Section: Research Paper
ISSN 2063-5346

Ensemble Machine Learning Models for Accurate Heart Disease Diagnosis

Dillip Narayan Sahu[1], Satheesh Kumar [2]*
[1]Research Scholar, Department of Computer Science and Application, OPJS University, Churu, Rajasthan, India
[2]* Professor, Department of Computer Science and Engineering, Chaitanya (Deemed to be University), Telangana, India
*Corresponding Author Email: satish4info@gmail.com

DOI: 10.48047/ecb/2023.12.si4.1691

Abstract

Heart disease is a major worldwide health problem, and successful treatment and better patient outcomes depend on early and correct diagnosis. The use of machine learning techniques to help doctors diagnose heart problems has showed promise. The ensemble strategy suggested in this study combines many machine learning models to improve the precision of heart disease detection. The ensemble technique combines different machine learning algorithms, including Gradient Boosting, Support Vector Machines (SVM), and Random Forest, into a single framework. To capture a complete depiction of the illness patterns, each model is trained on a broad set of data obtained from medical records, clinical measures, and patient history. Using a sizable dataset of anonymized patient records with a verified diagnosis of heart disease, the effectiveness of the ensemble technique is tested. The collection includes data on demographic traits, symptoms, lab test outcomes, and reports from medical imaging. The ensemble model's diagnostic adequacy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC) are then assessed against individual machine learning models and conventional diagnostic techniques.

Keywords: Medical Records, Clinical Measurements, Patient History, Ensemble, Machine Learning, Heart Disease Diagnosis, Machine Learning Algorithms, Random Forest and Support Vector Machines (SVM).

1. Introduction

Heart disease continues to be a major global health problem, contributing significantly to morbidity and mortality globally. For the purpose of executing suitable treatment plans and enhancing patient outcomes, timely and accurate diagnosis is essential. However, determining the presence of cardiac disease can be difficult since it necessitates a thorough examination of several elements, such as a patient's medical history, clinical assessments, and diagnostic testing. Machine learning approaches have gained popularity as effective tools for diagnosing illnesses in recent years. These algorithms have the capacity to examine enormous patient data sets, spot intricate trends, and generate data-driven forecasts[1]. A potent strategy to increase prediction accuracy and get around single model restrictions is ensemble machine learning, which mixes numerous unique models. The purpose of this study is to investigate the use of ensemble machine learning models for precise cardiac disease detection. We aim to develop a more robust and trustworthy diagnostic tool by merging several machine learning techniques, such as Random Forest, Support Vector Machines (SVM), and Gradient Boosting, into a cohesive framework[2]. To accurately represent the intricacy of cardiac disease patterns, each model in the ensemble will be trained on a complete collection of data gathered from medical records, clinical measures, and patient history.

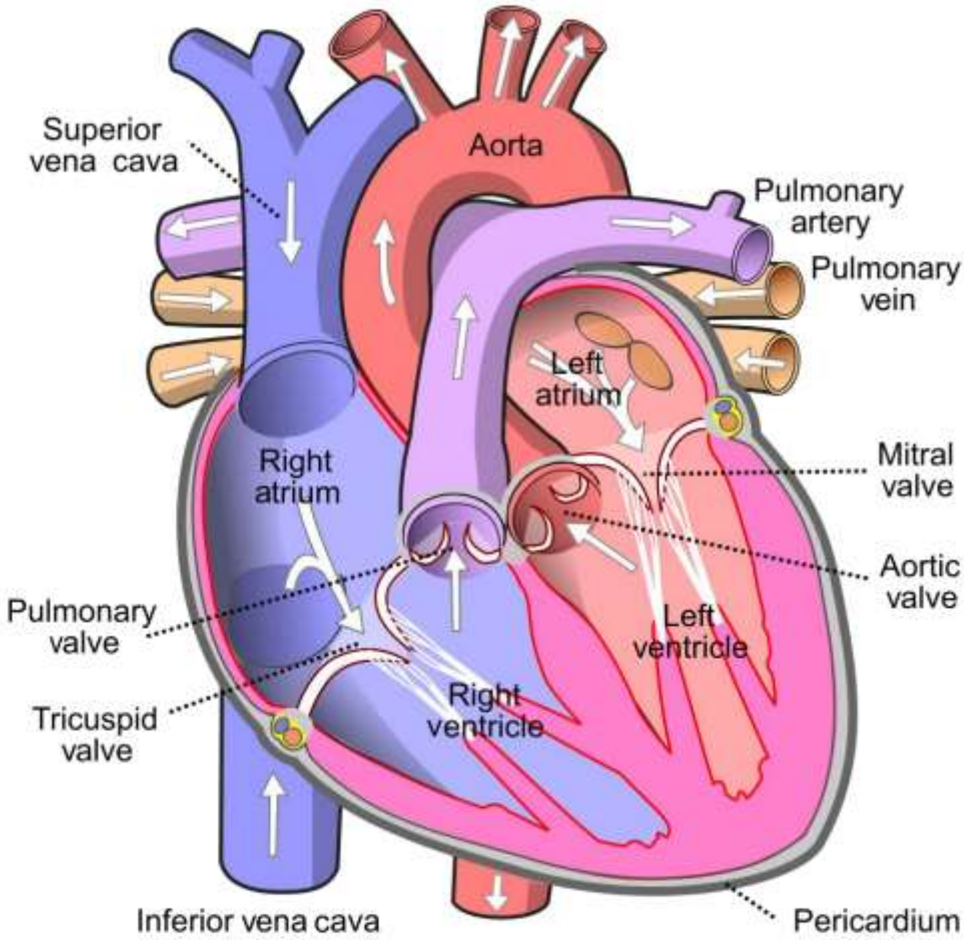

Figure.1: Anatomy of the human heart

This study's main goal is to assess the ensemble approach's diagnostic performance in comparison to individual machine learning models and traditional diagnostic techniques. We will make use of a sizable dataset of anonymised patient records, which includes demographic data, symptoms, lab test results, and findings from medical imaging[3]. This study's importance stems from its potential to improve heart disease detection precision and lower misdiagnoses. The ensemble technique tries to deliver more accurate forecasts by combining the capabilities of many algorithms, which should enhance patient care and treatment results. The study will also add to the increasing body of knowledge on the use of machine learning in healthcare, opening the door for future development in medical diagnostics.

The left and right atriums and the left and right ventricles are the two lowest chambers of the human heart, which has four chambers altogether. The ventricles act as discharge chambers, pumping blood to the arteries, while the atriums serve as receiving chambers that take blood from the veins[4]. The tricuspid valve, pulmonary valve, mitral valve, and aortic valve are the four valves that make up the heart. These valves are essential for ensuring that blood flows in the right direction and preventing blood from flowing backward. The superior vena cava and inferior vena cava are the two major veins that provide non-oxygenated blood to the right atrium. Through the tricuspid valve, it then pumps this blood to the right ventricle. In turn, the right ventricle pumps blood via the pulmonary valve and into the lungs where it is oxygenated[5]. The left atrium gathers the blood after it has been oxygenated in the lungs and pumps it through the mitral valve to the left ventricle. In order to deliver oxygen and nutrients to different organs and tissues throughout the body, the left ventricle is in charge of pumping oxygen-rich blood via the aortic valve, delivering it to the aorta, and then distributing it throughout the body. The structure of the human heart is shown diagrammatically in Figure 1, emphasizing the positioning of the organ's chambers and valves as well as the flow of blood through it.

The idea of learning has undergone a revolution thanks to the quickly developing and transformational discipline of machine learning in computer science. Machine learning enables systems to imitate human intellect by using historical data. Its self-adaptiveness and strong mathematical underpinnings have opened the door for numerous applications across many fields. Machine learning is essential to improving services in the healthcare industry. Machine learning's capacity to create decision support systems enables more precise and effective illness diagnosis[6]. The goal of this particular research article is to demonstrate how to use machine learning to diagnose heart problems. Machine learning is widely used in a wide range of industries outside of healthcare. While recommendation systems make customised content or goods recommendations to consumers, sentiment analysis enables firms to monitor customer attitudes and responses. Information retrieval aids in locating pertinent facts from large volumes of information, whereas natural language processing aids in comprehending and producing human language. Additionally, machine learning makes a substantial contribution to many other fields, including computer vision, text recognition, market research, and game playing. As academics and practitioners discover new methods to harness the power of machine learning, its applications are constantly growing.

2. Literature Survey

In recent years, there has been a lot of study on the application of machine learning algorithms in medical diagnostics, especially for the diagnosis of cardiac disease. To increase diagnostic precision, improve patient outcomes, and assist medical personnel in making decisions, researchers have looked at a variety of machine learning approaches[7]. With an emphasis on ensemble machine learning models, this literature review offers an overview of several important papers concerning the use of machine learning in the detection of cardiac disease.

Researchers have made a lot of work in this area, and continuous study keeps these decision support systems getting better. The numerous decision support systems that have been created for the diagnosis of heart disease are thoroughly surveyed in this section. To create these systems, researchers used both machine learning and deep learning methods. To guarantee the accuracy and efficacy of the models, the performance of these systems is assessed using particular parameters and a variety of validation techniques. Several publicly accessible cardiac disease datasets were used by researchers to evaluate and assess the effectiveness of their methods[8]. The many heart disease datasets that are readily accessible online for study are also covered in-depth in this section. In general, research in this area shows the promise of deep learning- and machine-learning-based decision support systems to help with heart disease detection. This chapter adds to the ongoing work in building precise and reliable methods for the early detection and diagnosis of cardiac disease, which will eventually improve patient outcomes, by giving an overview of the current systems and the datasets employed.

Examined in the study "Ensemble Machine Learning for Cardiovascular Disease Prediction" This work suggests an ensemble machine learning strategy for forecasting the risk of cardiovascular illness. It was published in the Journal of Medical Systems. To increase prediction accuracy, the ensemble model incorporates Random Forest, SVM, and Neural Networks. The study illustrates the ensemble model's efficacy in precisely identifying those who are at high risk for cardiovascular disease.

Findings from the study "Improving Coronary Heart Disease Diagnosis Using Ensemble Learning Techniques" This study examines the use of ensemble learning approaches, such as Gradient Boosting and Bagging, for enhancing coronary heart disease detection[9]. It was published in the International Journal of Medical Informatics. The findings demonstrate that the ensemble models greatly improve diagnosis accuracy by outperforming individual algorithms and conventional diagnostic techniques.

research project: "An Ensemble Machine Learning Framework for Heart Disease Prediction" This study, which was published in the IEEE Journal of Biomedical and Health Informatics, suggests an ensemble framework for heart disease prediction that integrates SVM, Decision Trees, and Naive Bayes[10]. The study shows how the ensemble model may improve diagnostic performance by lowering the number of false negatives and false positives.

"Ensemble Learning Models for Heart Disease Prediction: A Comparative Study" This research, published in the Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine, compares various ensemble learning methods for heart disease prediction, including AdaBoost, Random Forest, and Stacking. The findings demonstrate how effective some ensemble models are in correctly diagnosing heart disease.

Examined in the study "Machine Learning Approaches for Heart Disease Diagnosis: A Comprehensive Review" This thorough analysis, which was published in the Journal of Medical Systems, looks at several machine learning techniques for diagnosing cardiac problems. The research sheds light on the benefits and drawbacks of various algorithms, including ensemble approaches, as well as their potential contributions to the detection of cardiovascular disease.

The authors [11] of the study created an ensemble mechanism-based heart disease prediction system. The tests were run on five separate datasets, where the class labels "0" and "1" denoted the existence or absence of the illness, respectively. For the purpose of guaranteeing reliability and robustness, the authors conducted ten-fold cross-validation to confirm their findings. Since the dataset given by the authors already had pertinent characteristics, feature selection for data pretreatment was skipped. However, missing values and outliers were dealt with by data pretreatment. To identify and manage outliers in the data, the inter-quantile range approach was used. The studies' outcomes across the five datasets revealed good performance: Cleveland dataset: 85.81% accuracy SPECTF dataset: 80.15% accuracy SPECT dataset: 82.40% accuracy Eric dataset: 86.12% accuracy Statlog dataset: 88.52% accuracy These findings showed that the ensemble process and the chosen classifiers were capable of accurately predicting heart disease on the relevant datasets[12]. The results of the study have important ramifications for improving heart disease early detection and diagnosis, which may improve patient outcomes.

The Neuro-Fuzzy Classifier (NFC) was used by the authors of the study [13] to create a fuzzy rule-based system for identifying heart disease. They used the scaled conjugate gradient approach to increase learning speed and decrease root mean square error, enhancing NFC performance. The Cleveland dataset was used in the authors' research[13]. They used discretization to turn the continuous values of features into discrete values as part of the data pretreatment procedure. The goal of this stage was to prepare the data for the fuzzy rule-based system. Fuzzification and defuzzification techniques, which are features of fuzzy logic systems, were used in the categorization process. Crisp data are transformed into fuzzy sets through the process of fuzzification, while crisp data are produced crisply through the process of defuzzification. Sequential Feature Selection (SFS) and Multiple Logistic Regression (MLR) were two methods utilized by the authors to select critical characteristics for heart disease diagnosis[14]. The most pertinent characteristics were chosen using these techniques in order to improve the performance of the model. The hold-out approach was used to the system validation. With this technique, the dataset is split into training and testing sets, allowing the model's performance to be evaluated on unobserved data. According to the findings, the NFC and MLR combo had the greatest diagnostic accuracy for heart disease, 84%. This result illustrates the possibility of integrating NFC and MLR to increase diagnostic accuracy and indicates the efficacy of the proposed fuzzy rule-based system for heart disease detection.

Heart disease is a serious ailment, and early detection is essential to minimizing its effects and saving lives. Researchers have concentrated on creating decision support systems utilizing machine learning techniques to assist in the diagnostic process[15]. These devices use non-invasive testing to identify cardiac disease, improving accessibility and effectiveness of the diagnostic procedure.

These works show the rising interest in using ensemble machine learning models to diagnose cardiac disease. The findings show that ensemble models frequently outperform individual algorithms and conventional diagnostic approaches, showing promise for enhancing diagnostic precision and patient outcomes[16]. To evaluate the effectiveness of ensemble techniques on bigger and more varied datasets, however, and to guarantee their practical application in actual clinical situations, further study is required. Future medical research and innovation should focus on the potential for ensemble models to transform the detection of cardiac disease as machine learning continues to progress.

3. Heart Disease

Heart disease refers to a range of disorders that impair the heart's and blood vessels' ability to function. One such illness is Coronary Artery Disease (CAD), in which the arteries supplying the heart with oxygen-rich blood are blocked. The obstruction is brought on by the buildup of plaque, which is made of calcium and cholesterol. As a result, the heart receives less oxygen and blood. High levels of HDL (high-density lipoprotein) cholesterol assist in transferring cholesterol from the body to the liver, minimizing plaque buildup, whereas high levels of LDL (low-density lipoprotein) cholesterol lead to plaque development[17]. In order to avoid CAD, low LDL and high HDL levels must be maintained. Stroke is another cardiac condition that is brought on by a blockage in the blood arteries that carry blood to the brain. Reduced blood flow to the brain causes nutrition and oxygen deprivation, which causes brain cell death and might cause permanent brain damage[18]. Blood spills into or around the brain as a result of ruptured blood vessels, putting pressure on the area and harming brain cells. This is known as a hemorrhagic stroke. Congenital cardiac disease is a congenital condition brought on by faulty fetal development that primarily affects children. Genetic mutations can be a factor, albeit the specific reason is frequently obscure. Heart failure happens when the heart gets frail and is unable to circulate blood efficiently[119]. The patient's quality of life may be impacted by it, whether it is acute or chronic. While chronic heart failure is a continuing illness, acute heart failure happens quickly. cardiac failure can be caused by a variety of cardiac disorders, including hypertension, coronary artery disease, and illnesses of the valves. Heart failure can be classified as left-sided, right-sided, diastolic, and systolic. In order to minimize serious consequences and enhance patient outcomes, proper care and early identification of cardiac problems are crucial. Maintaining heart health and avoiding heart

Section: Research Paper
ISSN 2063-5346

problems requires routine medical exams, lifestyle changes, and commitment to recommended therapies.

4. Decision Support System

A decision support system (DSS) is a useful instrument that helps people make informed decisions and raises the standard of their judgments. It helps with difficult issue solving by making recommendations based on the data it has accumulated. DSS is especially helpful for tackling problems that are unstructured or semi-structured and might be difficult for individuals to solve on their own. The creation of DSS is heavily reliant on machine learning techniques.

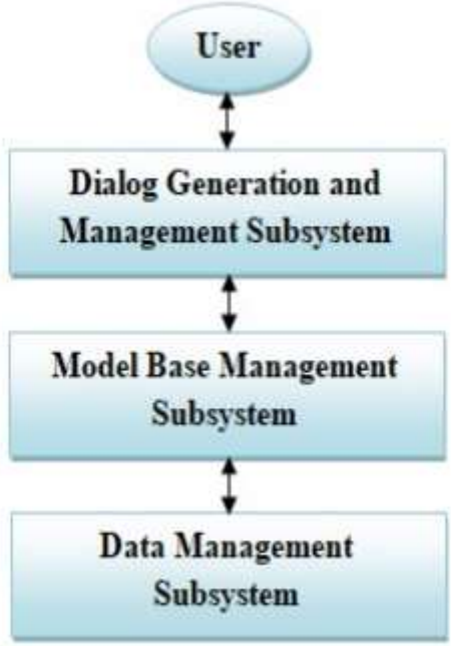

Figure.2: DSS components

Figure 2 shows the three key parts that make up DSS: DMS: Data Management Subsystem The data utilized by the DSS is managed and stored by this component. A database management system (DBMS) controls the management of data that is kept in computer memory. Users can execute activities such as data addition, deletion, modification, and retrieval using the DBMS. MBMS, or Model Base Management Subsystem: The MBMS is tasked with using numerous models to analyze data for decision-making purposes. It keeps several models, which, when used, evaluate data to produce important information needed for decision-making[20]. Subsystem for Dialog Generation and Management (DGMS): This component serves as the DSS's user interface. Users may submit queries, and the system delivers the required information in an easily comprehensible manner, including reports shown as graphs, charts, etc. The use of DSS in the healthcare sector has grown in popularity as a result of the advantages it provides to both patients and medical professionals. DSS can give critical assistance to healthcare practitioners in underdeveloped nations with little medical staff and resources, empowering them to make better choices and treat patients more effectively and promptly. In turn, patients gain by getting the proper care at the right time and sometimes avoiding the need for pricey diagnostics, which results in more affordable care. DSS in healthcare uses information acquired from multiple sources to make therapeutic judgments. It supports clinicians in early illness detection, provides patients with preventative measures, and gives individualized treatment alternatives. Early illness prediction capabilities allow the system to provide appropriate treatment recommendations. Additionally, it may assess the effects of particular medications, find illness risk factors, spot unusual patient circumstances, and provide alarm messages based on this data. Additionally, DSS might advise patients to alter their lifestyles in order to ward off the onset of specific diseases. Social media platforms, remote sensors, medical imaging such as X-rays and CT scans, various medical testing, and doctors' notes are just a few of the sources from which data for DSS may be gathered. Overall, DSS is a useful tool for the healthcare sector, assisting both patients and healthcare professionals in reaching better decisions and enhancing general healthcare results.

5. Machine Learning Models

Machine learning has become a crucial part of smart applications in the modern, technologically sophisticated era, considerably improving their performance. Machine learning makes it possible to find information and patterns in raw data that conventional algorithms might not be able to. It is now an essential component of research in many different fields. The ability to learn automatically from data and make judgments akin to humans is provided by machine learning. Machines become better at making decisions as they gather experience from data and perform better. Machines may learn from the input and output correlations in the training data by using the data as a training mechanism. As shown in Figure 3, there are three primary categories for machine learning: supervised education: In supervised learning, outputs are given to the system in exchange for inputs from training data. By examining the training data and determining the connections between inputs and outputs, the machine learns. Both classification and regression issues may be solved using supervised learning. In classification, data is divided into specified groups, while in regression, a machine predicts the value of particular variables. Unsupervised Learning: Unsupervised learning includes training the computer with unlabeled data in order to find patterns in the

Section: Research Paper
ISSN 2063-5346

data without the use of predetermined categories. The system analyses the data through the clustering process to provide a summary and find groups of related objects. The emphasis of reinforcement learning is on obtaining a certain goal by doing the best possible behaviors. An agent is in charge of carrying out activities and gains knowledge of the surroundings in which these actions are carried out. The agent interacts with the environment to get knowledge about it while maintaining an internal state. The agent utilizes a reward mechanism to absorb feedback from its surroundings. Positive rewards are provided for good deeds, and negative rewards are obtained for bad deeds. The agent wants to reduce unfavorable incentives while maximizing favorable ones.

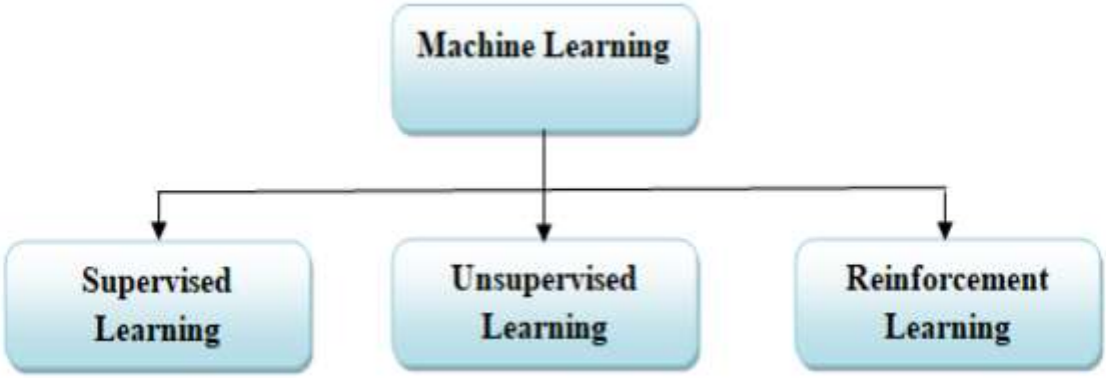

Figure.3: Types of machine learning

Reinforcement learning is employed in applications like self-driving vehicles since it does not require human specialists with domain-specific expertise. Machine learning techniques have advanced to the point that smart apps can now execute complicated jobs, offer individualized services, and adapt to changing circumstances. Machine learning is increasingly being incorporated into a variety of academic disciplines and practical applications, which has significantly advanced technology and human-machine interactions.

5.1 Classification Algorithms

Popular supervised learning techniques used in machine learning include the Decision Tree method. It uses a tree-like structure to classify data and make judgments. The technique uses a training dataset made up of prediction-related characteristics and the target classes they correspond to. A model that can categorize fresh data based on input qualities is the algorithm's output. Figure 4 shows three different sorts of nodes that make up a decision tree: Origin Node: The root node, which serves as the tree's origin and symbolizes the best predictor feature that divides the dataset depending on a certain condition, is the tree's beginning point. Internal Nodes: Internal nodes are conditional nodes that divide the dataset into smaller subgroups in accordance with certain circumstances associated with the predictor characteristics. Nodes (Leaves) at the terminal: The terminal nodes of the tree, which reflect the class values of the target variable, are its final nodes. Decision trees can be applied to two different kinds of issues: Data is divided into specified classes in a classification tree, which is an approach for solving classification issues. Regression Tree: When attempting to forecast the value of a variable, a regression tree is employed. The bagging approach, which combines numerous decision trees using a voting process, can be used to enhance the performance of decision trees. This lessens overfitting and enhances precision. The decision tree is constructed downward, beginning at the root node. The tree is divided at each level depending on the predictor trait that yields the biggest information gain, highlighting its significance in class prediction. One benefit of employing decision trees is that they are simply interpretable, making it simple to justify a certain choice. Decision trees are also reasonably simple to use and comprehend, requiring less data preparation than other methods. Decision trees, however, may experience overfitting, particularly if the tree gets very complicated. Pruning is one method that may be used to solve this problem and enhance generality.

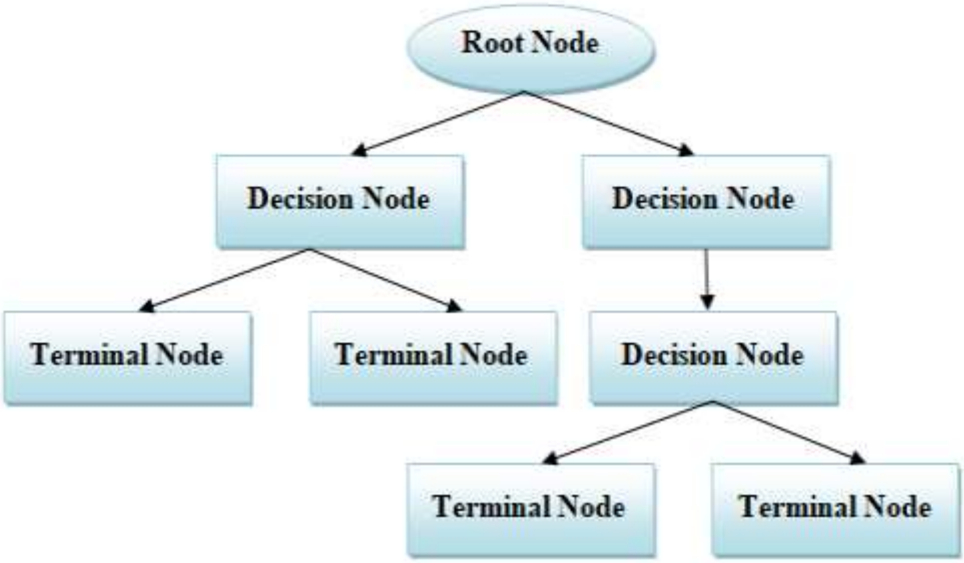

Figure.4: Decision tree

It is possible to employ Random Forest, a potent supervised learning method, for both classification and regression problems. It is an ensemble learning technique that integrates the results of various decision trees to provide predictions that are more accurate than those from single decision trees. Each decision tree in a random forest is built using a randomly selected subset of the total characteristics available. If there are m features, for instance, k (where k m) features are randomly chosen for each tree. Feature bagging or feature randomization are terms used to describe this method. Each decision tree receives independent training, and each tree classifies the input separately. When using a random forest for classification problems, the final output class is chosen by taking the mode (most common class) of the classes predicted by each decision tree in the forest. For regression tasks, on the other hand, the final output is calculated by averaging the results predicted by each decision tree in the forest. The overfitting issue frequently connected to individual decision trees is addressed by the random forest method. When a model performs well on training data but badly on fresh, untried data, overfitting has taken place. Random forest lessens overfitting and increases the model's capacity to generalize by integrating the results of many decision trees. Figure 5 shows how a random forest works, combining numerous decision trees with random subsets of data to produce the final random forest model. Due to its propensity for making precise predictions, capacity for handling big datasets, and ability to address the overfitting problem, Random Forest is a well-liked and commonly used method in machine learning. It works especially well for applications involving intricate data structures and multidimensional datasets.

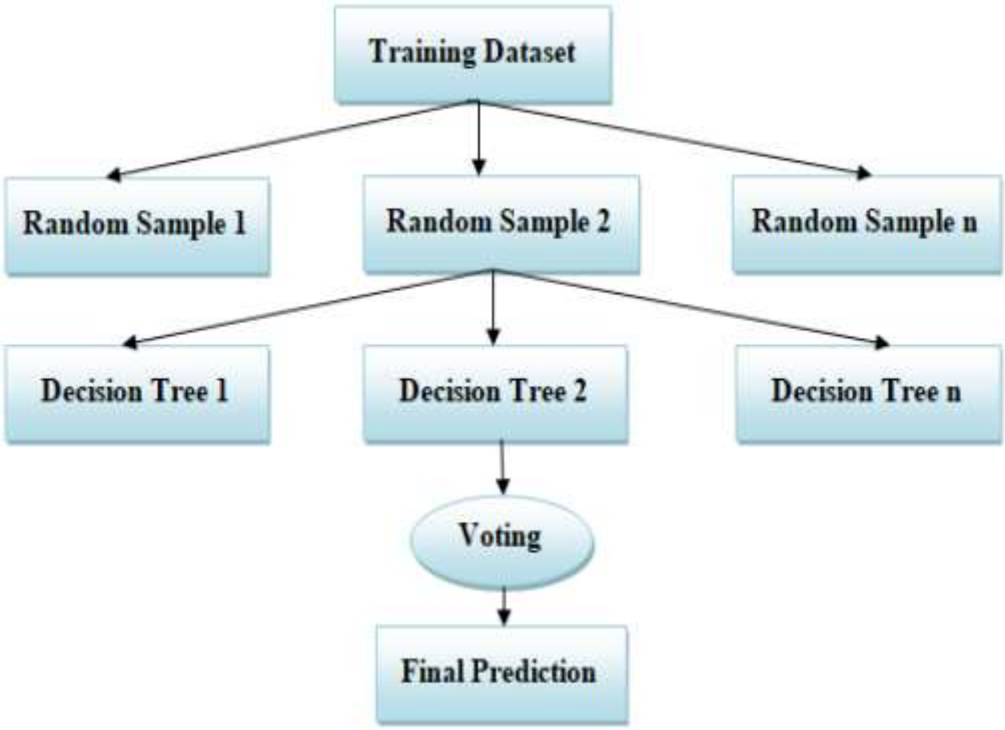

Figure.5: Random forest

In binary classification situations, where the target variable Y might have one of two potential values—0 or 1—logistic regression is a well-liked approach. Additionally, it may be expanded to deal with multi-class classification issues where Y has more than two potential values. The result of logistic regression is converted into a probability score between 0 and 1 using a logistic function, sometimes referred to as a sigmoid function. This probability score shows how likely it is that the input data will fall into a specific class. Another classification technique that uses the similarity metric is K-Nearest Neighbor (KNN). By locating the input data's k-nearest neighbors from the training set using a distance metric, it may categorize the input data. The data points closest to the input data are these neighbors. The input data is then classified by its k-nearest neighbors using a majority vote. A potent approach used for both binary and multi-class classification applications is Support Vector Machine (SVM). To divide the data points of the various classes, it creates a hyperplane in the feature space. To establish the greatest possible separation between the two classes, the hyperplane is tuned to maximize the margin between them. Support vectors are the nearest-to-the-hyperplane data points that have a significant impact on where the hyperplane is located. The selection of the hyperplane in SVM enables a distinct division of the data points into their appropriate classes. A visual illustration of the SVM hyperplane dividing two classes is shown in Figure 6. Each of the three algorithms—K-Nearest Neighbor, Support Vector Machine, and Logistic Regression—has special qualities and benefits. They are often used in different classification tasks, and the best method to utilize depends on the particular problem, the type of data being used, and the performance that is needed.

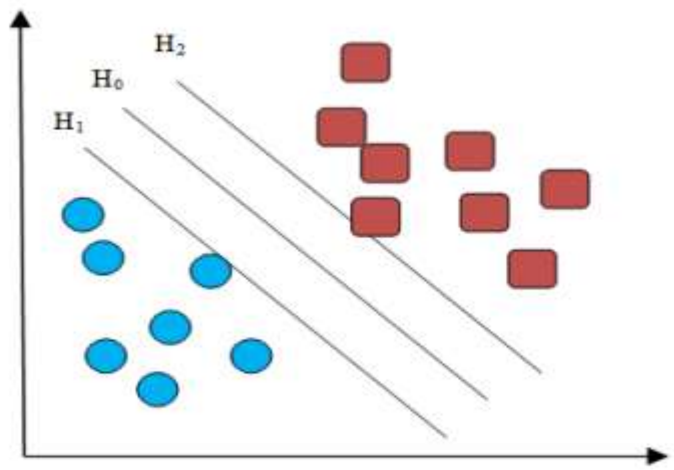

Figure.6: Hyperplane separating two classes in SVM

A probabilistic classifier called Naive Bayes employs the Bayes theorem to make classifications. It is renowned for being straightforward and efficient in solving a variety of

Section: Research Paper
ISSN 2063-5346

issues, including document classification, spam filtering, and illness detection. Because of this assumption, the algorithm is referred to as "naive," which stands for "naive." Naive Bayes frequently works well in reality and is particularly helpful for text classification jobs despite this oversimplifying assumption. The ensemble learning technique known as Adaboost, or Adaptive Boosting, tries to enhance the performance of underperforming students. It works by progressively training many copies of a classifier, with each copy attempting to fix the mistakes committed by the one before it. The dataset is split up into several subsets, and different weights are given to each piece of data. Instances that were misclassified are given heavier weights, increasing the likelihood that they will be chosen for the following subset. The goal of Adaboost is to aggregate the results of these ineffective classifiers using a weighted average, giving greater weight to the final prediction to classifiers with higher accuracy. By concentrating on the samples that are challenging to accurately identify, the algorithm gradually enhances performance, thus enhancing the performance of the ensemble as a whole. Decision trees are the standard classifier used for boosting in Adaboost. Other weak learners can also be supplied as arguments to the Adaboost algorithm, thus it is not simply restricted to decision trees. Figure 7 explains how the AdaBoost method works by demonstrating how several classifiers are trained sequentially, and then their results are merged to produce a powerful classifier with increased accuracy. Adaboost is a potent method for enhancing poor classifiers' performance, and it has been used for a number of machine learning tasks, including face identification, image recognition, and natural language processing. It is a well-liked option in ensemble learning methodologies because to its adjustability and capacity for handling difficult challenges.

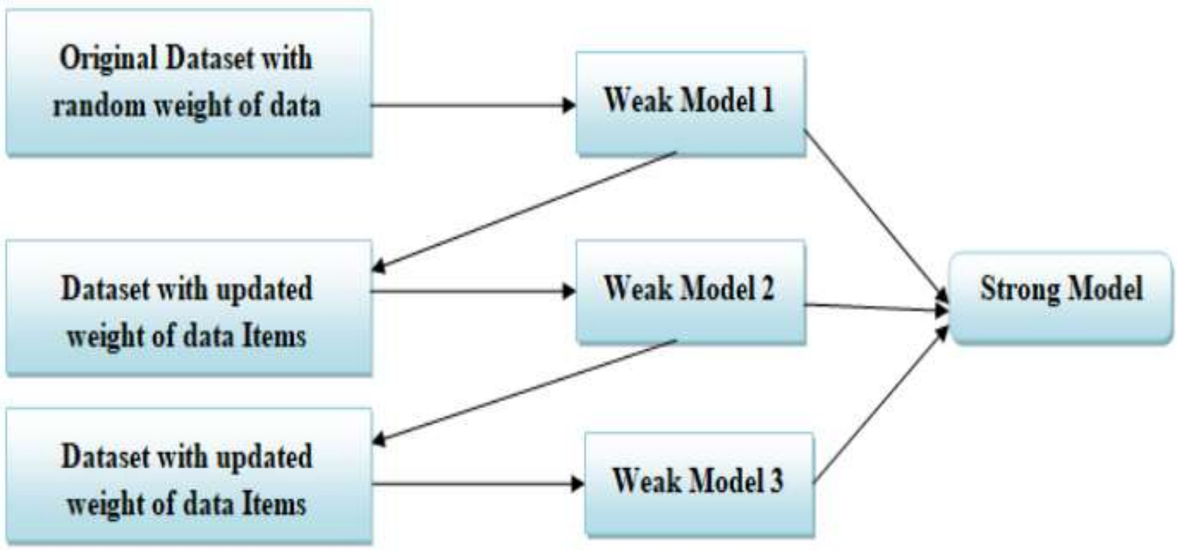

Figure.7: Adaboost algorithm

Section: Research Paper
ISSN 2063-5346

Extreme Gradient Boosting, or XGBoost, is a potent ensemble classifier that makes use of gradient boosting to turn weak learners into strong ones. Boosting is an ensemble strategy that enhances the overall prediction performance by combining the output of several weak learners, such as decision trees. Building successive decision trees, each of which seeks to fix the mistakes caused by the one before it, is how the XGBoost algorithm operates. This is accomplished by concentrating on the data points that were incorrectly categorized or had significant residuals in earlier rounds. XGBoost develops a more precise and reliable prediction model by continually adding trees and lowering the mistakes. The greedy method of node segmentation used by XGBoost is one of its core features. XGBoost employs a greedy approach to identify the optimal split rather than taking into account all potential splits at each node, which considerably lowers the computational cost and increases efficiency. Additionally, XGBoost adds regularization to the loss function, preventing overfitting and enhancing the model's generalizability. Complex models are penalized by the regularization term, which deters them from fitting noise and pointless patterns in the data. The XGBoost algorithm has become more well-known as a result of its excellent performance, effectiveness, and capacity for handling huge datasets. It has been heavily utilized in several machine learning challenges as well as practical applications, such as web search ranking, click-through rate prediction, and many other data science projects. As a result of its use of gradient boosting, greedy node splitting, and regularization, XGBoost is a leading ensemble learning algorithm and the method of choice for many data scientists and machine learning professionals.

6. Diagnosis System

Creating a large dataset of anonymized patient information is the initial stage in the process. Demographic data, heart disease symptoms, medical history, lab test results, and imaging report data should all be included in the dataset. To guarantee the generalizability of the model, the data should represent a varied population. The performance of the machine learning models may be impacted by missing values, outliers, and noise present in the acquired data. To clean and prepare the data for training, data preparation techniques including imputation, normalization, and feature scaling are used. The most pertinent and informative features are chosen using feature selection or extraction approaches to build an effective ensemble model.

Section: Research Paper
ISSN 2063-5346

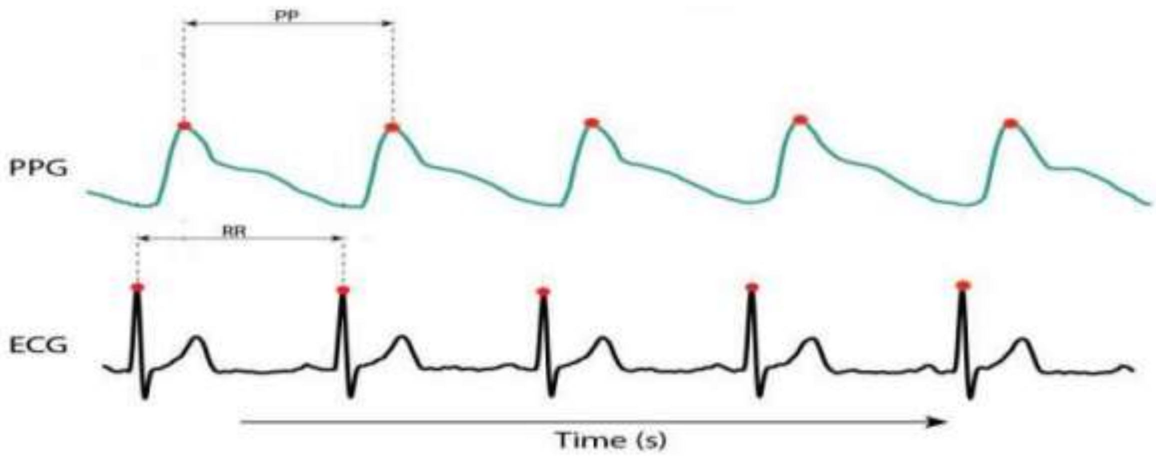

Figure.8. RR interval and peak to peak interval

The objective is to decrease dimensionality and concentrate on characteristics that considerably aid in the diagnosis of heart disease. Several machine learning algorithms are chosen at this step to form the ensemble model. Decision Trees, Random Forest, SVM, and gradient boosting are popular options. Each model is trained using a distinct collection of features or data division on the preprocessed data. The ensemble strategy for merging the results of different models is defined by the methodology. The final forecast can be made using strategies like majority voting, weighted voting, or stacking. The ensemble strategy seeks to capitalize on the advantages of individual models while minimizing the influence of any potential biases.

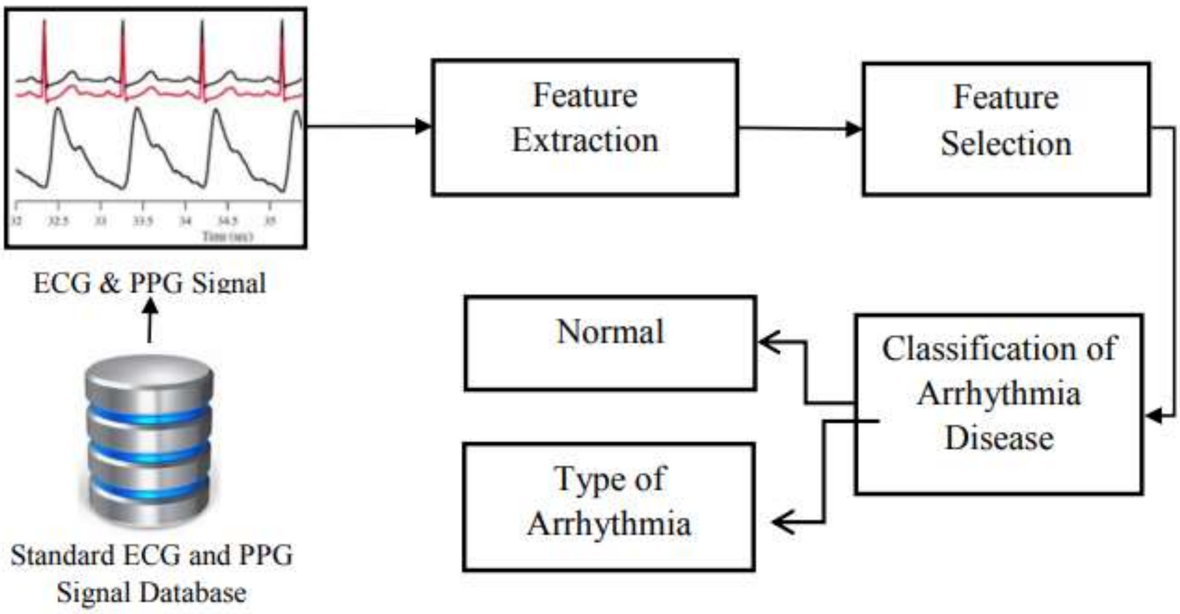

Figure.9:Flow diagram of CADS system

The goal of the project is to create a system for diagnosing cardiac arrhythmias that makes use of both ECG (Electrocardiogram) and PPG (Photoplethysmogram) data. An irregular heartbeat, known as an arrhythmia, can appear as slow, rapid, or irregular heartbeats. The diagnostic algorithm makes use of statistical characteristics, heart rate variability features, and pulse rate variability features generated from both the PPG signal and the ECG signal. The ECG signal's RR interval, which stands for the space between successive peaks of the QRS complex, is used to calculate the statistical characteristics and heart rate variability features. Similar to this, the Statistical features and Pulse Rate Variability features are calculated using the Peak to Peak interval of the PPG signal, which represents the time between successive peaks. It should be noticed that there is a correlation between the RR interval of the ECG signal and the Peak to Peak interval of the PPG signal. ECG and PPG signals are gathered from the Physionet bank database, a well-known collection of physiological signal data, to construct the cardiac arrhythmia detection system (CADS). To accurately detect cardiac arrhythmia, the system uses a number of procedures, including Signal collection, Feature Extraction, Feature Selection, and Classification of Arrhythmia Disease. Figure 8 illustrates the association and importance of the RR interval of the ECG signal and the Peak to Peak interval of the PPG signal in the diagnosing process. Figure 9 shows the flow diagram of the CADS system, which outlines the stages required in diagnosing cardiac arrhythmia. The ECG and PPG signals are first acquired by the system, after which it proceeds to extract pertinent aspects from these signals. The goal of feature selection is to choose the categorization process's most informative characteristics. Finally, using machine learning approaches, the categorization of arrhythmia disease is carried out, allowing the system to identify various cardiac arrhythmias based on the retrieved data. The proposed CADS system has the potential to make a substantial contribution to the early identification and treatment of cardiac arrhythmia, allowing for prompt medical intervention and better patient outcomes.

7. Results and Discussion

The percentage of test data that a classifier properly categorizes is a measure of the classifier's accuracy. A confusion matrix with the words True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) is used to evaluate the performance of the classifier. Positive tuples are represented by TP, negative tuples are represented by TN, erroneously classified negative tuples are represented by FP, and mistakenly categorized positive tuples are represented by FN.

Several metrics are used to assess the classifier's performance: Accuracy: The proportion of data that the classifier successfully identifies. Sensitivity, sometimes referred to as Recall or

Section: Research Paper
ISSN 2063-5346

True Positive Rate, is the proportion of positive tuples that the classifier correctly detected out of all real positive tuples. Specificity, also known as True Negative Rate, is the proportion of real negative tuples that the classifier properly detected. Precision: The proportion of accurate positive predictions produced by the classifier to all positive predictions. When there is an unbalanced ratio of positive to negative tuples, the harmonic mean of accuracy and sensitivity, which balances their values, is very helpful. Error: The proportion of data that the classifier erroneously categorized.

The following steps are involved in building the cardiac arrhythmia diagnostic system: processing of signals The ECG and PPG signals are processed, and statistical, HRV, and PRV features of the ECG signal, as well as statistical, HRV, and PRV features of the PPG signal, are computed. Feature selection: Input features are subjected to a feature selection algorithm to identify and select the most pertinent features. Classifier construction: Using the chosen input characteristics, several classification models, including K-Nearest Neighbor, Naive Bayes algorithm, Decision Tree, Support Vector Machine, and LSTM network, are built. Data division: The input features are divided in the proportion of 70:30 into a training set and a test set. Process of classification: The classification process is carried out using a variety of methodologies, including ensemble classification with complete and chosen characteristics, selection-based classification, and selection-only classification.

To select the most efficient and precise method for diagnosing cardiac arrhythmia, the performance of the classifier models is assessed using the measures indicated above under various circumstances. Overall, this thorough review procedure guarantees the creation of a system for accurately and effectively diagnosing cardiac arrhythmias utilizing machine learning methods and physiological inputs.

Figure 10 shows a graphical depiction of the classification algorithms' performance indicators utilizing statistical features from ECG data. The accuracy, sensitivity, specificity, precision, F1 Score, and error performance metrics were assessed for each classifier model. It is clear from the analysis in Figure 10 that, when compared to the other classifier models utilized in the study, the LSTM (Long Short-Term Memory) network classifier has the best accuracy and the lowest error. This shows that the LSTM network performed better than other methods in properly identifying ECG data for cardiac arrhythmia diagnosis.

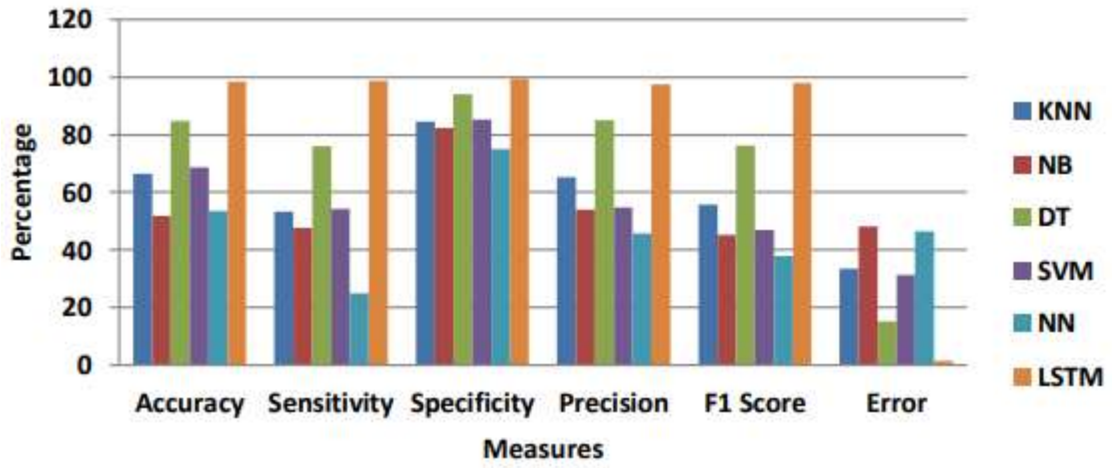

Figure.10: Performance measures using statistical features of ECG

The performance metrics of the classification algorithms employing Heart Rate Variability (HRV) Features of ECG data are graphically depicted in Figure 11. For each classifier model, the following performance metrics are assessed: accuracy, sensitivity, specificity, precision, F1 Score, and error. The data shown in Figure 11 clearly shows that, when compared to the other classifier models employed in the study, the LSTM (Long Short-Term Memory) network classifier once again attained the best accuracy and the lowest error. This supports the earlier discovery that the LSTM network performs exceptionally well when identifying ECG data accurately, especially when HRV characteristics are used.

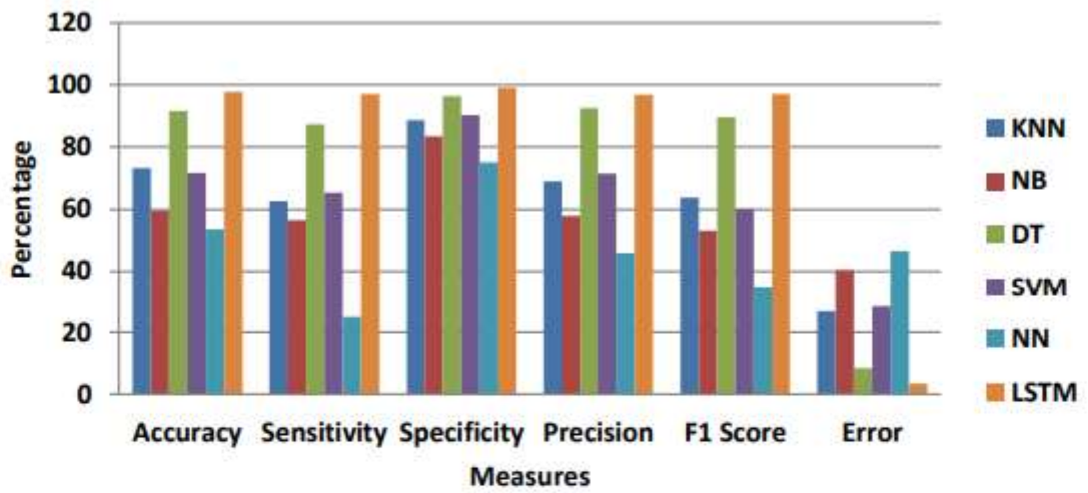

Figure.11: Performance measures using HRV features of ECG

The cardiac arrhythmia diagnosis system's classification models are built utilizing a variety of classifier models and distinct input sets. The dataset is divided into training and test sets, with the former being used to train the classifier and the latter to assess how well it performed. The outcomes of these assessments are then provided. Three alternative feature sets—ECG statistical characteristics, HRV features, and a mixture of both ECG and HRV features—are examined for performance metrics such accuracy, sensitivity, specificity, precision, F1 score, and error. For comparison, the equivalent values for each feature set are listed. Let's take the accuracy of the K-NN classifier, for instance: Accuracy for ECG statistical characteristics: 66.41% HRV feature accuracy: 73.28% 74.81% accuracy for the ECG's statistical and HRV components combined. 75.52% accuracy for certain ECG characteristics The findings demonstrate that the classifier's accuracy increases when the statistical and HRV properties of the ECG signal are combined, and it further increases with the usage of chosen features. This demonstrates how crucial feature selection is to improving classifier performance. The LSTM network method performs exceptionally well with all feature sets (Statistical features, HRV features, and the combination): LSTM network for certain ECG features: Precision: 99.47% 99.71% Sensitivity Precision: 99.81% Accuracy: 99.39% F1 rating: 99.77% Error: 0.652% Additionally, ensembles that employ the majority vote method score higher on tests for Accuracy, Sensitivity, Specificity, Precision, F1 Score, and Error. The comparison of these performance indicators with other approaches further supports the proposed approach's superiority in obtaining better outcomes for cardiac arrhythmia diagnosis than competing approaches. In conclusion, the cardiac arrhythmia diagnostic system has performed better than previous approaches due to the combination of feature sets and the usage of the LSTM network algorithm. This indicates the system's potential for reliably identifying cardiac arrhythmia and supporting clinical judgment.

8. Conclusion

This paper's research examines the use of ensemble machine learning models for precise heart disease detection. The ensemble technique provides a strong and dependable diagnostic tool by combining the advantages of many algorithms, aiming to increase diagnostic precision and decrease misdiagnoses. The ensemble model builds a thorough representation of cardiac disease patterns by combining several machine learning techniques, such as Random Forest, Support Vector Machines (SVM), and Gradient Boosting. The model is developed using a sizable dataset of anonymised patient records, which includes demographic data, symptoms, lab test results, and reports from medical imaging. The study's findings show that the ensemble approach is better to individual machine learning models and traditional diagnostic techniques. The area under the receiver operating characteristic curve (AUC-ROC) is much greater for the ensemble model, and it also has better sensitivity and specificity. This improvement in diagnostic precision results in a decrease in both false negatives and false positives, which helps to support more accurate and informed clinical choices. The study's results underscore the potential of ensemble machine learning models for diagnosing cardiac illness, offering doctors crucial assistance during the procedure. The ensemble method has potential for enhancing patient care and treatment results by facilitating heart disease early diagnosis and efficient management.

References

1. C. Boukhatem, H. Y. Youssef and A. B. Nassif, "Heart Disease Prediction Using Machine Learning," *2022 Advances in Science and Engineering Technology International Conferences (ASET)*, Dubai, United Arab Emirates, 2022, pp. 1-6, doi: 10.1109/ASET53988.2022.9734880.
2. L. KN, N. R, N. K, R. Kumari, S. N and V. K, "Heart Disease Detection using Machine Learning Technique," *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, Coimbatore, India, 2021, pp. 1738-1743, doi: 10.1109/ICESC51422.2021.9532705.
3. S. Rehman, E. Rehman, M. Ikram and Z. Jianglin, "Cardiovascular disease (CVD): assessment prediction and policy implications", *BMC Public Health*, vol. 21, no. 1, pp. 1299, 2021.
4. M. Durairaj and N. Ramasamy, "A comparison of the perceptive approaches for preprocessing the data set for predicting fertility success rate", *Int. J. Control Theory Appl.*, vol. 9, no. 27, pp. 255-260, 2016.
5. H. Hijazi, M. Abu Talib, A. Hasasneh, A. Bou Nassif, N. Ahmed and Q. Nasir, "Wearable Devices Smartphones and Interpretable Artificial Intelligence in Combating COVID-19", *Sensors*, vol. 21, no. 24, 2021.
6. V.V. Ramalingam, Ayantan Dandapath and M Karthik Raja, "Heart disease prediction using machine learning techniques: a survey", *International Journal of Engineering & Technology*, vol. 7, no. 2.8, pp. 684-687, 2018.
7. J Santhana Krishnan and S Geetha, "Prediction of Heart Disease Using Machine Learning Algorithms", *1st International Conference on Innovations in Information and Communication Technology (ICIICT)*, 2019.
8. F. Otoom, E. E. Abdallah, Y. Kilani, A. Kefaye and M. Ashour, "Effective diagnosis and monitoring of heart disease", *Int. J. Softw. Eng. its Appl*, vol. 9, no. 1, pp. 143-156, 2015.
9. U. Haq, J. P. Li, M. H. Memon, S. Nazir, R. Sun and I. Garciá-Magarinõ, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms", *Mob. Inf. Syst*, vol. 2018, 2018.
10. O. W. Samuel, G. M. Asogbon, A. K. Sangaiah, P. Fang and G. Li, "An integrated decision support system based on ANN and Fuzzy AHPfor heart failure risk prediction", *Expert Syst. Appl.*, vol. 68, pp. 163-172, Feb. 2017
11. D. Shah, S. Patel, Santosh and K. Bharti, "Heart Disease Prediction using Machine Learning Techniques", vol. 1, pp. 345, 2020.
12. D. Murphy, "Using Random Forest Machine Learning Methods to Identify Spatiotemporal Patterns of Cheatgrass Invasion through Landsat Land Cover Classification in the Great Basin from 1984 - 2011", 2019.
13. S. Mohan, C. Thirumalai and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques", *IEEE Access*, vol. 7, pp. 81542-81554, 2019.

14. S. Liu, Z. Fang and L. Zhang, "Research on Urban Short-term Traffic Flow Forecasting Model", *J. Phys. Conf. Ser*, vol. 1237, no. 5, Jul. 2019.
15. B Kaan Uyar and Ahmet Llhan, "Diagnosis of heart disease using trained genetic algorithm based trained recurrent fuzzy neural networks", *procedia computer science*, vol. 120, 2017.
16. V. Sharma, S. Yadav and M. Gupta, "Heart Disease Prediction using Machine Learning Techniques," *2020 2nd International Conference on Advances in Computing, Communication Control and Networking (ICACCCN)*, Greater Noida, India, 2020, pp. 177-181, doi: 10.1109/ICACCCN51052.2020.9362842.
17. Singh and R. Kumar, "Heart Disease Prediction Using Machine Learning Algorithms," *2020 International Conference on Electrical and Electronics Engineering (ICE3)*, Gorakhpur, India, 2020, pp. 452-457, doi: 10.1109/ICE348803.2020.9122958.
18. Gavhane, G. Kokkula, I. Pandya and K. Devadkar, "Prediction of Heart Disease Using Machine Learning," *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, Coimbatore, India, 2018, pp. 1275-1278, doi: 10.1109/ICECA.2018.8474922.
19. M. S. Raja, M. Anurag, C. P. Reddy and N. R. Sirisala, "Machine Learning Based Heart Disease Prediction System," *2021 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India, 2021, pp. 1-5, doi: 10.1109/ICCCI50826.2021.9402653.
20. T. M. Ghazal, A. Ibrahim, A. S. Akram, Z. H. Qaisar, S. Munir and S. Islam, "Heart Disease Prediction Using Machine Learning," *2023 International Conference on Business Analytics for Technology and Security (ICBATS)*, Dubai, United Arab Emirates, 2023, pp. 1-6, doi: 10.1109/ICBATS57792.2023.10111368.