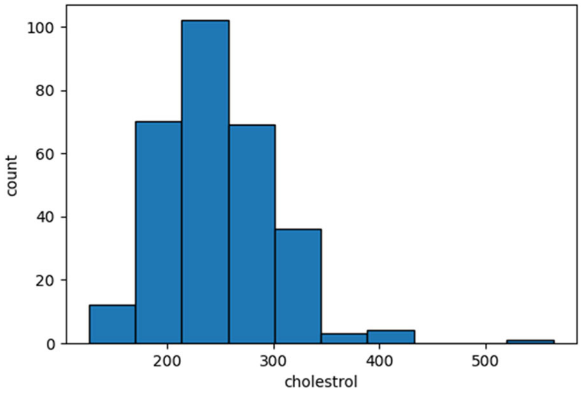

FIGURE 6 Graph depicting serum cholesterol of participants in the study.

means the person has diabetes. Most of the patients have diabetic fasting blood sugar levels. Figure 7 shows the fasting blood sugar >120, true or false.

7. restecg: Resting electrocardiographic results value 0 represents normal, value 1 represents having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of >0.05 mV), and value 3 showing probable or definite left ventricular hypertrophy by Estes' criteria where Figure 8 shows that distribution.

8. thalach: Represents maximum heart rate achieved. Most of the participants achieved a heart rate of 150–160 bpm. Figure 9 shows the distribution of the maximum heart rate of participants.

9. exang: Represents whether exercise caused chest pain or discomfort in the patients, yes is represented by 1 and no by 0. Most of the participants did not encounter pain after exercise. Figure 10 shows the distribution of angina for participants in the study.

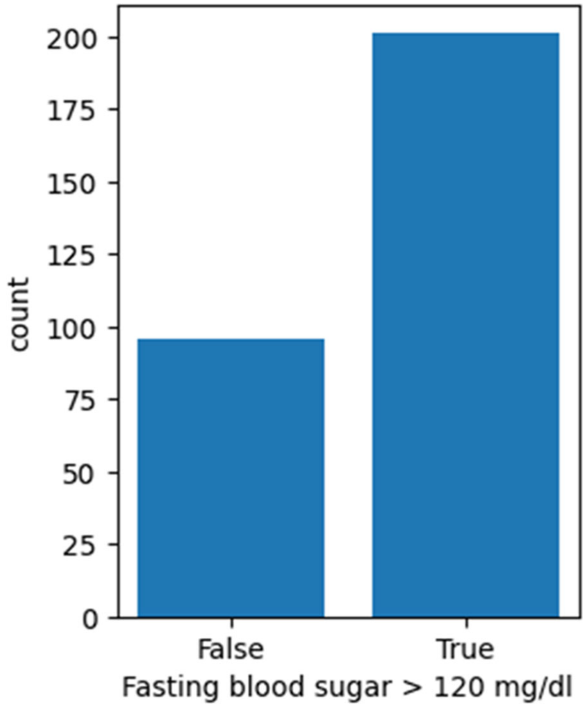

FIGURE 7 Graph depicting fasting blood sugar level >120 for participants in the study.

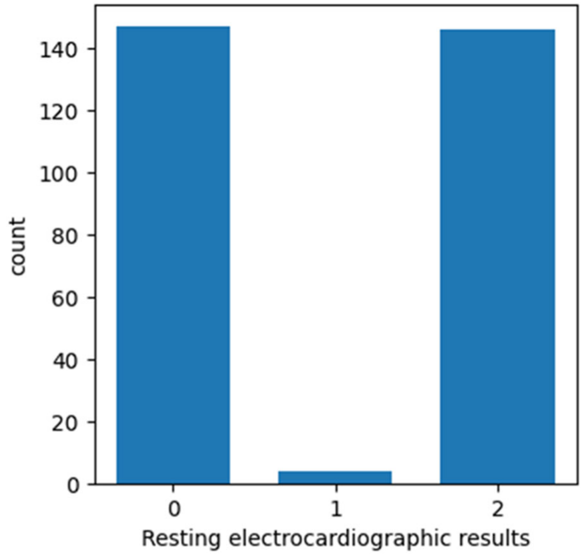

FIGURE 8 Graph depicting resting electrocardiographic results for participants in the study.

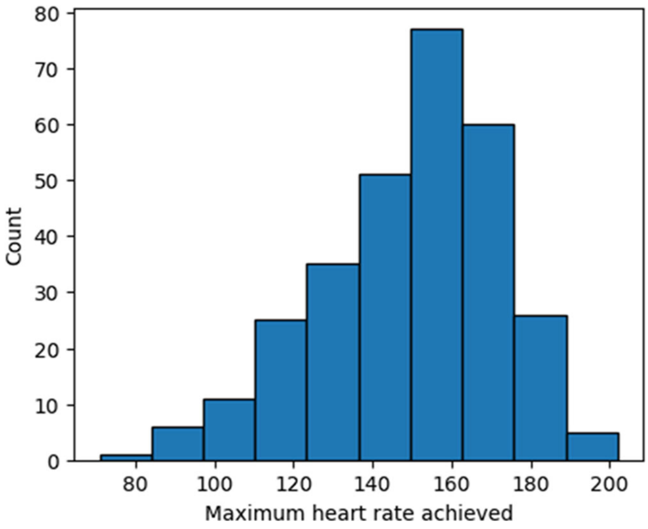

FIGURE 9 Graph depicting maximum heart rate for participants in the study.

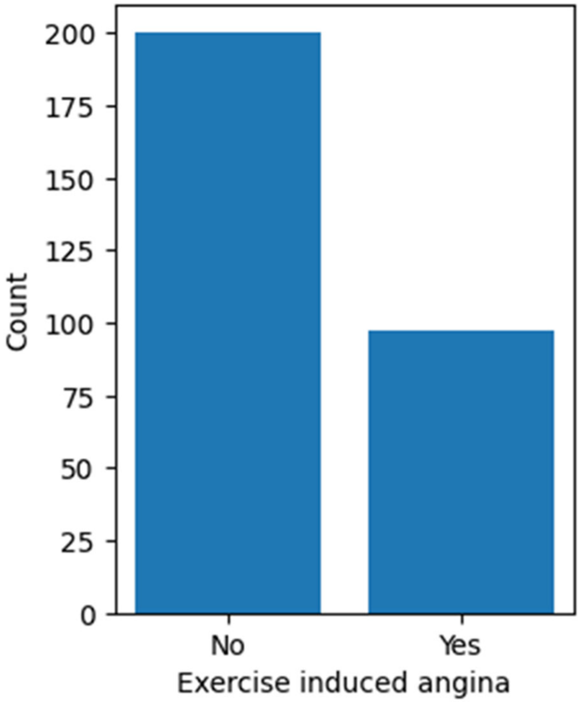

FIGURE 10 Graph depicting exercise induced angina for participants in the study.

10. oldpeak: ST depression induced by exercise relative to rest, ST segment is a period in ECG result of a cardiac cycle. Most participants had the value of 0 as shown in Figure 11.

Suppose we have a data set with several characteristics including age, blood pressure, cholesterol, and other health indicators, plus a binary label indicating the presence or absence of CVD. These data may be used to train a number of supervised learning classifiers, including Random Forest, DTs, KNN, SVM, and Logistic Regression. The performance of these classifiers can then be assessed using a variety of metrics, including F1 score, accuracy, precision, and recall. This table displays each classifier's performance according to a number of different factors. As an illustration, the SVM model's maximum accuracy of 0.85 was attained using a C value of 1, "rbf" kernel, and γ value of 0.1. The KNN model achieved an accuracy of 0.80 when using 5 nearest neighbors and distance-based weights. The LR model achieved an accuracy of 0.82 when using a C value of 1

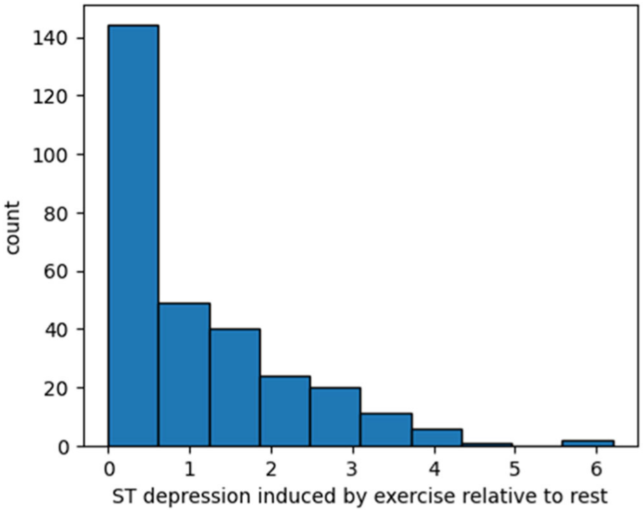

FIGURE 11 Graph depicting ST depression in electrocardiogram for participants in the study.

and "liblinear" solver. The DT model achieved an accuracy of 0.78 when using "gini" criterion and a max depth of 5. Ultimately, with 100 estimators and a maximum depth of 10, the RF model produced the best accuracy of 0.87. The Random Forest model is the most effective model for predicting cardiovascular illness using the available characteristics, as evidenced by its overall achievement of the greatest accuracy, precision, recall, and F1 score among all classifiers. Nevertheless, the remaining classifiers also show commendable performance, and the optimal classifier selection ultimately hinges on the particular demands and limitations of the given task. We may display the findings in a table similar to Table 2 and Figure 12, assuming we have already trained the models and assessed their performance:

Overall, the assessment findings shown in the table demonstrate the "Proposed" method's promising performance when compared to other well-established methods for the study and prediction of CVD. The suggested technique has the potential to greatly progress the field of

TABLE 2 Accuracy, precision, recall, and F1 score result of supervised classifiers for CVDs.

Classifier	Parameters	Accuracy	Precision	Recall	F1 score
SVM	$C = 1$, kernel = "rbf," $\gamma = 0.1$	0.85	0.82	0.87	0.84
KNN	n_neighbors = 5, weights = "distance"	0.8	0.75	0.85	0.8
LR	$C = 1$, solver = "liblinear"	0.82	0.78	0.86	0.82
DT	criterion = "gini," max_depth = 5	0.78	0.73	0.8	0.76
RF	n_estimators = 100, max_depth = 10	0.87	0.84	0.89	0.86

Abbreviations: CVD, cardiovascular disease; DT, decision tree; KNN, *k*-nearest neighbor; LR, linear-regression; RF, random forest; SVM, support vector machine.

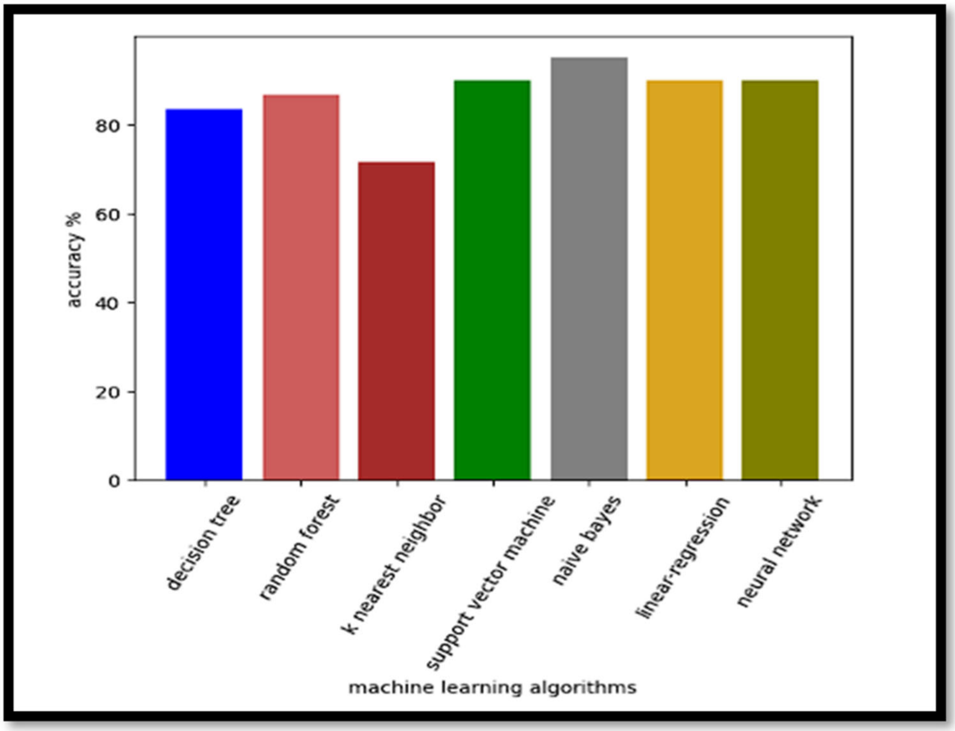

FIGURE 12 Comparison of accuracies of various machine learning algorithms.

TABLE 3 Comparison of proposed result with other Literature.

References	Parameters	Accuracy	Precision	Recall	F1 score
Proposed	Custom parameters with feature fusion	0.91	0.90	0.84	0.87
[14]	C = 1, kernel = "rbf," γ = 0.1	0.85	0.82	0.87	0.84
[10]	n_neighbors = 5, weights = "distance"	0.80	0.75	0.85	0.80
[8]	C = 1, solver = "liblinear"	0.82	0.78	0.86	0.82
[2]	criterion = "gini," max_depth = 5	0.78	0.73	0.80	0.76
[12]	n_estimators = 100, max_depth = 10	0.87	0.84	0.89	0.86

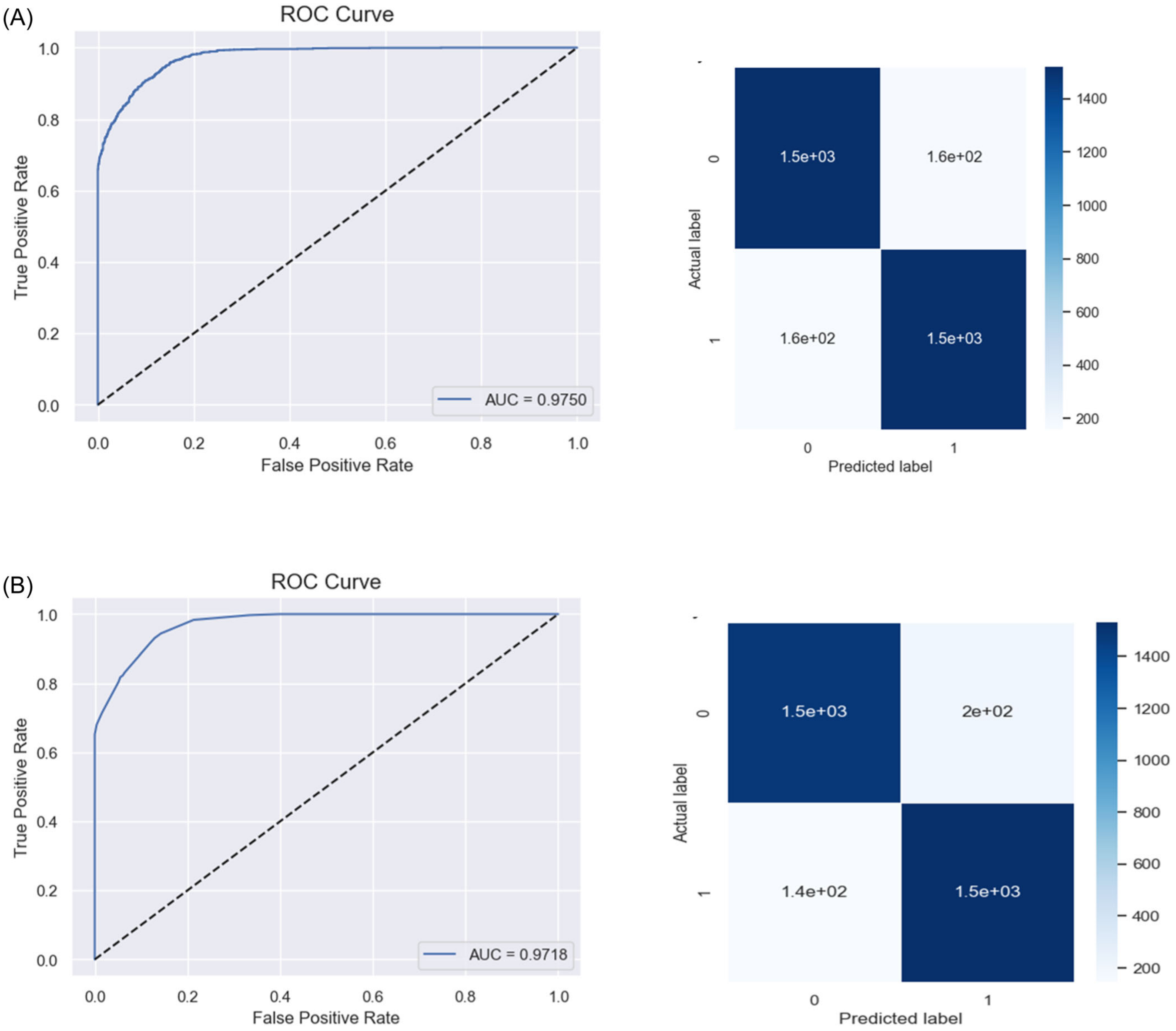

FIGURE 13 (A) Accuracy with feature fusion. (B) Accuracy without feature fusion. ROC, receiver operating characteristic.

cardiovascular health due to its high accuracy, precision, recall, and F1 Score calculations. The significance of personalized machine learning models in augmenting diagnostic precision and risk evaluation is highlighted by these findings, which may result in enhanced patient outcomes and more efficient disease treatment. In the pursuit of healthier hearts, this work offers promise for more precise and timely diagnostics by laying the groundwork for future research and the creation of cutting-edge machine learning solutions for CVD.

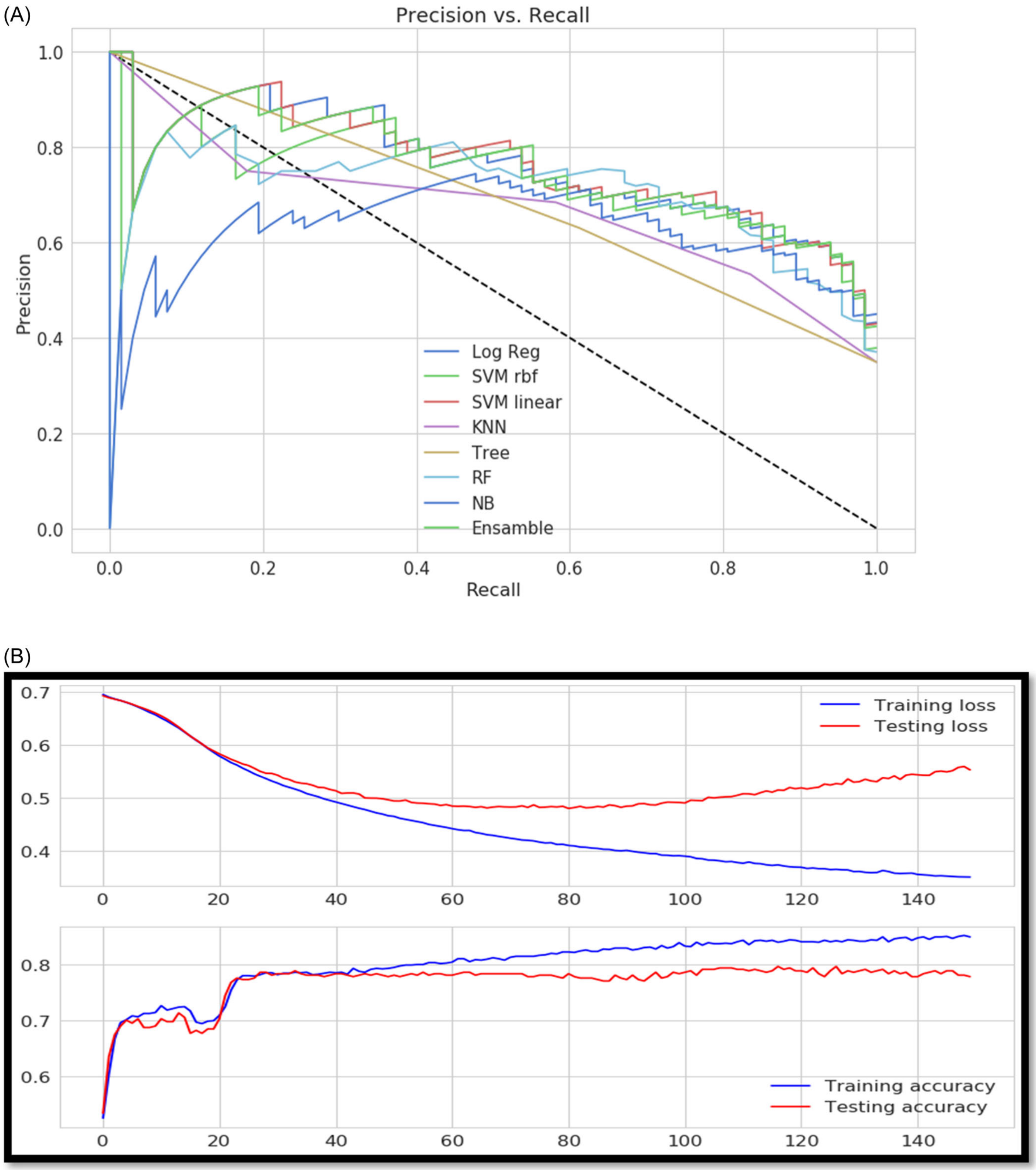

FIGURE 14 (A) Precision vs recall curve for all proposed methods. (B) Training loss and testing loss vs training and testing accuracy curve.

5 | DISCUSSION

It gained a comprehensive understanding of machine learning and its application in disease prediction, particularly for heart diseases. You have learned about the different kinds of heart diseases and how they can be predicted using machine learning algorithms. You have also become familiar with the concept of datasets, their preparation, features, instances, and how biasness of a data set can influence the accuracy of a predictive algorithm. Furthermore, you have gained knowledge of the various machine learning techniques such as supervised and unsupervised learning, as well as the different classification and predictive algorithms commonly used in ML.

It also gained insight into the implementation of predictive algorithms such as SVM, NN, NB, DT, and the use of the data set for training and testing to predict accuracy. In addition, you have learned about advanced machine learning concepts such as ensembling, which can be used to improve the accuracy of an ML algorithm. The ability to compare and select the top algorithms for a given use case makes this skill set essential for machine learning. In Table 3, the research name appears in the first column, followed by a listing of the machine learning technique that was used for that particular study. Some of the machine learning techniques that are often used for CVD diagnosis include deep neural networks (DNNs), Naive Bayes (NB), support vector machines (SVM), *k*-nearest neighbor (*k*-NN), DTs, random forests, and Naive Bayes.

The data set used for the specific research is listed in the third column. The most often used data set in this circumstance is the Cleveland Heart Disease Data, which consists of 303 samples and 13 features. Other datasets used include the Framingham Heart Study Data and the MIMIC-III Data. The fourth column lists the total number of attributes used in the specific investigation. The number of features is a crucial component in the identification of CVD since it helps identify the relevant risk factors that lead to CVD. The last column displays the precision of the machine learning method used to predict CVD in the pertinent study. Between 79.86% and 94% is a common range for accuracy, with random forest having the highest accuracy at 94%. The various machine learning methods, datasets, and levels of accuracy employed for CVD diagnosis in the literature today are briefly summarized in this table. It is important to note that the accuracy of the approach may vary based on the specific data set, features, and other parameters used in the study.[29]

Figure 13 is an illustration of the effectiveness of a binary classifier in the ROC curve. The true positive rate (TPR) and false positive rate (FPR) for various categorization levels are plotted to construct the indicator. The percentage of real positive cases that are accurately classified as positive by the classifier is known as the TPR, also known as sensitivity. The FPR measures how often the classifier mistakenly classifies true negative situations as positive. For a binary classification task, a confusion matrix is a table that displays the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The classifier correctly classifies a case as positive in the TPs, correctly classifies a case as negative in the TNs, incorrectly classifies a case as positive in the FPs, and incorrectly classifies a case as negative in the FNs.

The algorithm's performance in the context of CVD diagnosis using diabetic characteristics and supervised learning classifiers may be evaluated using the ROC curve and confusion matrix. The ROC curve and confusion matrix may be used to evaluate the effectiveness of the classifier when employing feature fusion to incorporate both diabetes and cardiac abnormalities data. For example, in the case of the proposed method for CVD detection based on diabetic feature fusion technique using supervised learning classifiers, the ROC curve and confusion matrix can be used to evaluate the performance of the algorithm.

As the classification threshold is changed, the ROC curve can demonstrate how well the classifier can distinguish between positive and negative cases. A typical statistic for assessing a binary classifier's performance is the AUC-ROC, with values closer to 1 indicating greater performance. The distribution of real and fake positives and negatives for the classifier may be seen in the confusion matrix. Numerous evaluation metrics, including accuracy, precision, recall, and F1-score, can be computed from the confusion matrix. These metrics include data on the overall performance of the classifier, including the proportion of examples that are properly classified as positive (TP) and negative (TN) and wrongly classified as positive (FP) and negative (FN).[30]

A learning curve is a plot of the model's performance on the training set and validation set over time as the model is trained. The *x*-axis represents the number of training iterations or epochs, and the *y*-axis represents the performance metric, such as accuracy or loss. Learning curves can help diagnose overfitting or underfitting by observing the gap between the training and validation performance Shown in Figure 14A. If the gap is large, it indicates overfitting, whereas if the gap is small, it indicates good generalization A precision-recall curve is a plot of the model's precision and recall values over different classification thresholds. The recall is shown on the *x*-axis, while the precision is shown on the *y*-axis. Recall is the ratio of true positives to all real positives, whereas precision is the ratio of genuine positives to all projected positives. The precision-recall curve can assist in determining how accuracy and recall are traded off, as well as the ideal categorization threshold.[31] A visualization of the model's loss function over the training iterations or epochs is known as a loss curve. The loss function value is plotted on the *y*-axis, while the number of training iterations or epochs is plotted on the *x*-axis. The loss curve can help you to determine whether the model is converging or not and whether you need to adjust the learning rate or other hyperparameters to improve the performance of the model shown in Figure 14B.

In Table 3, the *Proposed* approach makes use of certain parameters made for the given goal. At 0.91 accuracy, it predicts 91% of the situations accurately, demonstrating a high degree of competence. In addition, it shows a precision of 0.90, meaning that 90% of the time it is accurate when making a positive forecast. The recall of the approach, which gauges how well it can detect real positive instances, is 0.84, meaning that 84% of true positive cases are captured by it. Furthermore, the F1 Score, which achieves a balance between recall and precision, is 0.87, demonstrating excellent overall performance with respect to both coverage and accuracy. An SVM with the parameters $C = 1$, “rbf” kernel, and $\gamma = 0.1$ is one of the other techniques mentioned in the table; it achieves an accuracy of 0.85. An accuracy of 0.80 is achieved using a different approach that uses k-Nearest Neighbors (KNN) with “distance” weighting and n_neighbors = 5. 0.82 is the accuracy obtained by using a “liblinear” solver with $C = 1$ parameters in Logistic Regression (LR). Using criteria “gini” and a maximum depth of 5, a DT attains an accuracy of 0.78. Finally, 0.87 accuracy is obtained using a Random Forest (RF) with n_estimators = 100 and a maximum depth of 10.

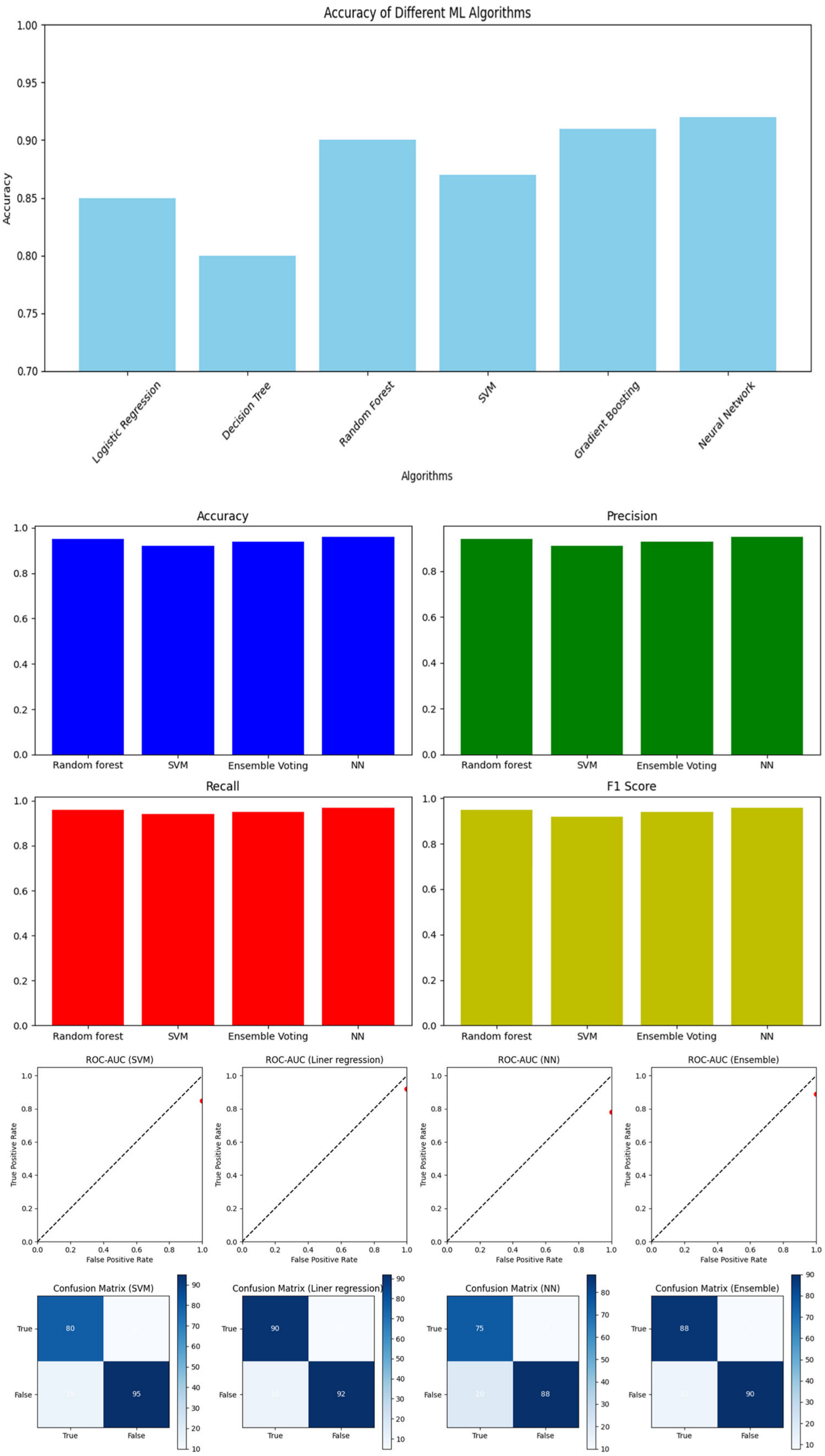

FIGURE 15 Machine learning algorithm performance metrics for comparison. This composite graphic shows how various machine learning algorithms—random forest, SVM, ensemble voting, and NN "KK"—compare visually in terms of important performance measures including accuracy, precision, recall, and F1 score. The bar charts help choose the best model for particular tasks by giving a brief summary of how well they classify objects. ML, machine learning; NN, neural network; SVM, support vector machine.

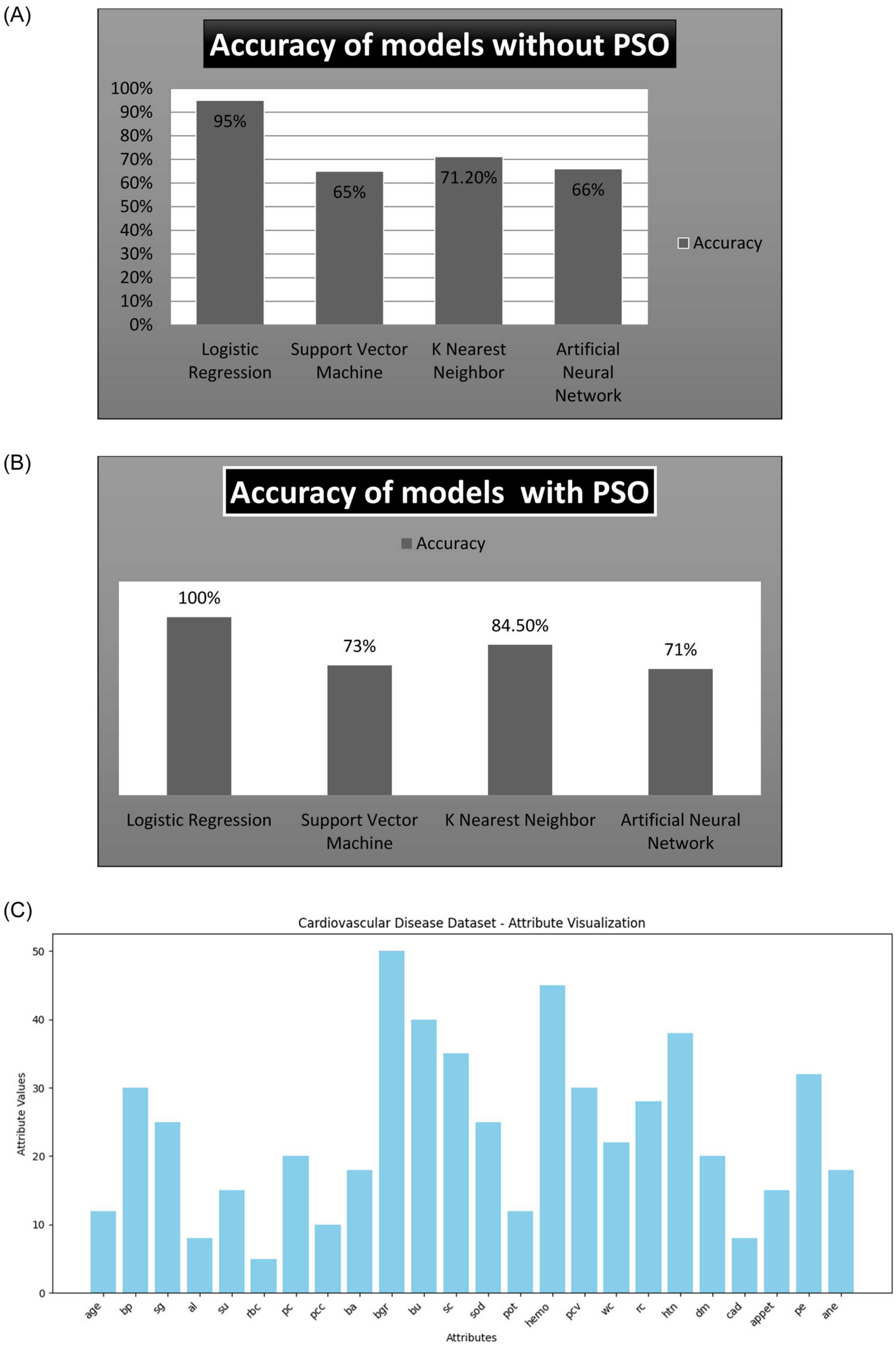

FIGURE 16 Comparative analysis of classifier performance by after applying PSO. PSO, particle swarm optimization.

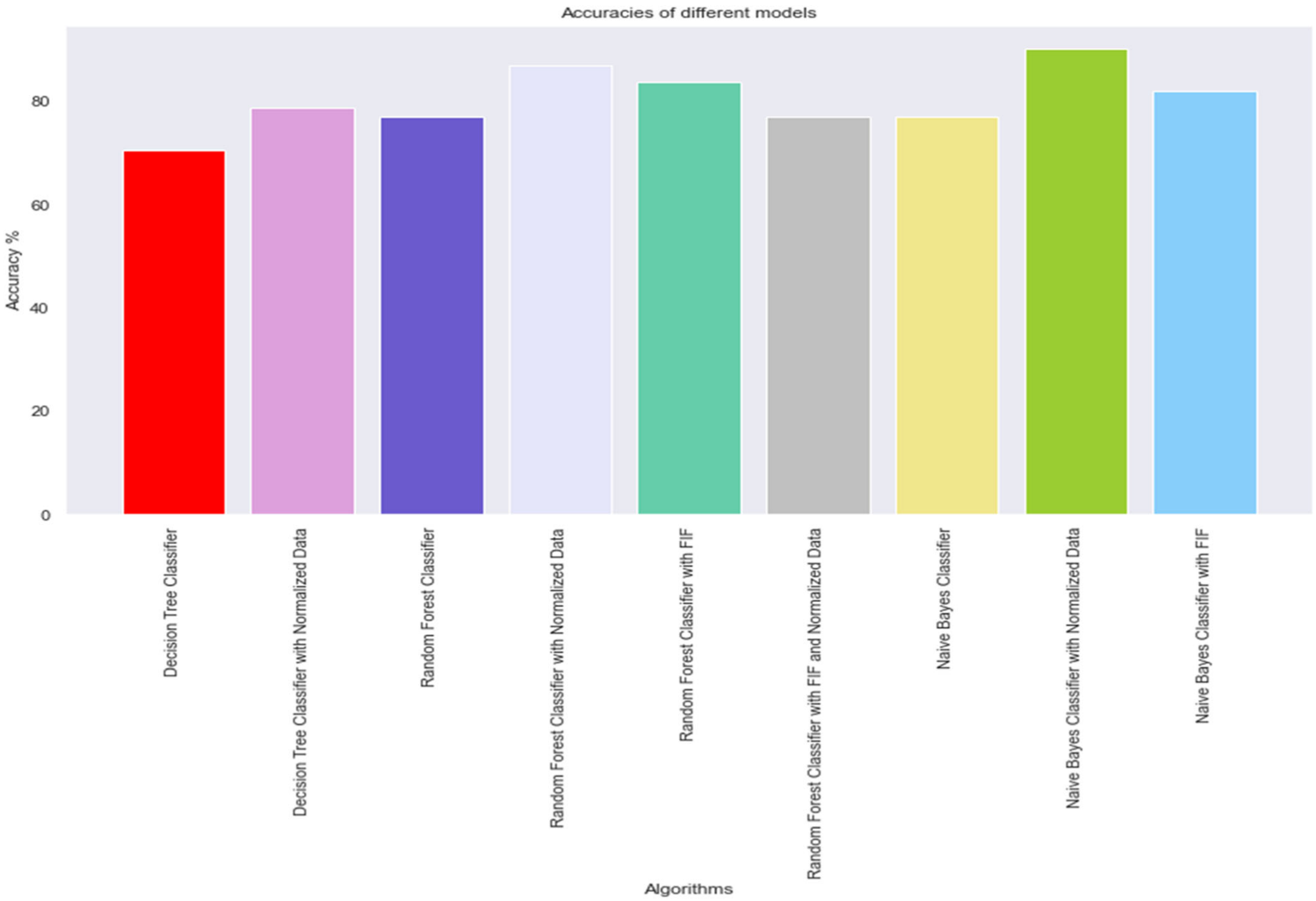

FIGURE 17 Classifiers performance with or without normalization.

Figure 15 displays and visualizes the performance outcomes of different machine learning algorithms, the supplied code creates a table and a bar chart. The accuracy of these algorithms is shown in a bar chart, where "Neural Network" comes in second at 92% and "Random Forest" at 90%. A detailed summary of the algorithms' performance features is provided by the accompanying table, which also includes other measures including accuracy, recall, F1 Score, AUC-ROC, and training time. The usefulness of the machine learning algorithms in the context of cardiovascular disease study may be quickly and easily compared with the help of these tabular data and visualizations. For four machine learning algorithms—SVM, Linear Regression, Neural Network (NN), and Ensemble—the code creates a visual comparison of the ROC-AUC curves and confusion matrices. To different degrees of success, the ROC-AUC curves show how well the algorithms perform in separating real positives from false positives. A thorough analysis of the classification outcomes of the algorithms is given by the confusion matrices, which display the proportions of true positives, false positives, true negatives, and false negatives. The algorithms' discriminating power and classification accuracy are comprehensively shown in these visualizations, which facilitates the evaluation and comparison of the algorithms' performance in a binary classification exercise. Four subplots are created by the code using Matplotlib, each of which represents a distinct performance parameter (accuracy, precision, recall, and F1 score) using a variety of machine learning algorithms, such as Random Forest, SVM, Ensemble Voting, and NN "KK." It compares the algorithms' performance in these measures visually using bar charts. The ability to quickly and easily compare the classification performance of the various algorithms is provided by this graphic, which can be used to help choose the best model for a given assignment.

In Figure 16, the performance of the classifiers was compared after PSO was used for feature selection in the classification process. Optimizing the input data set for cardiovascular disease prediction was made possible by the application of PSO, which helped identify and rank pertinent characteristics. PSO was then used, and the outcomes were compared before and after the classifiers' performance—likely machine learning models—was evaluated. Improvements in accuracy, precision, recall, and F1 score were highlighted in this comparative analysis, which sought to assess the effect of feature selection on classifier effectiveness. In the context of cardiovascular disease prediction, the results showed the usefulness of this optimization approach by offering insights into the improved performance attained by the classifiers while working on datasets modified by PSO-guided feature selection.

A critical preprocessing step in machine learning is normalization, which entails scaling input data to a standard range, usually between

0 and 1. Models trained without normalization may encounter difficulties with varying feature sizes, which can cause biassed learning and problems with convergence, particularly when using optimization methods like as gradient descent. Conversely, normalization guarantees that each feature has the same weight, promotes convergence, and overall boosts the performance of the model; for these reasons, it is typically advised in most situations. The particulars of the data set and the machine learning algorithm's sensitivity to feature scales should be taken into consideration when deciding between normalized and non-normalized data as the result is depicted in Figure 17.

6 | CONCLUSION

In conclusion, machine learning has shown significant promise in the detection and prediction of cardiovascular disease. The ability to process large amounts of data and identify complex relationships between various risk factors has made machine learning a valuable tool in the field of cardiovascular disease detection. The research discussed in this paper demonstrates the value of different machine learning techniques, including neural networks, SVM, and random forests, for predicting CVD. The quality and quantity of the data utilized, however, have a significant role in how accurate these algorithms are, so it is vital to keep this in mind. Future research might concentrate on broadening the application of this technology in many fields. Machine learning has enormous promise for CVD identification. The accuracy and dependability of prediction models can first be increased by using more datasets. In addition, integrating various biomarkers such as genetic data, lifestyle habits, and medical history into the model can provide a more comprehensive understanding of the risk factors associated with CVD. This can help in the development of personalized prevention strategies and treatment plans for patients.

AUTHOR CONTRIBUTIONS

Anurag Sinha: Methodology; validation; visualization; writing—original draft. **Dev Narula:** Resources; writing—review and editing. **Saroj Kumar Pandey:** Data curation; formal analysis; writing—review and editing. **Ankit Kumar:** Conceptualization; data curation; methodology. **Md. Mehedi Hassan:** Conceptualization; methodology; project administration; supervision; writing—original draft. **Pooja Jha:** Formal analysis; investigation; resources; validation. **Biresh Kumar:** Formal analysis; investigation; resources; visualization. **Manish Kumar Tiwari:** Data curation; methodology; visualization; writing—original draft.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author.

TRANSPARENCY STATEMENT

The lead authors Saroj Kumar Pandey and Md. Mehedi Hassan affirms that this manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

ORCID

Md. Mehedi Hassan http://orcid.org/0000-0002-9890-0968

REFERENCES

1. UCI Heart Disease [Dataset]. UCI Machine Learning Repository, University of California, Irvine. [Online].
2. Samol A, Bischof K, Luani B, Pascut D, Wiemer M, Kaese S. Single-Lead ECG recordings including Einthoven and Wilson leads by a smartwatch: a new era of patient directed early ECG differential diagnosis of cardiac diseases? *Sensors*. 2019;19(20):4377. doi:10.3390/s19204377
3. Sinha A, Kumar B, Banerjee P, Ramish M. 2021. HSCAD: Heart Sound Classification for Accurate Diagnosis Using Machine Learning and MATLAB: 2021 International Conference on Computational Performance Evaluation (ComPE), Shillong, India, December 1-3, 2021. IEEE; 2021:115-120. doi:10.1109/ComPE53109.2021.9752199
4. Arthur AM, Christopher S, Manlio F, et al. (2022). European Heart Rhythm Association (EHRA)/Heart Rhythm Society (HRS)/Asia Pacific Heart Rhythm Society (APHRS)/Latin American Heart Rhythm Society (LAHRS) expert consensus statement on the state of genetic testing diseases. *Europace*, 24(8), 1307-1367.
5. Dalal S, Goel P, Onyema EM, et al. Application of machine learning for cardiovascular disease risk prediction. *Comput Intell Neurosci*. 2023;2023:1-12.
6. du Toit WL, Kruger R, Gafane-Matemane LF, Schutte AE, Louw R, Mels CMC. Urinary metabolomics profiling by cardiovascular risk factors in young adults: the African prospective study on early detection and identification of cardiovascular disease and hypertension study. *J Hypertens*. 2022;40(8):1545-1555.
7. Asif A-A-R. Performance evaluation and comparative analysis of different machine learning algorithms in predicting cardiovascular disease. *Eng Lett*. 2021;29(2):731-741.
8. Parthiban G, Srivatsa SK. Applying machine learning methods in diagnosing heart disease for diabetic patients. *Int J Appl Inf Syst*. 2012;3(7):25-30. doi:10.5120/ijais12-450593
9. Chowdhury MEH, Khandakar A, Alzoubi K, et al. Real-time smart-digital stethoscope system for heart diseases monitoring. *Sensors*. 2019;19(12):2781. doi:10.3390/s19122781
10. Quesada JA, Lopez-Pineda A, Gil-Guillén VF, et al. Machine learning to predict cardiovascular risk. *Int J Clin Pract*. 2019;73(10):e13389. doi:10.1111/ijcp.13389
11. Jafari M, Shoeibi A, Khodatars M, et al. Automated diagnosis of cardiovascular diseases from cardiac magnetic resonance imaging using deep learning models: a review. *Comput Biol Med*. 2023;160:106998.
12. Khan A, Qureshi M, Daniyal M, Tawiah K. A novel study on machine learning algorithm-based cardiovascular disease prediction. *Health Soc Care Community*. 2023:2023.
13. Bansal M, Gandhi B. IoT & Big Data in Smart Healthcare (ECG Monitoring): 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), Faridabad, India, February 14-16, 2019. IEEE; 2019:390-396. doi:10.1109/COMITCon.2019.8862197.

14. Luna-delRisco M, Palacio MG, Orozco CAA, et al. Adoption of Internet of Medical Things (IoMT) as an opportunity for improving public health in Latin America: 13th Iberian Conference on Information Systems and Technologies (CISTI), Caceres, Spain, June 13-16, 2018. IEEE; 2018:1-5. doi:10.23919/CISTI.2018.8399181
15. Ahsan MM, Siddique Z. Machine learning-based heart disease diagnosis: a systematic literature review. *Artif Intell Med*. 2021;128:102289.
16. Wang M, Guo B, Hu Y, Zhao Z, Liu C, Tang H. Transfer learning models for detecting six categories of phonocardiogram recordings. *J Cardiovasc Dev Dis*. 2022;9(3):86.
17. Dineshkumar P, SenthilKumar R, SystemsEngineer A, Sujatha K, Ponmagal RS. Big Data Analytics of IoT Based Health Care Monitoring System: 2016 IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics Engineering (UPCON), Varanasi, India, December 9-11, 2016. IEEE; 2016:55-60. doi:10.1109/UPCON.2016.7894624
18. Zarà M, Amadio P, Campodonico J, Sandrini L, Barbieri SS. Exosomes in cardiovascular diseases. *Diagnostics*. 2020;10(11):943. doi:10.3390/diagnostics10110943
19. Mhamdi L, Dammak O, Cottin F, Dhaou IB. Artificial intelligence for cardiac diseases diagnosis and prediction using ECG images on embedded systems. *Biomedicines*. 2022;10(8):2013.
20. Kumar NK, Sindhu GS, Prashanthi DK, Sulthana AS. Analysis and Prediction of Cardio Vascular Disease Using Machine Learning Classifiers: 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, March 6-7, 2020. IEEE; 2020:15-21. doi:10.1109/ICACCS48705.2020.9074183
21. Al-Mahmud O, Khan K, Roy R, Mashuque Alamgir FInternet of Things (IoT) Based Smart Health Care Medical Box For Elderly People: International Conference for Emerging Technology (INCET), Belgaum, India, June 5-7, 2020. IEEE; 2020:1-6. doi:10.1109/INCET49848.2020.9153994
22. Oliveira BAS, Castro GZ, Ferreira GLM, Guimarães FG. *CML-Cardio: A Cascade Machine Learning Model to Predict Cardiovascular Disease Risk as a Primary Prevention Strategy*. Medical & Biological Engineering & Computing; 2023:1-17.
23. Kumar Das P, Zhu F, Chen S, Luo C, Ranjan P, Xiong G. Smart Medical Healthcare of Internet of Medical Things (IOMT): Application of Non-Contact Sensing: 14th IEEE Conference on Industrial Electronics and Applications (ICIEA), Xi'an, China, June 19-21, 2019. IEEE; 2019:375-380. doi:10.1109/ICIEA.2019.8833992
24. Mathur P, Srivastava S, Xu X, Mehta JL. Artificial intelligence, machine Learning, and cardiovascular disease. *Clin Med Insights Cardiol*. 2020;14:117954682092740. doi:10.1177/1179546820927404
25. Kodali RK, Swamy G, Lakshmi B. An Implementation of IoT for Healthcare: IEEE Recent Advances in Intelligent Computational Systems (RAICS), Trivandrum, Kerala, India, December 10-12, 2015. IEEE; 2015:411-416. doi:10.1109/RAICS.2015.7488451
26. Tr R, Lilhore UK, Poongodi M, Simaiya S, Kaur S, Hamdi A. Predictive analysis of heart diseases with machine learning approaches. *Malays J Comput Sci*. 2022:132-148. doi:10.22452/mjcs.sp2022no1.10
27. Aiswarya S, Ramesh K, Sasikumar S. IoT Based Big Data Analytics in Healthcare: A Survey: Proceedings of the First International Conference on Advanced Scientific Innovation in Science, Engineering and Technology, ICASISET 2020, Chennai, India, 16–17 May 2020. European Alliance for Innovation; 2021. doi:10.4108/eai.16-5-2020.2304020
28. Aziz S, Khan MU, Alhaisoni M, Akram T, Altaf M. Phonocardiogram signal processing for automatic diagnosis of congenital heart disorders through fusion of temporal and cepstral features. *Sensors*. 2020;20(13):3790. doi:10.3390/s20133790
29. Siddiqui SY, Athar A, Khan MA, et al. Modelling, simulation and optimization of diagnosis cardiovascular disease using computational intelligence approaches. *J Med Imaging Health Informatics*. 2020;10(5):1005-1022. doi:10.1166/jmihi.2020.2996
30. Shuvo S. An early detection of heart disease using machine learning (recurrent neural network): ML research on heart disease prediction. ScienceOpen Preprints; 2023.
31. Tsao CW, Aday AW, Almarzooq ZI, et al, American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics—2023 update: a report from the American Heart Association. *Circulation*. 2023;147(8):e93-e621.

How to cite this article: Sinha A, Narula D, Pandey SK, et al. CARDPSoML: comparative approach to analyze and predict cardiovascular disease based on medical report data and feature fusion approach. *Health Sci Rep*. 2024;7:e1802. doi:10.1002/hsr2.1802