ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 9, September 2024

A Novel Machine Learning-Based Model for Cardiovascular Disease Prediction

[1] Dr. Itti Hooda, [2] Mr. Vikas Hooda

[1] Assistant Professor Maharaja Surajmal Institute
[2] Manager DXC Tech Gurgaon
Corresponding Author Email: [1] ittihooda01@gmail.coml, [2] vikashooda01@gmail.com

Abstract*— This study addresses the urgent global health problem of heart disease (HD) by proposing an ensemble machine learning architecture for cardiovascular disease prediction. Typical HD symptoms include frailty, difficulty of breath, and swollen feet, but traditional diagnostic methods lack efficiency and precision. Accuracy is improved by using feature selection techniques like Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) inside the suggested model, which is an amalgamation of the Elastic Net, Logistic Regression, Gradient Boosting, and Extreme Gradient Boosting (XG Boost) algorithms. It makes use of all obtainable features for prediction using parallel processing. Compared to state-of-the-art approaches, experimental data show that the diagnosis accuracy is significantly higher. The precision of 92.86%, recall of 85.22%, F1-Score of 88.87%, and total accuracy of 88.30% are just some of the impressive metrics demonstrated by the models in the present research. In addition, the robustness of the model is validated by an Area Under the Curve (AUC) value of 0.95. The suggested ensemble model demonstrates superiority in precision when compared to other methods, which makes it a potentially useful technique for the prediction of cardiovascular disease.

Index Terms*— Cardiovascular, Disease, Feature. Diagnosis, Heart, Machine Learning.

I. INTRODUCTION

Heart disease (HD) affects many people, making it a major global health concern [1]. Weakness in the body, shortness of breath, and swollen feet are the HD symptoms that are most frequently experienced [2]. Slow execution and faulty Machine Learning (ML) models are two reasons why today's methods for diagnosing heart disease fall short. Researchers have made many recent attempts to bring new ML algorithms for early detection; however, there are still issues with efficiency and effectiveness [3]. Without a trained specialist and state-of-the-art equipment, diagnosing and treating heart disease could be difficult [4]. Numerous people's lives can be saved with the use of modern diagnostic tools. There are 3.6 million new HD patients per year worldwide, as reported by the European Society of Cardiology [5,6]. Heart disease accounts for about 3% of the healthcare budget; however, only about half of people diagnosed with the condition live for more than two years [7]. Congestive heart failure (CHF) is a progressive disease that takes time to infest the body and causes significant disruptions to the cardiovascular system's ability to perform its functions. Fig 1 depicts a picture of a patient with congestive heart failure and a normal heart.

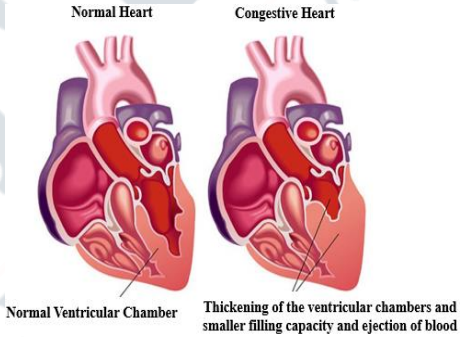

Fig. 1. Congestive Heart Failure [8].

Traditionally, the diagnosis of cardiac disease is based on the patient's symptoms, the doctor's knowledge of the patient's medical history, and the results of a physical examination. Identifying HD patients cannot be done with any degree of efficiency or precision using the results acquired from these procedures. In addition to this, these approaches are computationally challenging as well as costly [9]. The accurate diagnosis of HD requires developing non-invasive diagnostic tools based on machine learning [10,11,12]. The mortality rate for heart disease patients can be lowered with the help of expert decision systems that use machine learning and artificial fuzzy logic (AFL) [13,14]. The predictive machine learning models necessitate the correct data for their training and testing to function properly [15]. The effectiveness of the machine learning model can be significantly improved by using data sets that are evenly distributed for training and testing. In addition, the predictive skills of the model can be improved by selecting the appropriate and connected features from the data set. Therefore, picking the right characteristics for the model and

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 9, September 2024

ensuring the data is distributed evenly are two of the most important elements. The scientific literature is replete with proposals for new diagnostic approaches based on machine learning. These include the Neuro Fuzzy, Artificial Neural Network (ANN), Support vector machine (SVM), Decision tree (DT), Naive Bayes (NB), and other similar methods. However, these methods aren't without drawbacks, such as limited training data, inaccurate predictions, improper data distribution, etc. In addition, cardiac disease cannot be correctly diagnosed using these approaches.

Data homogeneity at the data processing layer improves the predictive abilities of machine learning models. The model's enhanced efficiency results from additional preprocessing tactics like Min-Max Scalar, eliminating missing features from the dataset, and employing a standard scalar [16]. The most important variables can be isolated by employing a variety of feature selection methods, such as the Greedy Algorithm (GA), Local Learning-Based Features Selection (LLBFS), Principal Component Analysis (PCA), etc. A plethora of optimization techniques also aids feature improvement for use in upcoming machine learning model training. These methods include Ant Colony Optimization (ACO), Bacterial Foraging Optimization (BFO), and others [17]. In addition, many Internets of Things (IoT)-based systems now incorporate a variety of machine learning algorithms for prediction and classification, including ANN, SVM, K-Nearest Neighbor (KNN), etc [18]. Unsupervised machine learning algorithms classify data collected by various Internet of Things sensors. ML algorithms produce more accurate findings on tagged data than people do. In addition, Neural network-based technologies have advanced to the point where they may accurately anticipate neurological and cardiovascular problems. Carotid artery stenting (CAS) is a popular medical procedure today. CAS techniques provide an early snapshot of HD patients' Major Adverse Cardiovascular Events (MACE). Results from the ANN are more reliable than those from the straightforward CAS approach [19]. Not only do the suggested ANN-based methods generate values from a variety of preceding techniques, but they also combine posterior probabilities. The outcomes from the ANN-based methods were significantly higher than those from the previous methods [20].

Predicting cardiovascular disease is the focus of this study, and to do so, an ensemble machine learning architecture is proposed, bolstered by the feature selection methods of Recursive Feature Elimination (RFE) and PCA. The suggested architecture focuses on enhancing the precision of ML-based solutions for cardiac illness diagnostics. The proposed model uses a parallel computation methodology, allowing us to use all available features rather than being constrained by the selection of features strategy in the preprocessing stage. The experimental results show that the proposed architecture outperforms the state-of-the-art machine learning methods for cardiovascular disease in terms of diagnostic accuracy. Here are the remaining parts of the paper: A similar investigation is discussed in section 2. Section 3 covers diagnostic techniques that are unique to cardiac problems. Section 4 discusses the results of the proposed architecture. Conclusion is presented in section 5.

II. LITERATURE OF REVIEW

Several studies have successfully used an ML-based system to diagnose heart disease. This literature review would explain the design and development of decision support utilizing ML approaches to accurately predict cardiovascular illness.

Arsalan et al. (2023) [21] examined all algorithms through exploratory and experimental output analyses. Recursive operating characteristic curves and confusion matrix estimates were calculated for each method. Many criteria were used to estimate the ML method's performance and zero in on the best model-class ML algorithm. Prediction accuracy for cardiovascular diseases (CVD) was 85.01%, sensitivity was 92.11%, and the recursive operating characteristic curve accuracy was 87.73% when using Random Forest (RF). It had the lowest specificity (43.44%) and misclassification rate (8.50%) for CVD. RF is the best option for classifying and forecasting CVDs. Classification and prediction of diseases are useful tools for healthcare practitioners everywhere.

Pham et al. (2023) [22] introduced unique, multi-faceted methods for Electrocardiography (ECG)-based heart disease classification. The first method proposes using deep learning-based image classifiers (DenseNet121 and ResNet50 were trained on Poincare diagrams), which demonstrated encouraging results for diagnosing AF (atrial fibrillation) but no other arrhythmias. Though inference was time-consuming due to the significant calculations required in the preprocessing phase, the gradient-boosting model XGBoost worked satisfactorily on long-term data. Finally, on both the CinC 2017 and CinC 2020 datasets examined, the 1D convolutional model and the 1D ResNet demonstrated the highest performance, achieving an F1 score of 85% and 71%, respectively; both scores were greater than the top-ranked response in their respective challenges. 1D CNN and 1D ResNet were shown to be the most efficient models in terms of both energy consumption and CO2 emissions. The model interpretation research found that although 1DResNet utilized raw ECG signals, DenseNet detected AF using heart rate variability.

Burak et al. (2023) [23] examined CAD diagnosis classification algorithms, as well as seven computational feature selection (FS) methods. In-depth and uncertain ensemble FS techniques were also made available. The suggested method is assessed on three open-source CAD data sets using six categorization methods and four voting procedures. Measures of performance have been compared using a wide variety of classifier-FS combinations. On three

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 9, September 2024

different datasets, the multilayer perceptron classifier showed promising results. The suggested method achieved an accuracy of 91.78 percent on the Z-Alizadeh Sani data set, 85.55 percent on the Statlog data set, and 85.47 percent on the Cleveland data set.

Ibrahim M et al. (2022) [24] intended that MMC, Random, Adaptive, QUIRE, and AUDI selection techniques for multi-label active learning were utilized to reduce labeling costs by iteratively picking the most relevant data to query labels. A grid search optimizes hyperparameters for label ranking classifier selection approaches to perform predictive modeling in each heart disease dataset scenario. Experiments include accuracy and F-score with/without hyperparameter tuning. The selection approach is more accurate for generalizing the learning model outside the optimized label ranking model's data. However, the optimum F-score selection was highlighted.

Wang et al. (2022) [25] developed a novel wireless ECG patch and implemented a deep learning framework using the Long Short-term Memory (LSTM) and Convolutional Neural Network (CNN) models. The newly collected data shows that models trained with existing methods perform poorly (only 58.0% accuracy) when asked to distinguish between two primary types of heartbeats (Supraventricular premature beat and Atrial fibrillation). The study presented a semi-supervised approach relying on confidence-level training to deal with poorly annotated data samples. Compared to the accuracy of traditional ECG classification methods, the suggested method is approximately 5.4% more accurate, as shown by the experimental findings.

Rustam et al. (2022) [26] presented a novel strategy for fixing this problem by mining a CNN for features. A soft-voting-based ensemble model trains linear models, including the CNN model, stochastic gradient descent classifier, logistic regression, and the support vector machine. Various feature set-to-training-dataset ratios are tested with large-scale trials. Four datasets are used for the performance study, and the results are compared to more modern CVD methods. The suggested model outperforms the others, with an accuracy of 0.93 and precision, recall, and F1 scores of 0.92 each. The findings prove that the proposed approach works and that the ensemble model may be expanded to accommodate larger datasets.

Saikumar et al. (2022) [27] studied that ML methods enhance a physician's ability to make treatment and diagnosis decisions with the help of AI. This study takes a deep dive into the fundamentals of systems and the important theories behind them, including the Decision Tree model, the Gaussian Navies Bayes model, the K-NN model, and the RCNN model. Data mining and AI create accurate outcomes with minimal mistake rates. This study's accuracy of 99.173%, sensitivity of 98.3%, precision of 99.164%, recall of 98.69%, and specificity of 0.0009 enhance a unique CAD risk prediction model. Follow-up findings outperform methodologies and compete with current technologies.

Hossen et al. (2021) [28] focused on several cardiac disease features and a model based on supervised learning methods, including RF, DT, and logistic regression (LR). It uses the UCI Cleveland heart disease database. The dataset has 303 instances and 76 attributes. Testing is conducted to validate the effectiveness of different approaches, but only on 14 of these 76 attributes. This study aims to create a system for calculating an individual's risk of developing heart disease. Results show that LR achieves the highest accuracy (92.10 %).

Prerna et al. (2020) [29] utilized the fingertip video dataset to predict whether a person has coronary heart disease and used the MAPO algorithm as the appropriate feature selection alongside other ML methods. After removing background noise from the videos, MAPO can make accurate predictions of heart rate with a Pearson correlation of 0.9541 and a Standard Error Estimate of 2.418. Optimizing features of two datasets based on expected heart rate, MAPO is employed once more. ML techniques are applied to the enhanced dataset to generate heart disease forecasts. MAPO reduces the dimensionality to the essentials, with a maximum of 81.25 percent, while maintaining accuracy parity among machine learning models. MAPO outperforms other optimizers in terms of precision.

Reddy et al. (2019) [30] analyzed that feature selection approaches could be useful tools for cutting diagnostic expenses by zeroing in on relevant characteristics. Using the Cleveland and stat log project heart datasets, this study aims to make predictions about the classification model and identify the elements that play a pivotal role in making those predictions about heart disease. Based on three distinct percentage splits, the random forest approach has been observed to have an accuracy of 90–95% in the classification and feature selection model. The 8 and 6 chosen features are the minimum needed to construct a superior performance model. Thus, further elimination of the 8 or 6 chosen features might not improve the prediction model's performance.

III. BACKGROUND STUDY

The study emphasizes the role of ML in cardiovascular disease diagnosis and prognosis, which can improve patient care. This study uses data-driven pattern recognition to create a novel model using Huang initialization and k-mode clustering to increase classification accuracy. Parameter optimization is accomplished with the help of GridSearchCV, which employs several ML techniques, including a random forest, a decision tree classifier, a multilayer perceptron, and an eXtreme Gradient Boosting (XGBoost). When the model's performance is measured against a sizable Kaggle dataset, numerous popular algorithms, including XGBoost (86.87% with cross-validation, 87.02% without), random forests (87.05%

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 9, September 2024

with cross-validation, 86.92%), decision trees (86.37% with cross-validation, 86.53% without), and multilayer perceptron (87.28% with cross-validation, 86.94%), achieve very respectable accuracy scores. Multilayer perceptron using cross-validation performed best regarding Area Under the Curve (AUC) value, with an impressive 87.28 percent accuracy. The study suggests that machine learning could significantly improve cardiovascular disease diagnosis and prediction [31].

IV. PROBLEM FORMULATION

Construct a study based on machine learning using patient demographics, lifestyle factors, and health records to forecast the likelihood of cardiovascular disease. The difficulty lies in developing a new predictive model that can outperform current approaches in terms of both accuracy and readability. Predicting cardiovascular disease risk and providing useful insights into its causes requires feature engineering, innovative ensemble techniques, or complicated deep-learning architectures. The purpose is to advance machine learning in medicine and the ability to detect diseases early on. An ensemble of Elastic Net, Logistic Regression, Gradient Boosting, and XG Boost algorithms are used to predict cardiovascular disease. The data is gathered from a trustworthy source and preprocessed to ensure data quality, and features are selected using PCA and RFE. This approach reliably and accurately predicts cardiovascular illness, aiding doctors in making better patient treatment choices.

V. RESEARCH OBJECTIVES

- To develop an innovative machine learning model for cardiovascular disease prediction.
- To design an ensemble approach to predict cardiovascular disease, aiming to produce robust and highly accurate predictions.
- To investigate the use of patient demographics, lifestyle, and historical health data as features to enhance the model's predictive power.
- To assess the predictive model's accuracy, robustness, and potential impact on healthcare decision-making for patient care and treatment.

VI. RESEARCH METHODOLOGY

Research strategies are discussed concerning the idea of planned architecture.

A. Dataset Description

The dataset consists of three types of input features collected during medical examinations: Objective features, including age (in days), height (in centimeters), weight (in kilograms), and gender (categorical code); Examination features, comprising systolic blood pressure (ap_hi), diastolic blood pressure (ap_lo), cholesterol levels (categorized as 1 for normal, 2 for above normal, and 3 for well above normal), and glucose levels (also categorized as 1 for normal, 2 for above normal, and 3 for well above normal); and Subjective features, encompassing smoking (binary), alcohol intake (binary), and physical activity (binary). These details were documented during the checkup. Cardiovascular disease is either present (1) or absent (0) and is represented by the dependent variable cardio. Table 1 provides a quick overview of the features, variable names, and data types found in the dataset.

Table I: Dataset

Feature	Abbreviation	Type	Data Type
Age	age	Objective Feature	int (days)
Gender	gender	Objective Feature	categorical code
Weight	weight	Objective Feature	float (kg)
Height	height	Objective Feature	int (cm)
Diastolic BP	ap_lo	Examination Feature	int
Systolic BP	ap_hi	Examination Feature	int
Glucose	gluc	Examination Feature	1: normal, 2: above normal, 3: well above normal
Cholesterol	cholesterol	Examination Feature	1: normal, 2: above normal, 3: well above normal
Alcohol Intake	alco	Subjective Feature	binary
Smoking	smoke	Subjective Feature	binary
Cardiovascular Disease	cardio	Target Variable	binary
Physical Activity	active	Subjective Feature	binary

B. Technique Used

A powerful method is included within the suggested approach:

a. XGBoost

The recommended method employs an eXtreme Gradient Boosting (XGBoost) classifier to identify botnet attacks when certain criteria are met. XGBoost (or just XGB) is a novel tree-based ensemble learning classifier that has recently attracted much attention. This is the best

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 9, September 2024

implementation to date for gradient-boosted decision trees. Each successive decision tree in a gradient-boosted decision tree series contributes to the improvement of the model and influences the subsequent tree in the series [32]. Combining weak classifiers with XGBoost may strengthen them. Instead of pre-existing data, XGBoost uses validated decision trees. Gradient boosting iteratively improves the baseline loss function. The prior phase's residue should be minimal.

The residual is the difference between desired and actual values. The model is ready for deployment when the residual value drops below a threshold. If many decision trees reach a value before the residual, training ends, and the final model is chosen. The XGB model varies from gradient boosting in its acceptance of regularisation, parallel processing, and execution speed. A short calculation to remember XGBoost's aim [33]. Utilizing this function, the model's overall performance may be evaluated [34].

$$P(\theta) = t(\theta) + r(\theta) \quad (1)$$

where denotes the parameters, r the regularisation period, and t the number of discarded training iterations.

b. Elastic Net

The elastic net (EN) expands the lasso to retain correlated variables. The elastic net was created to increase prediction accuracy when variables are highly correlated. The elastic net preserves correlated covariates if they improve prediction, unlike the lasso, which includes one variable but excludes the other when two variables are linked. The EN feature selection method uses convex Lasso-Ridge regression. These two regularisation procedures limit regression model coefficients using L1 and L2 norms to minimize the sum of squared residuals. Continuous lasso regression selects features and shrinks predictor coefficients. Ridge regression works best with more observations than features and strong correlations, unlike Lasso regression, which chooses features automatically. The EN, created by Zou and Hastie, is superior to other methods because it may pick the traits independently. Here is an example of such a formulation [35]:

$$\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\left[\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j|\right]\right\} \quad (2)$$

where β is the prediction F-measure for each model. Therefore, the F-measure indicates the degree to which one should trust the model's prediction. When computing the expected variable, a bigger value is given more consideration.

c. Logistic regression

Optimal input values for a logistic model can be determined using a statistical technique called logistic regression [36]. Logistic regression is a discriminative classification method that requires a vector of values that can be expressed as real numbers as input. Features or predictors are input vector characteristics that are used for classification. Logistic regression may be advantageous for the investigation of multiclass classification data. Commonly presumed to have originated from the Bernoulli trial, the use of the probability P of a dichotomous event in logistic regression is frequently tied to investigating the properties of the events themselves [37]. The Logit function is mathematically defined as the natural log of the probability that Y belongs to one of the categories [38]. If p represents the probability, the logit function for p can be expressed as [39]:

$$\text{Logit (p)} = 1\ \text{n}\ (\text{p}/1 - \text{p}) \quad (3)$$

d. Gradient Boosting

Gradient boosting is a form of supervised learning in which an effort is made to estimate a target characteristic by combining the estimates obtained via several simpler and weaker mode techniques. Learning of this kind is intended to make ML better [40]. Non-linearity is common in classification and regression models (also called decision and regression trees); hence, gradient boosting improves learning. The addition of learners is modeled sequentially to simulate the weak prediction models, such as regression decision trees. The nodes in the middle represent decisions, and the branches and leaves represent possible outcomes [41]. Regression trees perform poorly alone, but they improve greatly in ensembles. The ensembles are built incrementally, with each new one fixing imperfections of the previous one (as seen in Eq [4]).

$$f_k(x) = \sum_{m=1}^{k}\gamma_m\, h_m(x) \quad (4)$$

e. Recursive Feature Elimination (RFE)

A large dataset typically contains irrelevant features. Recurring features affect the classification algorithm's inefficiency. This may lower prediction accuracy. Isolating the most important variables for reliable predictions using RFE. This reduces the dataset's dimensionality while preserving the useful features developed from the more refined feature sets. RFE, an iterative approach that eliminates features in decreasing order of importance, was used extensively in this investigation. It is speculated that RFE can aid in creating more precise RF models, which can then be used in intrusion detection [42].

This effective feature-selection method might be applied to a training dataset. Input feature sets can be constructed without compromising classification accuracy if irrelevant features are omitted. The term "recursive" describes a method that iteratively searches for a target number of attributes. The RF classifier begins by determining the significance of each attribute, after which it ranks those characteristics in descending order of importance. After that, the model is retrained with the improved feature set to enhance classification accuracy while removing less significant features. This loop would continue so long as there are new features to consider adding. Fig 2 depicts the workflow

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 9, September 2024

diagram of recursive feature elimination (RFE).

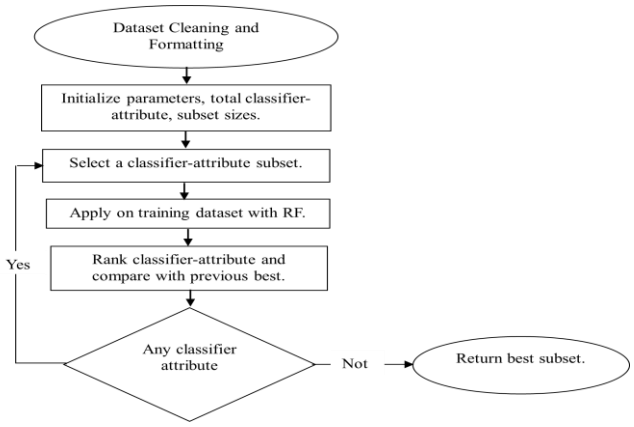

Fig. 2. Workflow diagram of Recursive Feature Elimination (RFE) [43]

f. Principal Component Analysis (PCA)

PCA is a popular dimensionality reduction technique in ML [44]. Reducing the number of dimensions in a dataset without compromising its quality is the main objective of dimensionality reduction [45]. This is essential because data with a high dimension can be difficult to perceive and analyze, and they can also induce overfitting in ML algorithms. The principal components, or linear combinations of the original features, are what the PCA generates [46].

In the feature space, the first principal component is responsible for capturing the greatest amount of variation [47]. The greatest amount of variation may be seen running parallel to the second principal component, which runs in a direction that is perpendicular to the first principal component. In the context of ML models, some potential applications for this include data compression and visualization, as well as the extraction of features. The PCA is depicted in Fig 3.

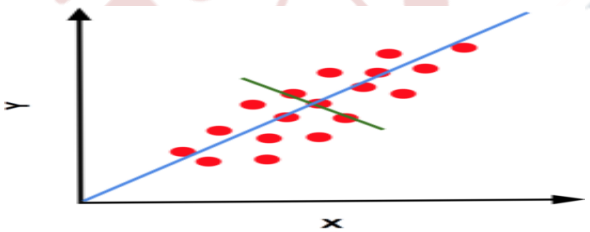

Fig. 3. PCA [48].

C. Proposed Methodology

The model's architectural components are depicted in Fig 4. The following procedures are required to develop a model capable of predicting heart disease:

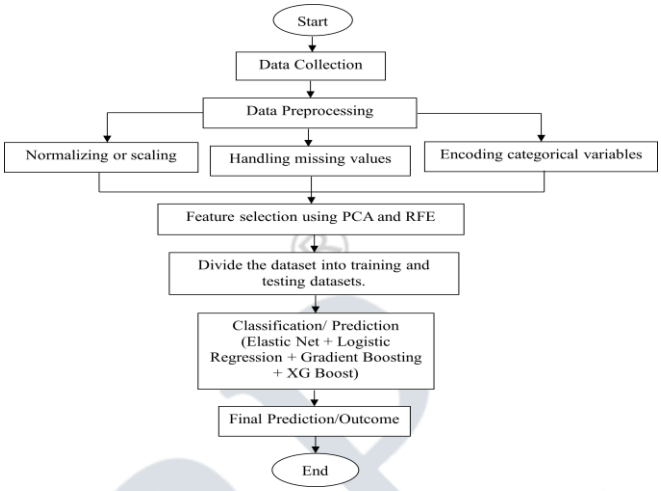

Fig. 4. Proposed Methodology

Step 1: Data Collection

Data collection involves gathering essential data from reputable sources, such as the UCI repository, a well-known repository for machine learning datasets. This data would likely include patient information and their cardiovascular disease status.

Step 2: Data Preprocessing

Preprocessing the data after collection is essential for ensuring it is of high enough quality and suitable for machine learning. Data cleansing (eliminating mistakes and outliers), data transformation (e.g., scaling or normalizing features), addressing missing values (imputation), and picking important attributes (feature selection) all fall under this category. The goal of these procedures is to get the data ready for modeling.

Step 3: Feature Selection

Feature selection is crucial for optimizing the model's performance and reducing dimensionality. In this step, PCA and RFE techniques are employed to identify the most important features for modeling. This helps streamline the dataset and focus on the most informative variables.

Step 4: Data Splitting

Once the most vital features have been identified, the dataset is divided into training and test sets. To ensure the model is trained and tested on separate subsets, a preset ratio is often utilized (e.g., 80% for training and 20% for testing).

Step 5: Classification/Prediction

This step applies an ensemble model to classify and predict cardiovascular disease. The ensemble model combines multiple machine learning algorithms, specifically Elastic Net, Logistic Regression, Gradient Boosting, and XG Boost, to leverage their strengths and enhance predictive accuracy. Each of these algorithms contributes to the final prediction, and their outputs are combined in some manner, often

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 9, September 2024

through voting or weighted averaging.

Step 6: Outcome

Finally, the performance of the ensemble model is evaluated and contrasted to that of alternative algorithms and techniques. The predictive power of a model may be evaluated using several different measures, including accuracy, precision, recall, F1-score, and perhaps AUC. Various algorithms provide various findings, so comparing them can help in choosing the one that would give you the most accurate predictions of cardiovascular disease.

D. Proposed Algorithm

Here's a mathematical algorithm for the above steps:

Step 1: Data Collection

Let $\boldsymbol{D}$ be the dataset collected from a reputable source, containing patient information and their cardiovascular disease status. $\boldsymbol{D}$ is structured as a set of instances.

$(x_1, y_1), (x_2, y_2),\ldots,(x_n,y_n),$ where x_i represents the features of patient i and y_i represents the binary cardiovascular disease status (1 for presence, 0 for absence).

Step 2: Data Preprocessing

a. Data Cleansing:

- Remove outliers and errors: $D_{clean} = \{(x_i, y_i) | x_i \; is \; within \; acceptable \; range\}$

b. Data Transformation:

- Scale/Normalize features: $D_{scaled} = \{(x_i, y_i) | x_i \; normalized\}$

c. Missing Value Management:

- Handle missing values: $\mathbf{D_{imputed}} = \{(\mathbf{x_i},\mathbf{y_i})|\mathbf{x_i}$with imputed missing values$\}$

d. Feature Selection:

- Utilize PCA or RFE to select relevant features: $D_{selected} = \{(x_i, y_i) | x_i$ with selected features$\}$

Step 3: Data Splitting

Split $D_{selected}$ into two subsets: a training dataset D_{train} and a testing dataset D_{test}, with a fixed ratio (e.g., 80% for training and 20% for testing).

Step 4: Classification/Prediction

a. Train multiple machine learning models:

- Elastic Net: $M_{EN} \leftarrow TrainModel(D_{train})$
- Logistic Regression:$\mathbf{M_{LR}} \leftarrow \mathbf{TrainModel}(\mathbf{D_{train}})$
- Gradient Boosting:$\mathbf{M_{GB}} \leftarrow \mathbf{TrainModel}(D_{train})$
- XG Boost:$\mathbf{M_{XGB}} \leftarrow \mathbf{TrainModel}(\mathbf{D_{train}})$

b. Predict disease status for the testing dataset:

- $Y_{EN} \leftarrow Pridict\,(M_{EN}, D_{test})$
- $Y_{LR} \leftarrow Pridict\,(M_{LR}, D_{test})$
- $Y_{GB} \leftarrow Pridict\,(M_{GB}, D_{test})$
- $Y_{XGB} \leftarrow Pridict\,(M_{XGB}, D_{test})$

Step 5: Ensemble Model

Combine the predictions of the individual models using an ensemble approach (e.g., majority voting or weighted averaging):

- $Y_{ensemble} = Ensemble\,(Y_{EN}, Y_{LR}, Y_{GB}, Y_{XGB})$

Step 6: Outcome

Evaluate the performance of the ensemble model $Y_{ensemble}$ using various metrics such as accuracy, F1-score, precision, recall, and possibly AUC. Compare these metrics with other algorithms or methods to determine the best approach for predicting cardiovascular disease.

VII. RESULT AND DISCUSSION

A confusion matrix, which is a performance evaluation tool used in machine learning and classification tasks, is included in the findings that have been supplied here. The confusion matrix may be seen in this context in Fig 5, which can be seen below.

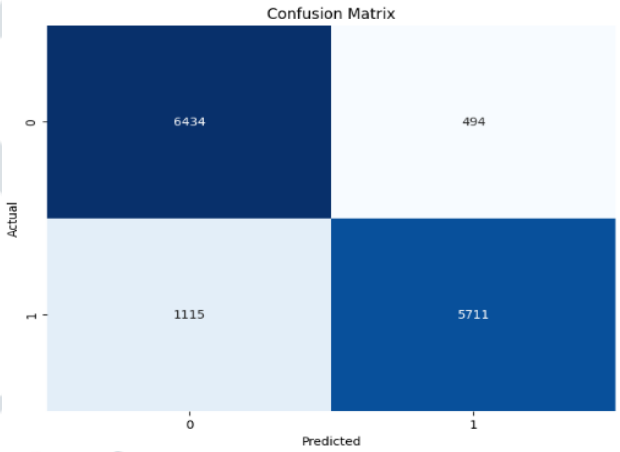

Fig. 4. Confusion Matrix

The confusion matrix contains the following values:

- True Positives (TP): 6434
- False Positives (FP): 494
- False Negatives (FN): 1115
- True Negatives (TN): 5711

The model's ability to distinguish between positive and negative cases may be evaluated using these measures. The model's high accuracy here indicates a low proportion of false positives, while the model's reasonably high recall indicates it captures many true positives. The accuracy measures how well the model predicts, whereas the F-1 Score strikes a good middle ground between precision and recall. Table 2 shows how well the suggested ensemble performs.

Table II: Performance of Proposed Ensemble Model

Precision	Recall	F-1 Score	Accuracy
92.86	85.22	88.87	88.30

Precision, a statistic of model performance that looks at how well the model makes correct predictions, is at 92.86

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Vol 11, Issue 9, September 2024

percent. The percentage of affirmative instances that can be identified is 85.22 percent, measured by recall. The F-1 Score, which considers both accuracy and recall, is 88.87%. Overall, 88.30% of forecasts have been true, as measured by accuracy. The combination of a high F-1 Score and Accuracy indicates that the ensemble model performs well in predicting positive situations while striking a good balance with recall. The graphical performance of the ensemble model is shown in Fig 6.

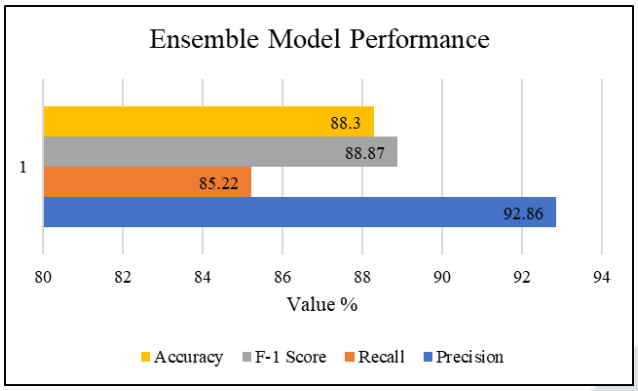

Fig. 6. Ensemble Model Performance

The graphical representation of the ROC curve represents the performance of a binary classifier. It displays the true positive rate (TPR) vs the false positive rate (FPR) for a variety of classification thresholds. The area under the receiver operating characteristic curve (AUC) is a quantitative assessment of a classifier's overall performance that accounts for its sensitivity and specificity. Fig 7 shows that the AUC of the ensemble models is 0.95.

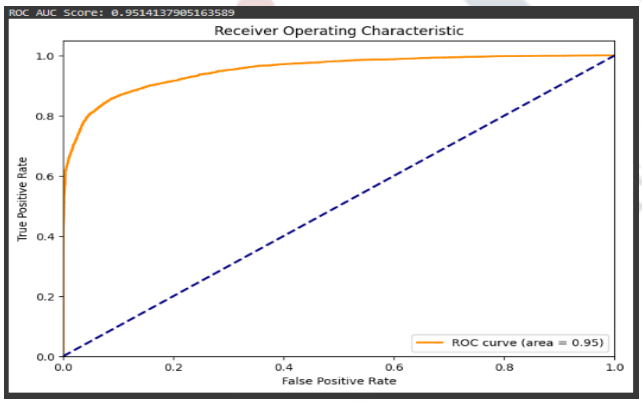

Fig. 7. An AUC is shown for the ensemble model.

A. Comparative Study

Table 3 summarizes the findings of a comparative research study on several approaches to the prediction of cardiovascular disease. This study highlights the performance metrics of several different approaches. The work carried out by Chintan M. and colleagues [31] is shown in this table with the support of the Random Forest (RF), XGBoost (XGB), Decision Trees (DT), and Multi-Layer Perceptron (MLP) algorithms. These techniques yield accuracy ranging from 86.37% to 87.28%, precision scores between 88.70% and 89.58%, recall rates from 81.61% to 84.28%, F1-Scores spanning 85.42% to 86.71%, and an Area Under the Curve (AUC) value of 0.94 to 0.95. In comparison, the present study introduces a proposed ensemble model, which achieves an accuracy of 88.30%, precision of 92.86%, recall of 85.22%, F1-Score of 88.87%, and an AUC of 0.95. Based on these findings, it appears that the suggested ensemble model performs better than the other approaches in terms of precision, which demonstrates that it is a potential strategy for the prediction of cardiovascular illness.

Table III: Comparative Study

Technique	Accuracy	Precision	Recall	F1-Score	AUC
RF [31]	87.05	89.42	83.43	86.32	0.95
XGB [31]	86.87	88.93	83.57	86.16	0.95
DT [31]	86.37	89.58	81.61	85.42	0.94
MLP [31]	87.28	88.70	84.28	86.71	0.95
Proposed Ensemble model	88.30	92.86	85.22	88.87	0.95

Combining Elastic Net, Logistic Regression, Gradient Boosting, and XG Boost techniques, the present study introduces a novel Ensemble model. This proposed model obtains an Accuracy of 88.30%, which surpasses all individual techniques evaluated by Chintan M. et al. In addition, the proposed model demonstrates superior Precision, Recall, F1-Score, and AUC, demonstrating its ability to accurately predict cardiovascular disease.

VIII. CONCLUSION AND FUTURE SCOPE

Weakness, shortness of breath, and swollen feet are just a few of the signs of heart disease (HD), which is still a major worldwide health problem. Unfortunately, the identification of HD patients using conventional diagnostic approaches that focus on symptoms, medical history, and physical examination is generally ineffective. For the purpose of cardiovascular disease prediction, this work proposes an ensemble machine learning architecture using RFE and PCA for enhanced feature selection. The proposed model uses a parallel computing strategy to make the most of all available information, with a focus on accuracy in ML-based cardiac diagnostics. Results from experiments show that the suggested design improves upon the diagnosis accuracy of current state-of-the-art machine learning algorithms for cardiovascular illness. In terms of diagnostic accuracy, the suggested ensemble model beats methods that are considered to be state-of-the-art, obtaining an outstanding 88.30%. The model incorporates techniques such as Elastic Net, Logistic Regression, Gradient Boosting, and XG Boost. In conclusion, the suggested approach is a major step forward in the precise prediction of cardiovascular illness, which might lead to

better treatment options for individual patients. Improving the suggested ensemble model is a future goal of this research, as is the integration of cutting-edge machine learning techniques and the examination of other pertinent characteristics. The model's applicability in real-world clinical settings may also be improved by the incorporation of real-time data streams and the use of upcoming technologies like wearable devices and telemedicine.

REFERENCES

[1] M. W. Nadeem et al., "Fusion-based machine learning architecture for heart disease prediction," *Comput. Mater. Continua*, vol. 67, no. 2, 2021.

[2] M. Durairaj and N. Ramasamy, "A comparison of the perceptive approaches for preprocessing the data set for predicting fertility success rate," *Int. J. Control Theor. Appl.*, vol. 9, no. 27, pp. 255-260, 2016.

[3] K. Taunk et al., "A brief review of nearest neighbor algorithm for learning and classification" in International Conference on Intelligent Computing and Control. Systems (ICCS). IEEE, 2019, pp. 1255-1260 doi:10.1109/ICCS45141.2019.9065747.

[4] M. T. Donofrio et al., "Diagnosis and treatment of fetal cardiac disease: A scientific statement from the American Heart Association," *Circulation*, vol. 129, no. 21, pp. 2183-2242, 2014 doi:10.1161/01.cir.0000437597.44550.5d.

[5] A. J. Coats, "Stewart". "Ageing, demographics, and heart failure.", *Eur. Heart J. Suppl.*, vol. 21, no. Supplement_L, pp. L4-L7, 2019.

[6] I. Spoletini and M. Lainscak, "Epidemiology and prognosis of heart failure" in *Int. Cardiovasc. Forum J.*, vol. 10, 2017 doi:10.17987/icfj.v10i0.420.

[7] M. H. Li, *Haq*. Amin Ul.: Jian Ping Memon, Shah Nazir, and Ruinan Sun. "A hybrid intelligent system framework for predicting heart disease using machine learning algorithms." Mobile Information Systems 2018 (2018): 1-21.

[8] Available at: https://www.healthspectra.com/congestive-heart-failure/.

[9] M. Elhoseny et al., "A new multi-agent feature wrapper machine learning approach for heart disease diagnosis," *Comput. Mater. Continua*, vol. 67, no. 1, 51-71, 2021 doi:10.32604/cmc.2021.012632.

[10] A. H. Gonsalves et al., "Prediction of coronary heart disease using machine learning: An experimental analysis" in *Proc. 2019 3rd International Conference on Deep Learning Tech nologies*, 2019, pp. 51-56 doi:10.1145/3342999.3343015.

[11] Y. Khan et al., "Machine learning techniques for heart disease datasets: A survey" in *Proc. 2019 11th International Conference on Machine Learning and Computing*, 2019, pp. 27-35 doi:10.1145/3318299.3318343.

[12] Y. Meng et al., "A machine learning approach to classifying self-reported health status in a cohort of patients with heart disease using activity tracker data," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 3, pp. 878-884, 2020 doi:10.1109/JBHI.2019.2922178.

[13] S. Ansarullah, "Immamul, and Pradeep Kumar," A Systematic Literature Review on Cardiovascular Disorder Identification Using Knowledge Mining and Machine Learning Method Int. J. Recent Technol. Eng 7, vol. 6s, 2019, pp. 1009-1015.

[14] S. Nazir et al., "Fuzzy logic-based decision support system for component security evaluation," *Int. Arab J. Inf. Technol.*, vol. 15, no. 2, pp. 224-231, 2018.

[15] D. Justus et al., "Predicting the computational cost of deep learning models" in IEEE International Conference on big data (Big Data). IEEE, 2018, pp. 3873-3882 doi:10.1109/BigData.2018.8622396.

[16] S. Ambesange et al., "Optimizing Liver disease prediction with Random Forest by various Data balancing Techniques" in IEEE international conference on cloud computing in emerging markets (CCEM). IEEE, 2020, pp. 98-102 doi:10.1109/CCEM50674.2020.00030.

[17] R. El-Bialy et al., "Feature analysis of coronary artery heart disease data sets," *Procedia Comput. Sci.*, vol. 65, pp. 459-468, 2015 doi:10.1016/j.procs.2015.09.132.

[18] J. P. Li et al., "Heart disease identification method using machine learning classification in e-healthcare," *IEEE Access*, vol. 8, 107562-107582, 2020 doi:10.1109/ACCESS.2020.3001149.

[19] L. Baccour. "Amended fused TOPSIS-VIKOR for classification (ATOVIC) applied to some UCI data sets." Expert Systems with Applications 99, 2018, pp. 115-125.

[20] S. Mohan et al., "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542-81554, 2019 doi:10.1109/ACCESS.2019.2923707.

[21] A. Khan et al., "A novel study on machine learning algorithm-based cardiovascular disease prediction," *Health Soc. Care Community*, vol. 2023, 1-10, 2023 doi:10.1155/2023/1406060.

[22] Pham et al., *'Machine learning-based detection of cardiovascular disease using ECG signals: performance vs. complexity.' arXiv Preprint ArXiv:2303.11429*, 2023.

[23] B. Kolukisa and B. Bakir-Gungor, "Ensemble feature selection and classification methods for machine learning-based coronary artery disease diagnosis," *Comput. Stand. Interfaces*, vol. 84, p. 103706, 2023 doi:10.1016/j.csi.2022.103706.

[24] I. M. El-Hasnony et al., "Multi-label active learning-based machine learning model for heart disease prediction," *Sensors (Basel)*, vol. 22, no. 3, p. 1184, 2022 doi:10.3390/s22031184.

[25] P. Wang et al., *'A wearable ECG monitor for deep learning based real-time cardiovascular disease detection.' arXiv Preprint ArXiv:2201.10083*, 2022.

[26] F. Rustam et al., "Incorporating CNN features for optimizing performance of ensemble classifier for cardiovascular disease prediction," *Diagnostics (Basel)*, vol. 12, no. 6, p. 1474, 2022 doi:10.3390/diagnostics12061474.

[27] K. Saikumar and V. Rajesh, "A machine intelligence technique for predicting cardiovascular disease (CVD) using Radiology Dataset," *Int. J. Syst. Assur. Eng. Manag.*, pp. 1-17, 2022 doi:10.1007/s13198-022-01681-7.

[28] Hossen, MD Amzad, Tahia Tazin, Sumiaya Khan, Evan Alam, Hossain Ahmed Sojib, Mohammad Monirujjaman Khan, and Abdulmajeed Alsufyani. "Supervised machine learning-based cardiovascular disease analysis and prediction." Mathematical Problems in Engineering 2021 (2021): 1-10.

[29] P. Sharma et al., "Artificial plant optimization algorithm to detect heart rate & presence of heart disease using machine

learning," *Artif. Intell. Med.*, vol. 102, p. 101752, 2020 doi:10.1016/j.artmed.2019.101752.

[30] N. Reddy et al., "Classification and feature selection approaches by machine learning techniques: Heart disease prediction," *Int. J. Innov. Comput.*, vol. 9, no. 1, 2019.

[31] C. M. Bhatt et al., "Effective heart disease prediction using machine learning techniques," *Algorithms*, vol. 16, no. 2, p. 88, 2023 doi:10.3390/a16020088.

[32] A. Parsa, "Bahador, Ali Movahedi", Homa Taghipour, Sybil Derrible, and Abolfazl Kouros Mohammadian. "Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis." Accident Analysis & Prevention 136 (2020), p. 105405.

[33] M. A. Awal et al., "A novel Bayesian optimization-based machine learning framework for COVID-19 detection from inpatient facility data," *IEEE Access*, vol. 9, pp. 10263-10281, 2021 doi:10.1109/ACCESS.2021.3050852.

[34] J. A. Faysal et al., "XGB-RF: A hybrid machine learning approach for IoT intrusion detection" in *Telecom*. MDPI, vol. 3, no. 1, pp. 52-69, 2022 doi:10.3390/telecom3010003.

[35] K. Topuz et al., "Predicting graft survival among kidney transplant recipients: A Bayesian decision support model," *Decis. Support Syst.*, vol. 106, pp. 97-109, 2018 doi:10.1016/j.dss.2017.12.004.

[36] S. Bashir et al., "Improving heart disease prediction using feature selection approaches" in 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST). IEEE, 2019, pp. 619-623 doi:10.1109/IBCAST.2019.8667106.

[37] A. K. Dwivedi, "Performance evaluation of different machine learning techniques for heart disease prediction," *Neural Comput. Appl.*, vol. 29, no. 10, pp. 685-693, 2018 doi:10.1007/s00521-016-2604-1.

[38] S. Sperandeo, *Lessons in Biostatistics Understanding Logistic Regression Analysis, (February)*, 2014.

[39] P. S. Kohli and S. Arora, "Application of machine learning in disease prediction" in 4th International Conference on Computing Communication and Automation (ICCCA). IEEE, 2018, pp. 1-4 doi:10.1109/CCAA.2018.8777449.

[40] G. N. Ahmad et al., "Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques With and Without GridSearchCV" *IEEE Access*, vol. 10, pp. 80151-80173, 2022 doi:10.1109/ACCESS.2022.3165792.

[41] N. Chakrabarty et al., *'Flight arrival delay prediction using gradient boosting classifier.' Emerging Technologies in Data Mining and Information Security*. Singapore: Springer, 2019, pp. 651-659.

[42] S. Ustebay et al., "Intrusion detection system with recursive feature elimination using random forest and deep learning classifier" in *The International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT)*. IEEE, 2018, pp. 71-76.

[43] J. A. Faysal et al., "XGB-RF: A hybrid machine learning approach for IoT intrusion detection" in *Telecom*. MDPI, vol. 3, no. 1, pp. 52-69, 2022 doi:10.3390/telecom3010003.

[44] B. M. S. Hasan and A. Mohsin Abdulazeez, "A review of principal component analysis algorithm for dimensionality reduction," *J. Soft Comput. Data Min.*, vol. 2, no. 1, pp. 20-30, 2021.

[45] A. Jamal et al., "Dimensionality reduction using pca and k-means clustering for breast cancer prediction," *LKJITI*, vol. 9, no. 3, pp. 192-201, 2018 doi:10.24843/LKJITI.2018.v09.i03.p08.

[46] H. Shafizadeh-Moghadam, "Fully component selection: An efficient combination of feature selection and principal component analysis to increase model performance," *Expert Syst. Appl.*, vol. 186, p. 115678, 2021 doi:10.1016/j.eswa.2021.115678.

[47] S. Gajjar et al., "Real-time fault detection and diagnosis using sparse principal component analysis," *J. Process Control*, vol. 67, pp. 112-128, 2018 doi:10.1016/j.jprocont.2017.03.005.

[48] Available at: https://learnopencv.com/principal-component-analysis/.