

AN ENSEMBLE-BASED MACHINE LEARNING FRAMEWORK FOR EARLY PREDICTION OF HEART

¹ Vandana Agarawal, ²Dr. Virendra Kumar Swarnkar (Associate Professor)

¹Research Scholar, ²Supervisor

¹⁻² Department of Computer Science and Engineering, Bharti Vishwavidyalaya, Durg, Chhattisgarh

Abstract : Heart disease continues to be a leading cause of death worldwide. Early and reliable diagnosis using traditional medical procedures is frequently difficult due to the complexity and diversity of clinical symptoms. This paper offers an ensemble-based machine learning framework for cardiac disease prediction in order to address this problem. The suggested method improves diagnostic robustness and accuracy by combining several classifiers. The model is trained and validated using clinical characteristics like age, blood pressure, cholesterol levels, and type of chest discomfort. According to experimental results, the ensemble model performs better than individual classifiers in terms of precision, sensitivity, and total prediction accuracy. As a result, it provides healthcare professionals with an effective and data-driven tool to help with early detection and treatment planning. These classifiers include decision trees, support vector machines (SVM), and ensemble approaches such as random forest and gradient boosting. A number of feature selection approaches and optimisation tactics are utilised by the hybrid model in order to enhance the performance of the model. For the purpose of ensuring robustness and minimising overfitting, cross-validation is utilised. It has been demonstrated through the results of the experiments that the hybrid machine learning technique achieves superior levels of accuracy and precision when compared to individual classifiers. As a consequence, it provides a trustworthy instrument that may assist physicians in the early and precise detection of heart disease.

Keywords: Medical Data Classification, Healthcare Analytics, Data-Driven Decision Making, Diagnostic Accuracy, Hybrid Model, Clinical Data Analysis,

1. Introduction

According to the records of the World Health Organization (WHO), cardiovascular diseases are responsible for 31% of all deaths that occur throughout the globe, resulting in the deaths of around 17.9 million people annually. It is thus essential to recognize cardiac sickness in a timely and accurate manner in order to reduce the number of fatalities and improve the outcomes for patients. Clinical judgment, human interpretation of diagnostic methods, and laboratory data have all been essential components in the process of diagnosing cardiac diseases over the course of medical history. The use of such procedures often results in issues such as time, human error, and the amount of competence possessed by the doctor. Machine learning (ML) is one example of cutting-edge technology that has been incorporated into the healthcare industry in recent years, which has resulted in the creation of new potential for boosting diagnostic efficiency and accuracy. The intricacy and multidimensional character of cardiac illness has led to the development of hybrid strategies that integrate machine learning with traditional methods or combine a variety of machine learning algorithms. These techniques have shown impressive promise.

It is possible for machine learning, which is a kind of artificial intelligence, to sift through mounds of medical data, discover patterns that were not previously recognized, and produce highly precise forecasts. In order to detect potential risk factors for cardiovascular disease, it is feasible to use machine learning techniques to evaluate data from patients. This data may include the patients' ages, cholesterol levels, blood pressure, electrocardiogram (ECG) readings, and lifestyle variables. Single-algorithm models, on the other hand, may not be up to the task when it comes to dealing with noisy data from the real world or capturing the nuances of different datasets. In this setting, hybrid models become a significant consideration. The purpose of hybrid systems is to mix a variety of machine learning techniques in order to achieve maximum performance. In addition to neural networks, decision trees, support vector machines,

k-nearest neighbors, random forests, and other algorithms, these methods also contain random forests. It is possible, for instance, that some models are better able to recognize non-linear patterns, whilst other models may be more adept at recognizing linear connections. Using a hybrid method results in higher prediction performance and more generalizability across patient profiles. This, in turn, leads to the development of diagnostic tools that are more reliable. The figure1, represent the respective model.

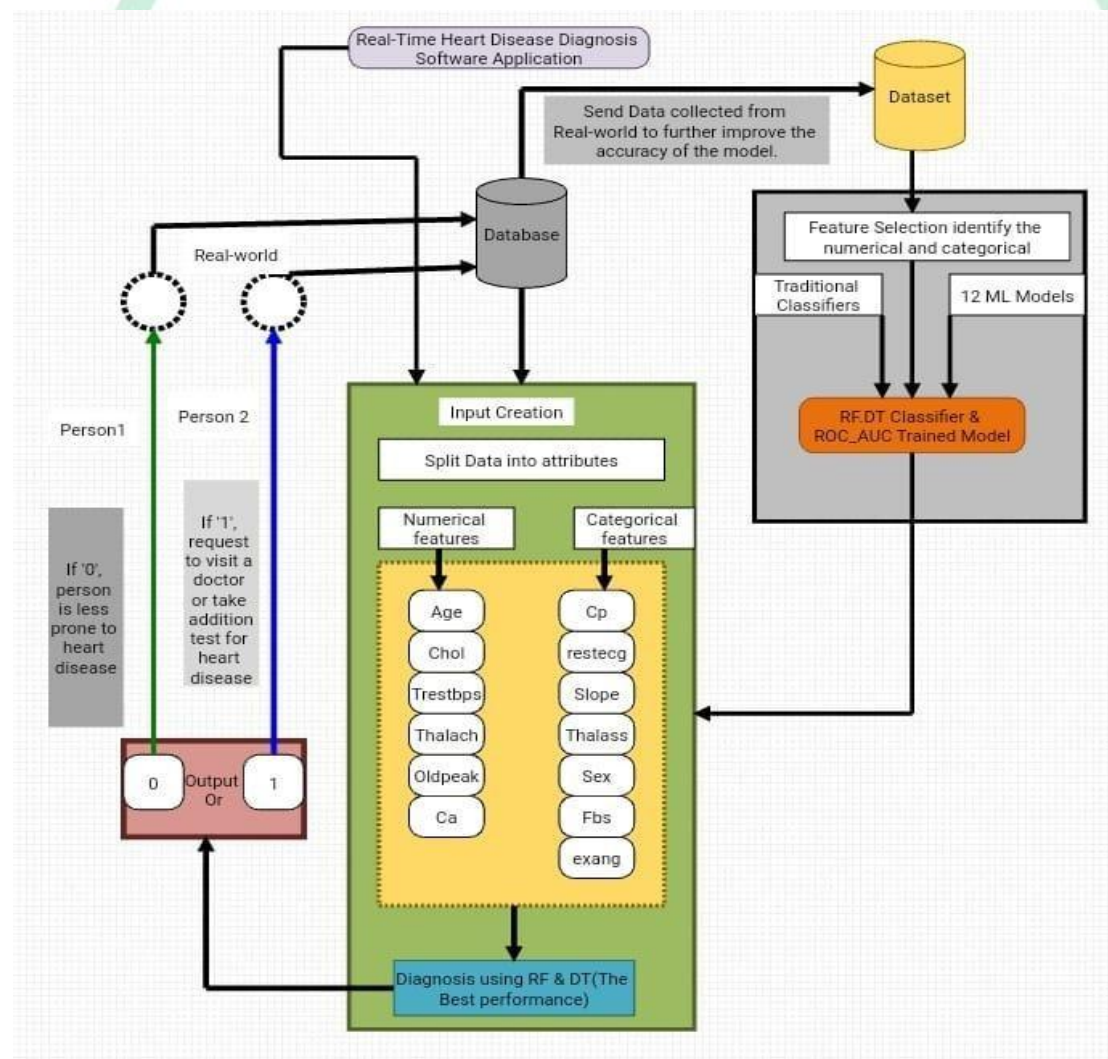


Figure 1: Application of Proposed Model.

2. Lecture Survey

Negi, Kumar, and Mishra (2016)[1], used linear discriminant analysis (LDA) to present a work on feature extraction and classification for EMG signals at the 2nd International Conference on Advances in Computing, Communication, & Automation (ICACCA).

Nikhil and Karandikar's (2016) [2], study, which was published in the International Journal of Advanced Engineering, Management, and Science, examined the use of machine learning algorithms to predict cardiac disease. They evaluated several algorithms, such as Decision Trees, Naïve Bayes, and Neural Networks, and found that hybrid and ensemble approaches produce higher predictive accuracy.

Nissa, Jamwal, and Mohammad (2021)[3], looked at machine learning methods for predicting cardiac disease.

Several algorithms, including logistic regression, random forest, and SVM, are evaluated in their work, demonstrating that ensemble approaches perform better than individual models. The study highlights the increasing significance of AI in enhancing cardiovascular risk assessment and bolstering preventative healthcare tactics.

Olufsen (1999) [4], presented a structured tree outflow condition model for modeling blood flow in major systemic arteries. The model advances knowledge of hemodynamics and circulatory system modeling by offering a mathematical framework for characterizing vascular dynamics. Numerous other works on cardiovascular modeling and simulation have been influenced by this basic work.

Ornish et al. (1990) [5], to investigate the possibility of reversing coronary heart disease with lifestyle changes. Their findings showed that dietary modifications, physical activity, and stress reduction could significantly reduce atherosclerosis without the need for medication or surgery. The significant influence of behavioral therapies on cardiovascular health was demonstrated by this ground-breaking study.

Patel and Rana (2014)[6], and published in the International Journal of Engineering Development and Research. Their research offers comparative insights between ID3, C4.5, and CART while methodically reviewing the mechanics, benefits, and drawbacks of decision tree approaches. The poll emphasizes the ease of use, transparency, and robust applicability of decision trees in both academic and industrial contexts.

Patel, Upadhyay, and Patel (2015)[7], investigated the use of machine learning and data mining approaches for the prediction of heart disease. Their research shows that hybrid models increase accuracy by using classification methods like Naïve Bayes, Decision Trees, and Neural Networks to clinical information. The study emphasizes how data mining may turn unprocessed health data into useful insights for medical decision-making.

Quah et al. (2014) [8], and published in BioMed Research International. The findings highlight the need for more education and awareness initiatives by showing differing levels of public understanding. The results of the study support public health planning for prompt medical intervention and early symptom detection.

The Lancet by Reddy et al. (2011)[9], on behalf of the Lancet India Group for Universal Healthcare. The writers argued for equity, accessibility, and affordability in medical services while outlining strategic changes and issues facing India's healthcare sector. Their work continues to have a significant impact on national policy debates about healthcare delivery and infrastructure.

Reddy, Palagi, and Jaya (2017)[10], presented a model for predicting cardiac disease using artificial neural networks (ANNs). Their research shows how ANN-based models may identify intricate patterns in patient data, outperforming traditional methods in terms of diagnostic accuracy. The outcomes confirm that ANNs are an effective early detection tool.

Repaka et al. (2019[11],) employ the Naive Bayes algorithm, a probabilistic classifier based on Bayes' theorem with strong independence assumptions between features, to construct a predictive model for heart disease. The system was put into place by the authors to categorize patient data and forecast the risk of heart disease. A structured dataset including clinical factors like blood pressure, age, heart rate, and cholesterol was used to train the model. Naive Bayes is a great option for big datasets with continuous and categorical variables because of its ease of use and effectiveness. When compared to alternative categorization methods like decision trees and support vector machines, experimental evaluation demonstrated competitive accuracy.

Richter (1995)[12], with an emphasis on its physiological mechanisms, underlying causes, and epidemiology. The study emphasizes how changed vascular elasticity, decreased contractility, and alterations in cardiac structure all lead to cardiovascular disease. It describes how cumulative risk factors such hypertension, diabetes, coronary artery disease, and lifestyle variables cause the prevalence of CHF to rise sharply with age. Richter examines compensatory mechanisms such as ventricular remodeling and neurohormonal activation and gives a thorough explanation of the pathophysiology, differentiating between systolic and diastolic dysfunction.

Rich (1997)[13], discusses the multiple elements of congestive heart failure (CHF) in the elderly in this groundbreaking work, focusing on both environmental factors and physiological decline. According to the study, CHF is a major cause of morbidity and mortality in older persons and is frequently brought on by long-term hypertension or ischemic heart disease. Rich talks about how myocardial stiffness and decreased reactivity to β -adrenergic stimulation are two ways that aging impacts heart function.

Robinson, Factor, and Sonnenblick (1986) [14], who focus in particular on how the heart functions as a suction pump.

The article explains how the myocardium's muscular and elastic characteristics enable it to produce negative pressure during diastole, successfully pushing blood into the chambers without depending only on passive filling.

Roy et al. (2015)[15], looked at the use of the Least Absolute Shrinkage and Selection Operator (LASSO) linear regression model for stock market predictive analytics. By carrying out variable selection and regularization concurrently, the study shows how LASSO can effectively handle high-dimensional datasets, preventing overfitting and improving model interpretability. Despite its emphasis on financial forecasts, the paper's methodological significance also applies to healthcare analytics.

Sarangi, Mohanty, and Pattnaik (2015) [16], with the goal of improving the precision of cardiac illness identification. In order to assess clinical parameters and produce diagnostic results, the system combines expert system methodology with data mining techniques. Their study uses machine learning algorithms, such as neural networks and decision trees, to choose features and classify patient data.

In order to compare the effectiveness of Naive Bayes (NB) and Artificial Neural Networks (ANN) algorithms for data categorization tasks, Saritas and Yasar (2019)[17], conducted a study. The authors examined both algorithms' accuracy, computational efficiency, and scalability under various data circumstances using real-world datasets. According to their findings, ANN shows better accuracy when there is enough training data, whereas Naive Bayes delivers simplicity and speed.

Shah, Patel, and Bharti (2020)[18], concentrated on using structured medical information to apply various machine learning approaches to forecast the occurrence of cardiac disease. In order to determine the best classifier for clinical prediction, the study investigates a number of methods, including Random Forests, Decision Trees, Support Vector Machines, and Logistic Regression.

Shaji (2019)[19], introduced a novel method for using sophisticated data mining algorithms to anticipate and diagnose individuals with heart disease. The study extracted significant patterns from extensive patient datasets using classification techniques like Decision Trees, Naive Bayes, and K-Nearest Neighbors. In order to ensure computing efficiency, the methodology concentrated on increasing predictive accuracy through feature selection and dimensionality reduction. Real medical records were used to validate the suggested system, which showed a significant improvement over conventional statistical models.

Singh, Thakur, and Sharma (2016)[20], who focused on the algorithms' theoretical underpinnings, classification techniques, and practical uses. The study divides algorithms into three categories: ensemble models (like Random Forests and Boosting), kernel-based techniques (like SVM), and linear models (like Logistic Regression).

Sinha, Aggarwal, and Das (2007)[21], The study focuses on digital phonocardiograms as input data for non-invasive detection methods. By capturing important auditory characteristics including amplitude, frequency, and duration, the researchers used a backpropagation neural network to distinguish between normal and pathological cardiac sounds. Their method was successful in identifying minute changes that were suggestive of a malfunctioning valve.

A thorough textbook on mathematical strategies for carrying out linear regression analysis was written by Späth (2014)[22]. The book explores the computational and theoretical underpinnings of regression modeling, offering detailed algorithms for ridge regression, ordinary least squares, and other associated methods. Späth places a strong emphasis on model optimization for big datasets, computational efficiency, and numerical stability.

According to Spencer et al. (2013)[23], the American Society of Echocardiography has issued formal guidelines regarding the use of Focused Cardiac Ultrasound (FCU) as a quick bedside imaging method. Standardized guidelines for FCU integration into clinical processes, training, image acquisition, and interpretation are described in the paper. It sets FCU apart from comprehensive echocardiography by highlighting its function in the rapid assessment of pericardial effusion, volume status, and heart function.

A large-scale population-based investigation by Sturgeon et al. (2019) [24], examined the connection between cancer and mortality from cardiovascular disease (CVD) among individuals in the United States. The study, which used data from national health registries, discovered that cancer survivors had a markedly increased risk of dying from cardiovascular disease (CVD), particularly if they have received cardiotoxic chemotherapy or radiation. Age is determined by analyzing various cancer types and demographic groups.

Sun and Huang (2010) [25], suggested an adaptive modification of the conventional K-Nearest Neighbor (KNN) algorithm. Their adaptive method ensures more dependable neighborhood selection by dynamically modifying the

value of k based on local data density and feature distribution. In order to address uncertainty and overlapping class boundaries, which frequently impair standard KNN performance, the model integrates fuzzy logic principles. The adaptive KNN performed better than static models in terms of accuracy and robustness, according to experimental results on benchmark datasets. The authors propose possible uses in medical diagnosis, especially in conditions with heterogeneous data, like heart disease, where classification is made more difficult by symptom overlap.

Tan et al. (2009) [26], A hybrid evolutionary algorithm was presented by for attribute selection in data mining applications. In order to improve classification accuracy and computational efficiency, their method optimizes the selection of pertinent attributes by combining genetic algorithms with local search techniques. Compared to conventional techniques, the algorithm achieves better convergence and less overfitting by combining exploitative refinement and evolutionary exploration.

Vonderer et al. (2014)[27], A thorough analysis of physiological alterations during the shift from intrauterine to extrauterine life was carried out by Important metrics like heart rate, oxygen saturation, and respiratory effort were measured in their study as soon as the baby was born. They offered insights into how newborns adjust to breathing and circulation on their own by using sophisticated neonatal monitoring equipment.

Saal and Hughes (2005)[28], demonstrated the shortcomings of current risk assessment approaches. They gathered data showing BPA's endocrine-disruptive effects at exposure levels much below previously set safety thresholds. Their results highlight the need for updated safety regulations and cast doubt on accepted toxicological hypotheses.

Venkatalakshmi and Shivsankar (2014)[29], investigated predictive data mining methods. To find important risk indicators and improve diagnosis accuracy, their approach combines a number of data mining technologies, such as decision trees and neural networks. The study shows how predictive analytics might reveal hidden relationships between symptoms and clinical indicators by examining patient information.

Virani et al. (2020) [30], offered a thorough data update on heart disease and stroke. The article provides an overview of risk factors, preventative health practices, and national and international trends in cardiovascular morbidity and death. The research illustrates the increasing global burden of cardiovascular illnesses and demographic inequities using extensive epidemiological data.

Voleti and Reddi (2016)[31], suggested an optimal framework for disease prediction. To improve prediction accuracy and scalability, their methodology uses hybrid algorithms that integrate clustering and classification techniques. The study reveals notable gains in recognizing disease patterns after testing several models using medical datasets. The framework's goals are to speed up computation, eliminate superfluous characteristics, and offer trustworthy diagnostic insights.

Wu, Huynh, and Souvenir (2014)[32], presented a motion factorization method for classifying echocardiograms at the IEEE International Symposium on Biomedical Imaging. To more accurately describe heart dynamics, their method breaks down cardiac motion into discrete temporal and spatial components. They increased echocardiographic sequence classification performance by using machine learning-based feature extraction.

Xing, Wang, and Zhao (2007)[33], used recently obtained medical records in conjunction with data mining techniques to forecast the course of coronary heart disease. Their research, which was presented at the IEEE International Conference on Convergence Information Technology, combines logistic regression, neural networks, and decision tree methods to improve prediction performance. The hybrid model increases classification accuracy while handling noisy data well.

3. Research Methodology

The primary emphasis of our research agenda was on developing issues that are associated with existing technology. With the use of contemporary computer methods, the problem that was associated with heart diseases has been resolved. Through the use of the tactics and methods that were created by early mathematicians, these issues have been found and resolved. Through the use of analytical procedures, such as various machine learning techniques, it is possible to determine the numerical results and the domain in which they have significance. Through the use of computer software that plots the pertinent graphs and tables, the results and conclusions are presented to the interested parties. In the course of our research, we made use of Anaconda 3, which is based on the Python programming code.

3.1. Grid Search (GS)

Grid search is the method that is often used to optimize hyper parameters. This method is also referred to as a parameter rout in certain environments. The process involves doing a comprehensive search throughout a subset of the hyper parameter space of a learning algorithm that has been manually chosen. It is necessary to have some kind of performance indicator in order to direct a grid search algorithm. It is represented on figure 2.

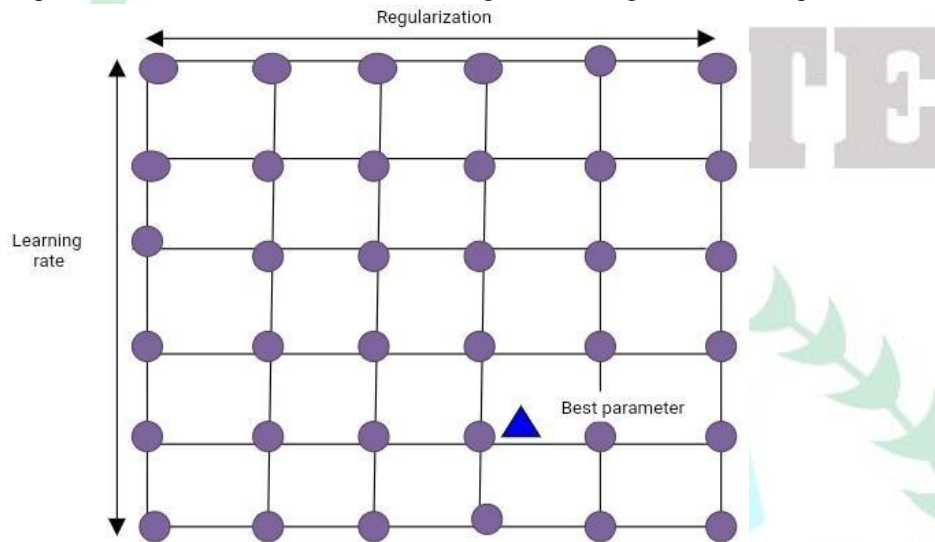


Figure 2: Grid Search Best Parameter Optimization

Learning rate and regularisation are the two hyperparameters that are used for a single family of trained models, and the purple dots indicate a simple grid search using these two parameters. While the left and centre panels represent the first two phases of the search, the right panel provides the final results of the search.

3.2. K-Fold Cross Validation

Cross-validation is a method that involves resampling, and it may be used to identify machine learning models so long as the data sample is relatively small.

This method figure3, has just one parameter, which is denoted by the letter k , and it represents the number of groups that a data sample is to be divided into. In the process of describing this approach, the phrase "k fold cross-validation" is often used. If a particular number is selected for k , that number may be used in lieu of k when referring to the model. For instance, $k=5$ indicates that the model is being tested using a 5-fold cross-validation.



Figure 3: Five -Fold Cross-Validation

In applied machine learning, cross-validation is often used as a method for evaluating the effectiveness of a machine learning model on a dataset that has not been trained. The purpose of this evaluation is to assess the overall performance of the model on data that was not employed during the training process by utilising a small sample size. Its popularity may be attributed to the fact that it is easy to comprehend, and in comparison, to other approaches, such as the conventional train/test split, it often results in a more accurate and less biased assessment of the performance of the model. This is the whole process all laid out:

- Randomly shuffle dataset
- Divide dataset into k groups.
- For distinct groups:
- Use the group as a test or holdout data set.
- Training data may come from the remaining categories.
- Apply a model to training data and evaluate it against testing data.
- Retain assessment results but discard model.

In light of the evaluation rating sample, please provide a concise overview of the performance of the selected model. The fact that every single data point in the sample is given its own unique category and remains in that category throughout the whole process is very important. In other words, this indicates that there is a 1 in the holdout set for each and every sample, and there are k-1 possibilities to use it for the purpose of model training.

3.3. Stacking Method

The stacking technique, which includes predicting a large number of nodes all at once, is one of the most popular ensemble machine learning algorithms. Its purpose is to enhance the performance of models that are already in existence. It is possible to train many models to handle situations that are comparable, and then combine the results of those models to create a new model that is more efficient. The technique in question is known as stacking. One method for putting together models for classification or regression is called stacking, and it makes use of estimators with two layers representing in figure 4.

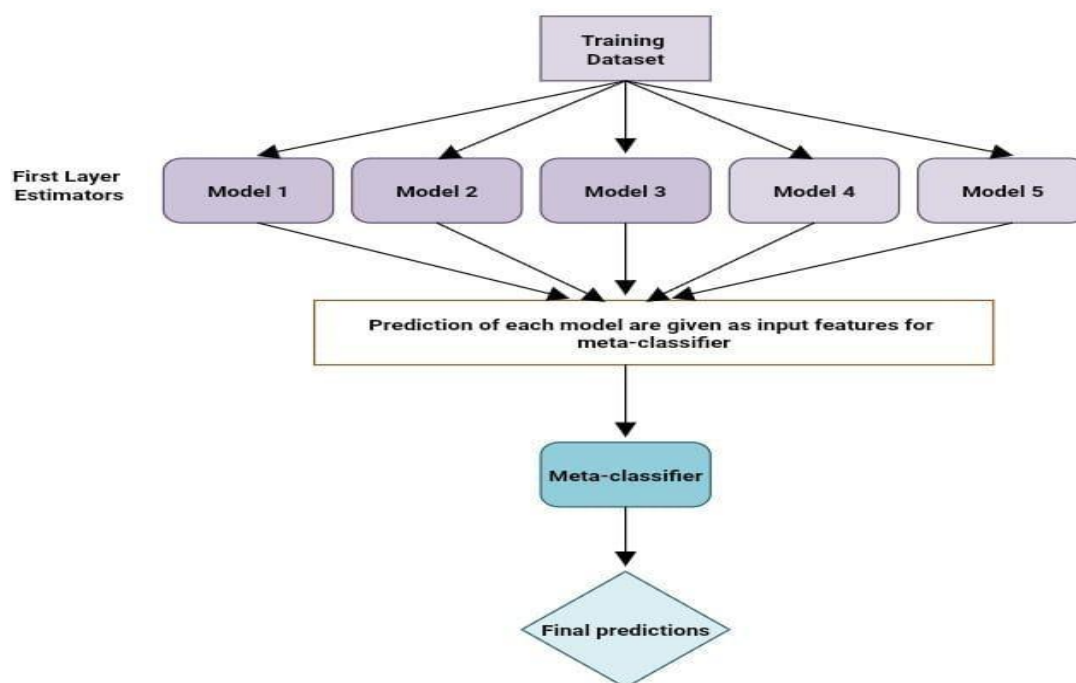


Figure 4: Stacking Meta Classifier Structure

- Original Data: This data is divided into n-folds and is also known as test data or training data.
- Base Models: These models generate predictions (level 0) using the training data and are classified as level-0 models.
- Level-0 Prediction: When applied to a dataset for training, each base model generates a unique set of level-0 predictions.
- Meta Models: The stacking model is structured around a single meta-model, which allows for the most optimal combination of predictions from the base models. One other way to refer to the meta-model is as the level-1 model.
- Level-1 Prediction: In order to complete the training of the meta-model, many base models will make predictions. It is the meta-model that is responsible for making predictions, and it receives the data that was not utilised to train the underlying models of the system. The input and output pairs in the training dataset are formed by the combinations of the anticipated outcomes and these predictions. The meta-model then utilises these pairs to fit itself. Using the information that it has gathered, the meta-model determines how to combine the predictions that were made by its predecessors.

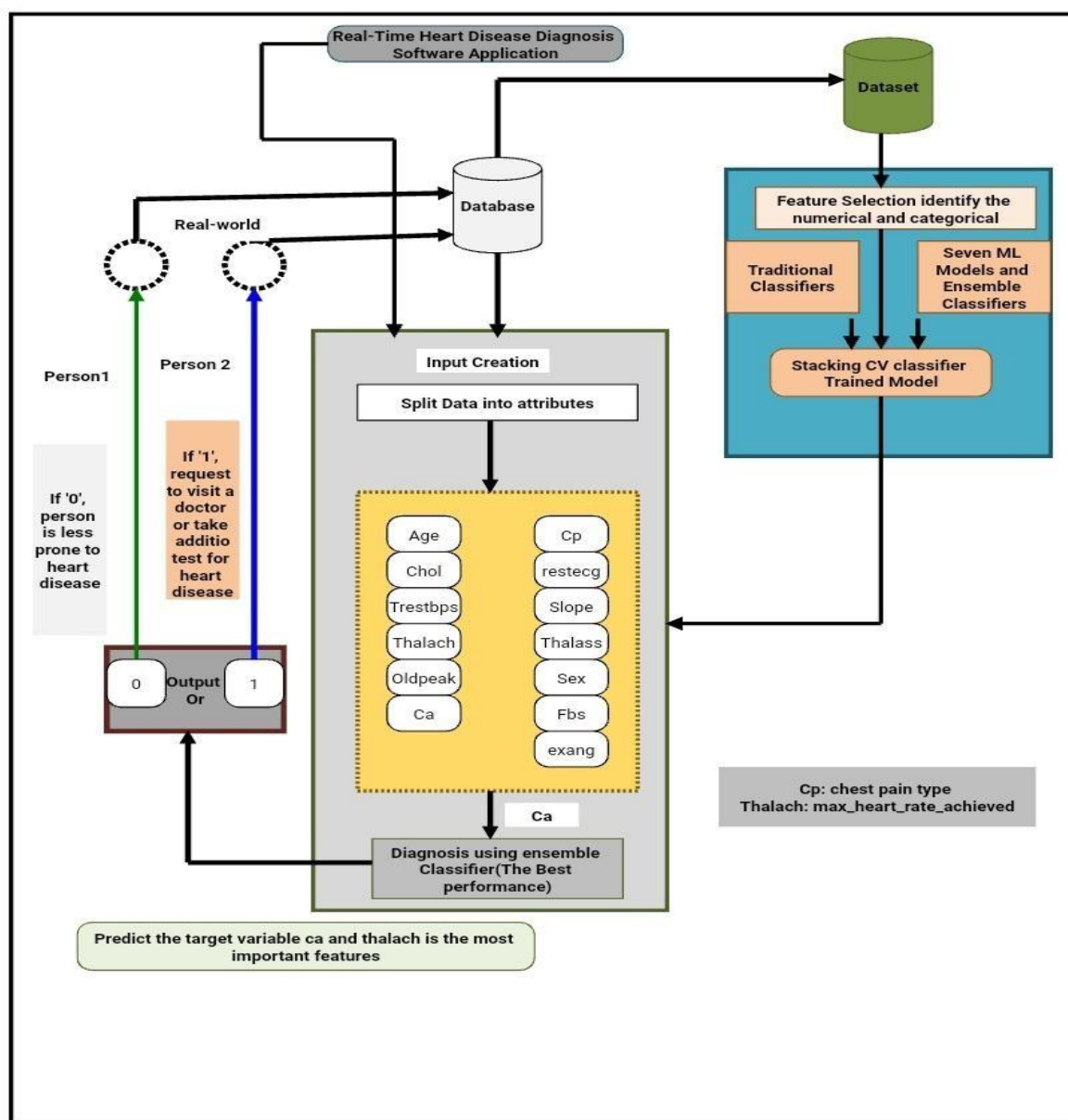


Figure 5: Proposed Model of Suitable Applications

There are a variety of factors that influence the risk of a person having a heart attack. These factors include the individual's age, gender, smoking status, and the role that they play in working. The dataset is used to train many machine learning models, and the results of these models are analysed in order to choose the most appropriate model for HD prediction. The accuracy scores that each machine learning algorithm attained have been presented in this article for the purpose of comparing and evaluating the numerous available approaches. Figure 5. The practice of pre-processing data after it has been collected results in the production of a clean dataset that is devoid of null or duplicate values; this enhances both accuracy and training.

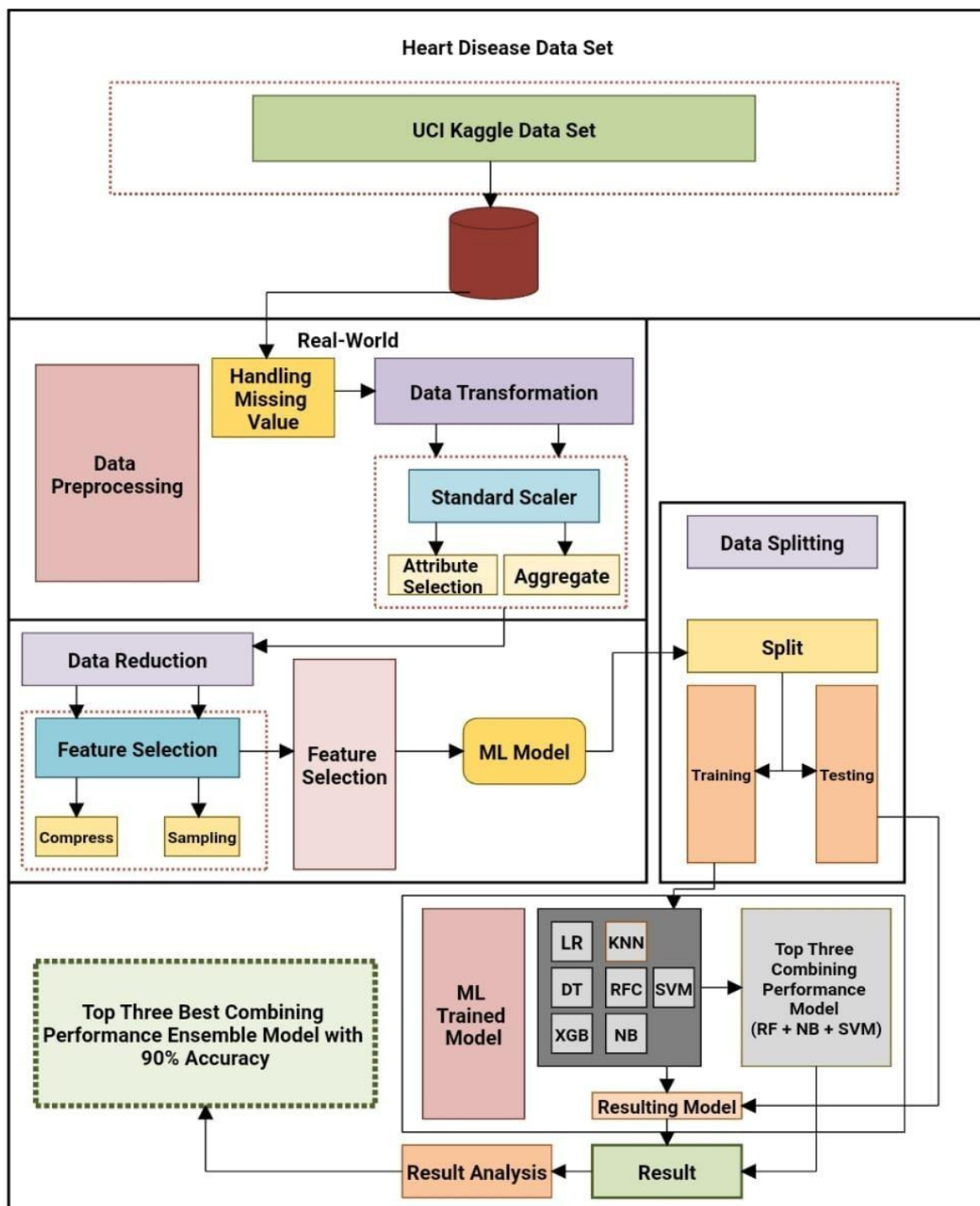


Figure 6 : Proposed Model of Flow Char

Proposed Model: Stacking Algorithm ,highlighting in figure 6.

- Input training data
- Output ensemble classifier E
- Step 1: Learn base classifiers
- For $r = 1$ to do

- LearnHr based on P
- End For
- Step 2: construct new dataset of prediction
- For $i=1$ to m do
- $P_h = x_i \{ , y , x_{i1} \}$ $i = \{h(x_i) \dots \dots h_R(x_i)\}$
- End for
- Learn H based on P_h dataset
- Return H

3.4. LDDS-Net

This section gives the detailed analysis of proposed LDDS-Net and the proposed model performs four major operations such as NLTF based preprocessing, HFM segmentation, LPD based feature extraction, GOA based feature selection, and DLCNN classification. Figure 1 presents the proposed LDDS-Net framework and Table 1 presents the algorithm of proposed LDDS-Net. DLCNN classification. Figure 7. presents the proposed LDDS-Net framework and Table 1 presents the algorithm of proposed LDDS-Net.

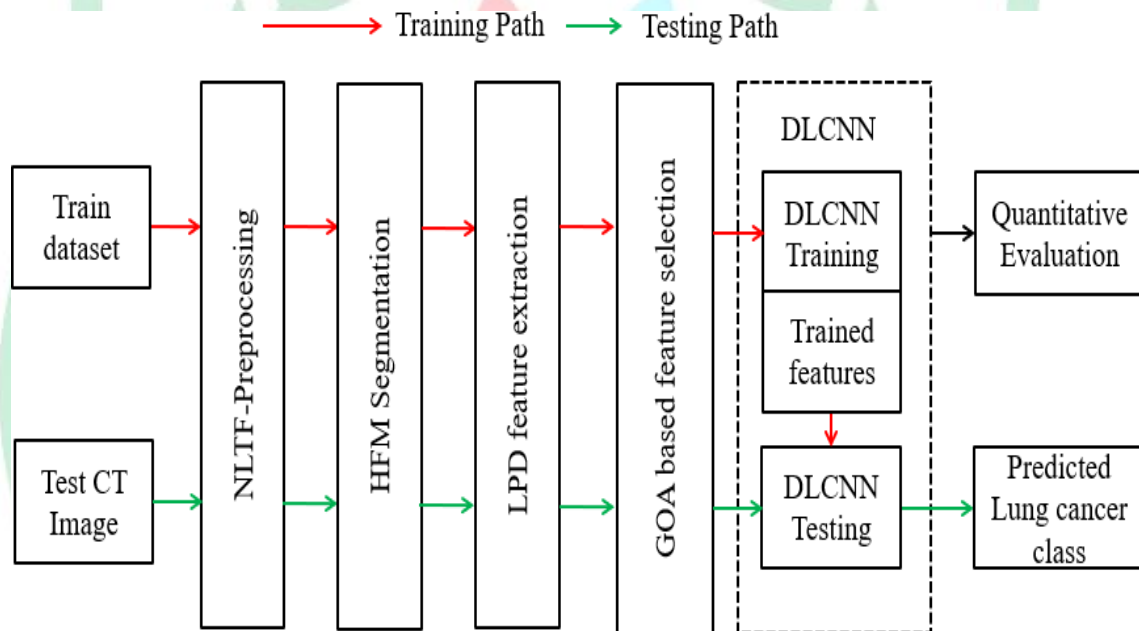


Figure 7: . Proposed LDDS-Net model.

Table 1. Proposed LDDS-Net model.

Input: LIDC-IDRI dataset, test CT image	
Output: Segmented image, classified lung cancer class.	
Performance measures: Preprocessing, segmentation, classification metrics.	
Step 1:	Consider the LIDC-IDRI training dataset and apply it to NLTF preprocessing for noise removal, enhancement of CT images.
Step 2:	Apply HFM segmentation for localization of lung cancer effected region.
Step 3:	Appl LPD on HFM segmented outcomes, which extracts the detailed disease specific features.
Step 4:	Apply GOA based bio-optimization approach for selection of optimal features.

Step 5:	Perform DLCNN based training and extracts the trained features and evaluate the segmentation, classification performance.
Step 6:	Consider the test CT image and repeat the step 1 to step5 for extracting the test features.
Step 7:	Apply the DLCNN model for testing, which identifies the lung cancer diseases through prediction.
Step 8:	Evaluate the preprocessing, segmentation, classification metrics and compare with state of art approaches.

The Table 1, DLCNN (Deep Local Convolutional Neural Network) for image classification in more detail:

1. Input Layer:
 - The input layer represents the input image or data.
 - It has dimensions corresponding to the size of the input image, typically in the form of height, width, and color channels (e.g., RGB).
2. Convolutional Layers:
 - Convolutional layers consist of a set of learnable filters or kernels.
 - Each filter performs convolution operations on the input image to extract different features.
 - The output of each filter is called a feature map, which represents the presence of specific features in the input image.
 - Multiple convolutional layers are typically stacked to capture increasingly complex features.
 - Activation functions (e.g., ReLU) are often applied element-wise to introduce non-linearity.
3. Local Connection Layers:
 - Local connection layers enhance the ability to capture local dependencies within the convolutional layers.
 - They achieve this by connecting each neuron in the convolutional layer to a larger receptive field.
 - The specifics of implementing local connections can vary depending on the DLCNN architecture, as it may involve modifying the connectivity patterns or incorporating skip connections.
4. Pooling Layers:
 - Pooling layers down sample the spatial dimensions of the feature maps generated by the convolutional layers.
 - Common pooling operations include max pooling or average pooling, where a sliding window extracts the maximum or average value within its region.
 - Pooling helps reduce the computational burden and makes the network more robust to variations in spatial location.
5. Fully Connected Layers:
 - After the convolutional and pooling layers, the features are flattened to a one-dimensional vector.
 - This flattened vector is then fed into one or more fully connected layers.
 - Each neuron in a fully connected layer is connected to every neuron in the previous and subsequent layers.
 - The fully connected layers enable the network to learn complex relationships between features.
 - These layers act as a classifier, combining the learned features to make predictions about the input image's class or category.
 - Activation functions, such as ReLU or softmax, are typically applied to the fully connected layers.
6. Output Layer:
 - The output layer is the final layer of the network.
 - It consists of one or more neurons, with each neuron representing a class or category.
 - The activation function applied to the output layer depends on the task.
 - For multi-class classification, a softmax activation function is commonly used to produce a probability distribution over the classes.

- The predicted class is typically determined based on the neuron with the highest probability.

3.5. Optimized ML-Based Heart Disease Diagnosis via Numerical Feature Analysis

HD, or heart disease, is a global health issue that can affect many people. Thus, early heart illness detection may aid treatment. It is crucial to develop technology that speed up diagnosis. Because the number of heart disease patients is rising rapidly. The heart disease system can characterise the most important characteristics of heart patients, identify high-risk patients, and create a model that allows distinguishing between them easy. The number of primary blood arteries, age, cholesterol, resting blood pressure, Stephenson, and other characteristics are utilised to train and compare machine learning algorithms. This research aims to create a basic machine learning model to improve heart disease diagnosis and therapy. Cardiovascular illness can impair heart function. The World Health Organisation estimated that cardiovascular disease killed ten million people. The healthcare industry now struggles to diagnose early sickness after chemical exposure. Due to the large volume of data, medical records and data may be insufficient and inconsistent when applied to real patient situations.

Machine learning (ML) experts from AI and CS work together to make better machines. Learning is the foundation of intelligence, hence machine learning evolved within artificial intelligence. It has industrial, healthcare, and other purposes. Healthcare requires data analysis, therefore ML is useful. Medicine generates tonnes of data. Recently, the digital revolution has enabled unparalleled data storage and collection. Modern hospitals acquire large amounts of data daily through monitoring and other methods. Machine learning (ML) is increasingly utilised in nursing and healthcare to analyse data and identify issues because manually finding and extracting meaningful information from vast amounts of data is difficult. Embolisation, thrombosis Early HD screening predictions require noninvasive machine learning over angiography. HD is predicted using historical data in model-based learning. We want to design the best classification system for anticipating the mentioned situation. Machine learning statistics and unique applications may help choose the optimal machine learning approach. The comparison of HD prediction methods RFC, LRC, SVM, NB, DTC, XGB, and Stacking CV Classifier supports this. After completion, the 303-item dataset was named Cleveland heart disease. The second Erbil heart disease dataset had 333 patients. All research studies utilise this classification to diagnose HD. The UCIKC and Erbil HDDs are notable datasets for this investigation the figure 8.

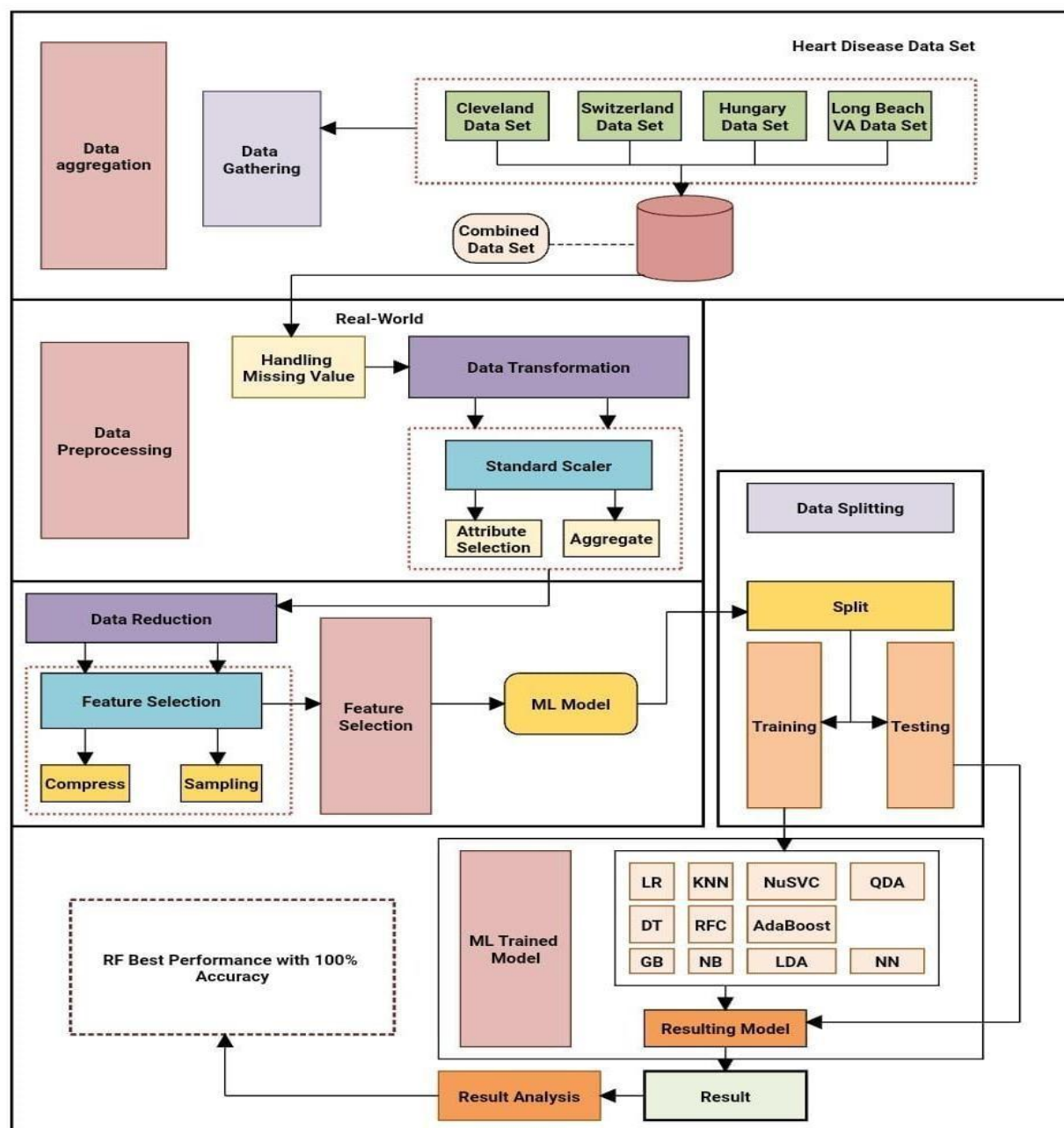


Figure 8: Working Diagram of Proposed Model.

4. Result Analysis

During this experiment, the machine learning algorithms are implemented using the parameters that are set to their default settings. In Table 2, the outcomes of the system are shown. Over the course of the fitting and running phases, it was discovered that the LR model had an accuracy of 84%, a precision of 83%, a recall of 84%, and an F1 score of 84%, respectively. As a consequence of the KNN model, the following findings were obtained: Accuracy of 82%, precision of 81%, recall of 82%, and F1 score of 82%, respectively.

Table 2 : Respective Types I and II and the Confusion Matrix Error of several models using F1-Score Dataset-I, Accuracy, Precision, and Recall

Models	Confusion matrix	I-Type error	II-Type error	Accuracy	Precision	Recall	F1 score
LR	[[21 6][4 30]]	51(correct)	10(incorrect)	84%	83%	83%	84%
KNN	[[20 7][4 30]]	50(correct)	11(incorrect)	82%	81%	81%	82%
DT	[[21 6][6 28]]	49(correct)	12(incorrect)	80%	80%	80%	80%
RFC	[[23 4][4 30]]	53(correct)	8(incorrect)	87%	87%	87%	87%
XGB	[[20 7][4 30]]	50(correct)	11(incorrect)	82%	81%	81%	82%
NB	[[22 5][3 31]]	53(correct)	8(incorrect)	87%	86%	87%	87%
SVM	[[22 5][2 32]]	54(correct)	7(incorrect)	89%	88%	88%	89%

In the order shown in above figure 9, the DT model obtained a score of 80% for accuracy, precision, recall, and F1 score. The accuracy, precision, recall, and F1 score for the RFC model were all determined to be a combined 87% of the total. The XGB model was shown to have an accuracy rate of 82%, a precision rate of 81%, a recall rate of 81%, and an F1 score of 82%, according to the observed results. The results of the analysis revealed that the NB model had an accuracy of 87%, a precision of 86%, a recall of 87%, and an F1 score of 87%, respectively. According to the findings of the SVM model, the accuracy was 89%, the precision was 88%, the recall was 88%, and the F1 score of the model was 89%.

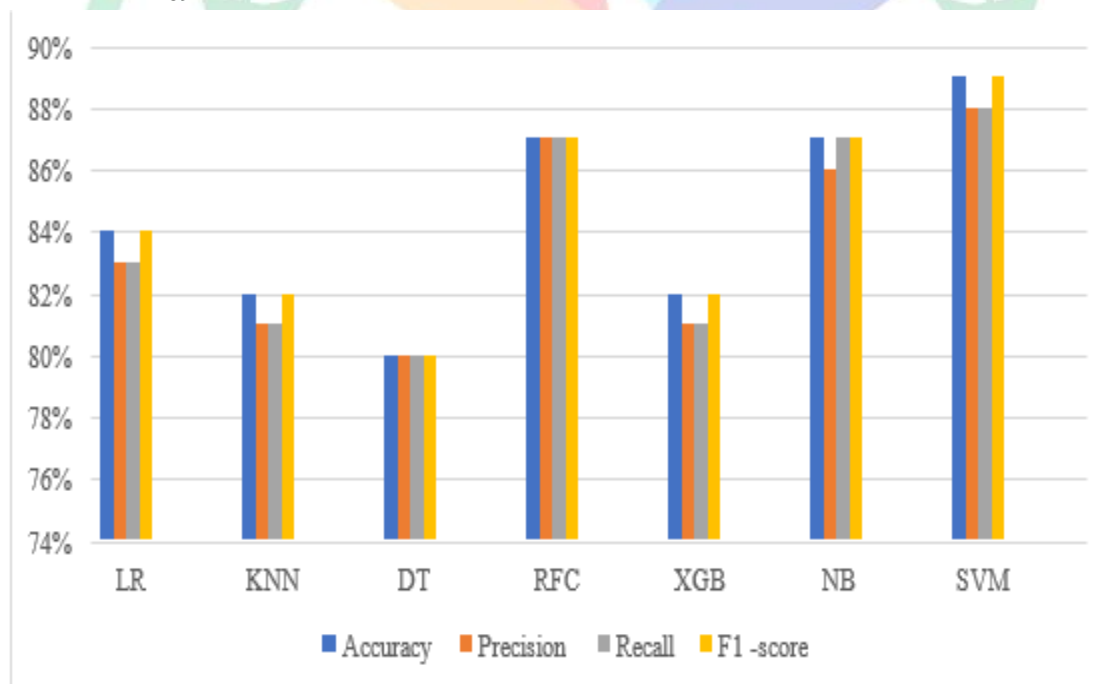


Figure 9: Basic Model of Accuracy, Precision, Recall, F1-score

Table 3 shows system output. The classification report of the LR model showed that during fitting and running, it had an 83% recall rate, 88% accuracy rate, and 86% F1 score for heart disease prediction. Without cardiac disease, the F1 score was 81%, recall 78%, and accuracy 84%. The KNN model classification report on actual testing and

heart disease prediction has 81% accuracy, 88% recall, and 85% F1 score. Data analysis yielded these figures. Additional note: The accuracy, recall, and F1 score for heart disease-free were 83%, 74%, and 78%. The classification report for the DT model showed an 82% accuracy rate for genuine testing and heart disease prediction, 82% recall rate, and 82% F1 score. The results showed no cardiovascular disease with 78% accuracy, 78% recall, and 78% F1 score. The RFC model's classification report showed that it predicted cardiac illness with an 88% recall rate, accuracy rate, and F1 score. Additionally, the F1 score, recall, and precision for heart disease absence were 85%, 85%.

Table 3 : Classification Report (y-test, y-predicted model) for Dataset I

S.No	Models	Predicted value Actual Value	Heart disease (1)	No heart disease (0)	Support	Precision	Recall	f-1 score
1.	LR	Heart disease (1)	30	4	34	83%	88%	86%
		No heart disease (0)	6	21	27	84%	78%	81%
2.	KNN	Heart disease (1)	30	4	34	81%	88%	85%
		No heart disease (0)	7	20	27	83%	74%	78%
3.	DT	Heart disease (1)	28	6	34	82%	82%	82%
		No heart disease (0)	6	21	27	78%	78%	78%
4.	RFC	Heart disease (1)	30	4	34	88%	88%	88%
		No heart disease (0)	4	23	27	85%	85%	85%
5.	XGB	Heart disease (1)	30	4	34	81%	88%	85%
		No heart disease (0)	7	20	27	83%	74%	78%
6.	NB	Heart disease (1)	31	3	34	86%	91%	89%
		No heart disease (0)	5	22	27	88%	81%	85%
7.	SVM	Heart disease (1)	32	2	34	86%	94%	90%
		No heart disease (0)	5	22	27	92%	81%	86%

The models' accuracy and F1 scores are displayed. Since the class distribution is stable, we used accuracy as our categorisation evaluation system. True Positive and True Negative rates are another sign of cardiac disease. Ensemble approaches include building and combining many models. Ensemble approaches often yield better results than individual models. You may find more information on ensemble techniques, although there are others. A stacking cv classifier, which uses multiple classification models, prepared the input data for the first level classifier. Table 5.8 shows system outcomes from the trial. During LR model fitting and running, accuracy, precision, and recall F1 scores were 84%, 83%, and 83%, respectively. The LR model achieved these values. The data showed that the KNN model has 82% accuracy, 82% precision, 81% recall, and 81% F1 score. The DT model's accuracy, precision, recall, and F1 score were 82% in that order. The RFC model scored 85% in accuracy, precision, recall, and F1. The observed findings showed an 82% accuracy rate, 81% precision rate, 81% recall rate, and 82% F1 score for the XGB model. The NB model has 87% accuracy, 87% precision, 86% recall, and 87% beginning score. The SVM model solved the

job with 89% accuracy, 89% precision, 88% recall, and 88% F1 score.

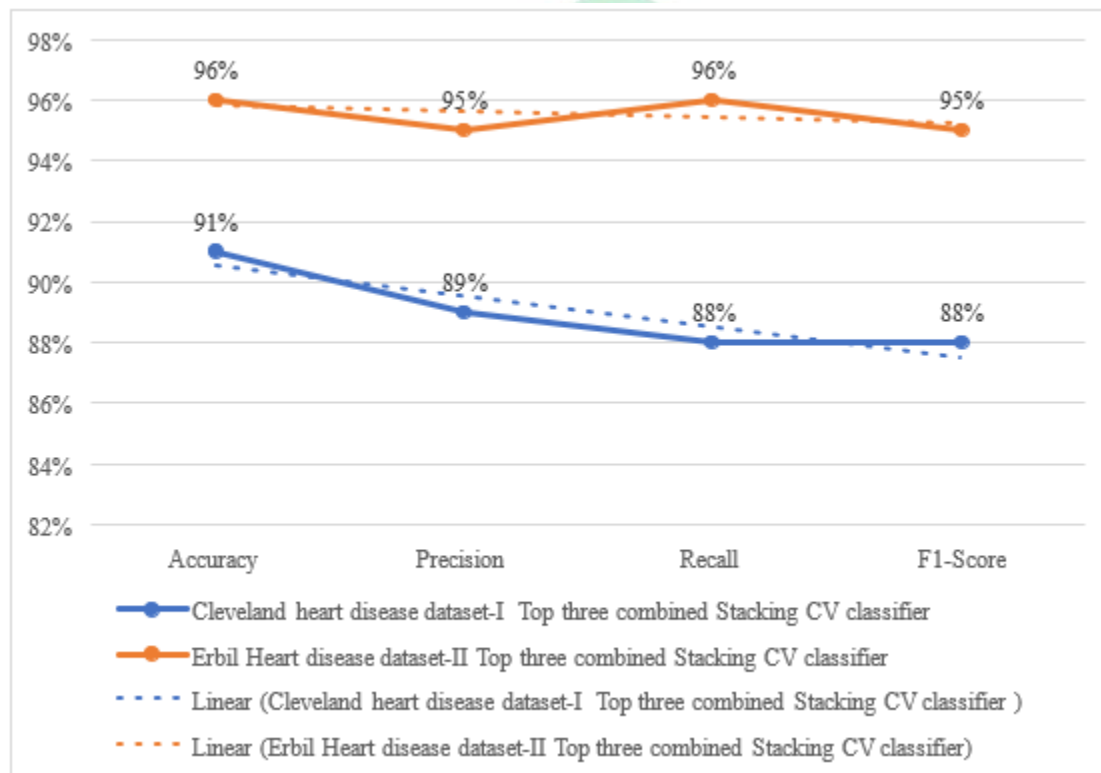


Figure 10: Comparison of System Performance on the Diagnostic Accuracy in Two Datasets

For the purpose of classification figure 10, , feature processing and ordering, and pre-treatment, the datasets pertaining to heart illness in Erbil and Cleveland were processed using the same methodologies. The diagnostic systems were able to obtain a greater degree of accuracy when compared to the analysis of the Erbil heart disease dataset that was carried out during the diagnosis phase. This was accomplished by applying the analyses that were discussed previously to the Cleveland heart disease dataset. The findings of the investigation that compared the performance of machine learning algorithms on the two datasets are shown in Table 4.12 below. During the testing phase, the top three model combined stacking cv classifiers achieved the following results: first, an accuracy of 96%, precision of 95%, recall of 95%, and F1-score of 95% on the Erbil heart disease dataset; second, an accuracy of 91%, precision of 89%, recall of 88%, and F1-score of 88% on the Cleveland heart disease dataset; and third, an accuracy of 98%, precision of 89%, recall of 88%, and F1-score of 88% on the Cleveland heart disease network.

4.1 Checking The Distribution Of The Data

When it comes to categorising issues and making forecasts, data distribution is essential. Recently, it was brought to everyone's attention that half of the sample had cases of heart ailment, while the other half contained none of these occurrences. Without the dataset being balanced, it is impossible for overfitting to occur. Using this, the model will be able to differentiate between data patterns that suggest the existence of cardiac disease and those that do not indicate the presence of cardiac disease, as shown in Figure 11.

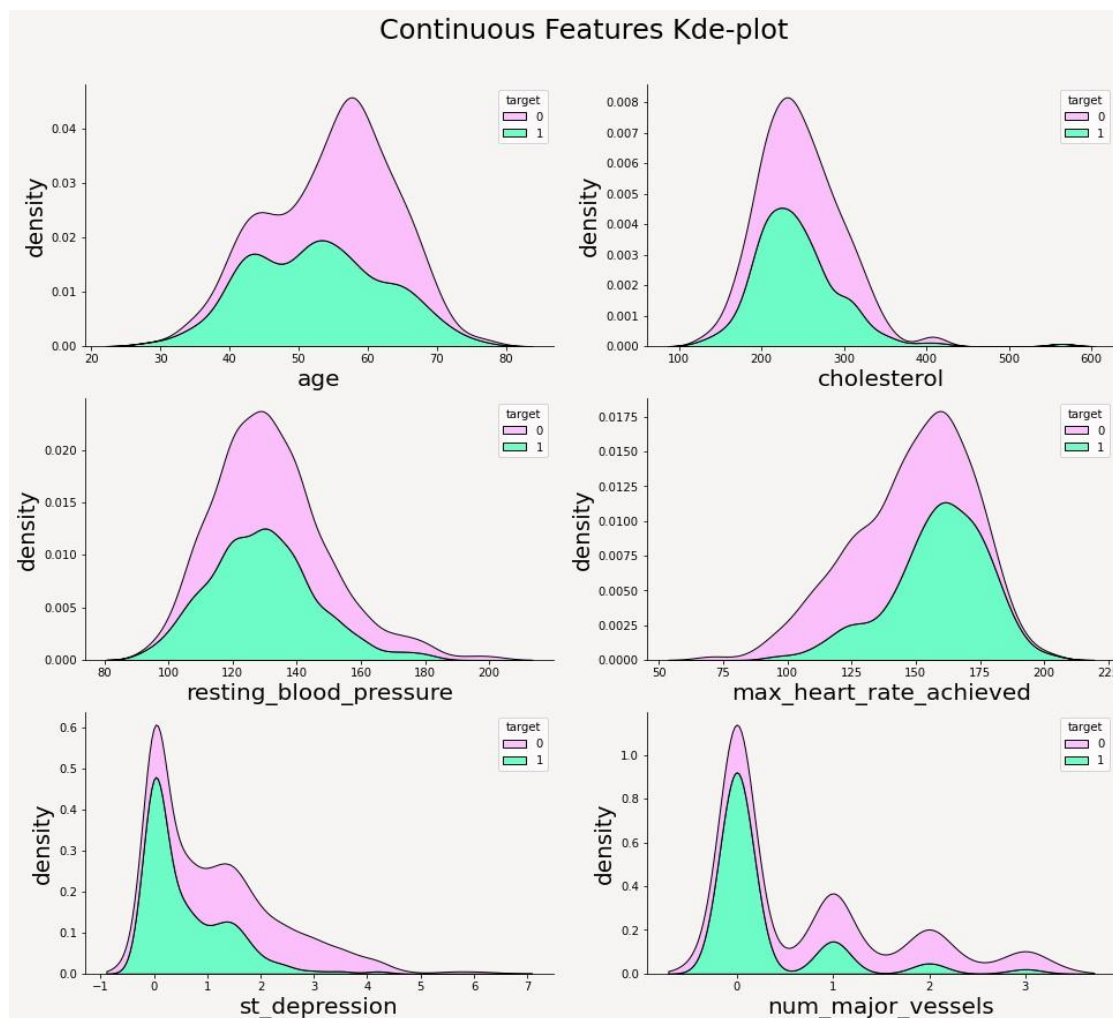


Figure 11 : Continuous Feature Kernel Density Estimate Plot.

The data are either over fitted or under fitted, as seen by the distribution plots. The target variable distribution for abnormal heart disease is 50.8%, as illustrated in Figure 5, whereas for normal heart disease it is 49.2%. Most people with defective hearts are between the ages of 55 and 65. Cholesterol and maximum heart rate people who have experienced this also seem to have attained a balance..

Figure 12. Shows that persons with a resting ECG cp of 1, 2, or 3 are more likely to develop cardiac disease than those with 0. People with a grade of 1 are more likely to have heart disease, which causes irregular heartbeat and moderate to severe symptoms. Myalgia from effort cardiovascular disease risk increased with a slope of 0 on the peak exercise ST segment, indicating no angina. Unlike 1, which suggests exercise causes angina. Those with slope values of 2 (indicating an unhealthy heart) had a higher rate of heart disease than those with slope values of 0 or 1 (indicating minimal change, a normal, healthy heart).

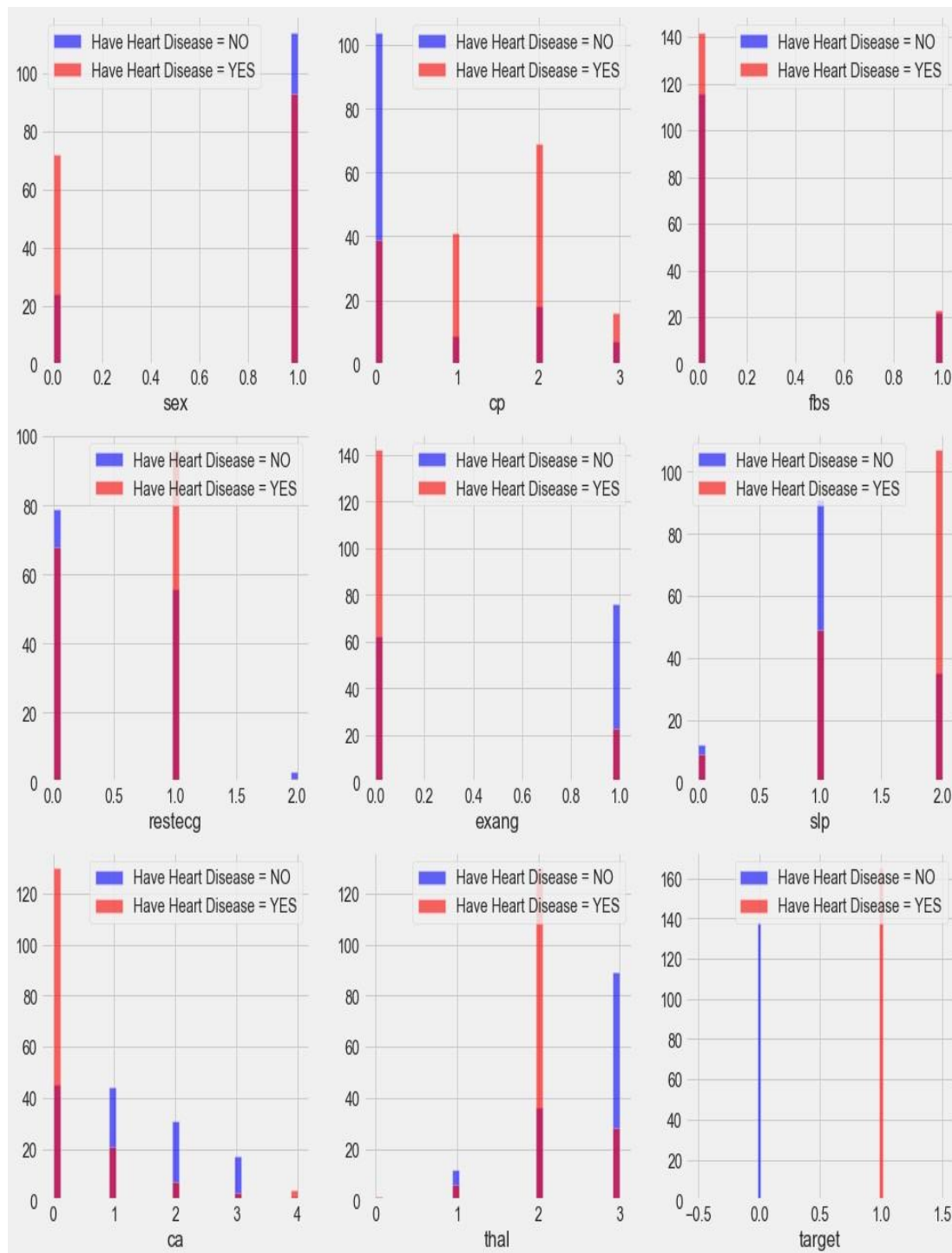


Figure 12: A Histogram of Characteristics with Categorical Values

Individuals who are capable of reaching a maximal heart rate of 140 beats per minute are at a greater risk of getting heart disease (Figure 13). The measurement of a resting blood pressure that falls between 130- and 140-mm Hg is generally considered to be reason for worry, whilst a serum cholesterol level that is 200 or above is often considered to be quite concerning. When compared to rest, activity-induced ST depression leads a heart that is already sick to experience more stress during physical activity. A scatter plot is shown here. When attempting to get a better

understanding of the distribution of characteristics, it may be beneficial to search for trends and patterns within the data.

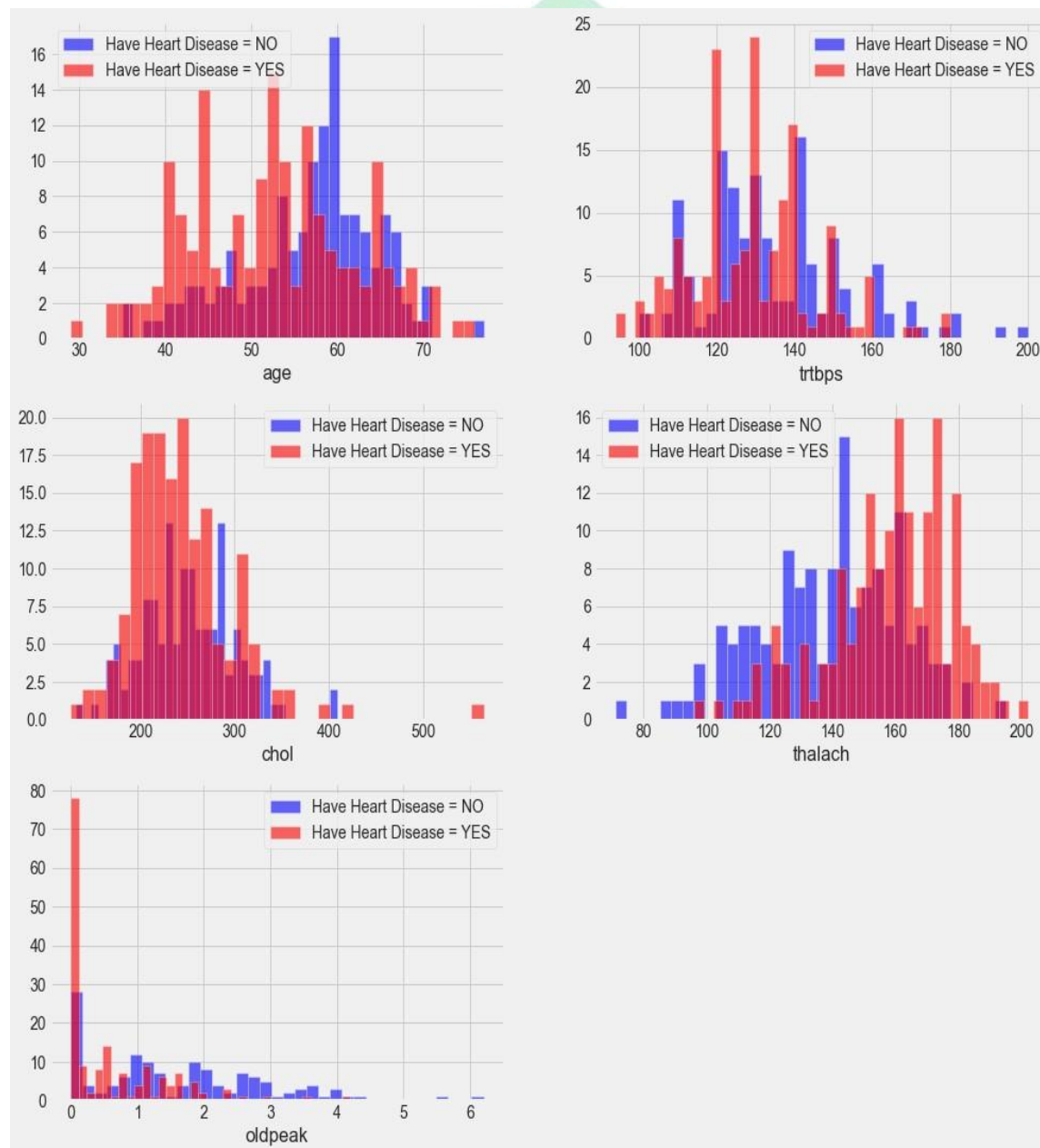


Figure 13 : The Histogram of Continuous-Valued Attribute

5. Conclusion

In the field of medicine, machine learning algorithms (MLAs) are an essential component because they have the potential to forecast human cardiac disease via the analysis of medical data. Cardiovascular diseases (CVDs) are a significant public health problem that need the efforts of academics and medical professionals to ameliorate. This research takes into consideration twelve various methodologies in order to enable it to carry out comparative analysis and get satisfactory outcomes. In comparison to learning algorithms that are directed by humans, the results of this research showed that machine learning algorithms performed much better. The comparison techniques that are used include, but are not limited to, accuracy, recall, precision, F1-score, and ROC-AUC. On the basis of the 14-parameter

dataset, the performance of the random forest classifier in the machine learning approach was increased by the use of data pre-processing. Using MLAs, a dataset of clinical data pertaining to cardiovascular disease (CVD) is analysed in order to figure out the primary risk factors that influence the progression of CVD. To find a solution to the issue, this is done. Additionally, the decision tree and random forest are used in order to improve the training and testing procedures for the algorithm. In the next step, we will examine the accuracy of

The analysis of heart illness will become more reasonable with the implementation of a decision support network design strategy that is based on machine learning. Further, the performance of the model is significantly impacted, and the amount of time required for calculations is increased due to the presence of certain attributes that are unimportant. Using feature determination calculations to find the most important characteristics, this study contributes yet another new component to the overall picture. These top features have the potential to increase classification accuracy while also reducing the amount of time required for the classification model to be executed. These examples of crucial classifiers for diagnosing cardiac diseases will be the outcome of additional inspections. These classifiers will use a variety of characteristics, such as feature selection and simplified techniques, to identify cardiac illnesses.

REFERENCES

1. Classification for EMG signals using linear discriminant analysis. In 2016 2nd International Conference on Advances in Computing, Communication, & Automation (ICACCA)(Fall) (pp. 1-6). IEEE.
2. Nikhar, S. and Karandikar, A.M., 2016. Prediction of heart disease using machine learning algorithms. *International Journal of Advanced Engineering, Management and Science*, 2(6), pp.239484 .
3. Nissa, N., Jamwal, S. and Mohammad, S., 2021. Heart Disease Prediction using Machine Learning Techniques. *Wesleyan Journal of Research*, 13(67).
4. Olufsen, M.S., 1999. Structured tree outflow condition for blood flow in larger systemic arteries. *American journal of physiology-Heart and circulatory physiology*, 276(1), pp.H257-H268.
5. Ornish, D., Brown, S.E., Billings, J.H., Scherwitz, L.W., Armstrong, W.T., Ports, T.A., McLanahan, S.M., Kirkeeide, R.L., Gould, K.L. and Brand, R.J., 1990. Can lifestyle changes reverse coronary heart disease?: The Lifestyle Heart Trial. *The Lancet*, 336(8708), pp.129-133.
6. Patel Brijain, R. and Rana, K.K., 2014. A survey on decision tree algorithm for classification. *Int. J. Eng. Dev. Res*, 2(1), pp.1-5.
7. Patel, J., TejalUpadhyay, D., & Patel, S., 2015. Heart disease prediction using machine learning and data mining technique. *Heart Disease*, 7(1), 129-137.
8. Quah, J.L.J., Yap, S., Cheah, S.O., Ng, Y.Y., Goh, E.S., Doctor, N., Leong, B.S.H., Tiah, L., Chia, M.Y.C. and Ong, M.E.H., 2014. Knowledge of signs and symptoms of heart attack and stroke among Singapore residents. *BioMed research international*, 2014.
9. Reddy, K.S., Patel, V., Jha, P., Paul, V.K., Kumar, A.S., Dandona, L. and Lancet India Group for Universal Healthcare, 2011. Towards achievement of universal health care in India by 2020: a call to action. *The Lancet*, 377(9767), pp.760-768.
10. Reddy, M.P.S.C., Palagi, M.P. and Jaya, S., 2017. Heart disease prediction using ANN algorithm in data mining. *International journal of computer science and mobile computing*, 6(4), pp.168-172.
11. Repaka, A.N., Ravikanti, S.D. and Franklin, R.G., 2019, April. Design and implementing heart disease prediction using naives Bayesian. In 2019 3rd International conference on trends in electronics and informatics (ICOEI) (pp. 292-297). IEEE.
12. Richter, M.W., 1995. Congestive heart failure in older adults*: epidemiology, pathophysiology, and etiology of congestive heart failure in older adults. *Journal of the American Geriatrics Society*, 45(8), pp.968-974.
13. Rich, M.W., 1997. Congestive heart failure in older adults*: epidemiology, pathophysiology, and etiology of congestive heart failure in older adults. *Journal of the American Geriatrics Society*, 45(8), pp.968-974.
14. Robinson, T.F., Factor, S.M. and Sonnenblick, E.H., 1986. The heart as a suction pump. *Scientific American*, 254(6), pp.84-91.

15. Roy, S.S., Mittal, D., Basu, A. and Abraham, A., 2015. Stock market forecasting using LASSO linear regression model. In *Afro-European Conference for Industrial Advancement* (pp. 371-381). Springer, Cham.
16. Sarangi, L., Mohanty, M.N. and Pattnaik, S., 2015. An intelligent decision support system for cardiac disease detection. *International Journal of Control Theory and Applications*, 8(5), pp.2137-2143.
17. Saritas, M.M. and Yasar, A., 2019. Performance analysis of ANN and Naive Bayes classification algorithm for data classification. *International Journal of Intelligent Systems and Applications in Engineering*, 7(2), pp.88-91.
18. Shah, D., Patel, S. and Bharti, S.K., 2020. Heart disease prediction using machine learning techniques. *SN Computer Science*, 1(6), pp.1-6.
19. Shaji, S.P., 2019, April. Prediction and diagnosis of heart disease patients using data mining technique. In *2019 international conference on communication and signal processing (ICCSP)* (pp. 0848-0852). IEEE.
20. Singh, A., Thakur, N. and Sharma, A., 2016, March. A review of supervised machine learning algorithms. In *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 1310-1315). IEEE.
21. Sinha, R.K., Aggarwal, Y. and Das, B.N., 2007. Backpropagation artificial neural network classifier to detect changes in heart sound due to mitral valve regurgitation. *Journal of Medical Systems*, 31(3), pp.205-209.
22. Späth, H., 2014. *Mathematical algorithms for linear regression*. Academic Press.
23. Spencer, K.T., Kimura, B.J., Korcarz, C.E., Pellicka, P.A., Rahko, P.S. and Siegel, R.J., 2013. Focused cardiac ultrasound: recommendations from the American Society of Echocardiography. *Journal of the American Society of Echocardiography*, 26(6), pp.567-581.
24. Sturgeon, K.M., Deng, L., Bluethmann, S.M., Zhou, S., Trifiletti, D.M., Jiang, C., Kelly, S.P. and Zaorsky, N.G., 2019. A population-based study of cardiovascular disease mortality risk in US cancer patients. *European heart journal*, 40(48), pp.3889-3897.
25. Sun, S. and Huang, R., 2010, August. An adaptive k-nearest neighbor algorithm. In *2010 seventh international conference on fuzzy systems and knowledge discovery (Vol. 1)*, pp. 91-94). IEEE.
26. Tan, K.C., Teoh, E.J., Yu, Q. and Goh, K.C., 2009. A hybrid evolutionary algorithm for attribute selection in data mining. *Expert Systems with Applications*, 36(4), pp.8616-8630.
27. Vonderen, J.J., Roest, A.A., Siew, M.L., Walther, F.J., Hooper, S.B. and Te Pas, A.B., 2014. Measuring physiological changes during the transition to life after birth. *Neonatology*, 105(3), pp.230-242.
28. Vom Saal, F.S. and Hughes, C., 2005. An extensive new literature concerning low-dose effects of bisphenol A shows the need for a new risk assessment. *Environmental health perspectives*, 113(8), pp.926-933.
29. Venkatalakshmi, B. and Shivsankar, M.V., 2014. Heart disease diagnosis using predictive data mining. *International Journal of Innovative Research in Science, Engineering and Technology*, 3(3), pp.1873-7.
30. Virani, S.S., Alonso, A., Benjamin, E.J., Bittencourt, M.S., Callaway, C.W., Carson, A.P., Chamberlain, A.M., Chang, A.R., Cheng, S., Delling, F.N. and Djousse, L., 2020. Heart disease and stroke statistics—2020 update: a report from the American Heart Association. *Circulation*, 141(9), pp. e139-e596.
31. Voleti, S.R. and Reddi, K.K., 2016. Design of an optimal method for disease prediction using data mining techniques. *International Journal of Advanced Research in Computer Science and Software Engineering*, 6(12), pp.328-337.
32. Wu, H., Huynh, T.T. and Souvenir, R., 2014, April. Motion factorization for echocardiogram classification. In *2014 IEEE 11th International Symposium on Biomedical Imaging (ISBI)* (pp. 445-448). IEEE.
33. Xing, Y., Wang, J. and Zhao, Z., 2007, November. Combination data mining methods with new medical data to predicting outcome of coronary heart disease. In *2007 International Conference on Convergence Information Technology (ICCIT 2007)* (pp. 868-872). IEEE.