Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 15s (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

An Ensemble Machine Learning-Based Classification for Cardiovascular Disease Prediction Using PCA and SVM with Bagging

S Anthony Mariya Kumari[*1], Viji Vinod[2]

[1]Research Scholar, Department of Computer Applications, Dr. M.G.R Educational and Research Institute, Chennai-95, Tamil Nadu, India. (

Email ID: selvarajanthony94@gmail.com

[2]Professor & Head, Department of Computer Applications, Dr. M.G.R Educational and Research Institute, Chennai-95, Tamil Nadu, India.

Email ID: vijivino@gmail.com, Email ID: hod-mca@drmgrdu.ac.in

Cite this paper as: S Anthony Mariya Kumari, Viji Vinod, (2025) An Ensemble Machine Learning-Based Classification for Cardiovascular Disease Prediction Using PCA and SVM with Bagging. *Journal of Neonatal Surgery*, 14 (15s), 1749-1759.

ABSTRACT

Background: Cardiovascular diseases (CVD) remain the leading cause of mortality worldwide, highlighting the need for accurate and timely diagnosis. Machine learning has emerged as a promising tool for enhancing predictive accuracy in medical diagnosis.

Objective: This research aims to predict cardiovascular diseases using various machine learning algorithms applied to the Erbil Cardiovascular Health Dataset (Clinical data) from Mendeley. The objective is to improve prediction accuracy through ensemble learning and to extract the feature for better results and by calculating the feature importance.

Methods: The study involves data pre-processing using regression based filling by calculating mean, median for missing values, feature extraction using Principal Component Analysis (PCA), and for classification we are using the feature extracted dataset and applying Random Forest, Support Vector Machine (SVM), and Logistic Regression (LR). An ensemble method, bagging, was introduced to enhance model robustness and accuracy.

Results: The ensemble models demonstrated improved accuracy compared to independent models. Support Vector Machine with Bagging achieved 97% accuracy, Random Forest with Bagging reached 92%, and Logistic Regression with Bagging achieved 95%. Independent models without bagging showed lower accuracies: Support Vector Machine at 91.04%, Logistic Regression at 88.06%, and Random Forest at 83.58%. The effective of the ensemble method is evaluated using accuracy, precision, recall, log loss f1score.

Conclusion: The results indicate that machine learning models, especially ensemble methods, Support Vector Machine with Bagging achieved highest accuracy of 97% among the other algorithms and it can significantly enhance the early diagnosis and management of cardiovascular diseases, thereby improving patient care and outcomes.

Keywords: *Cardiovascular disease, Machine Learning, PCA, Bagging, Random Forest, SVM, Logistic Regression.*

1. INTRODUCTION

Heart disease remains the world's leading cause of death, claiming almost 18 million lives annually. Rheumatic heart disease and coronary heart disease are among the most frequent causes of diseases affecting the blood vessels in the brain. An important number of cardiovascular disease deaths occur before age 70 and this large percentage comprises over 40% of all such fatalities. Poor diet, physical inactivity, smoking and high alcohol consumption are important behavioral risk factors and these factors are many. Air pollution causes ecological hazards. Obesity, dyslipidemia and hypertension are intermediate risk factors and hyperglycemia is also an intermediate risk factor.

In a society where cardiovascular illnesses have become a silent pandemic, accounting for at least 27% of fatalities, the need of maintaining good heart health has never been greater. World Heart Day is commemorated annually on September 29. On this day, several healthcare providers collaborate with the government and international organizations to raise awareness about various types of CVD, as well as the numerous lifestyle habits and poor diets that contribute to the deterioration of heart health, resulting in thousands of such illnesses and deaths.[19].The early diagnosis and prediction of CVD are essential to the wholesome measures of prevention and cure. Conventional methods open disillusioning challenges in tackling the

complexities of medical data of great volume and variety. ML is an extremely important tool in healthcare, which collects large data sets, recognizes patterns, and makes accurate predictions. Thus, it improves the future prediction of heart disease through algorithms while taking decisions by the doctors and acting on them quickly. The introduction of ML in heart care may also enhance not only the accuracy of diagnosis but also the very management of end-users.

With advancements in AI, ML, and DL, newer ways of diagnosing and predicting heart diseases have opened a new avenue for improvement [1-4]. These advances are promising to improve the accuracy and efficiency of CVD diagnosis and risk assessment [5-7]. Researchers have searched many ML and DL models, such as ensemble learning and feature selection, to enhance CVD detection systems reliability in predictions [8-13]. The intersection of AI, ML, and DL has improved treatments addressing neurological and cardiovascular conditions. [2,4]. However, it does raise ethical concerns. The progress in earlier detection with better diagnoses may not yet allow individualized treatment plans, but such changes can improve patient outcomes. Gradually but surely, these also make their implementation complicated. This research would develop the complete pipeline building in predicting heart disease utilizing advanced technological methodologies for its improved efficiency.

The study proposes an approach for heart disease prediction using feature extraction and an ensemble method. First, Data preprocessing is done by removing the missing values or replacing the values by calculating mean, median using regression based filling. For dimensionality reduction and to calculate the feature importance's in this study the researcher is using Principle Component Analysis. After calculating feature important we extract ten most important features. Using that featured dataset they are doing classification by using RF, SVM, and LR and created an ensemble method to significantly enhance the prediction by combining with bagging with the existing algorithm. They got the highest accuracy in ensemble model.

2. LITERATURE REVIEW

The authors propose a DL model for early detection of CVD based on data from individuals who underwent screening. The model predicts if a person has CVD and offers awareness or diagnosis. The model outperformed other methods such as long- and short-time memory, feed forward, cascade forward, and Elman neural networks, achieving 98.45% accuracy. This early diagnosis technique can help medical practitioners diagnose CVD more easily [7]. The study is aimed at improving the prediction of cardiovascular risk through irregularly spaced electronic health records with machine learning methods. Their work is spurred by the fact that the traditional models do not always effectively deal with such unpredictable time periods in EHR data and thus can impair accuracy. The authors applied a machine learning pipeline for the extraction of relevant temporal aspect from still-imbalanced medical records in terms of time points. The applied algorithms also included different preprocessing methods to compensate for either missing or sparse records. The trained algorithms predict future incidents of cardiovascular events. The analysis sets the premise for considering temporal data in health predictive and often makes the risk profile more accurate and dependable. This view asserts that machine learning can offer more accurate and reliable assessments in predicting risk for cardiovascular disease, thereby triggering timely and effective action by physicians [8].

Increased mortality rates from CVD make prediction difficult. Bringing in ML in the health area could use CVD prevention effectively. Investigations were carried out to improve addressing class imbalance in the Behavioral Risk Factor Surveillance System 2021 heart disease dataset by resampling techniques such as Synthetic Minority Oversampling Technique, Adaptive Synthetic Sampling, SMOTE-Tomek, and SMOTE-Edited Nearest Neighbor. Then the researchers trained, tested, and evaluated a number of classifiers (including LR, DT, RF, GB, XGBoost, CB, and ANNs) capable of performing excellently on maximizing sensitivity in prediction for CVD risk. In some hybrid resampling regimes, superior to other sampling methods, provided that the proposed implementation considers SMOTE-ENN as combined with Optuna-maintained CatBoost, which hence facilitates notable recall of 88% (sensitivity) and area under receiver operating characteristic curve of 82%. Moreover, the use of ANNs showed their applicability in the structured imbalanced data scenario, which further improved the detection of positive cases therein in the health sector [15].

This study seeks to propose the use of AI for accelerating the early detection of CVDs through training AI models using patient data sets including ECGs, wearable measurements, and medical history. Accuracy in prediction of CVD was presented with machine-learning techniques using RF, SVM, and NN. The models showed advantages over traditional methods, as neural networks detected high-risk patients with an accuracy of 92%. Such AI models for CVD would be used in communities, where CVD assessments would happen, and people would be introduced to early-targeted health interventions. But questions relating to data protection, ethical use, and clinical validation have hovered around [17].

ML has helped in diagnosing heart diseases and taking care of patients in such a manner that helps doctors make wise judgments. The three methods which are put to trail in the research include LR, DT, and SVM. The trial was done by using them with the Boruta feature selection tool and without it. The Heart Disease Dataset was used by the researchers to study the 14 attributes derived from the 303 patients' records at Cleveland Clinic. Out of the six factors selected by Boruta, the models improved after using them. The results show that the most successful classifier was Logistic Regression 88.52%. For overall accuracy, it remains a better classifier. However, Decision Tree and SVM improved with Boruta feature selection [10]. For a strong pipeline of ML in forecasting epidemiological cardiovascular diseases with the Erbil Cardiovascular Health

Dataset (ECHD), for the importance it bears on the Middle East trends. The methodology entailed feature selection, normalization, and ensemble learning techniques that improve performance by combining the selected features and the methods of bagging. This study is mainly into the cardiovascular health of Western population.

3. MATERIAL AND METHODS

To evaluate the effectiveness of these ensemble models, 80% of the data was used for training, and 20% was held out for testing. This split enabled unbiased evaluation of the accuracy with which the models would predict and generalize to new data, ensuring the results were valid and truly represented performance. The overall aim of the research should be to amalgamate the strength of each algorithm used here as RF, SVM, LR. And ensemble classification model is created with Bagging, to overcome its limitations, and build better-performing and robust predictive models.

3.1 Dataset Description

This descriptive statistics Table: 2 provides a detailed overview of the dataset's attributes, summarizing key statistics such as **Count, Mean, Standard Deviation (S.D), Minimum (Min),** and **Maximum (Max)** values. Each attribute represents a feature related to cardiovascular health.

Descriptive Statistics Table: 1

Attribute	Count	Mean	S.D	Min	Max
Age	**333**	**55.12**	**14.16**	**20**	**90**
Sex	**333**	**0.53**	**0.5**	**0**	**1**
HT	**333**	**162.1**	**11.3**	**128**	**192**
WT	**333**	**82.16**	**15.39**	**41**	**134**
LIFESTYLE	**333**	**1.63**	**0.72**	**1**	**3**
FH	**333**	**0.24**	**0.43**	**0**	**1**
PHY_ACT	**333**	**0.37**	**0.48**	**0**	**1**
SMOKING	**333**	**0.2**	**0.4**	**0**	**1**
YRS_SMK	**333**	**4.8**	**11.25**	**0**	**50**
DIAB	**333**	**0.23**	**0.42**	**0**	**1**
LDL_CHOL	**333**	**112.93**	**37.97**	**26**	**260**
CP	**333**	**2.89**	**1.03**	**1**	**4**
HRT_SURG	**333**	**0.26**	**0.44**	**0**	**1**
HR	**333**	**83.88**	**14.63**	**40**	**140**
BP_SYS	**333**	**123.62**	**21.34**	**80**	**220**
BP_DIA	**333**	**74.88**	**12.68**	**40**	**140**
HTN	**333**	**0.52**	**0.5**	**0**	**1**
IVS_DIA	**333**	**0.28**	**0.45**	**0**	**1**
ECG_PAT	**333**	**3.32**	**0.98**	**1**	**4**
Q_WAVE	**333**	**0.08**	**0.27**	**0**	**1**
TARGET	**333**	**0.35**	**0.48**	**0**	**1**

3.2 Class Balance

Class imbalance in the heart disease prediction of the Erbil Cardiovascular Health Dataset may lead to a higher rate of errors due to the wide variability of cases in individuals who are diagnosed and not diagnosed with heart disease. Figure 1 is a pie

chart representing the **distribution of heart disease patients** in a dataset

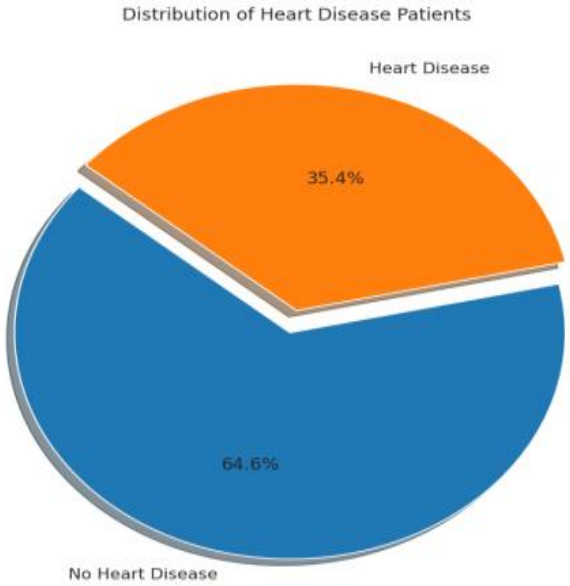

Fig 1: Class Diagram of Heart disease patients

3.3 Data Preprocessing

A step and very important in preparing data for analysis or ML is Data Preprocessing. This process takes care of ensuring dataset organization, consistency, and model readiness. It will cover various steps to solve problems of missing values, noise, and data inconsistencies. The first step here is to deal with the missing data. You can fill in gaps using methods like the mean or median, delete incomplete records, or employ a more advanced approach such as regression-based filling.

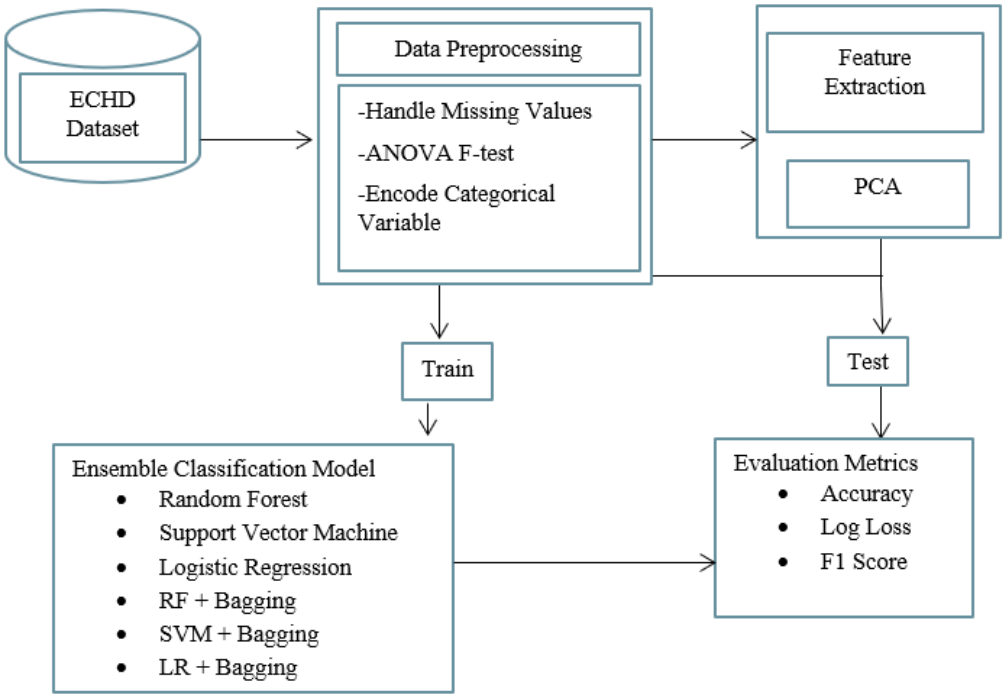

Fig: 2 Block Diagram of Ensemble Classification Model.

Secondly, normalize or standardize data to bring number-based features onto a similar scale.This ensures that attributes with different units or scales don't have too much impact on how well the model works. You then turn category-based variables

into number formats that machine learning algorithms can use. This happens through methods like one-hot encoding or label encoding, finding outliers are crucial. Extreme values can throw off analysis and model predictions.

Fig: 1 image represents a structured flowchart outlining a ML based approach for CVD prediction using the ECHD helps CVD using ML models. The dataset includes heart health information, carefully cleaned to remove errors and unnecessary details. To identify the best prediction model, three methods are applied: RF, SVM, and LR. Bagging is used to improve prediction accuracy by training multiple versions of these models on different parts of dataset. The models' performance is assessed using metrics like accuracy, precision, recall, and F1-score. This method helps detect and prevent heart problems early. The dataset was prepared by handling missing information and making sure all features were consistent. Numerical details, such as Age, Height (HT), and Weight (WT), had missing numbers replaced with the average. For categories like Sex and Smoking, the missing values were filled with the most common value. These categories were then converted into numbers through one-hot encoding, which is necessary for computer processing. We used StandardScaler to adjust the numerical features, making sure they were on the same scale. This step is important for accurate comparison when using machine learning techniques.

3.4 Feature Selection

Feature selection was performed using the SelectKBest method with ANOVA F-test to find the top 10 most significant features for predicting CVD. The selected features included Age, HT, WT, LDL_CHOL, HR, BP_SYS, BP_DIA, IVS_DIA, and encoded categorical variables. The feature importance fig 2 plot, ECG_PAT has an importance of around 0.6 in predicting cardiovascular disease.

Formula $\sum(n_i/N * \Delta i)$ (1)

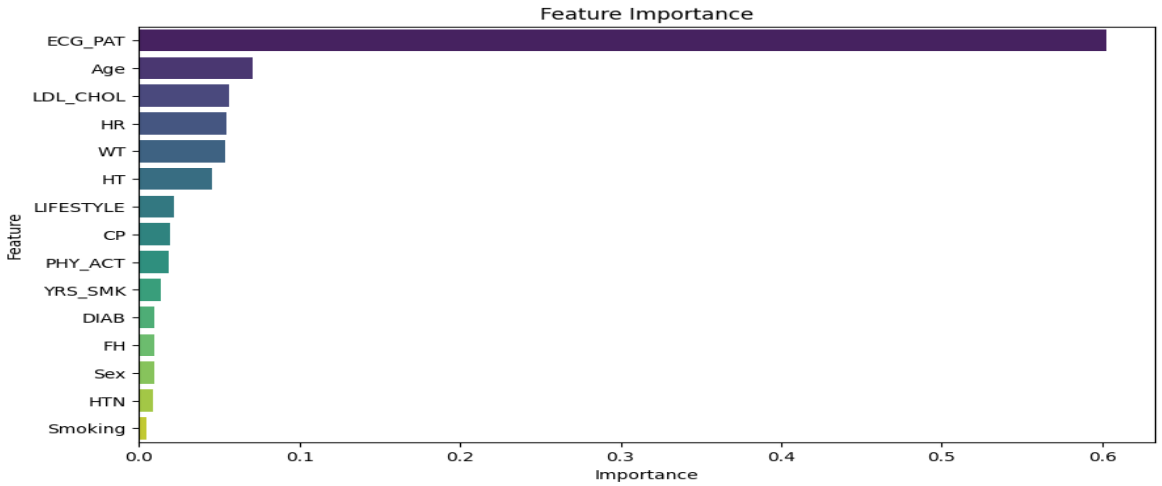

Figure 2: Feature Importance of ECHD Dataset

Many feature importance's are calculated using the Gini importance formula or mean decrease in impurity. The physiological parameters like age, LDL_CHOL, HR, WT, and HT are somewhat important; meanwhile, lifestyle and demographic variables such as smoking, HTN, and sex have lower relevance ratings. The prediction of the model relies mainly on ECG patterns and basic physiological measures.

PCA is a technique in ML aimed at simplifying large datasets. It focuses on identifying the parts of the data that vary the most, retaining essential information while eliminating less significant details. This makes data processing more efficient for computers and enhances the performance of ML models by cutting down on unnecessary data. The steps involved in PCA are as follows: Initially, the data is standardized so that all variables are on the same scale. A covariance matrix is then created to explore the relationships between different data points, helping to pinpoint the directions of maximum variation in the data.

The progressing gains in variance captured as more elements are included into the dataset are illustrated in this chart Fig 3. The cumulative explained variance ratio is used to determine the optimal number of components for efficient feature selection. CVD prediction, choosing too many components results in excessive noise while selecting too few components may exclude some crucial information. The best method is to determine the elbow point at which any additional component only accounts for the least variation improvement.

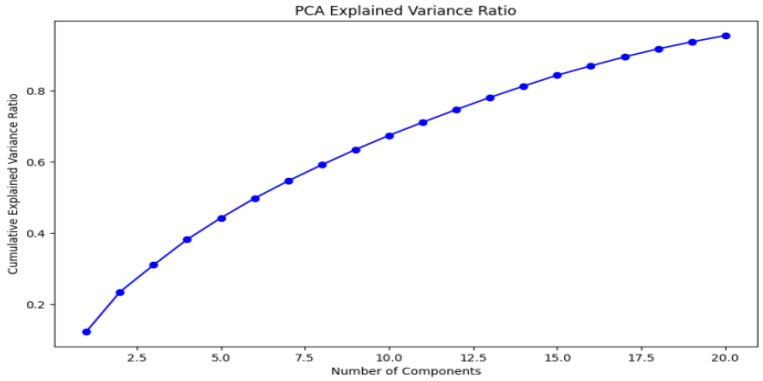

Fig 3: PCA Explained Variance Ratio

After that, PCA calculates eigenvalues and eigenvectors, selecting the largest eigenvalues to highlight the most critical features. The original data is then restructured based on these major components. It is extensively used in fields like image processing, where it aids in better management and analysis of images, in medical diagnostics for interpreting complex datasets, and in finance to analyze vast amounts of financial data. By refining data analysis, PCA provides clearer insights and enhances understanding of information across various domains

The number of primary components is reduced so as to make the process not complicated and therefore enhance the performance of the models with which they are applied. This increases the prediction accuracy of models to pattern-based automatic learning for CVD recognition but keeps clinical data integral from the dataset.

3.5 Traditional Methodology

3.5.1 Random Forest

Random forests are general methods for both classification and regression. It comprises building many decision-trees in the course of training and combining their outputs to improve accuracy while reducing overfitting. Each tree is constructed by sampling from a subset of the data randomly so as to guarantee a varied diversity in the learning process. Further improvement is realized on the diversity through the random selection of characteristics for each shared split. The majority voting is used to determine the final prediction for classification tasks or averaging for regression tasks.

One major advantage of random forest is that it handles such large data sets quite comfortably without generalizing poorly. Besides, it's robust to noise and missing values, which is an asset for real applications, e.g., medical diagnosis and financial forecasting. It randomly alleviates the problem of overfitting, which a single decision tree may cause. Also, it is interesting regarding functions that assist in determining the most important predictors for the forecast. Though it can be computationally expensive, particularly with many trees, the advantages usually surpass disadvantages. In general, the random forest is a sturdy and much-used automatic learning algorithm due to accuracy, versatility and resilience during the processing of complex datasets.

3.5.2 Support Vector Machine

In the last few decades, SVMs have remained the best in supervised learning for classification or regression applications. It works by finding the hyperplane that best separates the classes given in the dataset. The objective of SVM is to maximize the margin between the classes in order to provide a good classification for an unseen data. The points closest to this hyperplane are the support vectors that play a major role in defining the boundary of the classification. SVM works well with high-dimensional data and has been extensively used in fields like medical diagnostics, image recognition, and text classification. The other advantage is that it can be applied to both the simple and highly complex classification problems. For those datasets which can be separated linearly, it finds a hyperplane; for those which cannot be separated linearly, the SVM applies a technique called the kernel trick to transform the original feature space into a new feature space where the data becomes linearly separable. Types of commonly used kernels include linear, polynomial, RBF, and sigmoid kernels, and they differentiate in terms of the data pattern they support.

SVM attempted to maximize the distance between classes so as to minimize the possibility of misclassifying future events. Whenever those points are very close to such hyperplane, they are classified as support vectors, which remarkably give importance to the existence of the decision boundary. Application of SVM is very good in high-dimensional data and widely used in medical diagnosis, image recognition, and text classification. Manikandan G, Pragadeesh has used SVM with boruta feature selection for better accuracy model [10].

Regarding the adapters in the class, it is a classifier for easy and complex problems. When able, SVM will find a hyperplane

with a linear separation of the data; if not, some kernel trickery will be performed on the rest in whatever way transforms it to a data space there that is linearly separable. Some of the most commonly used kernel types are linear, polynomial, radial basis function, and sigmoid, each of which makes different assumptions about the underlying data. SVM can be compute-intensive, especially in the case of large datasets, because it attempts to solve a complicated optimization problem. Proper selection of the kernel and optimization of hyper-parameters are crucial for obtaining good results. Although the above obstacles exist, SVMs perform wonderfully when configured well and have become an essential ingredient for many machine learning tasks. They represent a strong powerful classification toolbest for finding the possible separating hyperplane between two distinct classes. The bagging technique aggregates the predictions of multiple SVM models, each trained on different bootstrap samples, by majority voting, ensuring that this ensemble approach is much more robust to outliers and increases the generalization ability of the classification system.

3.5.3 Logistic Regression

LR is a widely used method in statistics and ML for deciding between two options, like yes or no. Although its name includes "regression," it actually helps classify things rather than predicting numbers. It estimates the chance that something belongs to a specific category by using the logistic (sigmoid) function and a simple equation. The sigmoid function changes any number into values between 0 and 1, which helps in guessing probabilities. If this probability is more than a certain level, often 0.5, it is put into one category; if less, it goes into another.

A big benefit of LR is its simplicity and ease of use. It assigns weights to each feature, showing their impact on the result. This makes it easier to see which factors matter most for predictions. It works best when the relation between the factors and the outcome is almost straight. It's commonly used in fields like medical diagnosis, fraud detection, and marketing, where predicting one of two possible outcomes is important. It has some limitations. It assumes a straight line for making decisions, which means it doesn't do well with complex problems where the data doesn't follow a straight pattern. It can also be affected by outliers, which are data points that are much different from others, and these can mess up how well the model works. Also, it needs the features to be independent and not related to each other to give reliable results. To improve its work, you can use techniques like scaling features, regularization (L1 and L2), and creating polynomial features.

3.6 Ensemble Classification Model

In Ensemble classification model it combine the multiple machine learning algorithms as RF, SVM, LR used to address the class imbalance by combining the bagging to predict the multiple classifiers.

3.6.1 Random Forest + Bagging

A number of DT are used, being an ensemble learning methodology, to predict categorization. This is trained using multiple samples of the data set to ensure robustness. Data preparations include features like blood pressure, cholesterol, ECG. The ultimate prediction is made through majority voting. Random Forest + Bagging analyzed the cardiovascular features to reduce overfitting and consequently enhance generalization. In general, with respect to accuracy, precision, and recall, it has been proved successful in predicting heart disease, while bagging reduces overfitting and demonstrates generalization. When a new patients data is entered to test, each tree makes a prediction for heart disease.

$$y\hat{}=\text{mode}\{a1(X),a2(X),...,aT(X)\} \quad (2)$$

where $a_i(X)$ai(X) represents individual decision tree predictions.

For probability estimation, the mean of predicted probabilities is used:

$$P(y \mid X) = \frac{1}{T}\sum_{i=1}^{T} P_i\,(y \mid X) \text{P(y|X)=T1i=1∑TPi(y|X)} \quad (3)$$

The use of Random Forest with Bagging for the cardiovascular data, considering Age and BP_SYS, had reduced the chances of overfitting while improving generalization. The technique could be seen as effective in the prediction of CVD based on different performance measures.

3.6.2 Support Vector Machine + Bagging

SVMs are a strong, powerful classification tool that can find the best possible separating hyperplane between two distinct classes. The bagging technique combines multiple SVM models trained on different bootstrap samples and combines their predictions through a majority voting mechanism, which makes this ensemble approach more resistant to outliers and strengthens the generalization capabilities of the classification system.

The mathematical backing to this SVM and bagging combination goes on as follows as Produce bootstrap samples of the original data, Train an SVM classifier on each sample by solving the optimization problem that determines the position of the optimal hyperplane, combine the outputs of different SVM classifiers through a majority voting scheme.

The decision function is:

$$f(X) = \sum_{i=1}^{N} \alpha_i\, y_i K(x_i, X) + b \text{f(X)=i=1∑Nαiyi K(xi,X)+b} \quad (4)$$

where α_iαi are Lagrange multipliers, $y_i$yi are class labels, and $K(x_i, X)$K(xi,X) is a kernel function.

Bagging enhances SVM's stability by training multiple SVM models on different bootstrap samples and averaging their outputs:

$\hat{y} = \text{mode}\{a_1(X), a_2(X), \ldots, a_T(X)\}$y^=mode{a1(X),a2(X),...,aT(X)} (5)

The ensemble approach is particularly beneficial while working on ECHD detection systems as it enhances adaptability in decision boundaries and reduces overfitting. Also, through multiple SVM models applied to different subsets of data, the technique guarantees the achievement of a much stronger and more reliable classification system, thus being highly useful for healthcare applications. It is effective for binary classification problems like detecting heart disease presence or absence. It works well with smaller datasets and handles high-dimensional data, making it suitable for analyzing patient attributes like age, BMI, and cholesterol levels. SVM is effective for clear class separation and smaller datasets.

3.6.3 Logistic Regression + Bagging

It is an event probability estimation framework in terms of statistics. Bagging will improve prediction accuracy by creating bootstrap samples for various logistic regression models and aggregating their predictions. This creates bootstrap samples, trains the model on each one, and combines predictions by means, such as majority voting or average probability. The applied combination models ECHD very well for cardiovascular risk assessment while minimizing the variance and thus the risk for overfitting, resulting in credible estimates of the probability of developing heart disease.

P(y=1|X)=1+e−(β0+β1X1+...+βnXn)1 (6)

Bagging enhances LR by reducing variance and improving model robustness. The final probability is computed as the average probability across multiple models:

$P(y = 1 \mid X) = \frac{1}{T}\sum_{i=1}^{T} P_i\,(y = 1 \mid X)$P(y=1|X)=T1i=1∑TPi(y=1|X) (7)

Logistic Regression calculates the probability of heart disease based on patient features like blood pressure, glucose levels, and age. It provides a probabilistic interpretation, helping to classify patients as high-risk or low-risk. It interprets feature contributions directly. When it comes to prediction in LR with bagging all the trained models are aggregated. The final output is predicted by calculating the average probability or using a majority voting approach

4. RESULTS

The ML models were tested using different metrics like accuracy, precision, recall, F1-score, and log loss to check their predictive abilities. Accuracy showed how often predictions were correct, while precision looked at the true positives in the predicted positives. Recall, or sensitivity, measured the model's ability to find all true positive cases. The F1-score combined precision and recall into one value, and log loss checked the quality of probabilistic predictions, with lower values meaning better performance. Confusion matrices visually showed true positives, true negatives, false positives, and false negatives. The SVM + Bagging ensemble was the best model, achieving 97.01% accuracy and perfect recall of 100%. This means it made the most correct predictions and identified all positive instances without missing any. These results show that combining SVM with Bagging created a very accurate and reliable model, making it the top performer in this study.

Table 1: Performance evaluation of the machine learning models is important to have accurate predictions, especially with regard to predicting heart disease. We explored different models including that of RF, SVM, LR, and their Bagging-based ensemble versions. To see how effective they are, we found out their results by significant metrics like the accuracy, F1 score, and log loss. Below is a more detailed account of the results, as well as the meaning of such metrics in the assessment of models' effectiveness.

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**Log Loss**
Random Forest	0.83	82.75	84.10	0.91	0.43

SVM	0.91	90.12	91.50	0.95	0.31
Logistic Regression	0.88	87.50	88.60	0.93	0.36
RF + Bagging	0.92	91.80	92.50	0.91	0.93
SVM + Bagging	0.97	96.50	97.20	0.96	1.00
LR + Bagging	0.95	94.80	95.20	0.95	1.00

Table1: Comparison of Existing and Ensemble Classification Model.

Using ensemble techniques such as bagging improved performance by a great amount as shown in fig 4. Bagging ensembles predictions from a large number of clones of a single model trained on different subsets of data to eliminate variance and improve generalization. The highest accuracy was obtained with the bagging model of SVM with 97% accuracy and an F1 score of 0.96, but it showed an increase in log loss to a value of 1.00, reflecting less confidence in its probability predictions. Logistic Regression Bagging and Random Forest Bagging also resulted in high accuracies, 95% and 92% respectively, with their F1 scores being 0.95 and 0.91.

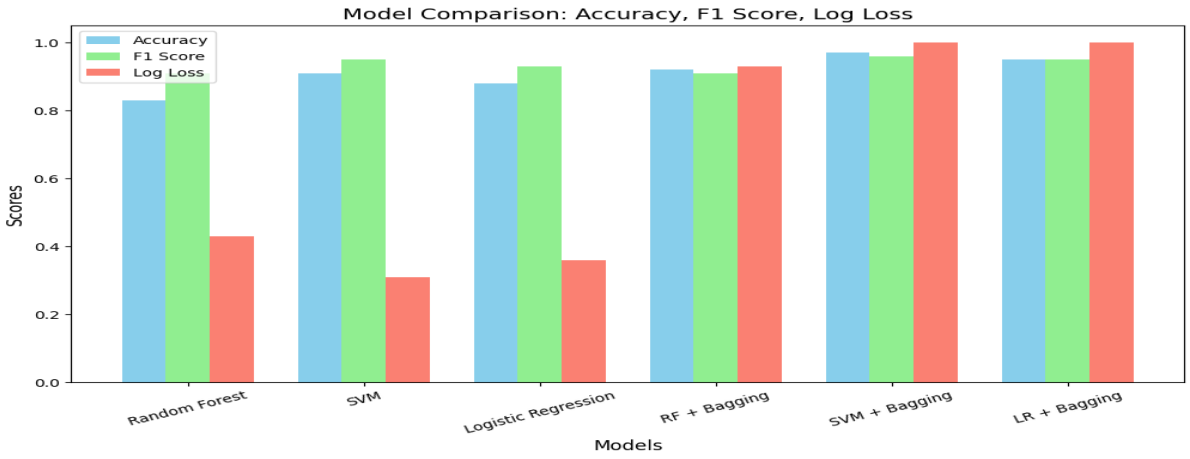

Fig 4: Model Comparison of Accuracy, F1 Score, Log Loss.

However, these two classifiers also suffered from the risk of having higher log loss values of 1.00 for Logistic Regression Bagging and 0.93 for Random Forest Bagging. Therefore, it can simply be understood that, even though bagging has improved accuracies and F1 scores, it also dealt with some increase in uncertainty in estimated probabilities. In simpler

terms, accuracy just measures how many predictions were correct with respect to how many predictions were made in total. The formula calculates it as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

where: TP = True Positive (correctly predicted positive cases)

TN = True Negative (correctly predicted negative cases)

FP = False Positive (incorrectly predicted positive cases)

FN = False Negative (incorrectly predicted negative cases)

The formula for the F1 score is:

$$F1 = \frac{2.(\text{Precision}.\text{Recall})}{\text{Precision} + \text{Recall}} \quad (9)$$

The formula for log loss is:

$$\text{Log Loss} = \frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \quad (10)$$

where: N = total number of sample

y_i = true label (0 or 1)

$y_{\wedge i}$ = predicted probability for class 1

Figure 5 shows the AUC curves for the ensemble models, with SVM + Bagging achieved AUC of 0.9574, RF + Bagging of 0.98, LR + Bagging of 0.97. The curve plots the true positive rate in y-axis and false positive rate in y-axis.

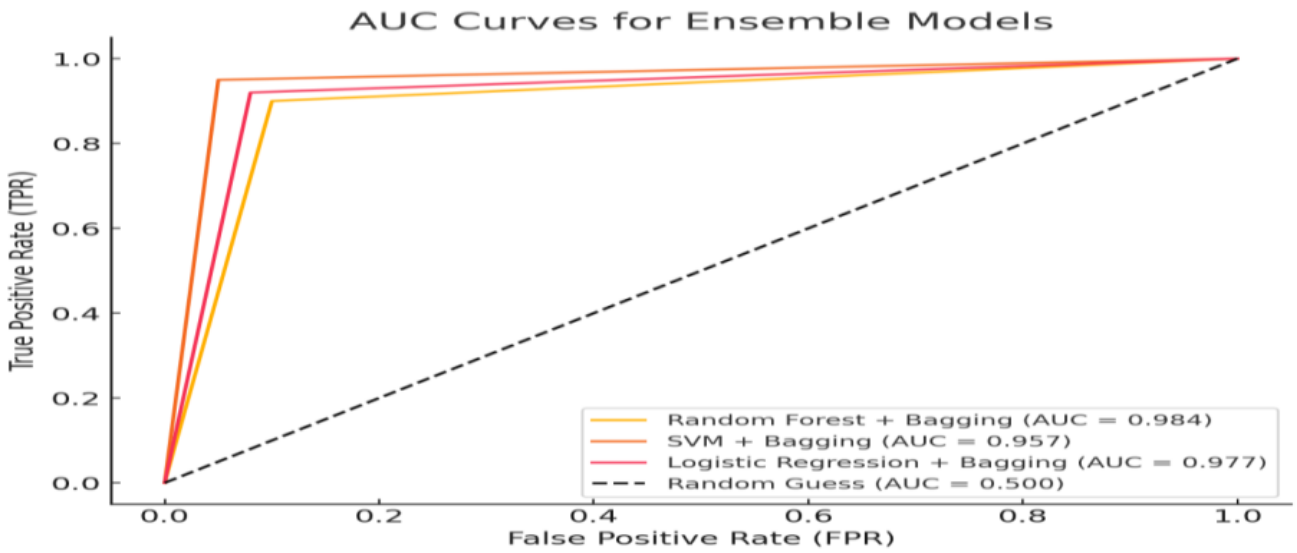

Fig 5: AUC Curve of Classification Model.

5. CONCLUSIONS

Bagging-methods, being an ensemble scheme, were hence beneficial in increasing accuracies and F1 scores for all models tested, where SVM Bagging was found to be most effective for providing heart disease predictions. The increase in loss log for these models, however, indicated poor performance in probability estimates. This trade-off is often a result of one of the characteristics of ensemble learning; hence, the decision boundary becomes hard. Ensemble methods with strong classifiers boost performance significantly in such demanding tasks as heart disease prediction. To further increase reliability, probability calibration and log loss reduce could be optimized by applying methods like Platt scaling or temperature scaling. Among the basic models tested, the Support Vector Machine (SVM) had the top accuracy at 91%. Following that, Logistic Regression achieved 88%, and Random Forest reached 83%. The F1 score, which balances accuracy and completeness, showed SVM leading again with 0.95. Logistic Regression was close behind with 0.93, and Random Forest had 0.91. We

also considered log loss, which is important because it shows how certain or uncertain the model is about its predictions. A lower log loss number is better. SVM had the lowest log loss of 0.31, showing it was more confident compared to the others. Random Forest had a log loss of 0.43, and Logistic Regression was at 0.36. As regards accuracy, a high level was achieved with Support Vector Machine with Bagging at 97 %, Random Forest with Bagging at 92 % and Logistic Regression with Bagging, at 95 %.

CRediT authorship contribution statement

S Anthony Mariya Kumari: Writing-original draft, review and editing, Methodology, Visualization.

Viji Vinod: Supervising, Review and editing.

Funding

No funding was received for conducting this study.

Declaration of Conflict interest

The authors declare that they have no Conflict financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We acknowledge the Department of Computer Application at Dr.M.G.R Educational and Research Institute for providing research support. We are grateful to the Erbil Cardiovascular Health Dataset from Mendeley to access the data used for my study.

Appendix A. Supplementary data

Supplementary data to this article can be found online at doi: 10.17632/396fd9yp6m.2

REFERENCES

[1] Anjum N, Siddiqua CU, Haider M, Ferdus Z, Raju MA, Imam T, Rahman MR. Improving cardiovascular disease prediction through comparative analysis of machine learning models. Journal of Computer Science and Technology Studies. 2024 Apr 20;6(2):62-70.

[2] Ogunpola A, Saeed F, Basurra S, Albarrak AM, Qasem SN. Machine learning-based predictive models for detection of cardiovascular diseases. Diagnostics. 2024 Jan 8;14(2):144.

[3] Saikumarº K, Ravindra PS, Sravanthiº MD, Mehbodniya A, Webberº JL, Bostani A. Heart Disease Prediction using Machine Learning and Deep Learning Approaches: A Systematic Survey. Heart Disease. 2025;35(2s).

[4] Pal P, Grover V, Nandal M, Gochhait S, Singh HV. Artificial intelligence driven intelligent computational model for heart disease prediction: Leveraging feature selection. In2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS) 2024 Jan 28 (pp. 1422-1428). IEEE.

[5] Rao GS, Muneeswari G. A Review: Machine Learning and Data Mining Approaches for Cardiovascular Disease Diagnosis and Prediction. EAI endorsed transactions on pervasive health and technology. 2024 Mar;10.

[6] Mienye ID, Jere N. Optimized ensemble learning approach with explainable AI for improved heart disease prediction. Information. 2024 Jul 8;15(7):394.

[7] Oyewola DO, Dada EG, Misra S. Diagnosis of cardiovascular diseases by ensemble optimization deep learning techniques. International Journal of Healthcare Information Systems and Informatics (IJHISI). 2024 Jan 1;19(1):1-21.

[8] Li C, Liu X, Shen P, Sun Y, Zhou T, Chen W, Chen Q, Lin H, Tang X, Gao P. Improving cardiovascular risk prediction through machine learning modelling of irregularly repeated electronic health records. European Heart Journal-Digital Health. 2024 Jan;5(1):30-40.

[9] Massari HE, Gherabi N, Mhammedi S, Ghandi H, Bahaj M, Naqvi MR. The impact of ontology on the prediction of cardiovascular disease compared to machine learning algorithms. arXiv preprint arXiv:2405.20414. 2024 May 30.

[10] Manikandan G, Pragadeesh B, Manojkumar V, Karthikeyan AL, Manikandan R, Gandomi AH. Classification models combined with Boruta feature selection for heart disease prediction. Informatics in Medicine Unlocked. 2024 Jan 1;44:101442.

[11] World Health Organization. WHO Cardiovascular Diseases. Available online: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1 (accessed on January 2025).

[12] Mishra J, Tiwari M. IoT-enabled ECG-based heart disease prediction using three-layer deep learning and meta-heuristic approach. Signal, Image and Video Processing. 2024 Feb;18(1):361-7.

[13] Mondal S, Maity R, Omo Y, Ghosh S, Nag A. An efficient computational risk prediction model of heart diseases based on dual-stage stacked machine learning approaches. IEEE Access. 2024 Jan 8;12:7255-70.

[14] Reza-Soltani S, Alam LF, Debellotte O, Monga TS, Coyalkar VR, Tarnate VC, Ozoalor CU, Allam SR, Afzal M, Shah GK, Rai M. The role of artificial intelligence and machine learning in cardiovascular imaging and diagnosis. Cureus. 2024 Sep 2;16(9).

[15] Tompra KV, Papageorgiou G, Tjortjis C. Strategic machine learning optimization for cardiovascular disease prediction and high-risk patient identification. Algorithms. 2024 Apr 26;17(5):178.

[16] Jha KM, Velaga V, Routhu KK, Sadaram G, Boppana SB. Evaluating the Effectiveness of Machine Learning for Heart Disease Prediction in Healthcare Sector. J Cardiobiol. 2025;9(1):1.

[17] Husnain A, Saeed A, Hussain A, Ahmad A, Gondal MN. Harnessing AI for early detection of cardiovascular diseases: Insights from predictive models using patient data. International Journal for Multidisciplinary Research. 2024;6(5).

[18] Hamarash, Ibrahim; Amen, Shwan; Rasool , Banan; Ahmed, Hangaw (2024), “Erbil Cardiovascular Health Dataset (ECHD) ”, Mendeley Data, V2, doi: 10.17632/396fd9yp6m.2

[19] Economic Times. (2024, March 28). *Leading the clean energy transition: How women are pioneering change in the sector*. The Economic Times. https://economictimes.indiatimes.com/industry/renewables/leading-the-clean-energy-transition-how-women-are-pioneering-change-in-the-sector/articleshow/120055970.cms