Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 01, January-2025

A Predictive Paradigm: Proposed Ensemble Frame Work for Cardiovascular Disease

Jaya Pal, Anand Prakash, Sunny Kumar
Department of Computer Science and Engineering,
Birla Institute of Technology Mesra Ranchi (Jharkhand), India

Abstract
Heart failure, a non-transmissible illness, is the main cause of mortality globally. As a whole, there are four main kinds of heart illness: inheritable, coronary artery disease, cardiac failure, and artery. Timely and accurate identification of cardiac disease is crucial for patient survival and preventing additional injury. Having a predictive system that can forecast the development of cardiovascular disease before it deteriorates is of utmost importance. To predict cardiac disease, researchers employ a variety of machine learning techniques and algorithms. Machine learning has garnered interest in several domains, including the realm of medical sciences. Scientists employ several machine learning algorithms and methodologies to predict cardiac disease. This study employs data from IEEE Data Port, a comprehensive, publicly accessible online dataset particularly tailored for persons with cardiovascular ailments. The collection consists of vital data gathered from many sources, such as the Hungarian, Long Beach VA, Switzerland, and Statlog databases. The features encompass the highest achieved heart rate, serum cholesterol levels, the type of cardiac symptoms experienced, and fasting blood sugar levels. Performance metrics such as accuracy, precision, recall, F1-score, confusion matrix, and precision recall curve can evaluate the model's usefulness and robustness. The study presents a stacked ensemble classifier framework that incorporates many machine learning techniques, including random forest, K-nearest neighbour, logistic regression, support vector, and others. The approach we devised had a 94% accuracy rate, surpassing the current body of research.

Keywords: **Machine Learning Algorithms, Cross validation, Stack ensemble Technique, cardiovascular disease, Heart disease dataset, Performance measures**

I. INTRODUCTION

Disease is the leading cause of death worldwide [7]. The United Nations Health Organization (WHO) estimates that heart-related diseases accounted for 32% of the global fatalities in 2019 [11]. Cardiovascular disorders (CVD) are responsible for 28.1% of all fatalities in India, according to the Ministry of Health and Family Welfare [15–18]. At this time, ST-elevation myocardial infarction (MI) and acute coronary syndrome are more common in India than any other country in the globe. In 2013, 261,694 people lost their lives due to diseases in India. This number is up 138% from 1990 levels [15]. For an accurate assessment of the severity of cardiovascular disease, practitioners use blood tests, electrocardiograms (ECG or EKG), cardiac MRI, and cardiac computed tomography (CT). A sufficient number of trained medical experts should be available in developing nations to reliably diagnose cardiovascular disease. Errors in these tests, brought on by insufficient infrastructure, can further complicate matters and end longer patients' lives [25]. There is one doctor for every 91,000 people in India [24].

The early diagnosis of heart disease in patients, together with the right medical treatment, may reduce the occurrence of premature death [1]. Disease significantly impacts both the healthcare system and people's health. Furthermore, coronary heart disease is the first of many types of cardiovascular illnesses. Coronary heart disease, a well-known and common cardiovascular disease [36], affects many people. This disease narrows the coronary arteries, leading to their eventual death. The heart receives its oxygenated blood via the arteries. But it doesn't work very well when plaque, which contains cholesterol, is present. Cardiomyopathy, often known as congestive heart failure, is the subject of the second group. The heart is thus unable to circulate blood effectively to every area of the human body. A significant decline in the heart's pumping ability, resulting from a significant weakness of the heart muscles, characterizes this advanced symptom of coronary artery disease. Another example is congenital heart disease, a medical condition that is present from birth [37]. One may observe cardiac interatrial or interventricular communication, also referred to as septal defects. Cyanotic heart failure is defined by the presence of disruptive anomalies that either completely or partially block circulation to various parts of the heart or result in insufficient oxygenation throughout the body. Cardiomyopathy is the final medical advice. Cardiomyopathy is a pathological condition that impairs the heart's capacity to effectively pump blood. It causes dysfunction or changes in the structure of the cardiac muscles, potentially resulting in heart failure.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 01, January-2025

Using ensemble learning, which includes combining different model types and making changes to the architecture, could improve prediction tasks' accuracy and generalizability [9]. Therefore, utilising an ensemble learning approach is beneficial for addressing crop production forecasts. An ensemble learning method Sacking is used, which combines base and Meta models to make predictions using self-learning [14]. Using this strategy can significantly enhance the accuracy and relevance of forecasts [17]. The result section provides a comprehensive analysis and evaluation of previous research studies.

Therefore, the primary objectives of this research work are outlined below:

- Data preprocessing and data cleansing
- Examine the viability and precision of cardiovascular disease prediction models using various machine learning methodologies. Analyze and compare the Random Forest (RF), Support Vector Machine (SVM), and K-Nearest Neighborhood (KNN) foundation models (Layer 1).
- Test the suggested Stacking Logistic Regression Ensemble (LORENS) Meta model (Layer 2) to see how well it works at improving the accuracy of heart failure prediction. This study trains the ensemble model with a substantial dataset, enhancing its ability to generalize well.
- With a value of K=5, the K-fold cross validation technique is used by both the Base model and the Meta model.
- Measures including accuracy, precision, recall, F1 score, and ROC analysis will be used to evaluate the efficacy of the proposed framework in comparison to existing literature.

The proposed ensemble structure for the diagnosis of heart disease is illustrated in Fig. 1.

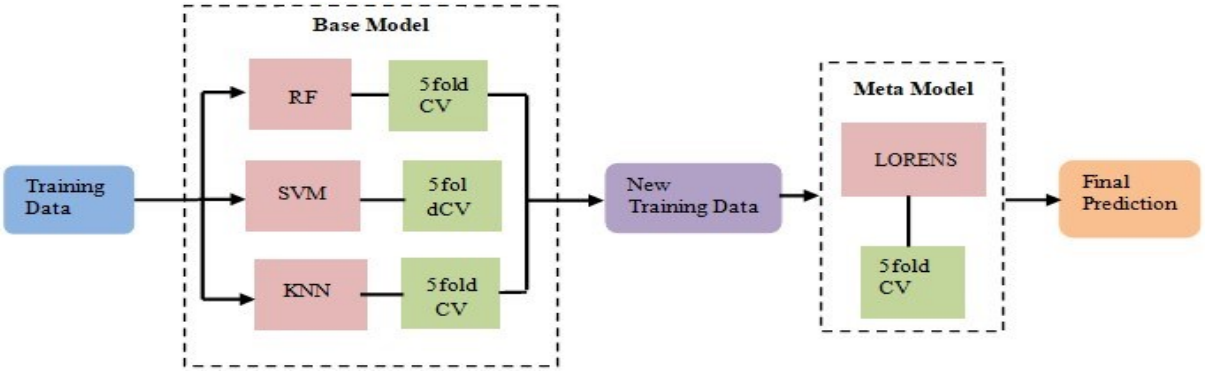

Fig. 1: Overview of proposed ensemble model of cardiovascular disease prediction, RF, Random Forest; SVC, Support Vector machine; KNN, K-Nearest Neighbor Classifier; CV, Cross validation

The subsequent phases of the research are outlined as follows: Section 2 comprises the literature review. Section 3 provides an extensive analysis of the various methodologies. Section 4 provides a comprehensive assessment of the proposed framework, including an in-depth analysis of model selection, experimental conditions, and the recommended approach. Section 5 conducts a comparative analysis based on the findings. Section 6 concludes the paper.

II. REVIEW OF LITERATURE

This section presents an analysis of 10 unique research publications, examining how previous researchers have addressed the same topic using diverse methodologies, as shown in Table 1. The incidence of cardiovascular disease has experienced a significant surge [7]. Scientists are utilising various approaches and algorithms to predict cardiovascular illness. Over time, researchers have carried out several investigations on the prognosis of cardiovascular disease. An overview of these investigations is provided below. While conducting their research, the researchers made use of a wide variety of approaches, such as multi-layer perceptron (MLP), decision trees, artificial neural networks, support vector machines (SVMs), K closest neighbors, decision trees, and random forests [12]. The authors originally employed the dataset at a granular level. Afterwards, the researches merge the dataset to create a unified representation.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 01, January-2025

Table 1: Comparison of existing methods for predicting heart disease

Sl. No.	Author(s)	Approach	Dataset	Accuracy
1	Atallah et al. (2019) [2]	(SGD)Classifier, K-Nearest Neighbor Classifier, Random Forest Classifier, Logistic Regression Classifier	Cleveland	90%
2	Bialy et al. (2016) [19]	Ensemble of FDT, C4.5, MLP, SVM, and Naive Bayes	Cleveland	85.30%
3	Bashir et al. (2014) [21]	Memory-based learner, DT-IG,DT-GI, Ensemble of Naive Bayes, and SVM	UCI Repository, ricco database	88.52%
4	Latha et al. (2019)[28]	Random Forest, Multilayer Perceptron, Bayes Net Naïve Bayes	Cleveland	84.49%
5	Modak et al.(2022) [27]	Multilayer perceptron	Cleveland, Hungarian, Switzerland, Long Beach, and Statlog	87.70%
6	Pawlovsky (2018) [4]	Weightedk-nearest neighbour	Cleveland	84.83%
7	Sarah et al.(2022)[3]	Logistic regression	Cleveland	85.25%
8	Miao et al. (2016)[22]	Adaptive boosting	UCI Repository	80.14%
9	Achyut et al. (2022)[10]	Ensemble Framework	IEEE Data Port	92.34%
10	Nguyen et al. (2021)[30]	Naïve Bayes, Logistic Regression, SVM and Decision Trees	Cleveland	83.5%

The researchers used classification matrices to reliably and accurately analyze several risk factors associated with coronary heart disease. In order to create a model for building an intelligent mixed structure, the authors employed many methodologies and used K-fold (10-fold) cross-validation techniques for both the full and specifically selected attributes. The authors used a total of four feature selection techniques, including the LASSO strategy [32], the relief feature selection methodology [33], and others. The researchers calculated the chi-square statistic and P-value used in choosing feature approaches and employed HGBDTLR, a stack-based algorithm. The HGBDTLR algorithm is built on a stack. The framework was constructed using a variety of machine learning techniques, including SVMs, DTs, LR, Adaboost, RF, GBDT, KNN, and an HF with a linear model. Precision, recall, accuracy, and the F1 score were among the several characteristics analyzed by researchers [34].

The authors employed a majority voting ensemble technique, including many algorithms, to achieve a peak accuracy of 90% [2]. The authors employed a consensus technique by aggregating the results of all the algorithms by a majority vote, improving the overall accuracy. In addition, the authors calculated the correlation between the target variable and another characteristic, which analyses the connection between the two variables when the data point is positive. The authors in [8] perform multilevel data splits using CHAID (Chi-Squared Automated Interaction Detector), a structural technique that bears similarities to a decision tree. The CHAID decision tree algorithm provides cardiologists with a comprehensive analysis of a patient's health situation, enabling them to effectively distinguish between various illnesses [35]. Subsequently, they used majority voting as a means to enhance the overall accuracy. The researchers primarily focus on using machine learning techniques to predict heart disease. Despite evaluating several aspects, the authors found that the dataset's size is generally moderate in most circumstances. The aim of this research was to predict cardiovascular disease by examining various algorithms and approaches. This research work focuses on algorithm creation using a dataset [26] with 1190 occurrences and 11 attributes; this work focuses on the creation of algorithms. The goal is to demonstrate these algorithms' massive-scale performance.

Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 01, January-2025

III. METHODOLOGY

The section explains how the suggested framework makes use of machine learning classification algorithms. Several classifier models were tested before the ensemble of best models was finalized. Five different classifiers were trained using the training data set. We chose three different classifiers as our foundational models for further analysis after the initial training. Their performance in Layer 1 is the deciding factor. Using the Logistic Regression Ensemble (LORENS) as the met model (Layer 2) is key to our strategy.

A. Random Forest Classifier (RFC) [20][38]

Random forest modeling is a classification approach that uses tree-based models as depicted in Fig. 2. The Random Forest classifier, also known as RFC, is a widely used learning technique in machine learning that involves supervision. In order to generate judgments, multiple decision regression trees may use branching to choose the optimal feature from a portion of the whole feature set [20]. This approach has the benefit of maintaining the autonomy and variety of each decision tree while mitigating the risk of excessive fitting [38]. The Random Forest technique does this by randomly picking a subset of features to use for splitting nodes. Unlike conventional decision trees, which determine the most probable saturation points, this approach employs optional thresholds for each characteristic to maximize the randomness of the decision trees.

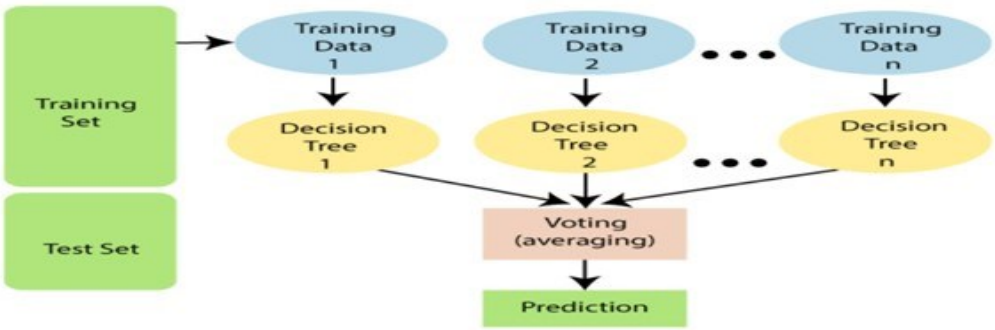

Fig. 2: Random Forest Classifier

B. K-Nearest Neighbor Classifier (KNC)

Applications, such as recognizing patterns, categorizing, and forecasting often use K-nearest neighbors (KNN) in machine learning. By averaging the values of the closest data points, the K-nearby neighbors (KNN) regression process determines the predicted value. Among its many benefits are its exceptional predictive accuracy, resistance to out-of-range values, and limitless capacity to process uncertain inputs. Using the standard geometrical distance measure, we determined the separations within the data values [23]. The calculation for the distance in terms of the Euclidean plane is:

$$d(p,q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2} \quad (1)$$

C. Support Vector Classifier

Supervised machine learning makes extensive use of the Support Vector Classifier (SVC) and Support Vector Machine (SVM) to tackle classification and regression problems. Finding the hyperplane that differentiates the two sets allows us to classify them. The exact locations of each observation are linked to support vectors. The Support Vector Classifier (SVC) uses mapping techniques to generate a high latitude estimate function from low latitude geographic information. This method achieves a balance between the computational cost and the accuracy of the regression model [6]. SVR's primary objective is to identify the optimal decision boundary. Vectors are the training sample points closest to the hyperplane and meet certain requirements.

D. Naive Bayes Classifiers

The Bayes method leads algorithms to believe that every pair of features is conditionally independent, depending on the current value of the class parameter. This assumption is referred to as "naive". Naive Bayes classifiers exhibit notable speed when contrasted with more advanced techniques. The division of the feature distributions based on class conditions enables the individual estimate of each distribution as a one-dimensional distribution. Consequently, this helps to mitigate the problems caused by high dimensionality [5].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 01, January-2025

E. Logistic Regression model

There are two potential solutions for the dependent factor, and Logistic Regression (LR) is the best method for regression to use[13]. Logistic regression is a form of predictive analysis among several other forms of regression analysis. The tool enables us to visually depict data and illustrate the correlation between one or more independent variables with varying measurement scales (nominal, ordinal, interval, or ratio) and a sole dependent binary variable. Logistic regression produces outcomes that are distinct and separate, as opposed to the continuous outcomes generated by linear regression. Logistic regression employs the logistic sigmoid function to provide distinct outcomes, whereas linear regression generates continuous numerical values. Multiple distinct groups may assign the resulting probability value.

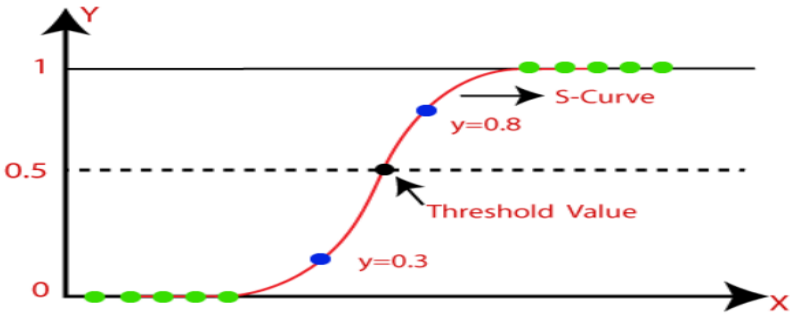

Fig. 3: Logistic Regression

F. K-Fold Cross Validation

The k-fold cross-validation technique is an essential method for evaluating the effectiveness of forecasting algorithms in the disciplines of machine learning and statistical analysis [39]. The procedure involves partitioning the dataset into k subgroups of equal size, also known as "folds." Fig. 4 demonstrates that the procedure iterates, training the model on the remaining k-1 folds while employing one fold as a set of validations.

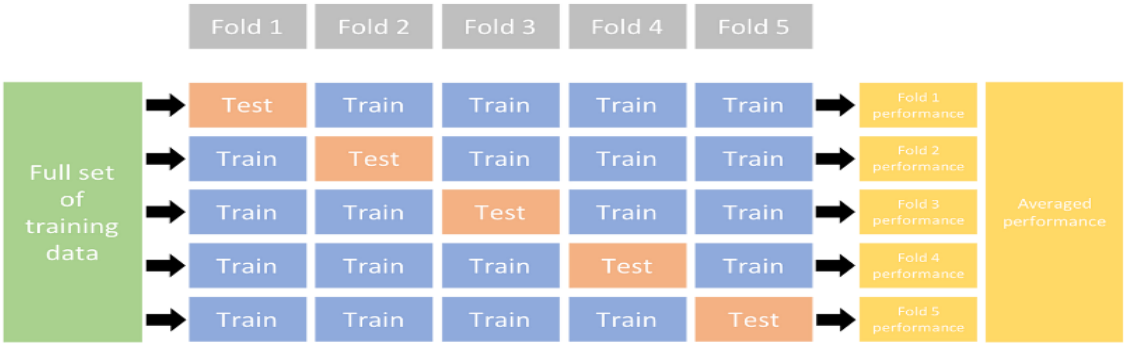

Fig.4: K-fold cross validation

In order to guarantee every fold is used as the validation set exactly once, we iterate this process k times. For each cycle, we compute performance metrics such as accuracy or mean square error. Subsequently, we combine these indicators to give a comprehensive evaluation of the model's performance [16].

G. Stacking Ensemble

Stacking is a technique that combines several base models with Meta models [31]. This technology's core is a progressively created multi-layer system of instruction. When merging models, stacking frameworks employ multiple base learners, a feature that sets them apart from the typical integrated framework-guided methods of clustering (bagging) and boosting techniques. The stacking approach begins with applying cross-validation to transform the primary features into secondary features. There are three parts to the training plan. The stacking ensemble learning method is employed to train a diverse group of learners, who subsequently acquire knowledge from the dataset and combine the training results of all classifiers into a single, distinct dataset prior to feeding it into the Meta classifier. According to the resulting value of the Meta learner, the secondary layer model determines the final result.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 01, January-2025

IV. EXPERIMENT

4.1 Dataset Description:

The port of IEEE Data is used to gather data on heart failure (CVD) [26, 13]. We integrated five recognized heart failure (CVD) programs to generate this set of data: the Statlog (Heart) data set, the Hungarian database, the Cleveland dataset, the Long Beach VA dataset, and the Swiss dataset. Previously, one could view these datasets independently without any linkage. The dataset includes 1190 instances and 11 variables that deal with patient characteristics and outcomes. Old maximum = ST (depressive disorder), years lived, blood pressure, serum cholesterol, and fastest heart rate (in the range of 71 and 202 bpm) are a few examples of data types. Here is a collection of statistical data that includes the following details: sex (0-1), type of chest discomfort (1-2, 3-4), baseline glucose level (1>120 milligrams), resting ECG results (0-1.2), exercising-induced angina (0-1), inclination of the peak of the exercise ST section (0, 1, 2), and goal (0-1). In Table 2, we can see the bare minimum of characteristics.

Table 2: Nominal Attributes descriptions

Attribute	Description
Sex	Gender of patients(1=male,0= female)
Resting electrocardiogram results	Results of ECG while at rest --Value0: normal
	-- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) -- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
Chest Pain Type	Types of chest pain experienced by patients categorized:- --Value1:typical angina --Value2:atypicalangina --Value3:non-anginalpain --Value4: asymptomatic
The Slope of the Peak Exercise ST Segment	ST segment measured in terms of slope during peak exercise --Value1: up sloping --Value2: flat --Value3:down sloping
Exercise induced angina	Angina induced by exercise(1= Yes;0= No)
Fasting Blood sugar	(Blood sugar levels on fasting > 120 mg/dl) (1 = true; 0 = false)
class	1=heartdisease,0= Normal

The correlation coefficient is computed to examine the dataset and determine the relationship between the target diagnosis and each of the parameters. Table 3 demonstrates the strongest link between the ST Slope and the target features of exercise-induced angina, the type of chest pain, and ST depression Figure 5 displays a heat map that illustrates the association between all traits, offering a distinct perspective on the connection between each element. Figure 6 displays a pie chart and histogram that depicts the distribution of patients in the Heart Disease Dataset (Comprehensive) based on gender and age. It is evident from the data that males account for 76% of the total while females account for only 24%. Furthermore, Figures 7–13 showcase histograms that we use to provide an early look at the data presented in the ongoing feature visualization.

Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 01, January-2025

4.2 Relationship between data attributes and data visualization.
The relation between data attributes and data visualization is represented by the corresponding Table and figures as shown below.

Table 3: Correlation with Target Interpretation

st_slope	0.505608
exercise_induced_angina	0.481467
chest_pain_type	0.460127
st_depression	0.398385
sex	0.311267
age	0.262029
fasting_blood_sugar	0.216695
resting_blood_pressure	0.121415
rest_ecg	0.073059
cholesterol	-0.198366
max_heart_rate_achievedsnip	-0.413278

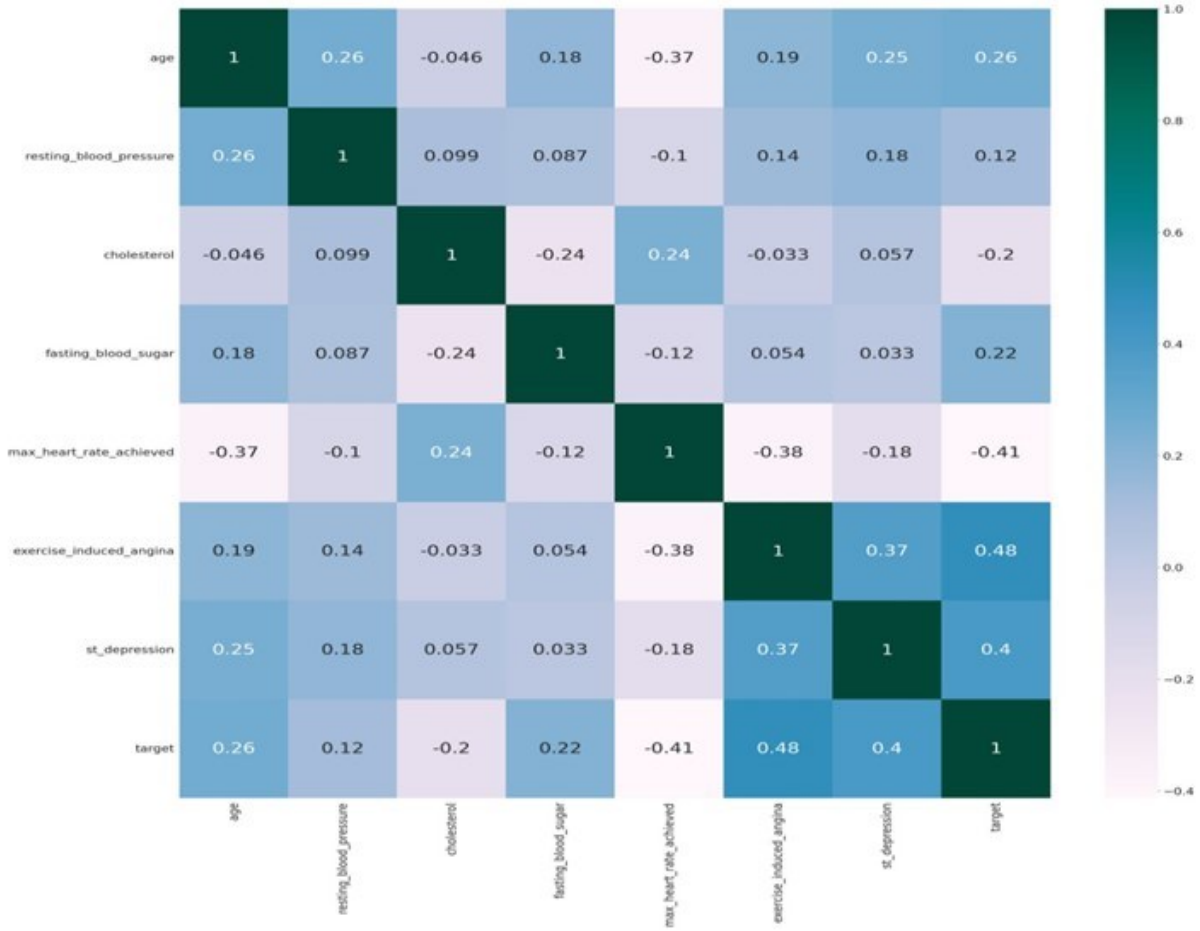

Fig. 5: Heat map of cross-correlation values

Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 01, January-2025

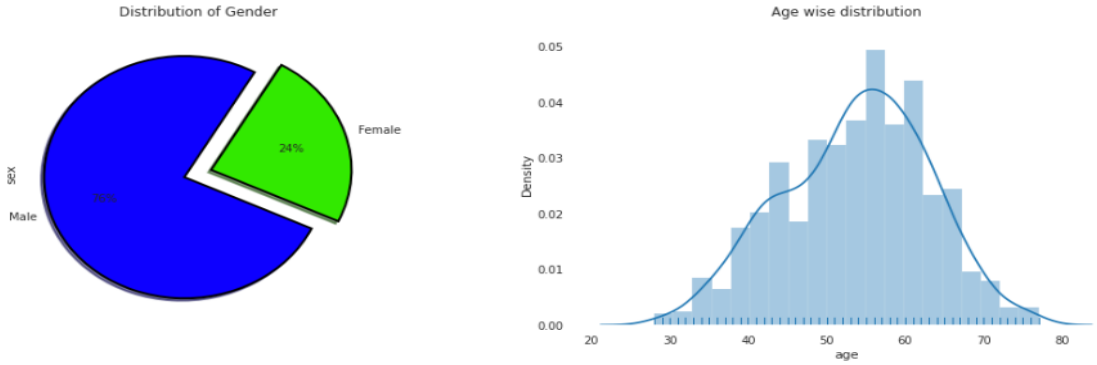

Fig. 6: Age and gender distribution within the dataset

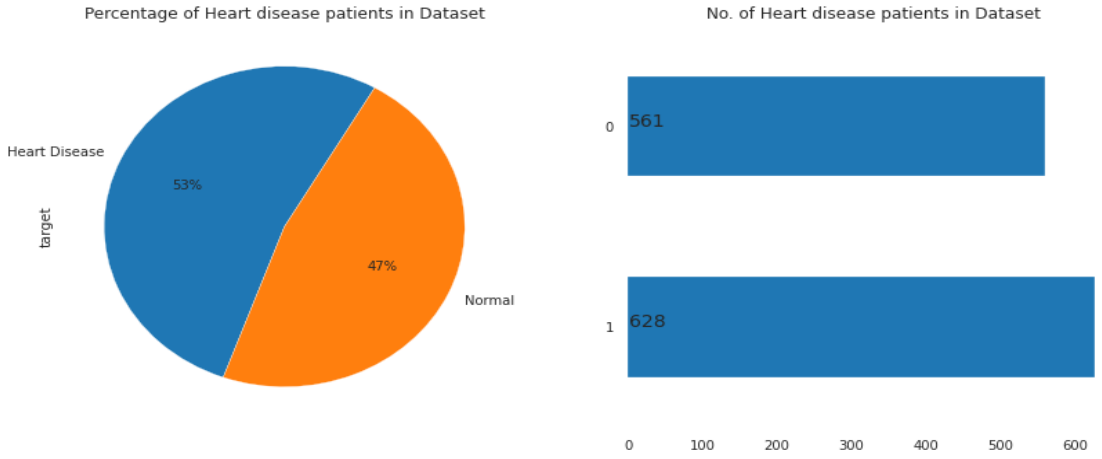

Fig. 7: Representation of the dataset's enumeration of patients with heart disease

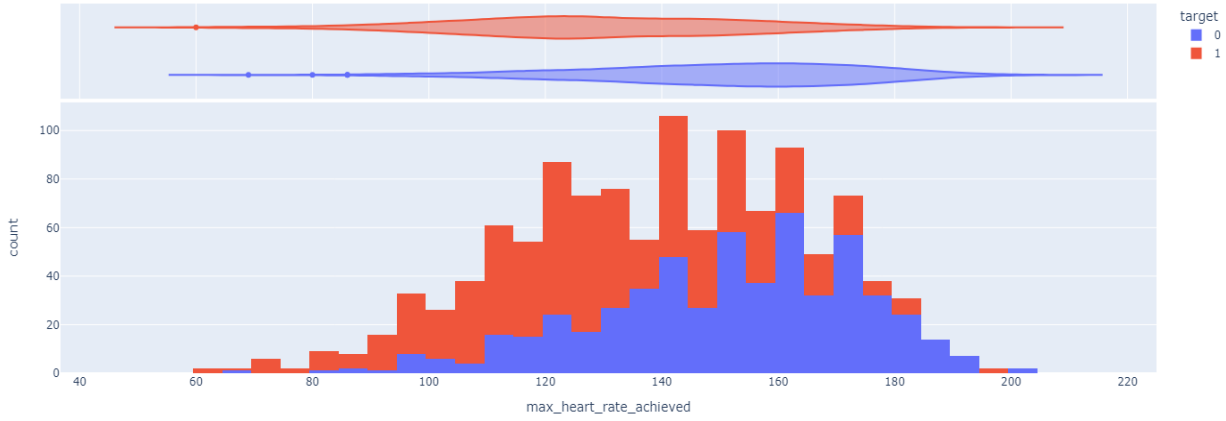

Fig. 8: Representation of the distribution of the maximum heart rate achieved

Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 01, January-2025

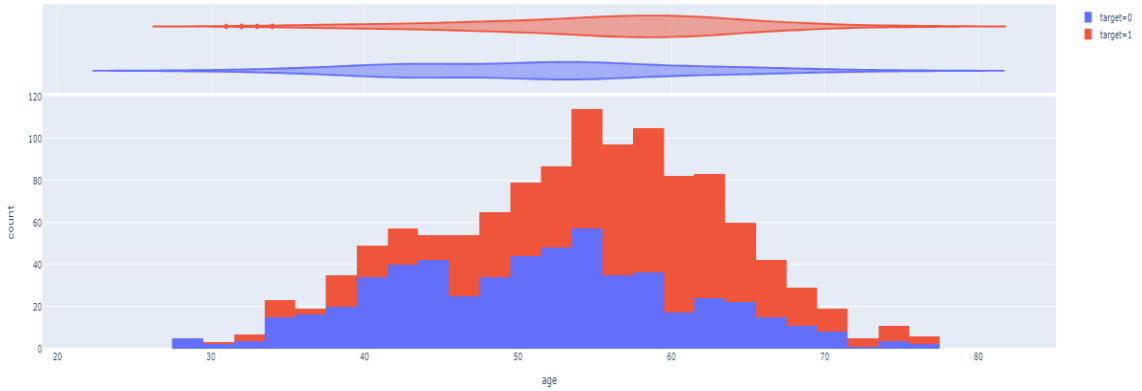

Fig.9: Representation of the relationship between age and the target variable

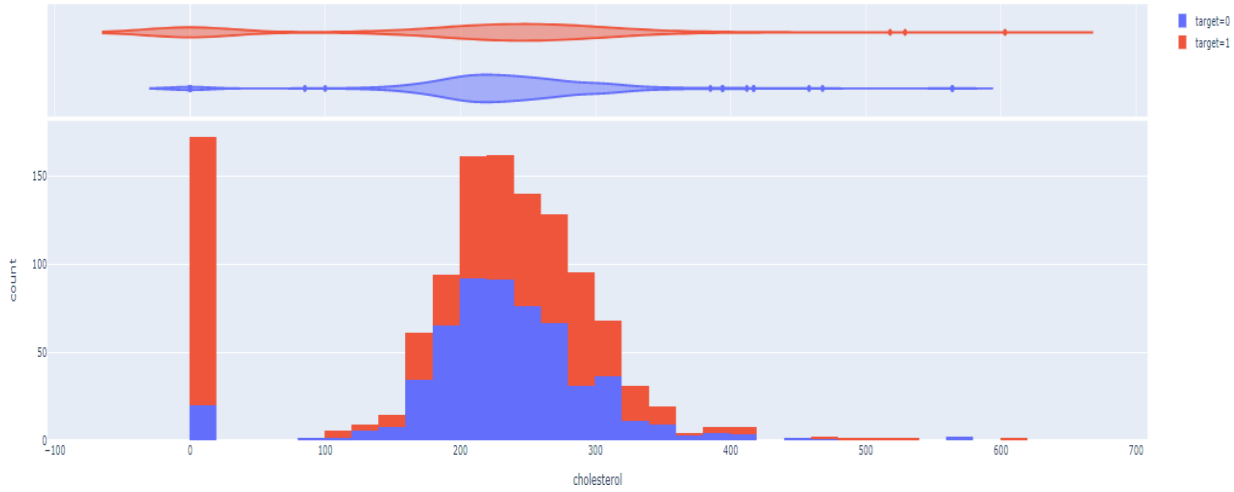

Fig.10: Distribution of cholesterol.

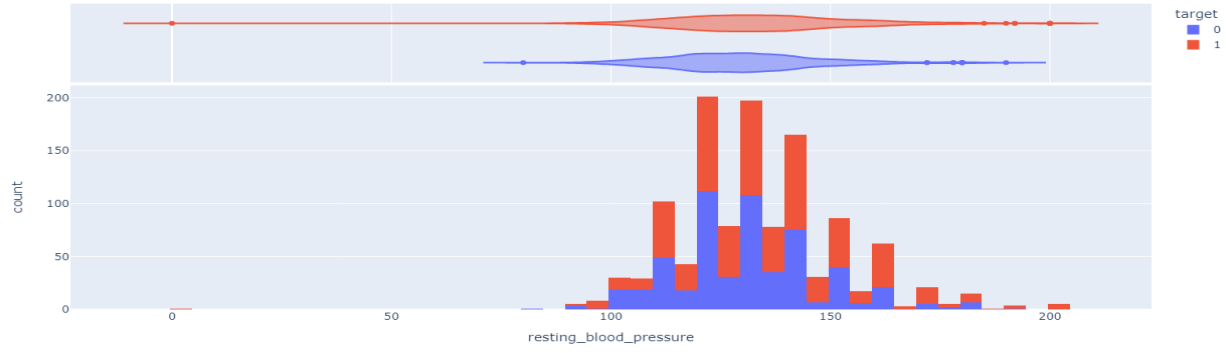

Fig 11: Distribution of resting Blood Pressure

Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 01, January-2025

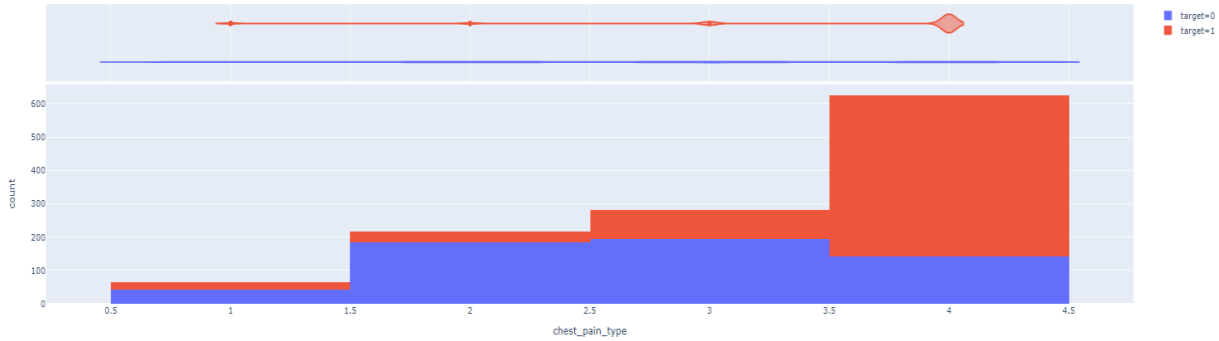

Fig. 12: Distribution of chest pain

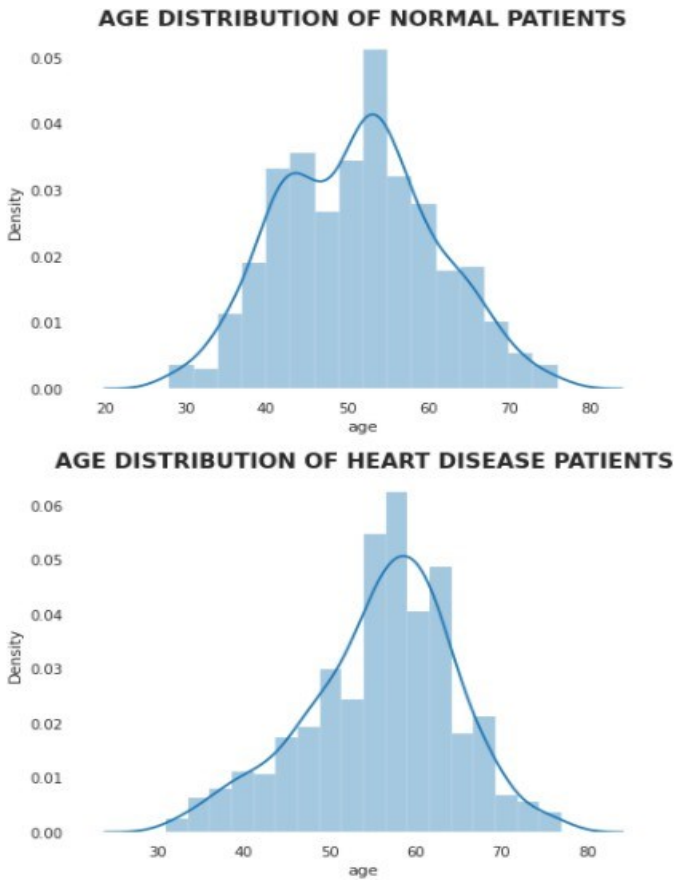

Fig. 13: Distribution of patients, categorized by age, between those who are normal and those who have heart disease.

Histograms in figures 7-13 show the frequency or density of data points within predetermined intervals, providing information on core tendencies, dispersion, skewness, and probable outliers. Analysts can use these histograms to spot patterns and potential abnormalities and make informed judgments about data pretreatment, modelling, or additional analysis. Furthermore, histograms allow for comparisons of distinct attributes, which aid in determining their relative value or contribution to the total dataset. Overall, using histograms in the visualization process improves data comprehension, allowing researchers to extract useful insights and draw meaningful conclusions from continuous attribute data.
Figure 14 plots the greatest related ongoing feature (ST Slopes) against age to determine the presence of a relationship. No matter what age they are, those with a ST inclination of 2 have an increased susceptibility to heart disease.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 01, January-2025

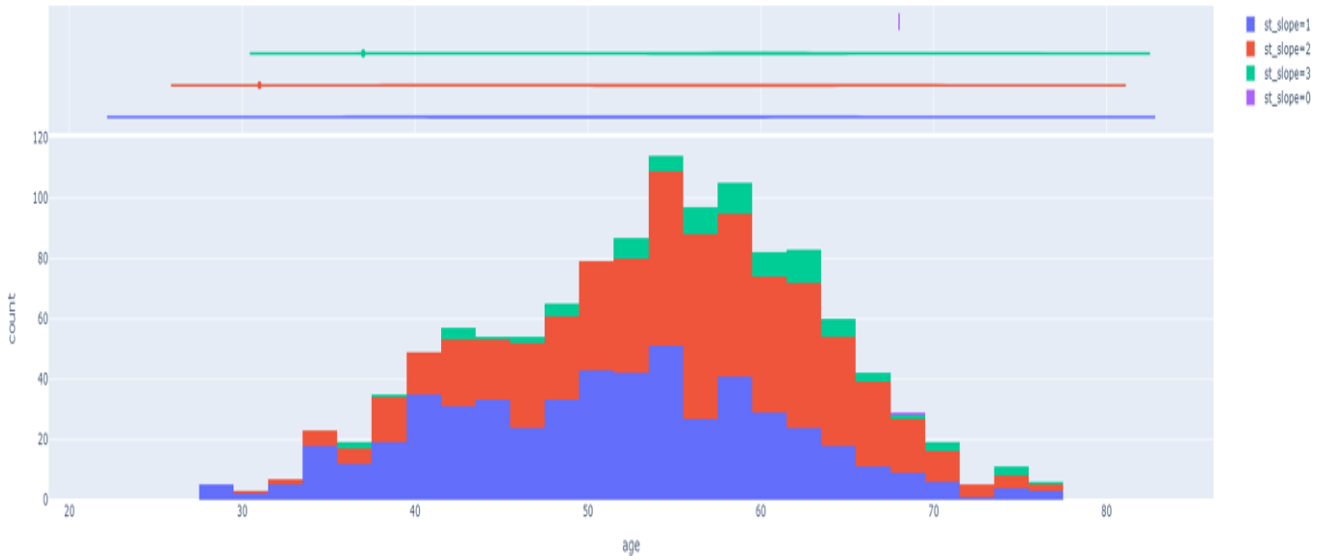

Fig. 14: Representation of the link between the continuous property "ST Slope" and age in terms of the likelihood of heart disease.

Electrocardiogram data most likely serves as the source of "ST Slope," a measure of heart function. The passage implies that people with a ST slope of 2 are at a higher risk of heart disease, regardless of age. This correlation emphasizes the potential clinical utility of ST slope patterns in detecting cardiac problems. However, the paragraph does not include information on the strength of the correlation, any confounding variables, or statistical analytic methodologies used. Nonetheless, recognizing such correlations may help clinicians recognize and treat cardiac disease earlier, thereby improving patient outcomes.

4.3 Proposed Methodology

A) Obtaining Initial Data Sets:

- Data Cleaning and Preprocessing: Remove outliers, use mean or median imputation for missing values, and manually validate and repair errors to ensure accuracy.
- Obtain the initial data sets (Xtrain and Ytrain), with Xtrain consisting of features and Ytrain including target labels.
- Divide the training set into five folds (Fold1-Fold5) for cross-validation.

B) Learning and Generating New Data Sets (Base Models):

- Choose three basic classifiers for the foundation layer: Random Forest (RF), Support Vector Classifier (SVC), and K-Nearest Neighbors (KNC).
- Use five-fold cross-validation to train each base classifier with training data. This technique entails dividing the data into five subsets, training the classifier on four of them, and then testing its performance on the remaining subset five times.
- Generate fresh data sets (Xtrain2 and Xtest2) for the second layer based on base classifier predictions
- Train all base classifiers on the training set, and then use them to predict class labels for both training and test data.
- Combine predicted class labels from base classifiers to create Xtrain2 and Xtest2 feature sets, with each column representing the expected class label.

C) Forecasting on Test Data (Base Models):

- Use trained base classifiers (RF, SVC, and KNC) to predict class labels for the test data set (Xtest).
- Create a new feature set Xtest2 by integrating predicted class labels from each base classifier, resulting in a matrix with three columns for RF, SVC, and KNC.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 01, January-2025

D) Training and Cross-Validating Secondary Layer Model (Meta Model):

- The Meta model chooses logistic regression as its secondary layer classifier.
- Train the Logistic Regression classifier using the new training set (Xtrain2) made up of base classifier predictions and true labels (Ytrain).
- Use cross-validation on the training set to assess the performance of the logistic regression classifier and adjust its parameters as needed.
- After training and validation, the logistic regression model predicts the new test set (Xtest2) using the predictions of the base classifiers.

E) Overall Evaluation and Conclusion:

- The technology improves predictive accuracy by combining predictions from multiple base classifiers in a two-layer ensemble.
- Both base and meta classifiers use cross-validation to evaluate performance and optimise hyper parameters, resulting in a robust and reliable model evaluation
- The methodology seeks to identify diverse patterns in data and make reliable predictions for classification problems by integrating different classifiers at both layers.

The recommended organization of this research work shown in Fig. 15 is divided into three sections:
1) Acquisition of data: We procured the CVD dataset via the IEEE Data Port.
2) Data processing: We quantitatively analyzed the heart disease data.
3) Creating a stacking model: To improve heart failure forecasting accuracy, it is necessary to make use of the strengths and features of each individual base model.

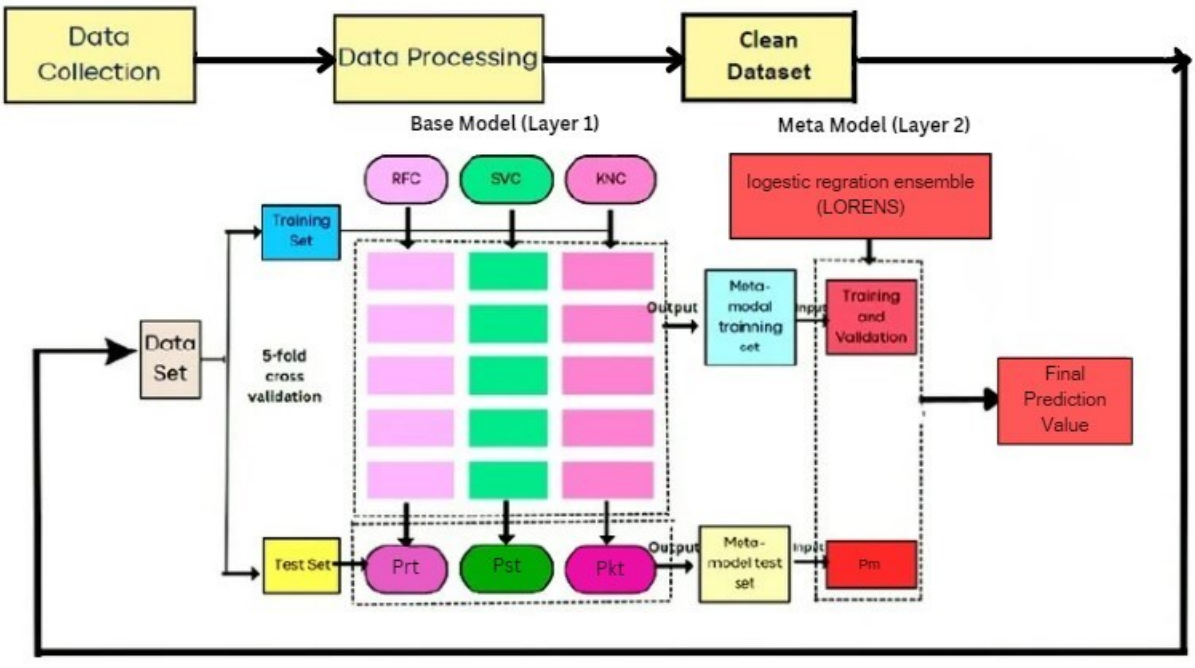

Figure 15: A Model for Stacking Ensembles for Learning

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 01, January-2025

V. RESULT AND COMPARISON ANALYSIS

This section presents the findings and analysis of our suggested framework. We assessed the algorithms using several performance measures. We conducted a comparative analysis of our model against existing models, evaluating their performance using criteria such as accuracy, precision, sensitivity, F1 Score, and AU-ROC. Additionally, the proposed framework contrasted with the other methods and models outlined in Section 2.

5.1 Evaluation metrics

This section examines performance metrics derived from the Truth table, including true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs).

Accuracy: It is a commonly employed metric for evaluating the effectiveness of a classifier. We derived the computation by determining the proportion of accurately recognized samples, and expressed it in the following manner:

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

Precision: It determines the accuracy of the model's positive predictions.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall: It helps to determine cost associated with false negatives.

$$\text{Recall} = \frac{TP}{TP + FN}$$

A classification system's recall refers to its ability to accurately identify and retrieve all of the samples it has categorized.

F1- score: It is a metric that quantifies a test's accuracy by measuring its purity. It evaluates both the precision and capacity of memory. A score of 1 signifies flawless performance, while a score of 0 signifies complete failure.

$$\text{F1-Score} = 2 * \frac{(\text{Precision}) x \, \text{Recall}}{\text{Precision} + \text{Recall}}$$

AU-ROC: The AU-ROC, or the area below the receiver operational characteristic curve, is a popular performance indicator that is particularly useful for evaluating categorization issues. Derive the displayed figure by evaluating the true positive rate (TPR) and the false positive rate (FPR) at various threshold settings. The AU-ROC is a useful tool for performance comparison, as it evaluates performance across a wide range of class variations and error levels.

5.2 Overall Performance Analysis:

A. Accuracy Comparison

We generated five foundational models and conducted cross-validation using a 5-fold approach to choose the most superior ones. The recommended approach for stacking ensembles in logistic regression involves using models with high levels of accuracy. The comparison of the accuracy of several machine learning methods and their relative accuracy is depicted in Figure 4.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 01, January-2025

Table 4: Comparative analysis of machine learning models and their corresponding accuracies

Sl. No.	Techniques	Accuracy
1.	Proposed Approach	0.94
2.	Random Forest	0.90
3.	SVM	0.85
4.	Logistic Regression	0.83
5.	Naive Bayes	0.83
6.	KNN	0.84

The comparison of the proposed framework with several machine learning techniques is depicted in Figure 16.

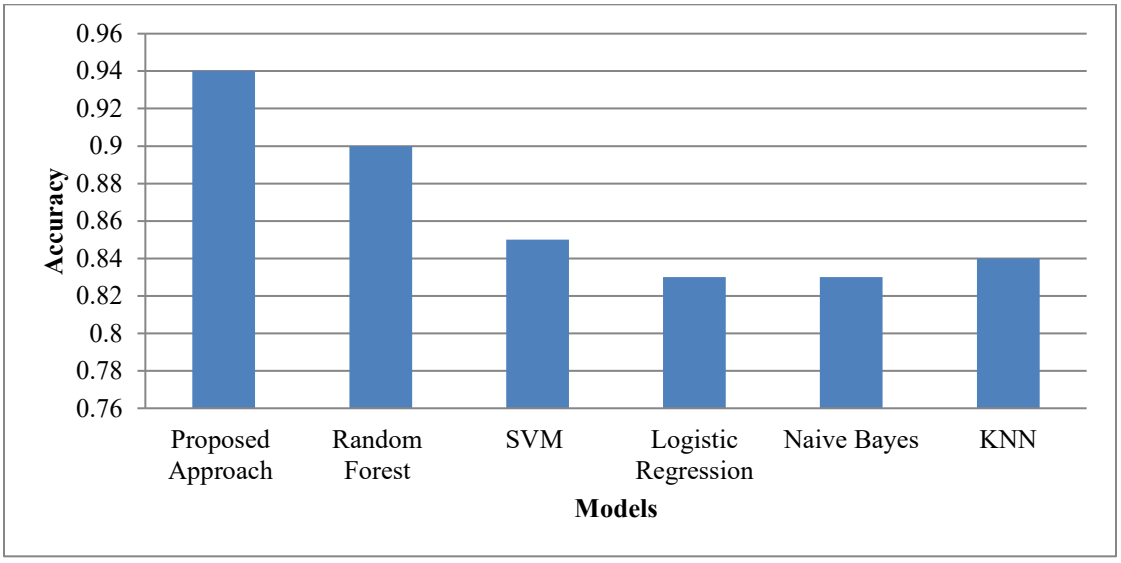

Fig.16: Comparison of proposed framework with different ML

The suggested technique has the best accuracy (0.94) of all the analyzed machine learning models, suggesting greater prediction accuracy. Individual models with lesser accuracies include Random Forest (0.90), SVM (0.85), Logistic Regression (0.83), KNN (0.84), and Naive Bayes (0.83). This highlights the benefits of the suggested technique, which combines the capabilities of numerous models using a complex ensemble strategy, resulting in improved prediction performance.

According to the in-depth study, the proposed approach makes use of the diversity of base models, such as Random Forest, SVM, and K-Nearest Neighbors, to capture varied data aspects and create a meta-model that optimally combines their predictions. This allows the recommended method to be more accurate than individual models. Overall, the suggested methodology emerges as the most successful method for the given job, demonstrating its greater predictive accuracy and showing its potential for real-world applications where precise forecasts are critical.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 01, January-2025

B. Comparison of Accuracy, Precision, Recall, and F1- Score

The F1- score as a measure of performance, along with accuracy, precision, and recall is evaluated in Table 5.

Table5.Comparison of the suggested model and established machine learning techniques

Model	Accuracy	Precision	Recall	F1-score
Proposed Approach	**0.94**	**0.94**	**0.93**	**0.94**
Random Forest	0.90	0.87	0.88	0.86
SVM	0.85	0.83	0.88	0.86
Logistic Regression	0.83	0.84	0.84	0.84
Naïve Bayes	0.83	0.84	0.83	0.84
KNN	0.84	0.84	0.87	0.85

The comparison of the proposed framework with the current ML models in terms of accuracy, precision, recall, and F1-Score is depicted in Figure 17.

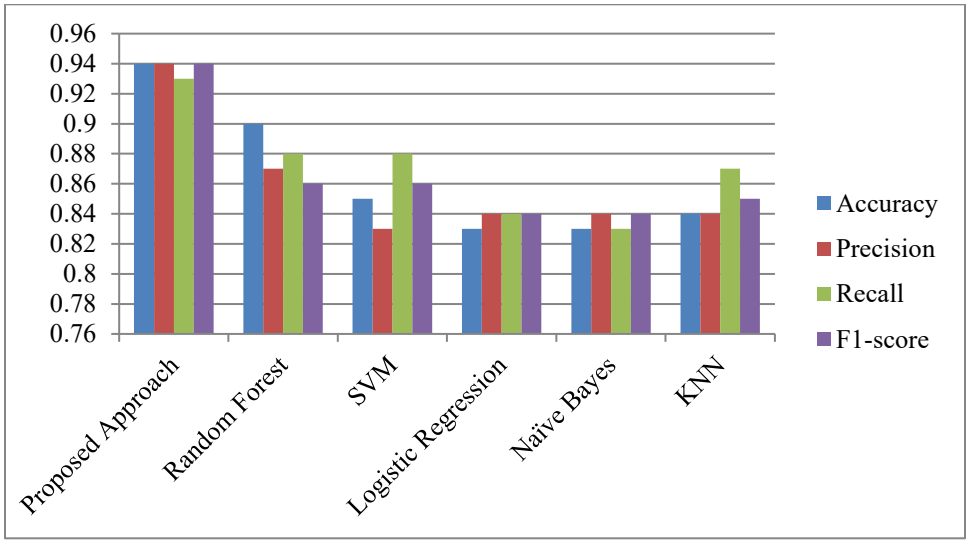

Fig 17: Comparison of the accuracy, precision, recall, and F1-Score of the proposed technique with existing machine learning models.

Among all the analyzed machine learning models, the suggested method has the greatest accuracy (0.94), confirming its efficacy in producing accurate predictions. It also achieves excellent accuracy (0.94), which suggests a high percentage of accurately detected positive cases and a low false positive rate. The suggested approach's recall (0.93) shows that it can catch a significant percentage of real positive situations without missing many. Furthermore, demonstrating the overall balance between the two and emphasizing the suggested approach's excellent performance across various assessment measures is the F1-score (0.94), which combines accuracy and recall into a single statistic.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 01, January-2025

In comparison to the ensemble techniques and the proposed framework, individual models like Random Forest (accuracy: 0.90, precision: 0.87, recall: 0.88, F1-score: 0.86) and SVM (accuracy: 0.85, precision: 0.83, recall: 0.88, F1-score: 0.86) perform pretty well but fall short. In comparison to the suggested strategy and alternative ensemble approaches, Logistic Regression (accuracy: 0.82, precision: 0.83, recall: 0.83, F1-score: 0.83) and Naive Bayes (accuracy: 0.84, precision: 0.85, recall: 0.84, F1-score: 0.85) perform worse on all measures.
The thorough study demonstrates how much better the suggested strategy is in terms of memory, F1-score, precision, and prediction accuracy. The suggested framework is a strong and dependable option for the job at hand, as it effectively captures both real positive situations and minimizes false positives by using a variety of models and an optimized ensemble method.

C. Precision-Recall Curve Comparison
The Precision-Recall curve and the Area Under the Curve (AUC) statistic provide useful information about how well classification models perform, especially when working with unbalanced datasets or where the goal is to maximize true positives while reducing false positives. Fig.18 displays the curve of precision-recall for the classifiers used to predict heart disease.

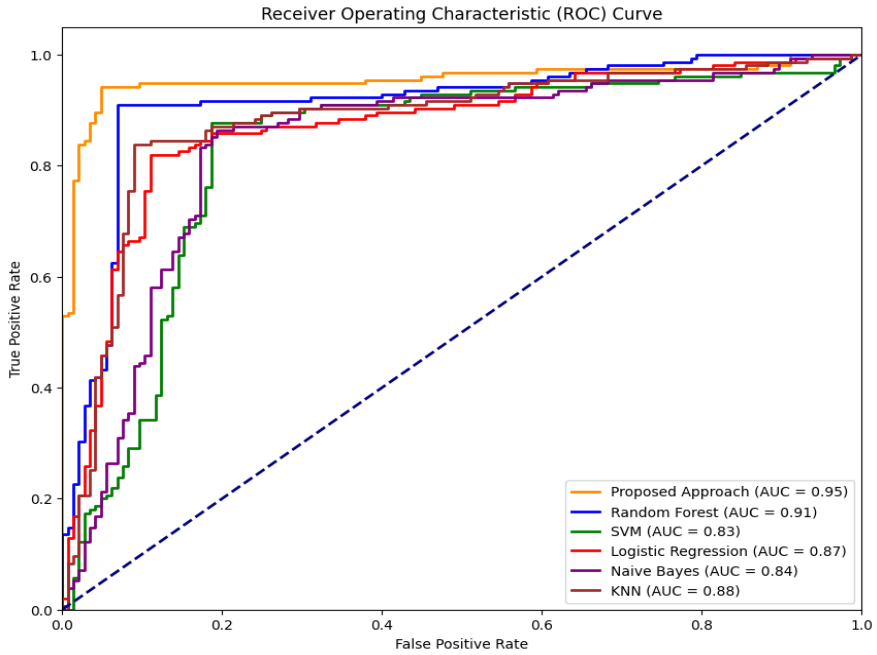

Fig. 18: Curve of precision-recall for the classifiers used to predict heart disease.

The suggested framework performs very well in accurately identifying positive instances (precision) while retaining a high recall (ability to catch actual positive cases), as seen by its high precision-recall AUC of 0.95. This shows that the suggested framework successfully minimizes false positives while maximizing the identification of real positives by striking a compromise between accuracy and recall. A high AUC score demonstrates strong overall model performance, demonstrating the framework's ability to predict positive instances across various decision thresholds.
In contrast, Random Forest, SVM, Naive Bayes, Logistic Regression, and KNN also demonstrate robust Precision-Recall AUC values of 0.91, 0.83, 0.84, and 0.87, 0.88. These models perform commendably in properly recognizing positive situations and maintaining a high recall rate, but they are somewhat worse than the suggested framework. The AUC results show that these models successfully balance recall and accuracy, but somewhat less so than the suggested method.
A close study reveals that the suggested strategy does better than single techniques such as Random Forest, SVM, Naive Bayes, Logistic Regression, and KNN with respect to precision-recall AUC. This suggests that the suggested framework's ensemble method leverages the strengths of multiple models, enhancing performance in accurately identifying positive situations while minimizing false positives.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 01, January-2025

E. Comparative analysis with the existing body of literature

Fig. 19 presents a comparison of the accuracy of the recommended technique with the existing body of literature.

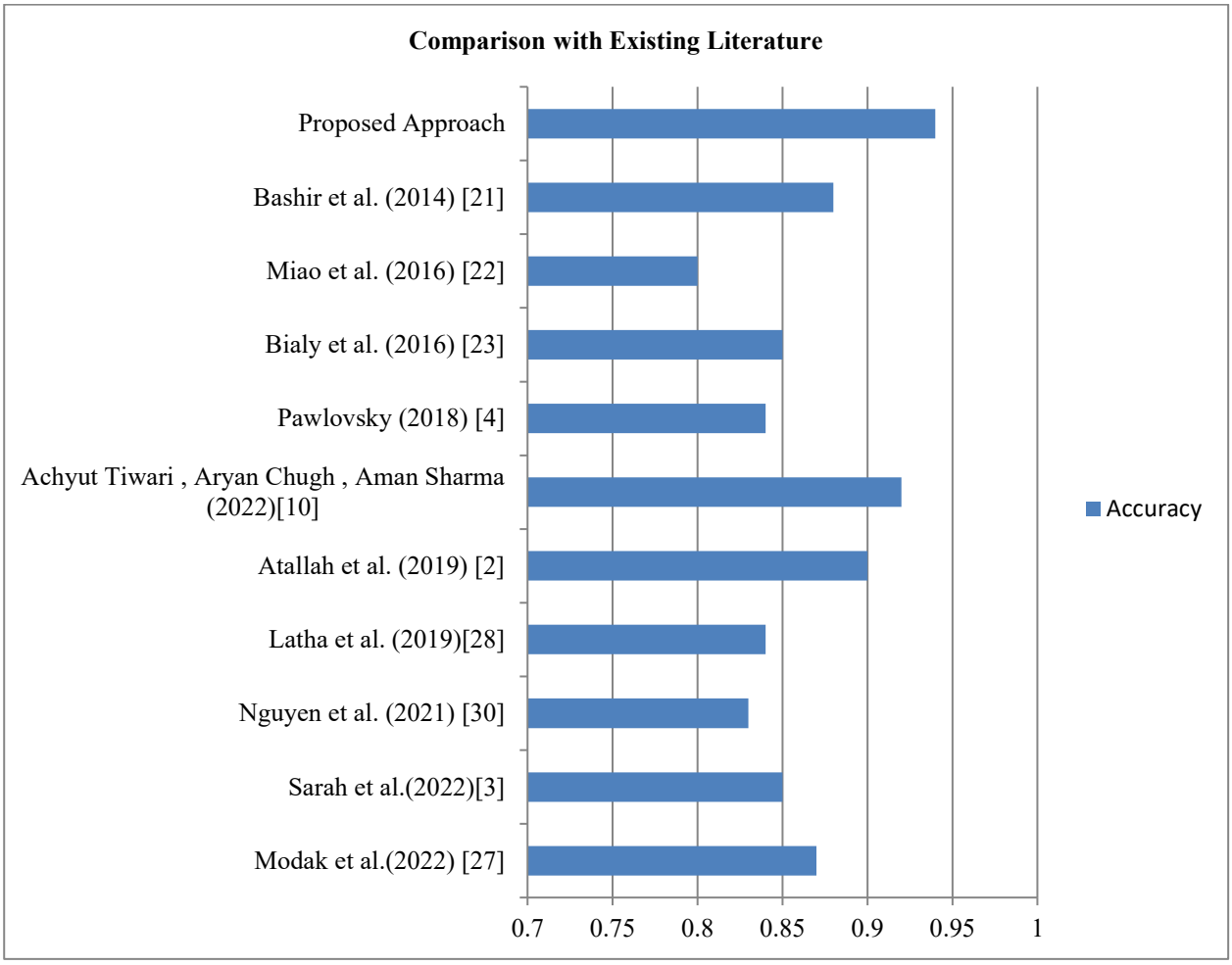

Fig. 19: Comparison of accuracy of proposed approach with existing literature

The suggested method's remarkable 94% accuracy shows that it constitutes a major advancement in machine learning research. This outperforms current methods documented in the literature, which typically achieve accuracy between 80% and 92%. The suggested method establishes a new benchmark for excellence in the industry and exhibits unmatched predictive ability by using a complex ensemble of machine learning algorithms.

VI. CONCLUSION AND FUTURE WORK

As a result, this research shows that the suggested strategy outperforms current machine learning techniques in terms of precision, recall, F1-score, precision-recall AUC, and prediction accuracy. The suggested framework performs very well at correctly recognizing affirmative instances while retaining a high recall rate. It achieves 94% accuracy and a precision recall AUC of 0.95.The ensemble technique, which integrates multiple models to enhance prediction capabilities, is responsible for this approach's efficacy.

In the context of recent pandemics such as COVID-19, nations such as India, where the population is rapidly expanding and healthcare resources are limited, are in dire need of improved healthcare solutions. The proposed strategy has the potential to fulfill this requirement and make a significant contribution to the healthcare industry by facilitating the early identification of conditions such as heart failure.

Future studies might focus on examining how well our method scales and applies to larger datasets using deep learning principles. This could increase the framework's efficacy even more and broaden its range of practical applications.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 01, January-2025

Conflict of Interest

The authors confirm that there is no conflict of interest to declare for this publication.

Data Availability Statement

The port of IEEE Data is used to gather data on heart failure (CVD) in this research work, available in https://pair-code.github.io/facets/ [13] and https://dx.doi.org/10.21227/dz4t-cm36 [26].

REFERENCES

[1] F. Babic and Z. Vantová, "Predictive and descriptive analysis for heart disease diagnosis", In Proceeding of Federated Conference on Computer Science and Information Systems (CSIS), pp. 155–163, 2017. DOI: 10.15439/2017F219

[2] R. Atallah and A. Al-Mousa, "Heart Disease Detection Using Machine Learning Majority Voting Ensemble Method",2nd International Conference on new Trends in Computing Sciences (ICTCS), pp. 1-6. 2019. DOI: 10.1109/ICTCS.2019.8923053

[3] S. Sarah, M. K. Gourisaria, S. Khare and H. Das, "Heart Disease Prediction Using Core Machine Learning Techniques—A Comparative Study", In: Tiwari S., Trivedi M.C., Kolhe M.L., Mishra K., Singh B.K. (eds) Advances in Data and Information Sciences. Lecture Notes in Networks and Systems, vol 318. Springer, Singapore, 2022. DOI:10.1007/978- 981-16-5689-7_22

[4] A. P. Pawlovsky, "An ensemble based on distances for a KNN method for heart disease diagnosis", International Conference on Electronics, Information, and Communication (ICEIC), Honolulu, HI, USA, pp. 1-4, 2018. DOI:10.23919/elinfocom.2018.8330570

[5] R. Kohavi, (1996). Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. Kdd. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, pp. 202 – 207, 1996. https://dl.acm.org/doi/10.5555/3001460.3001502

[6] 1.4. Support Vector Machines. Scikit. (n.d.). Retrieved March 11, 2022. https://scikit-learn.org/stable/modules/svm.html

[7] A. R. Gregory, A. M. George and F. Valentin, "The Global Burden of Cardiovascular Diseases and Risks", Journal of the American College of Cardiology, vol. 76, no. 25, pp. 2980-2981, 2020. https://www.sciencedirect.com/science/article/pii/S0735109720377755

[8] K. Raza (2019), "Improving the prediction accuracy of heart disease with ensemble learning and majority voting rule", U-Healthcare Monitoring System, vol 1, pp. 179-196, 2019. https://www.sciencedirect.com/science/article/pii/B9780128153703000086?via%3Dihub

[9] W. K. Mutlag, S. K. Ali, Z. M. Aydam and Bahaa H. Taher, "Feature Extraction Methods: A Review" Journal of Physics: Conference Series, vol. 1591, 2020. **DOI**: 10.1088/1742-6596/1591/1/012028

[10] A. Tiwari, A. Chugh and A. Sharma, "Ensemble Framework for Cardiovascular Disease Prediction", Journal of Biomedical Informatics, vol. 146, no. C, July 2022 https://doi.org/10.1016/j.compbiomed.2022.105624

[11] www.who.int, "Cardiovascular diseases (CVDS)", [online] June 2021, Available: https://www.who.int/en/news-room/factsheets/detail/cardiovascular-diseases-(cvds)

[12] R. Katarya, & S. K. Meena, "Machine learning techniques for heart disease prediction: A comparative study and analysis", Health Technology, vol.11, pp.87–97, 2021. https://link.springer.com/article/10.1007%2Fs12553-020-00505-7

[13] People + AI Research Initiative, Google. Facets - know your data. Facets – "Visualizations for ML datasets", [online] September 2021, Available: https://pair-code.github.io/facets/

[14] M. LaValley, "Logistic Regression Literature Review", Circulation, vol.117(18), pp.2395–2399, 2008. https://doi.org/10.1161/CIRCULATIONAHA.106.682658

[15] A. S. Kumar and N., Sinha, "Cardiovascular disease in India: A 360degree overview" Med J Armed Forces India, vol.76, no.1, pp. 1–3, Jan 2020. DOI: 10.1016/j.mjafi.2019.12.005

[16] D. Anguita, L. Ghelardoni, A. Ghio, L. Oneto and S. Ridella, "The 'K' in K-fold Cross Validation," European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), pp. 441-446, April 2012. http://www.i6doc.com/en/livre/?GCOI=28001100967420

[17] R. M. Terol, A. R. Reina, S. Ziaei and D. Gil, "A Machine Learning Approach to Reduce Dimensional Space in Large Datasets," IEEE Access , 8, 134658-134675, 2020. DOI: 10.1109/ACCESS.2020.3012836

[18] Ministry of Health and Family Welfare Government of India, "Health and Family Welfare Statistics in India," 2019-20. https://main.mohfw.gov.in/sites/default/files/HealthandFamilyWelfarestatisticsinIndia20 1920.pdf.

[19] R. EI. Bialy, R. M. A. Salama and O. Karam, "An ensemble model for Heart Disease Data Sets," Proceedings of the 10th International Conference on Informatics and Systems – (INFOS), 2016. https://doi.org/10.1145/2908446.2908482

[20] J. Rogers and S. Gunn, "Identifying feature relevance using a random forest," In International Statistical and Optimization Perspectives Workshop Subspace, Latent Structure and Feature Selection, pp. 173-184, Springer, Berlin, Heidelberg, February 2005.

[21] S. Bashir, U. Qamar, F. H. Khan, and M. Y. Javed, "MV5: A clinical decision support framework for heart disease prediction using majority vote based classifier ensemble," Arabian Journal for Science and Engineering, vol. 39, no. 11, 7771–7783, 2014. https://doi.org/10.1007/s13369-014-1315-0

[22] K. H. Miao, J. H. Miao and J. George , " Diagnosing coronary heart disease using ensemble machine learning," International Journal of Advanced Computer Science and Applications, vol.7 no.10, 2016 . https://doi.org/ 10.14569/IJACSA.2016.071004

[23] M. L. Zhang and Z. H. Zhou, "A k-nearest neighbor based algorithm for multilevel classification," IEEE International Conference on Granular Computing, vol. 2, pp. 718-721, 2005. DOI: 10.1109/GRC.2005.1547385

[24] R. Kumar, R. Pal, "India achieves WHO recommended doctor population ratio: A call for paradigm shift in public health discourse," J Family Med Prim Care, vol.7, no.5, pp.841-844 2018. DOI:10.4103/jfmpc.jfmpc_218_18

[25] V. Shorewala, "Early detection of coronary heart disease using ensemble techniques," Informatics in Medicine Unlocked 26(6):100655, vol. 26, no.6, June 2021. DOI: 10.1016/j.imu.2021.100655

[26] M. Siddhartha, "Heart Disease Dataset (Comprehensive)," IEEE Data port, November, 2020. DOI: https://dx.doi.org/10.21227/dz4t-cm36

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 01, January-2025

[27] S. Modak, E. A. Raheem and L. Rueda, "Heart Disease Prediction Using Adaptive Infinite Feature Selection and Deep Neural Networks," International Conference on Artificial Intelligence in Information and Communication (ICAIIC), pp. 235-240, 2022 DOI: 10.1109/ICAIIC54071.2022.9722652

[28] C. B. C. Latha, and S. C. Jeeva, "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," Informatics in Medicine Unlocked, 16, 100203, 2019. https://doi.org/10.1016/j.imu.2019.100203

[29] Y. Muhammad, M. Tahir, M. Hayat, and K. T. Chong, (2020, November 12). "Early and accurate detection and diagnosis of heart disease using intelligent computational model," Nature, November 2020 https://www.nature.com/articles/s41598-020- 76635-9

[30] K. Nguyen, J Lim, K. P. Lee, T. Lin, J. Tian, T. T. Trang, "Heart Disease Classification using Novel Heterogeneous Ensemble," IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), pp. 1-4, 2021. DOI: 10.1109/BHI50953.2021.9508516

[31] K. Yuan, L. Yang, Y. Huang and Z. Li, "Heart Disease Prediction Algorithm Based on Ensemble Learning," 7th International Conference on Dependable Systems and Their Applications (DSA), pp. 293-298, 2020. DOI: 10.1109/DSA51864.2020.00052

[32] R. Muthukrishnan and R. Rohini, "LASSO: A feature selection technique in predictive modeling for machine learning," IEEE International Conference on Advances in Computer Applications (ICACA), Coimbatore, India, pp. 18-20, 2016. DOI: 10.1109/ICACA.2016.7887916

[33] R. J. Urbanowicz, M. Meeker, W. L. Cava, R. S. Olson, J. H. Moore, "Relief-based feature selection: Introduction and review," Journal of Biomedical Informatics vol. 85, pp. 189-203 September 2018.

[34] B. P. Salmon, W. Kleynhans, C. P. Schwegmann and J. C. Olivier, "Proper comparison among methods using a confusion matrix," IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Milan, Italy, 2015, pp. 3057-3060, 2015. DOI: 10.1109/IGARSS.2015.7326461

[35] A. M. Elsayad, M. Al-Dhaifallah and A. M. Nassef, "Analysis and Diagnosis of Erythemato-Squamous Diseases Using CHAID Decision Trees," 15th International Multi-Conference on Systems, Signals & Devices (SSD), Yasmine Hammamet, Tunisia, 2018, pp. 252-262, 2018. DOI: 10.1109/SSD.2018.8570553

[36] D. Krishnani, A. Kumari, A. Dewangan, A. Singh and N. S. Naik, "Prediction of Coronary Heart Disease using Supervised Machine Learning Algorithms," TENCON IEEE Region 10 Conference (TENCON), Kochi, India, pp. 367-372, 2019. DOI: 10.1109/TENCON.2019.8929434

[37] L. Ji, Y. Gu, K. Sun, J. Yang and Y. Qiao, "Congenital heart disease (CHD) discrimination in fetal echocardiogram based on 3D feature fusion," IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 2016, pp. 3419-3423, 2016. DOI: 10.1109/ICIP.2016.7532994

[38] J. K. Jaiswal and R. Samikannu, "Application of Random Forest Algorithm on Feature Subset Selection and Classification and Regression," World Congress on Computing and Communication Technologies (WCCCT), Tiruchirappalli, India, pp. 65-68, 2017. DOI: 10.1109/WCCCT.2016.25

[39] K. Pal and B. V. Patel, "Data Classification with k-fold Cross Validation and Holdout Accuracy Estimation Methods with 5 Different Machine Learning Techniques," Fourth International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, pp. 83-87, 2020. DOI: 10.1109/ICCMC48092.2020.ICCMC-00016