

ARTICLE TYPE

An Improved Hybrid Model for Cardiovascular Disease Detection Using Machine Learning in IoT

Arslan Naseer¹ | Muhammad Muheet Khan¹ | Fahim Arif¹ | Waseem Iqbal*^{2,3} | Awais Ahmad⁴ | Ijaz Ahmad⁵

¹Department of Software Engineering,
National University of Sciences and
Technology, Islamabad-44000, Pakistan

²Department of Information Security,
National University of Sciences and
Technology, Islamabad-44000, Pakistan

³Department of Electrical and Computer
Engineering, College of Engineering, Sultan
Qaboos University, Muscat-123, Oman

⁴Information Systems Department, College
of Computer and Information Sciences,
Imam Mohammad Ibn Saud Islamic
University (IMSIU), Riyadh-11432, KSA

⁵University of Oulu, Finland

Correspondence

*Waseem Iqbal, Email:
waseem.iqbal@mcs.edu.pk

Abstract

Cardiovascular disease (CVD) believes to be a major cause of transience and indisposition worldwide. Early diagnosis and timely intervention are critical in preventing the progression of CVD and improving patient outcomes. Machine learning (ML) algorithms have emerged as powerful tools in CVD recognition, with the potential to assist physicians in making accurate and efficient diagnoses. This research paper explores the combination of multiple ML algorithms for CVD recognition, utilizing diverse datasets such as the Cleveland, Hungarian, Switzerland, statlog, and VA Long Beach datasets. Additionally, a CVD dataset comprising 12 attributes and 70,000 records is employed, demonstrating improved results through the proposed and trained model compared to previous prediction techniques for CVD. The performance of various ML techniques, including support vector machines (SVM), Naive Bayes (NB), K-Nearest Neighbor (KNN), Random Forest (RF), and Logistic Regression (LR), is evaluated and compared. The impact of feature selection and feature scaling on the models' performance is also examined. An ensemble bagging techniques is applied which is being embedded with other classifiers. LR classifier embedded with bagging techniques proved to be our proposed model. The findings reveal that the proposed Hybrid Linear Regression Bagging Model (HLRBM) outperforms other models. Furthermore, the study highlights the significance of data preprocessing techniques, such as data normalization and class balancing, which significantly enhance the performance of all models. To this end, standard scalar and Synthetic Minority Over-sampling Technique (SMOTE) are employed. The study emphasizes the importance of selecting an appropriate ensemble technique in conjunction with various ML algorithms and preprocessing methods for CVD prediction. Overall, the research provides valuable insights into the potential of ML in improving CVD risk assessment.

KEYWORDS:

Cardiovascular Disease; Machine Learning; Healthcare; IoT

1 | INTRODUCTION

The heart proves to be a vibrant organ that plays a vibrant role in a body's inclusive functionality. Indicators of CVD can fluctuate depending on the explicit situation; however it may include chest tightness or pain, discomfort, shortness of breath, fatigue, dizziness, and swelling in the legs or abdomen ¹. Without a properly functioning of heart, the body cannot survive. These situations can be instigated by various aspects, like drastic change in lipid profile which includes cholesterol, triglyceride and some other factors like diabetes, high blood pressure, smoking, sleeplessness, obesity, lack of physical activity, and family history ². Heart disease disturbs around 126 million individuals which is 1.72% of World's population ³ and it causes one-third of deaths worldwide. Some people may experience no symptoms at all, particularly in the early stages of the disease ⁴. Early detection of CVD is crucial for timely intervention and prevention of adverse outcomes. With the recent advancements in ML, there has been a growing interest in using ML algorithms for CVD recognition. ML algorithms have shown promising results in detecting CVD by analyzing various factors, such as demographic data, medical history, clinical examination results, and laboratory test results. In this context, multiple ML algorithms have been proposed, in particular SVM, NB, KNN, RF, and LR, among others. This approach involves training these algorithms with a dataset of CVD patients and non-CVD patients and using them to detect CVD in a new patient. The objective of our research work is to provide an overview of the numerous ML classifiers that have been proposed for CVD recognition and compare their performance based on numerous metrics, such as precision, recall, sensitivity, specificity and accuracy⁵. Ultimately, this research could lead to the development of a reliable and efficient CVD recognition system. This system could aid healthcare professionals in making timely and accurate diagnoses, improving patient outcomes and quality of life ^{6, 7}. Moreover, the efficiency of these algorithms is extremely reliant on the worth of input data. However, mobile health equipment can also be utilized to implant CVD recognition system using mobile gadgets. These mobile technologies will gather factual data of patients and will deliver proficient health services. Live monitoring of CVD patients can be controlled without making them visit to clinical health centers ^{8, 9, 10}. Fig.1, shows Architecture of CVD Prediction system in IoT.

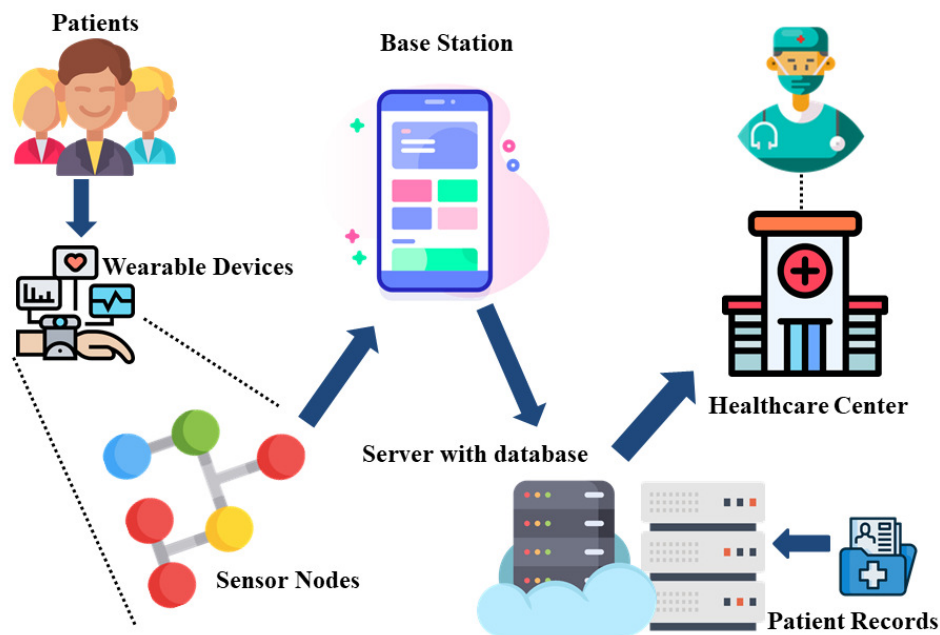


FIGURE 1 Illustration of overall architecture of CVD prediction model used in IoT

⁰Abbreviations: ANA, anti-nuclear antibodies; APC, antigen-presenting cells; IRF, interferon regulatory factor

Feature selection and feature scaling techniques can help to improve the quality of input data for ML algorithms. Feature selection methods aim to identify the most useful attributes that contribute to the accuracy of the model while reducing the dimensionality of the data¹¹. Feature scaling techniques, on the other hand, aim to standardize the data to improve the convergence rate of the ML algorithms. The combinations of multiple ML algorithms, feature selection techniques and feature scaling methods have the potential to further enhance accuracy and robustness of CVD recognition models. By selecting the most relevant features and normalizing the data, these models can reduce overfitting and improve generalization to new data. In addition, by combining the outputs of multiple ML algorithms, these models can capture complementary information and improve overall prediction performance. The main objective and motivation behind this research is to construct an effective model to detect CVD as accurate and precisely as possible. Fig. 2, represents block diagram of proposed hybrid model and required steps followed in this research are summarized as follow:-

- Five Datasets are combined to prepare an effective and mature dataset
- Data Preprocessing techniques standard scalar and SMOTE are used for normalizing and Class balancing of data
- A Comparison of results is drawn which indicates the difference between with and without preprocessing techniques applied on dataset
- Various ML algorithm like SVM, NB, KNN, RF, LR were applied on UCI
- After preprocessing of data, all five applied algorithm were tempted with bagging method to achieve better results. Proposed Hybrid Model (HLRBM) overtakes in achieving accuracy
- To endorse efficacy and performance, recommended model is applied to another CVD dataset having 7000 records with 12 attributes
- A comparison of results is drawn with existing results of former researchers

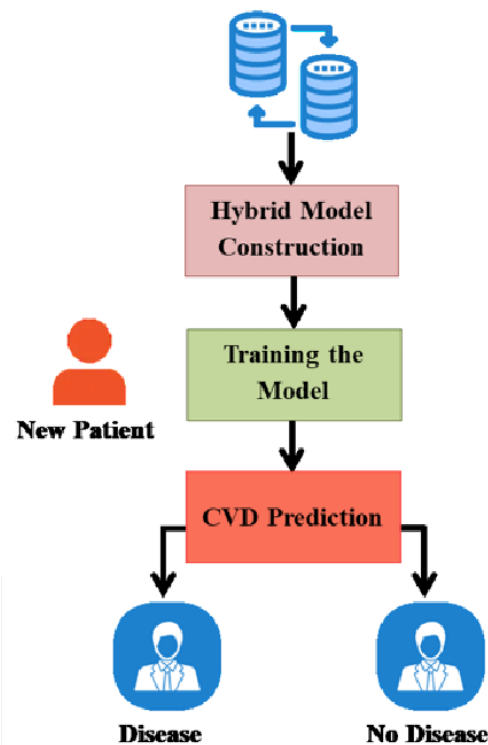


FIGURE 2 Block Diagram of CVD Prediction Model

1.1 | Layout of Paper

The Paper layout will be followed as: Section 2 describes the Literature review and section 3 covers the research methodology followed by implementation of proposed model in section 4. Results and analysis are discussed in section 5 and paper is concluded in section 6.

2 | LITERATURE REVIEW

CVD is a primary cause of mortality worldwide, proving the development of accurate and reliable predictive models an urgent task. In 2020, it has been reported that approximately 244.1 million people are surviving with heart diseases¹² whereas 19.1 million deaths were caused by CVD globally. ML techniques have shown great potential in predicting the risk of CVD, and several studies have been conducted to compare their performance. ML algorithms have emerged as favorable tools for predicting the risk of CVD based on a wide range of patient data. In this literature review, we summarize the recent research on CVD prediction using multiple ML algorithms. We examine the methodologies, datasets, and performance metrics used in various studies, and identify the strengths and limitations of different ML algorithms¹³ for CVD prediction. Some of the research works are addressed below:-

In this paper¹⁴, several ML techniques including RF, LR, SVM, NB and Adaboost (AB) were utilized to detect CVD. The Cleveland dataset was used from UCI repository in which missing values were accredited by MICE algorithm. The authors have improved results of classifiers with feature selection technique. Standard Scalar and SMOTE technique were also used for preprocessing of data. The suggested model by the authors has delivered accuracy of 86.6% which was better than all the applied techniques.

The system was established to diagnoses heart diseases¹⁵ based on ML classifiers which includes ANN, LR, KNN, SVM, NB and DT. For removing immaterial and unessential features, the authors have used feature selection technique like Minimal redundancy maximal relevance, Relief, least absolute shrinkage selection operator (LASSO) and Local learning. Outcomes indicated the suggested diagnosis system, fast conditional mutual information feasible with SVM (FCMIM-SVM) attained better accuracy as compared to formerly proposed methods.¹⁶ In this proposed research, Switzerland, Hungarian, Long Beach VA and Cleveland CVD datasets from UCI database were used. Dataset contained total of 920 records along with 76 features associated with CVD. Preprocessing of data techniques were applied like elimination of noisy data, redundant values, filling of omitted values and sorting of attributes were carried out. The authors have taken only 14 attributes based on which heart diseases were diagnosed through proposed model. Multiple ML classifiers were used including RF, NB, SVM, Gradient Boosting (GB) and LR. The performance of each selected algorithm was obtained which include accuracy, sensitivity and specificity analysis.

In¹⁷, A unique model was introduced in this paper by uniting five CVD datasets (Cleveland, Switzerland, Long Beach VA, Hungarian and Statlog) to make a larger, mature, and trustworthy dataset. For selection of suitable features LASSO and Relief techniques were utilized which also helped to overcome underfitting and overfitting glitches of machine learning. New hybrid classifiers were introduced by incorporating customary algorithms with bagging and boosting techniques which were used to train dataset.

¹⁸ Authors have formulated a prediction model by using hybrid RF with a linear model (HRFLM) to improve its performance level. UCI Cleveland dataset was used on which an analytical approach was applied with three connotation rules (apriori, predictive and Tertius) of mining to discover features of CVD. Authors have utilized R studio rattle to achieve CVD catalog. The proposed model HRFLM deliver improved results with an accuracy of 88.7%.

A fusion model was proposed to envisage CVD by utilizing DT and RF algorithm in¹⁹. Authors have applied both methods individually and later applied combination of these models which produced better results on Cleveland dataset collected from uci.edu. A basic GUI interface was formulated to predict heart disease by giving all required values as input that produced a binary classified calculation. Both models along with hybrid model were applied and mean square error, R-Squared parameter, mean absolute error, root mean square error and accuracy of applied models were calculated and plotted on a graph for a better comparison. Rahul and Sunit²⁰ suggested a comparative study and analysis in oct 2020, of all the ML techniques used for prediction of CVD. Authors have given a detailed analysis of each model and advantages / disadvantages were discussed for each machine learning algorithm used for prediction of CVD. Comparison was done to check performance level of each method based on accuracy, precision, sensitivity, recall and error²¹. Supervised ML classifiers like RF, NB, DT and KNN on UCI, Cleveland

dataset were applied in this research paper. In order to attain precise and effective outcomes data preprocessing was done to avoid issues related missing values and noisy data. Performance level of each algorithm was determined and plotted based on precision, recall and accuracy resultantly K-nearest neighbor produced a highest accuracy as compared to former methods used for Cleveland dataset. In this paper ²², CVD dataset was used which contained 7000 records with 12 attributes where 11 are input features and one output feature. The authors have applied various ML classifiers such as ANN, RF, DT, SVM, KNN and ANN. The optimal results were attained by ANN along with the use of Genetic algorithm (GA). The authors claimed to produce 5.08 percentage improvements of results as compared to other algorithms applied. The results were achieved by using GA-ANN with 3 layers as prominent parameters and 64 neurons. For better results, Softmax function and adagrad optimizer were used which gave 73.3% average accuracy. In ²³, the authors have recommended a hybrid model using KNN and ANN for CVD prediction. Cleveland dataset from UCI was used which have 14 features with 303 records. In this dataset 13 attributes are input features where as one feature is output indicating presence or absence of heart disease against data of each patient. Table 1, illustrates the overall existing works described in literature review.

TABLE 1 Comprehensive comparison of existing work of former researchers

Reference	Classifiers	Features Used	Accuracy (%)	Year of Publication	Dataset Used
Pooja Rani et al: ¹⁴	NB, SVM, LR, RF, AB & Hybrid Model	14	86.60	2021	UCI, Cleveland CVD Dataset
J.P Li et al: ¹⁵	LR, KNN, ANN, SVM, NB and DT	14	89.50	2020	-do-
P.Gosh et al: ¹⁷	DT, RF, KNN, AB and GB	10, 11, 13	89.07	2020	-do-
S.Mohan et al: ¹⁸	NB, LR, DT, RF, SVM and HRFLM	14	88.4	2019	-do-
M.Kavitha et al: ¹⁹	DT, RF and Hybrid (DT+RF)	14	88.7	2021	-do-
Devnsh et al: ²¹	RF, NB, DT and KNN	14	90.0	2020	-do-
Jan Carlo et al: ²²	RF, DT, SVM, KNN and ANN	12	73.3	2022	CVD Dataset with 70k records
D Kumar et al: ¹⁶	LR, RF, NB, GB and SVM	14	86.51	2021	5 CVD datasets Combined

3 | RESEARCH METHODOLOGY

3.1 | Overview of Proposed Model

In this study, a CVD prediction system has been developed by the authors. The CVD datasets used in this research work is publically available on UCI repository. Former researchers utilized the same datasets in which most appropriate thirteen features were selected by correlation coefficient technique for attaining better CVD prediction. Missing values were handled by Multiple Imputation by Chained Equations (MICE) as the primary method to address this issue. MICE is a well-established technique for imputing missing data in complex datasets, particularly in the context of healthcare and epidemiological studies ultimately enhancing the reliability and validity of our CVD prediction results. The three phases of this hybrid system are data gathering, data preprocessing and the model creation. Features are scaled by standard scalar, and class balancing is carried out through

SMOTE technique during the preprocessing stage. For better results before applying classifiers data must be standardized or normalized. The standard scalar is used to standardized the data, guaranteeing that each feature has a mean of 0 (μ) and a standard deviation (Σ) of 1. Conversion formula from ²⁴ is appended below:-

$$\text{Standardization, } X = \frac{X - \mu}{\sigma} \quad (1)$$

The SMOTE technique is also used for class balancing to handle imbalance data. ML traditional classifiers like SVM, NB, KNN, RF and LR were applied on selected features. After data preprocessing, the data was split into 80% of training and remaining 20% of data into test data. Various ensemble methods with traditional classifiers are imputed to create a combination over same dataset. Finally, the classifier determines if the person is CVD positive or negative. A remarkable difference is observed in results while applying these classifiers when used without data preprocessing and after preprocessing techniques. Different training methods are applied to check the performance of each model so that we can choose finest hybrid model for our trustworthy dataset ²⁵. However, our proposed model HLRBM resulted in providing more accurate results than other models. Fig. 3 depicts the framework and application technique of recommended hybrid model for heart disease prediction.

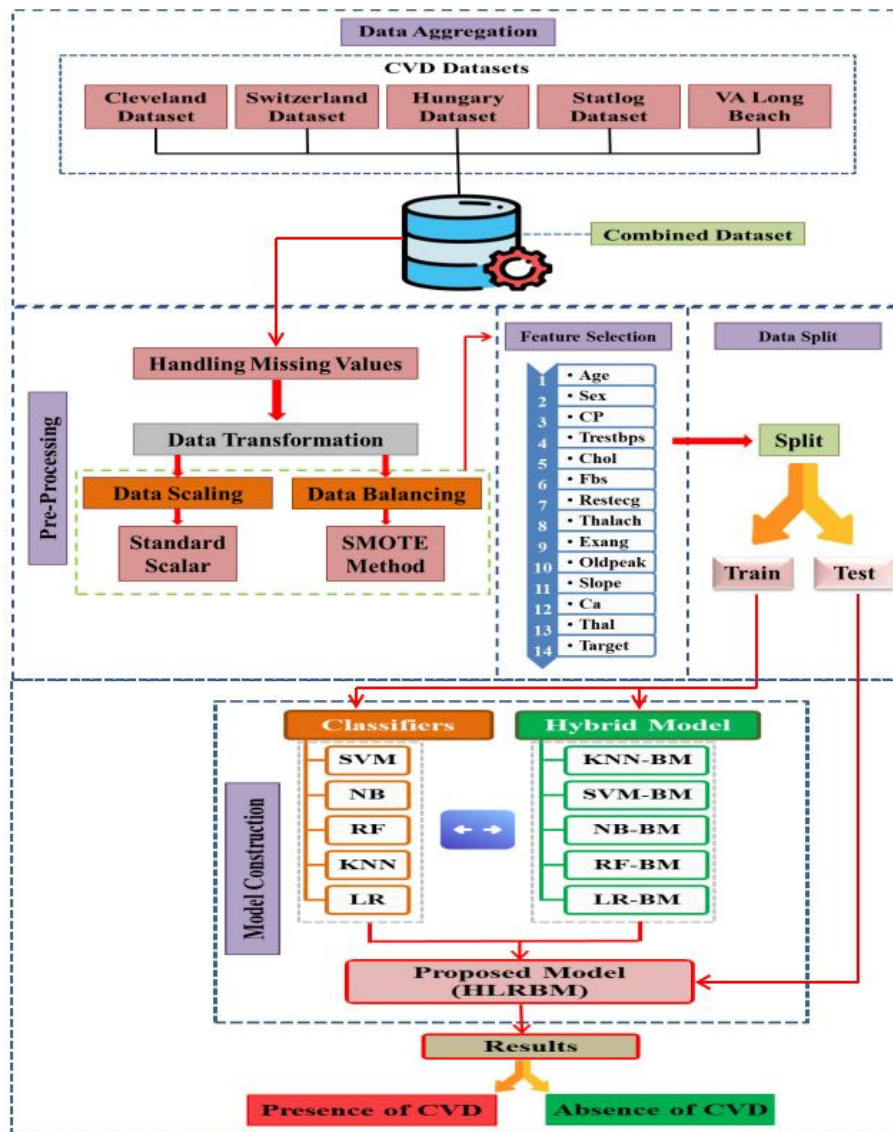


FIGURE 3 Suggested Architecture of proposed Model for CVD detection

3.2 | Dataset Description

The datasets available on UCI repository ^{26,27} were utilized to construct an effective dataset by merging five CVD datasets (Switzerland, VA long Beach, Cleveland, Statlog and Hungary). Table 2, describes the dataset's attributes along with their value ranges. The dataset used in this research work contains 303 records having 14 attributes out of which 13 are input features along with 1 output feature indicating presence or absence of CVD. Following are the comprehensive depiction of each attribute:-

- **Age:** The age of a person in completed years. This feature signifies the patient's age at the time of data collection
- **Sex:** This attribute represents the patient's gender and is encoded as 0 indicating female and 1 indicating male
- **Type of Chest Pain:** This attribute provides information about the chest pain encountered by the patient. It is classified into four categories: 1 denoting typical angina, 2 indicating atypical angina, 3 for non-angina pain and 4 indicating asymptomatic.
- **Resting Blood Pressure:** This feature symbolizes the patient's blood pressure while at rest, measured in mm Hg (millimeters of mercury) upon admission in hospital
- **Serum Cholesterol:** The cholesterol level is measured in milligrams per deciliter (mg/dl) indicates the amount of serum cholesterol present in the patient's blood
- **Blood Sugar in Fasting:** This attribute provides information about the patient's fasting blood sugar level, measured in mg/dl. A value of 1 indicates high blood sugar, defined as greater than 120 mg/dl, while a value of 0 indicates normal blood sugar, equal to or less than 120 mg/dl.
- **Resting Electrocardiographic Results:** This attribute represents the results of the patient's resting electrocardiogram (ECG) test. It is encoded as follows: 0 denoting normal, 1 indicating ST-T wave abnormality, or 2 presenting apparent or certain left ventricular hypertrophy.
- **Maximum Heart Rate:** This attribute indicates the peak heart rate attained by the patient during exercise.
- **Exercise-Induced Angina:** This attribute indicates whether the patient experienced angina (chest pain) while exercising. A value of 1 indicates the presence of induced angina, while a value of 0 signifies the absence of angina.
- **ST Depression Induced by Exercise Relative to Rest:** This feature measures the amount of ST depression observed during exercise compared to the patient's resting stage ²⁸.
- **Slope of the Peak Exercise ST Segment:** This attribute represents the shape of the ST segment during the peak exercise, classified as 1 denotes upsloping, 2 indicating flat or 3 for downsloping ^{28,29}.
- **Number of Major Vessels Colored by Fluoroscopy:** This attribute denotes the count of major blood vessels that have been visualized and colored through fluoroscopy. The possible values range from 0 to 3 ²⁸.
- **Thallium Stress Test:** The attribute shows the results of the thallium stress test, which measures blood flow to the heart. It is encoded as 3 indicating normal, 6 denoting fixed defect, and 7 represents reversible defect.
- **Target:** The target variable indicates the presence of cardiovascular disease. A 0 value means no presence of disease, while reading greater than 0 represents the presence of disease.

3.3 | Application of Proposed Hybrid Model

The suggested model proves to be an efficient model once its suitable application is justified and it also aid to deal with real world challenges. Fig. 4, illustrates the workflow of our proposed model. This intelligent devised model can be utilized in various health centers to predict CVD in an effective way. The following procedure can be followed to attain prediction of CVD.

- Data of each patient is collected and put together in a database
- The main features of patient's data will be selected as an input to our proposed model HLRBM to perform prediction
- Selected features will be handled in our trained model

TABLE 2 Dataset Description in details including value ranges and Data types

No.	Feature Name	Feature Code	Description	Value Ranges	Data Type
1	Age	Age	Age in years completed	between 29 and 77	Numeric
2	Sex	Sex	Male: 1, female: 0	0 or 1	Nominal
3	Type of chest pain	CP	Typical angina: 1, atypical angina: 2 non-angina pain: 3, asymptomatic: 4	1 to 4	Nominal
4	Resting blood pressure	Trestbps	Patient's Resting Blood Pressure Range	94 to 200 mm Hg	Numeric
5	Serum cholesterol	Chol	Cholesterol level in mg/dl	126 to 564 mg/dl	Numeric
6	Fasting blood sugar	Fbs	Fasting Blood Sugar >120 mg/dl (true:1, false: 0)	0 or 1	Nominal
7	Resting electrocardiographic results	Restecg	Normal: 0, ST-T wave abnormality:1, Hypertrophy: 2)	0, 1 and 2	Nominal
8	Maximum heart rate	Thalach	Heart Rate of Patients	71 to 202	Numeric
9	Exercise-induced angina	Exang	Patient experienced angina during exercise(Yes=1, No=0)	0 or 1	Nominal
10	ST depression induced by exercise relative to rest	Oldpeak	Depression caused by exercise, Up sloping: 1, Flat: 2, down sloping: 3	1 to 3	Numeric
11	The slope of the peak exercise	ST segment Slope	Slope of peak exercise	1, 2, 3	Nominal
12	Number of major vessels (0–3) colored by fluoroscopy	Ca	Major Vessels colored by fluoroscopy with range 0 to 3	0 to 3	Numeric
13	Thallium	Thal	Represents thallium stress test, Normal:3, fixed defect: 6, reversible defect: 7	3, 6, 7	Nominal
14	Target	Target	Output, Heart disease present: 1, heart disease absent: 0	0 or 1	Nominal

- As a result, binary output will be generated either 0 or 1. Result 0 identifies as negative (absence of CVD). 1 in case of CVD results are positive
- In case of 1 patient will be guided to visit Heart Specialists for further investigations
- Data of each patient will be preserved in a database to aid better CVD prediction in future

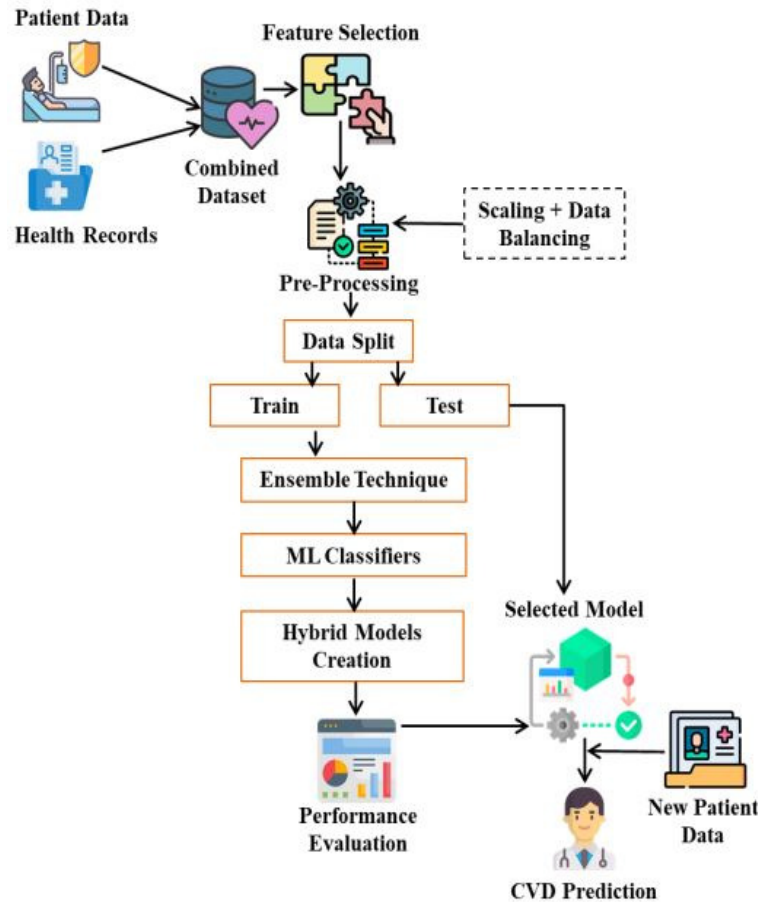


FIGURE 4 Workflow of Proposed CVD Prediction System

3.4 | Justification of Proposed Hybrid Model

The CVD datasets used in this research work is publically available on UCI repository ^{26,27}. Former researchers utilized the same datasets in which most appropriate thirteen features were selected by correlation coefficient technique for attaining better CVD prediction. Missing values were handled by Multiple Imputation by Chained Equations (MICE) as the primary method to address this issue. MICE is a well-established technique for imputing missing data in complex datasets, particularly in the context of healthcare and epidemiological studies ultimately enhancing the reliability and validity of our CVD prediction results. The unique hybrid model has been devised by using five various ML classifiers. Consequently, an ensemble Bagging technique is used to make this model efficient and persistent in achieving results. data preprocessing techniques used for data normalization and class balancing remains the hallmark difference of former results. After achieving improvements in result in data preprocessing phase, an ensemble bagging techniques is applied as base classifier which is being embedded with other classifiers. Logistic regression classifier embedded with Bagging techniques as based classifier proved to be our proposed model. All the five models

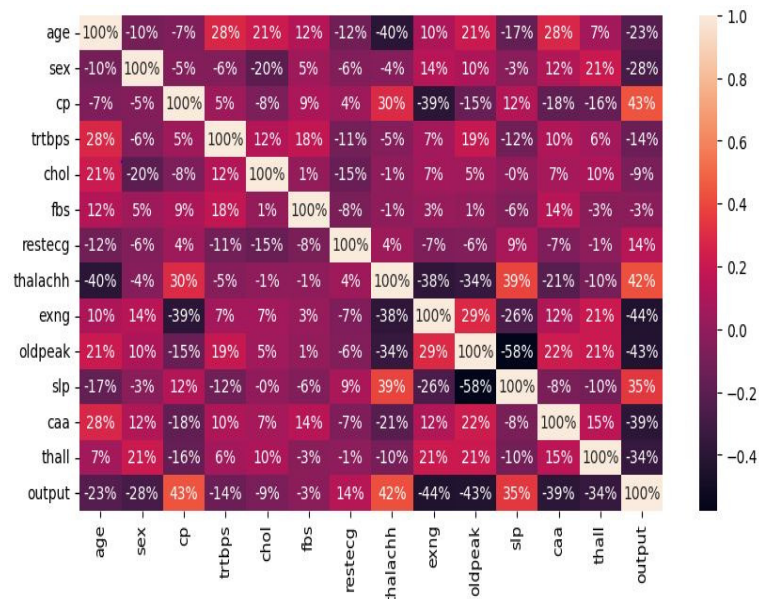


FIGURE 5 Representation of Co-related features of CVD

produced better results where as our proposed model HLRBM overtakes in achieving accuracy. Our paper highlights the former researchers results which have used ensemble technique with various ML classifiers and their results are shown in our comparison to our achieved results. Various studies have already been conducted by using various ML algorithm that contracts with the same dataset don't produced better results as expected. After an exhausted study we came to a conclusion that some models don't perform well as those systems don't identify most important and highly co-related features. Fig. 5, represents the co-relation of feature of CVD dataset. For that matter we tried to split highly co-related features and grouped together. Some features are having numeric values and some features are using nominal values according to their nature of readings. We tried to make it unique while using our proposed model to another CVD dataset publically available having 70000 records with 12 attributes. 11 attributes are serves as input features where the 12th feature is output feature giving result in binary form either 0 or 1 indicating presence or absence of CVD. A comparison is carried out of former outputs with improved results achieved by our proposed model on same dataset having 70000 rows (record of patients). Pseudo code of proposed model is illustrated below:-

4 | IMPLEMENTATION

In this section, implementation of proposed hybrid model HLRBM is described in details.

4.1 | Environment

All the experiments are executed using language Python 3 in google colab. 8GB RAM with windows 10 on Intel(R) Core(TM) i7-1065G7 CPU @ 1.30GHz 1.50 GHz is used in this research work.

4.2 | Data Preprocessing Techniques

In this modern era, a large amount of data can be collected via the internet, different valuable experiments and surveys etc. Most of the time data collected for research is noisy data and contain missing or null values. In this case, some popular techniques are applied like deletion and imputation that can deal with missing values. Furthermore, before applying any kind of ML algorithm data must be normalized or class balancing is executed to make an efficient dataset. standard scalar for scaling the dataset was used that normalized the data and furthermore SMOTE technique was used to balance the dataset. The techniques used for preprocessing of data are described as:-

Algorithm 1 Pseudo code of proposed Model**Require:** CVD Datasets, ML Classifiers**Ensure:** *CVDAccuracy, Precision, Recall, F1score, ROC – AUC*Datasets \leftarrow {UCI (Cleveland, Hungarian, Switzerland, statlog, VA long Beach)}Scalars \leftarrow {Standard scalar (min-max), SMOTE}Classifiers \leftarrow {SVM, NB, KNN, RF, LR}

SMOTE-SVM, SMOTE-NB, SMOTE-KNN, SMOTE-RF, SMOTE-LR

for model \in Classifiers **do** model \leftarrow TrainClassifier(ScaledX_train, X_trainTarget)

model.fit(X_train, Y_train)

 y_predict \leftarrow model.predict(X_test) Accuracy \leftarrow ComputeAccuracy(y_predict, X_TestTarget) Precision \leftarrow ComputePrecision(y_predict, X_TestTarget) Recall \leftarrow ComputeRecall(y_predict, X_TestTarget) F1score \leftarrow ComputeF1score(y_predict, X_TestTarget) ROC-AUC \leftarrow ComputeROC(y_predict, X_TestTarget)**end for****Return** {Accuracy, Precision, Recall, F1score, ROC-AUC score}

- **Standard Scalar:** Standard Scalar is a widely used scaling technique in ML. This normalization process guarantees that all features are uniformly scaled and prevents certain features from exerting excessive influence on the learning procedure. By standardizing the features, convergence of optimization algorithms is improved, facilitating faster and more stable training. It also enables fair comparison and interpretation of feature effects on the model. Standard Scalar is particularly beneficial for algorithms relying on distance metrics and is commonly applied in SVM, linear regression, LR, and neural networks.
- **SMOTE:** Synthetic Minority Over-sampling Technique (SMOTE) is a method used in ML to address imbalanced datasets. By interpolating between existing instances of the minority class, this technique generates mockups to augment the minority class. SMOTE helps in class balancing, reducing over fitting and improving model performance. It retains the inherent characteristics of the minority class while increasing the number of samples. However, it is important to be cautious of potential noise or overgeneralization introduced by the synthetic samples. Evaluating the performance with and without SMOTE is recommended to determine its suitability in a given scenario ³⁰.

After data preprocessing phase, the data is distributed into 80% of training and remaining 20% into Test data. The algorithms were trained and on training data followed by achievement of results from Test data. Five algorithms which include SVM, NB, KNN, RF and LR were applied on dataset for achieving results. A comparison is drawn of attained results of with and without using data preprocessing techniques.

4.3 | Ensemble Techniques of ML

Ensemble techniques are assorted with classifiers to achieve better results for achieving CVD detection. The main purpose behind this is that weak learners combined can work with strong learners to become an efficient combination ³¹. Fig. 5 illustrates the ensemble process ³¹. Mainly Bagging and Boosting techniques are used to produced more accurate and efficient results. In our research work we used Bagging method which is described below:-

- **Bagging Method:** Bagging (Bootstrap Aggregating) is a ML ensemble technique where multiple models are trained on various subsections of training data, created through bootstrapping (sampling with replacement). The main purpose is to generate numerous divisions of data out of training models. Arbitrarily elected pools of subset data are recycled to train their DT. Resultantly, we acquire an ensemble of various models ³². The individual models, typically decision trees, are then combined through voting or averaging to make predictions, resulting in improved accuracy and reduced variance. Bagging helps to reduce overfitting and improve generalization by leveraging the diversity among the models. It also helps resolving of missing values problems and preserves accuracy

- **Boosting Method:** Boosting is a ML ensemble technique which sequentially trains a series of weak models to create a robust model. Every fragile model are trained on a modified version of the training data, where the misclassified samples are given higher weights. The absolute prediction is made by joining the estimation of all the fragile models. Boosting iteratively focuses on the trials that are hard to categorize, continuously improving the model's performance. It helps to reduce bias and increase accuracy by emphasizing the learning from previously misclassified examples. Usually Boosting builds better predictive models ³³.

In our research work we utilized Bagging technique with five classifiers which include SVM, NB, KNN, RF and LR to formulate hybrid models. The hybrid Models: SVM-BM, NB-BM, KNN-BM, RF-BM and HLRBM are constructed and applied on training and Test data of our dataset. Resultantly, our proposed model HLRBM outperformed and produced better results among other models.

4.4 | Classification Modeling

Multiple classifiers used in our research work which are embedded with ensemble technique. Each model has its own impact on dataset. Specific description of used algorithms are given below :-

- **SVM:** A Support Vector Machine (SVM) is a supervised ML classifier that splits data into various modules by discovering the best hyper plane in a high-dimensional attribute space. It exploits the edges, the distance among the hyper plane and the nearest data points from each class, to improve generalization. SVM can handle non-linear boundaries using the "kernel trick" that transforms the feature space. They are trained by minimizing classification errors while maximizing the margin. SVMs are used for classification and regression tasks, and their applications include text categorization, image classification, and financial forecasting. They are powerful, versatile algorithms that excel in handling high-dimensional data and complex patterns.
- **Naïve Bayes:** The NB algorithm is a simple yet powerful ML classifier based on Bayes' hypothesis. It accepts that all attributes in a dataset are independent of each other, hence the "naive" assumption. The probability of a particular instance belonging to a specific class is calculated by multiplying the conditional probabilities of each feature, given that class the algorithm works well with large datasets and is computationally efficient. Despite its simplifications, NB often produces competitive results, making it a popular choice in machine learning applications.

$$\mathbb{P}(A | B) = \mathbb{P}(B | A) * \frac{\mathbb{P}(A)}{\mathbb{P}(B)} \quad (2)$$

Here, the probability we aim to calculate, $\mathbb{P}(A | B)$, is referred to as the posterior probability, while the prior probability of the event, $\mathbb{P}(A)$, is known as the marginal probability ³⁴.

- **Logistic Regression:** The LR algorithm is supervised ML technique utilized for binary classification related problems. It simulates the association among input features and the probability of belonging to a particular class by the logistic function. LR estimates the parameters by minimizing the logistic loss function, typically using gradient descent optimization. During training, for a given input, LR computes the possibility for each class and assigns the predicted class as the one with the highest probability. LR is popular due to its simplicity, interpretability, and efficiency. It is widely used in numerous applications, such as disease diagnosis, sentiment analysis and spam detection where the aim is to predict binary outcomes based on input features.
- **Random Forest:** The RF algorithm is a ML technique that utilizes an ensemble approach by combining the predictions of multiple decision trees. This combination allows for more accurate and reliable predictions. By randomly selecting subsets of the training data and features for each tree, the algorithm creates a "forest" of DT. During training, each tree learns patterns and makes predictions independently. When making predictions, the algorithm aggregates the predictions of all the trees to determine the final outcome through voting or averaging. They are used for classification and regression tasks, and their applications include areas such as finance, healthcare, and image recognition, where accurate predictions and interpretability are essential.
- **KNN:** This algorithm is most versatile and intuitive classification algorithm. To classify new instances, this approach identifies the K nearest neighbors in the training set and assigns the majority class label among them ³⁵. KNN measures

the similarity between instances using a distance metric, typically Euclidean distance. KNN relies on the supposition that instances in the similar class are adjacent to each other in attribute space. Being a non-parametric algorithm, it does not rely on explicit assumptions regarding the underlying data distribution. KNN has the ability to handle classification problems with multiple classes and is also applicable in regression tasks. While simple and easy to implement, KNN's performance can be sensitive to the choice of K and the feature scaling.

4.5 | Evaluation Parameters

On a scale of precision, recall, F1 score, accuracy and ROC-AUC classifiers performance were assessed. If a patient with the condition is anticipated the system determines the person to have cardiac disease, then the result is a true positive; in other case it's a false negative. Similar to the last example, a prediction that a healthy person will remain disease-free is said to be a true negative; and false positive in other case. These terms are precisely defined below ³⁶:-

- **True Positive (TP):** Instances correctly identified as positive when they are truly positive.
- **True Negative (TN):** Instances correctly identified as negative when they are truly negative.
- **False Positive (FP):** Instances incorrectly identified as positive when they are actually negative.
- **False Negative (FN):** Instances incorrectly identified as negative when they are actually positive.
- **Accuracy:** It is a performance metric that measures the system's ability to make accurate predictions.

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN} \quad (3)$$

- **Precision:** Precision quantifies the system's ability to generate only relevant results.

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

- **Recall:** Recall assesses the model's effectiveness in correctly identifying all positive instances, representing the ratio of true positives to all actual positives.

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

- **F-Measure:** It combines the outcomes of precision and recall by utilizing the harmonic mean.

$$F1Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (6)$$

4.6 | Experimental Work with proposed Hybrid Model

The technique that is used to improve accuracy of results are utilizing more time on data preprocessing techniques. Resultantly, better accuracies are achieved in each applied model. These 12 attributes provide a range of information about an individual's demographics, physical characteristics, lifestyle factors, and medical history. By analyzing and modeling these attributes, researchers and healthcare professionals can have perceptions of the risk factors linked with CVD and make strategies for anticipation and treatment. To check the efficacy of our recommended model we used another Dataset ³⁷ from kaggle ML repository. The dataset contains 70000 records with 12 attributes. Out of 12 attributes 1 to 11 attributes are considered to be input features and 12th attribute is output feature. The dataset ³⁷ includes attributes such as age in completed years, sex (male or female), height, weight, systolic blood pressure, diastolic blood pressure, cholesterol levels, glucose levels, smoking status, alcohol consumption, physical activity level and the presence or absence of CVD. A comparison of results is drawn in results and analysis section which indicates that result achieved by our recommended model are better than results achieved at ²².

5 | RESULTS AND ANALYSIS

A range of classification methods is employed to diagnose patients with heart disease, including SVM, NB, KNN, RF and LR. The UCI Cleveland dataset is used for the studies. The diagnosis of heart disease was made using several medical parameters extracted from the dataset. The classification was performed by utilizing these factors, where class 1 indicated the presence of an illness and class 0 indicated the absence of a disease ³⁸.

5.1 | Performance of classifiers without preprocessing

Firstly, experiments are performed on all selected 14 features without applying any kind of data preprocessing technique. System performance is evaluated using metrics such as accuracy, precision, recall, ROC-AUC, and F-measure. Table 3, Indicates the results of five classifiers applied on all 14 features without using any data preprocessing techniques:-

TABLE 3 Performance of classifiers without applying preprocessing techniques

Classifier	Accuracy	Precision	Recall	F1 Score	ROC-AUC
SVM	68.85	67	85	75	68
NB	81.96	81	88	85	82
KNN	63.93	71	59	65	64
RF	83.64	83	88	86	84
LR	82.01	81	88	85	82

5.2 | Performance of Classifiers with Scaling and Class Balancing technique

Scaling the features ensures that all features are treated equally by the algorithms, resulting in more effective and stable optimization processes. Consequently, this leads to enhanced performance and improved generalization on unseen data. However, class balancing enhances the representation of the minority class by generating synthetic samples, resulting in a more balanced dataset and improved learning by the classifier. Standard Scalar and SMOTE technique used in research work proved to be a best suitable on our dataset as standard scalar brings the features to a comparable scale, results in improved model performance, stability and interpretability. However, the efficacy of SMOTE relies on the characteristics of the particular dataset and the specific problem being addressed. The dataset utilized in this research demonstrated compatibility with the SMOTE technique being employed. Results achieved indicate that scaling and class balancing have positive influence on each classifier. Table 4, indicates improvements of results achieved by data preprocessing techniques are better than the former ones. Overall, accuracy of each

TABLE 4 Improvement in Performance with the use of preprocessing techniques

Classifier	Accuracy	Precision	Recall	F1 Score	ROC - AUC
SVM	86.88	84	94	89	87
NB	85.52	91	88	90	87
KNN	81.96	81	88	85	82
RF	85.24	88	85	87	85
LR	85.24	86	88	87	85

classifier is improved where specifically accuracy increased by SVM is 18.03%, NB increased by 3.56%, KNN improvement is 18.03%, RF increased by 1.6% and LR improved by 3.23%. Possible reasons for wide range improvement by SVM and KNN could be the effectiveness of SVM relies on various factors, including the selection of the kernel, fine-tuning of hyperparameters, and the inherent characteristics of the data. Whereas KNN demonstrates strong performance with small datasets, mitigating the risk of overfitting. Fig.[6], illustrates that with the use of scaling and classing balancing technique, results of each evaluation parameters in term of precision, accuracy, recall and F1 score are improved. Graphical representation of each parameter is indicated below:-

5.3 | Performance improvement using Ensemble Technique

This section describes that Ensemble Bagging technique is used and all five classifiers are treated as based classifiers and tested which produced further improvements in results. The main reason behind these improved results could be the primary advantage

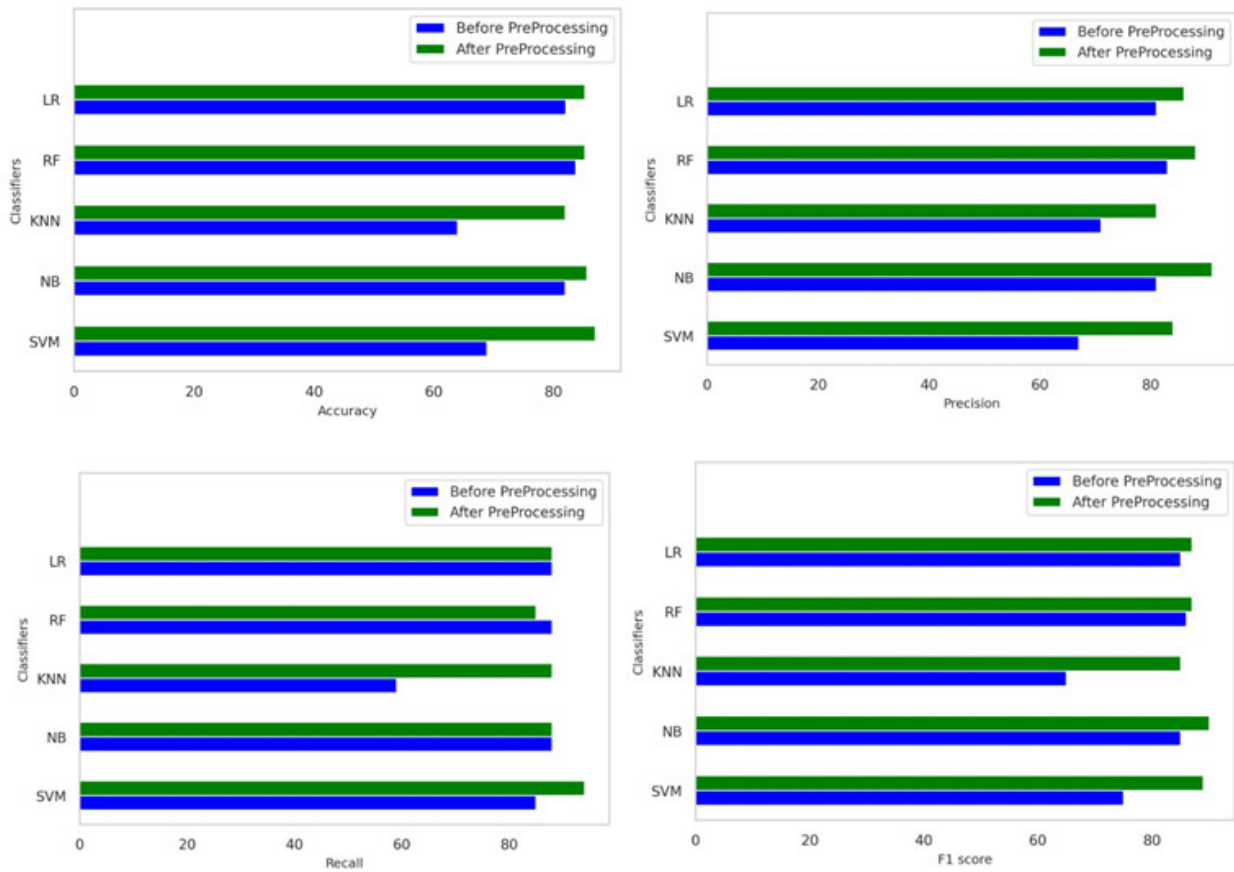


FIGURE 6 Performnce Improvement of classifiers using Standard Scalar and SMOTE Technique

of the bagging ensemble technique that reduces variance and overfitting by aggregating predictions from multiple base models, resulting in more robust and accurate predictions. The all five hybrid model embedded with ensemble bagging techniques proved to be fruitful on account of results achieved as precision, accuracy, recall and F1 score. Table 5, illustrates the improvements of results in each case where as LR embedded with bagging method produced better accuracy. The advantages offered by the

TABLE 5 Results of numerous Models with Proposed Hybrid Model

Classifier	Accuracy	Precision	Recall	F1 Score	ROC-AUC
SVM-BM	88.52	91	88	90	89
NB-BM	86.88	88	88	88	87
KNN-BM	86.93	84	94	89	87
RF-BM	86.88	86	91	89	87
HLRBM	90.16	93	88	90	90

bagging ensemble with LR make it a competitive choice, particularly when compared to individual classifiers with their unique limitations and performance trade-offs. However, its effectiveness may vary depending on the specific dataset and problem at hand. It worked better than other classifiers by mitigating overfitting to noise and outliers in the dataset, the ensemble ensures more accurate predictions on unseen data and significantly improves performance. A graphical presentation of achieved results is visualized in fig.7.

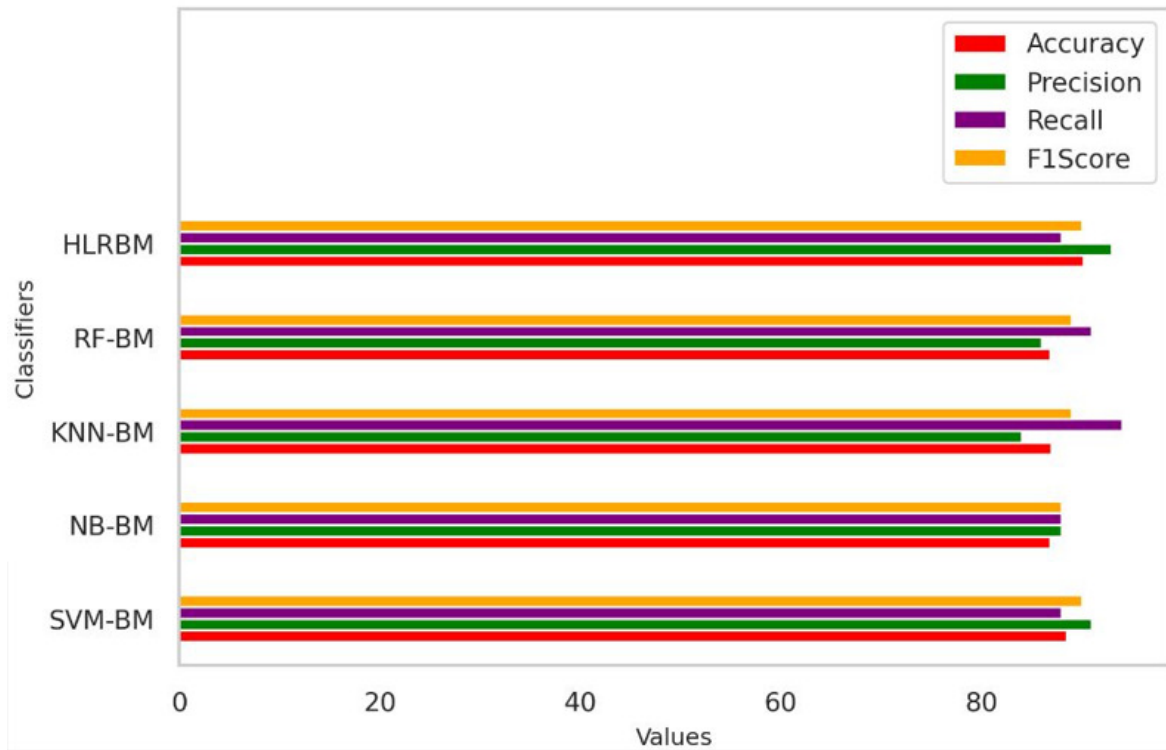


FIGURE 7 Graphical representation of each classifier with proposed Model

5.4 | Comparison of Results with former Researcher Results

A comparison of results achieved is drawn with existing researcher’s results. Our research work proves that with the use of better data preprocessing techniques and formulation of unique hybrid model overall satisfactory results can be achieved for prediction of CVD. Table 6. illustrates the overall comparison of application of numerous models on same dataset by various researchers. In fig. 8, A detailed comparison is visualized in graphical form to have better idea of results in term of accuracy of each classifiers applied on same dataset.

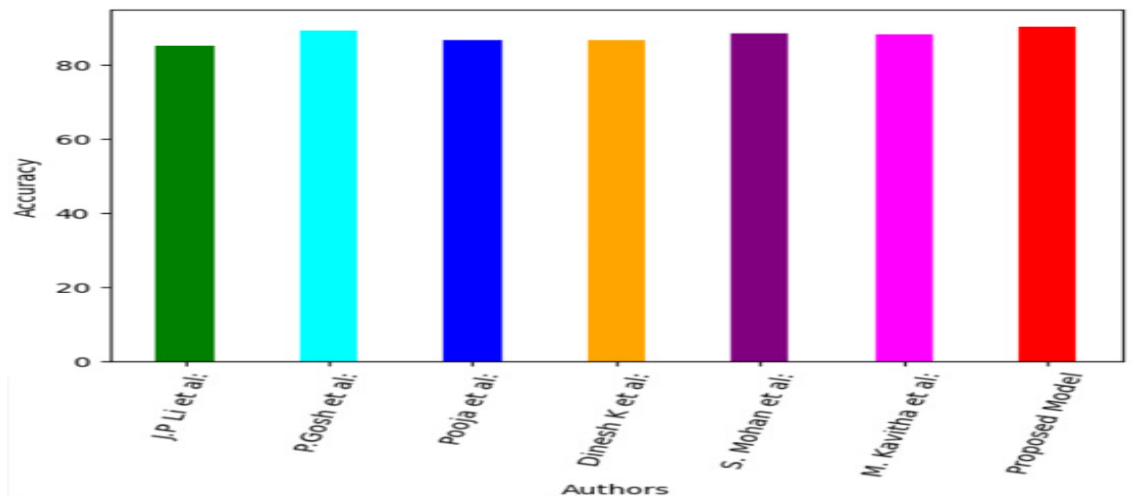


FIGURE 8 Improvement in accuracy of Proposed Model with existing work

TABLE 6 Comparison of proposed Model with various existing research work

Reference	Classifiers	Accuracy %	Precision	Recall/ (Sen%)	F1 Score
15	LR	83	95	75	-
	K-NN	69	70	64	-
	ANN	60	100	0	-
	SVM	85	95	75	-
	NB	75	90	78	-
	DT	70	72	83	-
17	DT	86.97	-	-	-
	RF	88.65	-	-	-
	K-NN	83.61	-	-	-
	AB	89.07	-	-	-
	GB	86.97	-	-	-
14	NB	85.07	84.37	82.31	83.33
	SVM	84.16	86.92	81.09	83.91
	LR	83.24	86.62	82.92	84.73
	RF	83.85	88.46	84.14	86.25
	AdaBoost	82.34	88.96	83.53	86.16
	Hybrid Model	86.60	-	-	-
16	LR	86.51	-	-	-
	RF	80.89	-	-	-
	NB	84.26	-	-	-
	GB	84.26	-	-	-
	SVM	79.77	-	-	-
18	NB	75.8	90.5	79.8	84.5
	LR	82.9	89.6	91.1	90.2
	DT	85	86.0	98.8	91.8
	RF	86.1	87.1	98.8	92.4
	SVM	86.1	86.1	100	92.5
	HRFLM (Hybrid)	88.4	90.1	92.8	90
19	DT	79	-	-	-
	RF	81	-	-	-
	Hybrid (DT+RF)	88	-	-	-
Our Work	SVM-BM	86.88	84	94	89
	NB-BM	88.52	91	98	90
	KNN-BM	86.93	89	91	90
	RF-BM	88.52	89	91	90
	HLRBM	90.16	93	88	90

5.5 | Results of additional experimental work

All the preprocessing techniques are applied to another dataset which is publically available ³⁷. The dataset used as an additional experiment to check the efficacy of our proposed model comprising 70000 records with 12 attributes which are huge in numbers as compared to CVD datasets utilized from UCI repository. Being a large datasets it took longer time to attain accuracy for each classification model applied. The proposed model took about 2 minutes and 42 seconds to attain its outcome in term of accuracy, precision, recall and F1 Score. Five algorithms were applied including recommended Model that gave the maximum results of the former results. The efficacy and efficiency of our research work is proved to be productive after having comparison of results on existing dataset. Table 7. shows the results achieved on datasets utilized to check performance of recommended

model on other than Cleveland dataset used in earlier part. A graphical visualization is illustrated in fig. 9, to compare evaluation parameters against proposed models.

TABLE 7 Results achieved by recommended models on different datasets.

Classifier	Accuracy	Precision	Recall	F1 Score	ROC - AUC
SVM-BM	76.66	77	86	81	77
NB-BM	72.55	66	93	77	73
KNN-BM	72.62	76	65	70	73
RF-BM	76.25	74	80	77	76
HLRBM	77.90	81	71	76	78

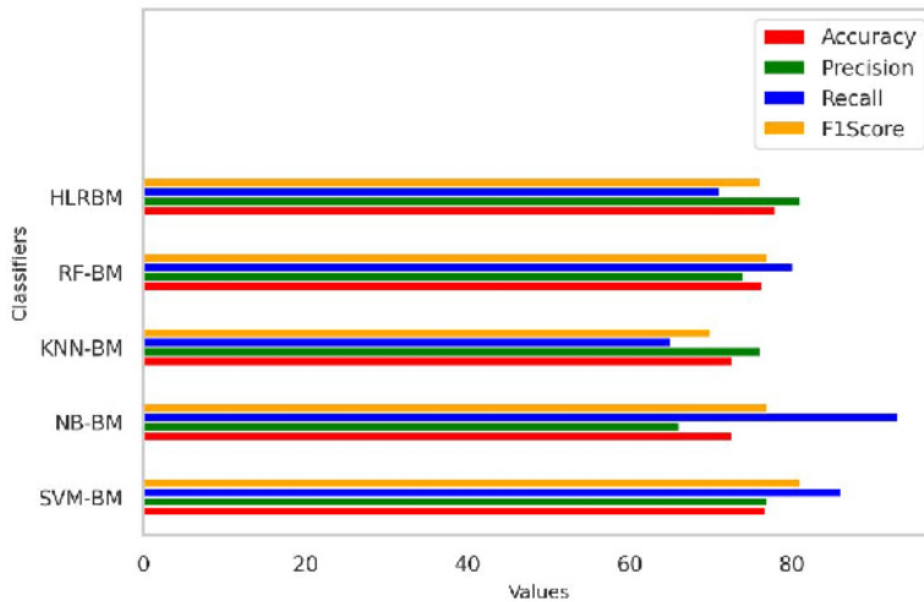


FIGURE 9 Graphical representation of results achieved by proposed model on different dataset

Experimental work is carried out to check efficacy of our proposed model. A comparison is drawn to show its performance on different dataset having 70000 records publically available. Table 8, Shows its performance against existing work on this dataset.

6 | CONCLUSION

The development of an improved hybrid model HLRBM for CVD detection using ML shows significant promising results. The model combines multiple ML classifiers and leverages relevant medical parameters to achieve enhanced accuracy and efficiency in identifying CVD. The results obtained from this study demonstrate the potential of ML techniques in improving the accuracy of CVD detection, which can aid in early diagnosis and intervention, leading to better patient outcomes. However, there are several avenues for future work in this area. Firstly, the model could benefit from incorporating additional patient data from diverse populations to improve its generalizability and robustness. This would help ensure that the model performs effectively across different demographic groups and can be widely applicable in real-world healthcare settings. Furthermore, the integration of advanced deep learning methods, such as recurrent neural networks or attention mechanisms, could potentially

TABLE 8 Comparison of results achieved by proposed model with former researcher's results

Reference	Algorithms	Accuracy (%)	Precision	Recall	F1 Score
22	ANN	68.35	-	-	-
	LR	72.35	-	-	-
	DT	61.72	-	-	-
	RF	68.94	-	-	-
	SVM	72.16	-	-	-
	KNN	68.34	-	-	-
	GA-ANN	73.43	-	-	-
Our Work	SVM-BM	76.66	77	81	77
	NB-BM	72.55	66	77	73
	KNN-BM	72.62	76	70	73
	RF-BM	76.25	74	77	76
	HLRBM	77.90	81	76	78

enhance the model's performance by capturing complex temporal dependencies and extracting more informative features from the data. Additionally, the model's interpretability can be further explored by incorporating explainable AI techniques, allowing healthcare professionals to understand the underlying reasons behind the model's predictions. This would increase trust and acceptance of the model within the medical community. Lastly, conducting extensive validation studies using large-scale clinical datasets and comparing the performance of the hybrid model against existing diagnostic methods would be crucial for assessing its clinical utility and effectiveness in real-world scenarios. Overall, the improved hybrid model for CVD detection shows promise in revolutionizing e-healthcare systems. Future research and development in this field can greatly contribute to advancing early detection and proactive management of CVD, ultimately improving patient care and reducing the burden on healthcare systems.

6.1 | Bibliography

References

- Heart Disease Symptoms and Causes - Mayo Clinic. <https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118>; .
- Subhadra K, Vikas B. Neural network based intelligent system for predicting heart disease. *International Journal of Innovative Technology and Exploring Engineering* 2019; 8(5): 484–487.
- Khan M, Hashim M, Mustafa H, others . Global Epidemiology of Ischemic Heart Disease: Results from the Global Burden of Disease Study. *Cureus* 2020; 12(7): e9349. doi: 10.7759/cureus.9349
- Coronary Heart Disease Symptoms - NHLBI, NIH. <https://www.nhlbi.nih.gov/health-topics/coronary-heart-disease-symptoms>; .
- Kausar F, Mesbah M, Iqbal W, Ahmad A, Sayyed I. Fall Detection in the Elderly using Different Machine Learning Algorithms with Optimal Window Size. *Mobile Networks and Applications* 2023: 1–11.
- Jain A, Tiwari S, Sapra V. Two-phase heart disease diagnosis system using deep learning. *International Journal of Control and Automation* 2019; 12(5): 558–573.
- Yaseen M, Iqbal W, Rashid I, et al. Marc: A novel framework for detecting mitm attacks in ehealthcare ble systems. *Journal of medical systems* 2019; 43: 1–18.
- Vithanwattana N, Mapp G, George C. Developing a comprehensive information security framework for mHealth: a detailed analysis. *Journal of Reliable Intelligent Environments* 2017; 3(1): 21–39. doi: 10.1007/s40860-017-0038-x
- Jusob FR, George C, Mapp G. Exploring the need for a suitable privacy framework for mHealth when managing chronic diseases. *Journal of Reliable Intelligent Environments* 2017; 3(4): 243–256. doi: 10.1007/s40860-017-0049-7

10. Ahmad A, Din S, Paul A, Jeon G, Aloqaily M, Ahmad M. Real-time route planning and data dissemination for urban scenarios using the Internet of Things. *IEEE Wireless Communications* 2019; 26(6): 50–55.
11. Ahmed I, Ahmad A, Piccialli F, Sangaiah AK, Jeon G. A robust features-based person tracker for overhead views in industrial environment. *IEEE Internet of Things Journal* 2017; 5(3): 1598–1605.
12. Tsao CW, Aday AW, Almarzooq ZI, others . Heart disease and stroke statistics—2022 update: a report from the American Heart Association. *Circulation* 2022. doi: 10.1161/CIR.0000000000001052
13. Muzammal M, Gohar M, Rahman AU, Qu Q, Ahmad A, Jeon G. Trajectory mining using uncertain sensor data. *IEEE Access* 2017; 6: 4895–4903.
14. Rani P, Kumar R, Ahmed NMOS, others . A decision support system for heart disease prediction based upon machine learning. *Journal of Reliable Intelligent Environments* 2021; 7: 263–275. doi: 10.1007/s40860-021-00133-6
15. Li JP, Haq AU, Din SU, Khan J, Khan A, Saboor A. Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare. *IEEE Access* 2020; 8: 107562–107582. doi: 10.1109/ACCESS.2020.3001149
16. Dinesh KG, Arumugaraj K, Santhosh KD, Mareeswari V. Prediction of Cardiovascular Disease Using Machine Learning Algorithms. In: ; 2018: 1–7
17. Ghosh P, others . Efficient Prediction of Cardiovascular Disease Using Machine Learning Algorithms With Relief and LASSO Feature Selection Techniques. *IEEE Access* 2021; 9: 19304–19326. doi: 10.1109/ACCESS.2021.3053759
18. Mohan S, Chandrasegar T. Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access* 2019; 7: 81542–81554.
19. Kavitha M, Gnaneswar G, Dinesh R, Rohith Sai Y, Sai Suraj R. Heart disease prediction using hybrid machine learning model. In: IEEE. ; 2021: 1329–1333.
20. Katarya R, Meena SK. Machine Learning Techniques for Heart Disease Prediction: A Comparative Study and Analysis. *Health Technology* 2021; 11: 87–97. doi: 10.1007/s12553-020-00505-7
21. Shah D, Patel S, Bharti SK. Heart Disease Prediction using Machine Learning Techniques. *SN Computer Science* 2020; 1: 345. doi: 10.1007/s42979-020-00365-y
22. Arroyo JCT, Delima AJP. An Optimized Neural Network Using Genetic Algorithm for Cardiovascular Disease Prediction. *Journal of Advances in Information Technology* 2022; 13(1): 95–99.
23. Malav A, Kadam K, Kamat P. Prediction of Heart Disease Using K-Means and Artificial Neural Network as Hybrid Approach to Improve Accuracy. *International Journal of Engineering and Technology* 2017; 9: 3081–3085. doi: 10.21817/ijet/2017/v9i4/170904101
24. Acharya A. Comparative study of machine learning algorithms for heart disease prediction. 2017.
25. Majumder S, Pratihari DK. Multi-sensors data fusion through fuzzy clustering and predictive tools. *Expert Systems with Applications* 2018; 107: 165–172.
26. Heart Disease Datasets From UCI Machine Learning Repository. <https://archive.ics.uci.edu/ml/datasets/Heart+Disease>; .
27. Heart Disease Statlog Dataset of UCI Machine Learning Repository. [http://archive.ics.uci.edu/ml/datasets/statlog+\(heart\)](http://archive.ics.uci.edu/ml/datasets/statlog+(heart)); .
28. Alneamy JSM, Alnaish RAH. Heart disease diagnosis utilizing hybrid fuzzy wavelet neural network and teaching learning based optimization algorithm. *Advances in Artificial Neural Systems* 2014: 6–6.
29. Marreiros G, Martins B, Paiva A, Ribeiro B, Sardinha A., eds., *Progress in Artificial Intelligence: 21st EPIA Conference on Artificial Intelligence, EPIA 2022, Lisbon, Portugal, August 31–September 2, 2022, Proceedings*. 13566; Springer Nature: 2022.

30. Hasanin T, Khoshgoftaar TM, Leevy JL, Seliya N. Examining characteristics of predictive models with imbalanced big data. *Journal of Big Data* 2019; 6(1): 1–21.
31. Latha CBC, Jeeva SC. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatcs in Medicine Unlocked* 2019; 16(2): 100203.
32. Ensemble Techniques of Bagging. <https://quantdare.com/what-is-the-differencebetweenBagging-and-Boosting/>; .
33. An Explanation of Ensemble Bagging Techniques. <https://towardsdatascience.com/ensemble-methods-Bagging-Boosting-and-stackingc9214a10a205/>; .
34. Posterior Probability: Definition + Example - Statology. <https://www.statology.org/posterior-probability/>; .
35. Lee W. *Python Machine Learning* . 2019.
36. Taamneh M. Investigating the role of socio-economic factors in comprehension of traffic signs using decision tree algorithm. *Journal of Safety Research* 2018; 66: 121–129.
37. Ulianova S. Cardiovascular Disease dataset. <https://www.kaggle.com/sulianova/cardiovascular-disease-dataset>; 2019.
38. Morita K, Tsuka H, Kuremoto KI, others . Association between buccal mucosa ridging and oral feature/symptom and its effects on occlusal function among dentate young adults in a cross-sectional study of Japan. *CRANIO®* 2019.

How to cite this article: Williams K., B. Hoskins, R. Lee, G. Masato, and T. Woollings (2016), A regime analysis of Atlantic winter jet variability applied to evaluate HadGEM3-GC2, *Q.J.R. Meteorol. Soc.*, 2017;00:1–6.