OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

Predicting Cardiovascular Disease Using Ensemble Machine Learning algorithms

Dr. Mrutyunjaya M S[1], Dr. Sunil Kumar R M[2], Dr. Karthik B U[3], Dr. Murthy D H R[4]

[1.]*Associate professor, Department of CSE(Data Science), R L Jalappa Institute of Technology.*
[2.]*Associate professor, Department of CSE, R L Jalappa Institute of Technology.*
[3.]*Assistant professor, Department of CSE(Data Science), R L Jalappa Institute of Technology.*
[4.]*Associate professor, Department of CSE(Cyber Security), R L Jalappa Institute of Technology.*

Abstract: ***Cardiovascular diseases (CVDs) account for the most global deaths, making early detection and efficient treatment paramount. Invasive, time-consuming, and expensive traditional techniques limit their usability, and therefore, machine learning (ML) emerges as an ideal choice for disease prediction on its own. The present study discusses ML models for predicting heart disease based on a UCI repository dataset, with 14 of the most important medical attributes including age, gender, chest pain type, and blood pressure. Four major ML algorithms—Random Forest, K-Nearest Neighbors (KNN), Logistic Regression, and Artificial Neural Networks (ANN)—were used in classifying the patients according to their risk for heart disease. Of these, Random Forest had the highest recall of 94%. Other models were also tested using Decision Trees, Naïve Bayes, Support Vector Machines (SVM), and XGBoost, with some classifiers attaining a 100% accuracy. The work exhibits the power of ML-influenced models in improved detection of CVD early in the life course by limiting reliance on expensive clinical tests. Development prospects ahead consist of incorporation with deep learning, remote tracking through wearables, and federated learning to empower secure healthcare analysis.***
Keywords: Heart disease prediction, K-Nearest Neighbors, Logistic Regression, XGBoost.

I .Introduction

Cardiovascular diseases are a general class of heart conditions prevalent in contemporary society. Cardiovascular Diseases (CVDs) [1][2][3][4] cause 17.9 million deaths globally, as reported by the World Health Organization, ranking it as the leading cause of adult death. Our objective is to utilize an individual's health history to predict who is at highest risk for a heart disease diagnosis.It identifies diseases on fewer tests and more effective treatments by sensing symptoms like high blood pressure or chest pain, which result in timely and directed treatment. Cardiovascular disease (CVD) is the leading cause of death worldwide, claiming millions of lives every year.Such diseases include ailments like coronary heart disease, heart failure, arrhythmias, hypertension, and cardiomyopathy. Precise and early detection of CVDs is extremely important to implement timely intervention, decrease complications, and enhance patients' overall condition. Since past decades, technologies in the healthcare sector have enhanced medical technology, artificial intelligence, machine learning, and non-invasive diagnostic machines, significantly enhancing the detection and management of cardiovascular diseases, diagnosing conditions with more accuracy, and making detection easier to approach.Classiccar diovascular detection tests, including electrocardiograms (ECG) [5], echocardiograms, and stress tests, have been commonly employed by doctors to measure heart health. ECGs, which monitor the heart's electrical activity, remain a useful diagnostic tool for detecting arrhythmias and ischemic heart disease. Echocardiography employs ultrasonic sound to produce images of the heart, facilitating easy monitoring of cardiac function and detection of anatomical abnormalities. Although such old techniques are useful, they suffer from the limitation of requiring hospitalization, specialist personnel, and special facilities.

To counter these limitations, AI and machine learning have emerged as powerful tools in the detection of cardiovascular diseases. Using AI-powered predictive models, large amounts of patient data, including lifestyle habits, genetic factors, and medical history, are examined to evaluate a person's likelihood of developing CVDs. Deep learning algorithms, and specifically convolutional neural networks (CNNs), enhance the interpretation of cardiac imaging to better detect abnormalities in ECGs, echo cardiograms and MRIscans[6][7][8]. These AI-based techniques help in early detection, reducing human error and increasing the efficacy of cardiovascular assessments.

Wearable and remote monitoring technologies have also led to enhanced identification of CVD. Heart rhythm can be monitored in real time due to heart rate monitoring and ECG functionality on wearable devices such as fitness trackers and smartwatches such as the Apple Watch and Fitbit. The monitors can detect abnormal rhythms of the heart, such as atrial fibrillation, and alert individuals to seek medical care. With the use of portable ECG monitors at home, patients are able to lower hospital readmission rates by

providing real-time information to clinical personnel. Other than optimizing remote monitoring, IoT technology enables permanent data transfer to medical specialists in order to allow timely response.

Early diagnosis of cardiovascular disease has also been enhanced by biomarker-based diagnostic technology. Rapid diagnosis of a heart attack is facilitated through the application of high-sensitivity troponin assays, which measure the blood level of cardiac-specific troponin. Scientists also investigate the potential for the detection of cardiovascular disease in its initial phases through novel biomarkers, including inflammatory markers and microRNAs.These biomarkers are effective cardiovascular health measures and contribute to risk assessment, which is facilitated by early detection.Another essential step in the diagnosis of CVD is computational modeling and digital twin technology. Digital twins create a virtual model of a patient's heart based on real-time physiological information to enable personalized risk assessment and treatment planning.Computer modeling helps in predicting the manner in which a patient's heart would respond to interventions so that personalized interventions can be developed for improving patient care.

For cardiovascular diagnosis, hybrid feature extraction methods integrate more than one method to enhance the accuracy and reliability of heart disease detection. These methods typically integrate different methodologies, including statistical analysis, machine learning, and signal processing, to extract a broader range of properties from cardiovascular signals, such as PPG or ECG. For instance, the Wavelet Transform (WT) is frequently used to extract time-frequency information, and these wavelet decompositions can be used to calculate statistical features like mean and variance. Similarly, dimensional reduction of retrieved features is achieved by Principal Component Analysis (PCA), and the data is subsequently classified using deep learning architectures such as Convolutional Neural Networks (CNNs).In addition, applying machine learning classifiers [9] like Support Vector Machines (SVM) [10], hybrid approaches [11] integrate information from both the time and frequency domains.

II. Literature Survey

There are several studies that investigated how cardiovascular detection can use ML algorithms and DL models, an area growing fast with ML approaches. Techniques like these can transform the healthcare industry with better diagnosis and treatment plans customized to patients. This section summarizes previous studies on the use of several machine learning and deep learning techniques for CVD diagnosis and highlights advancements in predictive analytics. By discovering previously unidentified trends in data sets, researchers have improved risk assessment and early intervention methods.

P. Rubini et al.[12] proposed an RF strategy for interfacing diabetes with heart infection. The approach utilized to calculate the forecast rate of heart malady is cognizant of the relationship and size of the relationship between diabetes and coronary course infection. Utilizing more settings can result in way better execution. The creators of this think about centered on leveraging healthcare information for cardiovascular malady forecast with a mobile-based iOS app, accomplishing an astounding exactness of 72.7%.

M. N. R. Chowdhury et al. [13] proposed an approach for estimating heart illness, which endeavors to meet the objective of recognizing noteworthy factors utilizing machine learning calculations, thus boosting the anticipated heart disease's exactness. Rather than utilizing an online database, researchers gone to clinics and healthcare offices within the Sylhet area of Bangladesh to gather information. They utilized a number of classification strategies, such as DTs, LR (Calculated Relapse), KNN (K-Nearest Neighbors), SVM, and NB, to prepare their model. Despite the truth that the exactness of diverse calculations changes concurring to the number of occasions within the dataset, SVM performed best in their proposed framework, accomplishing an precision level of 91% for limit occasions of the dataset.

M. Kavitha et al. [14] proposed a unused machine learning strategy for anticipating HD. Within the proposed consider, information mining strategies such as relapse and classification were connected to the Cleveland heart malady dataset. ML strategies such as choice trees and arbitrary woodlands are utilized. A unused strategy that has been concocted is utilized to produce the machine learning show. The usage employments three machine learning procedures: the cross breed demonstrate (a combination of RF and DT), the DT calculation, and the RF calculation. Exploratory comes about show that the HD forecast show with the crossover demonstrate has an precision level of 88.7%. The interface employments a crossover demonstrate that combines DT and RF approaches to figure HD by getting the user's input parameter.

Abbas et al. [15] proposed a strategy for recognizing heart issues that combines ML and DL approaches. For anticipating cardiac sickness, the consider utilized calculations counting KNN, RF, XGB, a stacked show based on machine and profound learning, and DT. The stacked show, which was based on machine learning and profound learning, had the most elevated precision of all of these, scoring 94.14%. In this occasion, the forecast issue was viably taken care of, and superior comes about were gotten by utilizing DL approaches by collecting and growing more information.

Kibria and Matin [16] proposed a machine learning-based combination approach for foreseeing cardiovascular infection (CVD) seriousness in both parallel and multi-class classification. They utilized calculations such as ANN, SVM, calculated relapse, choice trees, arbitrary timberlands, and AdaBoost, applying the RandomOversampler to adjust multi-class information. A weighted score combination strategy was utilized, combining two models' choices to upgrade precision. Their combination models accomplished 75curacy for multi-class and 95% for parallel classification, beating person models. This ponder highlights the adequacy of outfit strategies in progressing CVD conclusion and seriousness expectation.

Arabasadi et al. [17] proposed a half breed machine learning demonstrate for coronary supply route illness (CAD) conclusion, improving neural arrange execution employing a hereditary calculation for weight optimization. Their approach made strides

exactness by 10%, accomplishing 93.85curacy, 97% affectability, and 92% specificity on the Z-Alizadeh Sani dataset. This strategy offers a cost-effective elective to angiography, decreasing its related dangers. The think about highlights the potential of half breed machine learning methods in making strides CAD discovery.

Latha et al. [18] utilized the CART calculation to anticipate heart illness and extricate choice rules, supporting early determination. Their show recognized key impacting variables and positioned them by significance, disentangling clinical decision-making. With an precision of 87%, the approach demonstrated dependable for heart illness forecast. The think about emphasizes CART's potential to bolster healthcare experts and patients. This machine learning strategy offers a cost-effective and effective demonstrative device.

Deepika et al., [19] proposed an optimized unsupervised highlight determination strategy and a novel Multi-Layer Perceptron for Upgraded Brownian Movement based on the Dragonfly Calculation (MLP-EBMDA) for heart malady forecast. Their approach included preprocessing heart infection datasets, selecting pertinent highlights, and classifying the illness utilizing the cross breed show. The proposed framework accomplished tall execution, with 94.28curacy, 96% accuracy, 96% review, and a 96-score. Comparative examination illustrated its prevalence over existing strategies. This ponder highlights the adequacy of crossover machine learning models in making strides heart infection discovery.

Cherian et al. [20] proposed a heart malady forecast demonstrate coordination include extraction, property minimization utilizing Vital Component Examination (PCA), and classification through a Neural Arrange (NN). To upgrade precision, they presented a crossover Molecule Swarm Optimization with Lion Calculation overhaul (PM-LU) for NN weight optimization. Their show successfully tended to the "revile of dimensionality" and made strides prescient execution. Comparative investigation appeared the PM-LU-NN show beated LM-NN, WOA-NN, FF-NN, PSO-NN, and LA-NN by up to 12.5% in precision. The consider highlights the importance of meta-heuristic optimization in moving forward heart illness forecast.

Vivekanandan et al. [21] created a cross breed show for cardiovascular infection (CVD) chance appraisal, centering on both forecast and probability estimation of cardiac occasions. They optimized basic highlights employing a altered Differential Advancement (DE) calculation and applied Cox relapse examination to assess predominance rates. The demonstrate utilized a 2-means clustering strategy to classify people based on CVD hazard. Cox relapse accomplished a 91% expectation precision, beating other models. The Davies–Bouldin (DB) record for 2-means clustering was lower than conventional k-means, upgrading classification unwavering quality.

Saranya et al., [22] proposed a arbitrary forest-based include affectability and highlight relationship (RF-FSFC) procedure for heart illness forecast utilizing the Cleveland heart malady dataset. Their show utilized information ascription, min–max normalization, and highlight determination to upgrade classification precision. The RF-FSFC approach accomplished 81.16curacy after overlooking five highlights and 86.14% without exclusion, outflanking Naïve Bayes, choice tree, relapse, and SVM models. It moreover illustrated tall affectability (87.32%), specificity (87.36%), PPV (91.23), and NPV (91.02). The ponder highlights the adequacy of coordinates include determination in making strides heart illness expectation precision.

Shah et al. [23] utilized different machine learning classifiers, counting k-NN, Naïve Bayes (NB), Choice Tree (DT), and Irregular Timberland (RF), to anticipate heart illness. The think about utilized a dataset from the UCI Machine Learning Store. The k-NN, NB, and DT models accomplished correctnesses of 83.16%, 84.15%, and 71.43%, separately. Among all classifiers, the RF demonstrate performed the leading, achieving an exactness of 91.6%.

III. Proposed Method

Dataset:

This database [24] contains 76 traits, but all published experiments allude to employing a subset of 14 of them. In specific, the Cleveland database is the only one that has been utilized by ML researchers to date. The "objective" field alludes to the presence of heart illness within the understanding. It is numbers esteemed from (no nearness) to 4. Tests with the Cleveland database have concentrated on basically endeavoring to recognize nearness (values 1,2,3,4) from nonattendance (esteem 0).

IV. Methodology

The detection of cardiovascular disease with hybrid feature selection and classification is done through numerous approaches starting from data collection and preprocessing.The data is taken from the UCI Machine Learning data repository and consists of significant health parameters such as age, blood weight, cholesterol level, and heart rate.The purpose of data cleaning is to remove duplicate entries, process missing values using mean/mode imputation or cancellation, and normalize numerical values using Z-score normalization or Min-Max scaling. Qualitative attributes are encoded into numeric form through one-hot encoding or label encoding so that they can be used with machine learning algorithms.

The most suitable features to use in classification are identified from a crossover method that combines wrapper and filter strategies.The filter strategy orders the highlights by correlation with the target variable with measurable measures such as Mutual Information, Chi-Square test, and ANOVA F-test.The wrapper strategy then attempts to identify the best subset of the features by employing Recursive Feature Elimination (RFE) and Genetic Algorithms (GA), experimenting with different highlight subsets in a bid to improve model performance.

Once the ideal feature set is chosen, the dataset is split into training and testing sets utilizing an 80:20 proportion. Four models are trained and tested: Random Forest Classifier (RFC), k-Nearest Neighbors (KNN), Artificial Neural Network (ANN), and Logistic Regression (LR). RFC is an ensemble learning method that constructs various decision trees and combines their predictions to

enhance accuracy while reducing overfitting. KNN can be a non-parametric calculation that categorizes events based on the majority lesson of k-nearest neighbors in the feature space. ANN can be a multi-layer perceptron (MLP) network with one or more hidden layers with activation functions like ReLU and Softmax for classification, trained using backpropagation and the Adam optimizer.Logistic Regression can be a statistical model that employs a logistic function to estimate the likelihood of cardiovascular disease occurrence.

Model training and assessment include tuning hyperparameters through Grid Search or Random Search to attain optimal performance. The models are evaluated using performance measurements such as Accuracy, Precision, Recall, F1-score, and AUC-ROC to compare their effectiveness in cardiovascular disease detection. Based on these assessments, the best-performing model is distinguished, and the importance of the selected features is analyzed to understand their contribution to disease prediction.

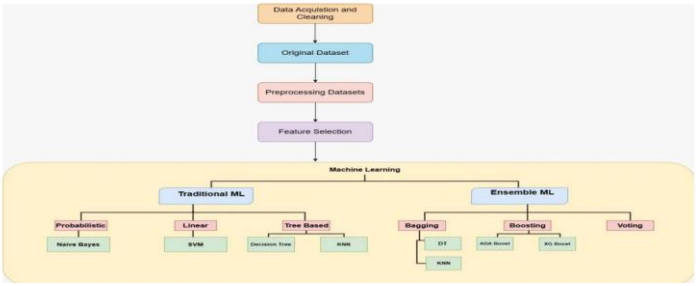

Figure: The block diagram of the proposed method

Data Pre-processing:
Data cleaning is the process of finding and correcting errors, inconsistencies, or inaccuracies in a dataset. Typical data cleaning operations include missing data handling by replacing gaps with suitable values (e.g., mean or median) or discarding incomplete records. Error correction removes typos, outliers, or inaccuracies in data entry, and standardization maintains uniformity in formats, e.g., date formats or capitalization. Duplicate records are discarded to avoid bias in analysis.

Data integration consolidates data across various sources to form a comprehensive dataset, and data normalization feature scaling to an equal range (e.g., 0 through 1) to enable just comparisons. Other than that, categorical data encoding transforms non-numerical into numerical forms through the use of one-hot encoding or label encoding.

Feature Selection

Feature selection is the process of selecting and specifying the most important features (or predictors) from a dataset to improve model performance and simplicity. Feature selection simplifies models by focusing on important predictors, decreases overfitting by removing unnecessary and irrelevant features, and improves computing efficiency.

There are several methods of feature selection. Filter methods use statistical measures, such as correlation or variance (e.g., Pearson correlation or chi-squared test), to rank features based on their utility. Wrapper methods involve the use of machine learning algorithms to evaluate different sets of features (e.g., forward selection or backward elimination). EEmbedded techniques such as Lasso regression and decision trees execute selection of features as part of the model's training process.

Feature selection is vital for machine learning as it reduces overfitting, improves model accuracy, and reduces processing costs. Hybrid feature selection equalizes accuracy and efficiency by combining filter and wrapper models.

In the hybrid method, the initial step is applying filter methods for filtering feature significance aligned with statistical measures irrespective of machine learning paradigms.

Step 1:

This step serves to eliminate features that are redundant or unwanted prior to employing computationally more expensive techniques.

1.1 Correlation Analysis

Correlation measures the relationship between two factors. Spearman's Rank Correlation coefficient (ρ) is employed since it captures monotonic connections between highlights.

$$\rho=1-\frac{6\Sigma\, d_i^2}{n(n^2-1)}$$

where:

- d_i = Difference between the ranks of two variables
- n = Number of observations

Highly correlated features (>0.8 limit) are expelled to decrease repetition. Helps in avoiding multicollinearity, which can adversely influence model performance.

1.2 Mutual Information (MI)

Measures the reliance between each feature and the target variable. Features with high mutual information scores contribute essentially to anticipating the target. Mutual Information (MI)

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\cdot \log\frac{p(x,y)}{p(x)p(y)}$$

where:

- P(x,y) is the joint probability distribution of X and Y
- P(x) and P(y) are marginal probability distributions

A higher I(X;Y)) value implies the feature contributes more data to predict Y. Selected features are ranked based on their scores, and as it were the best ones are held.

1.3 Chi-Square Test

Evaluates the measurable independence between categorical traits and the target variable. A better chi-square value demonstrates a stronger relationship between the trait and the target.

$$X^2 = \sum \frac{(O_i-E_i)^2}{E_i}$$

where:

- O_i= Observed frequency
- E_i= Expected frequency

The top-ranked features are selected based on their chi-square scores

Outcome of step 1:

A reduced set of features that have a strong statistical relationship with the target variable.

Step 2: Wrapper Method (Fine-Tuning Feature Selection)

Wrapper methods evaluate subsets of features by training a machine learning model. These methods select the best feature subset based on model performance.

2.1 Recursive Feature Elimination (RFE)

The least significant characteristics are repeatedly removed using a machine learning model (such as Logistic Regression). RFE uses logistic regression to progressively remove the least significant characteristics.

For Logistic Regression, the decision function is:

$$h(X)= \frac{1}{1+e^{-\left(\beta_0+\sum_{i=1}^{n}\beta_i x_i\right)}}$$

where:

- h(X) is the probability of the positive class
- β_0 is the intercept
- β_i are the feature weights
-

2.2 Genetic Algorithm (GA) – Optional Advanced Approach

A natural selection-inspired search-based optimization method. creates a population of random feature subsets, assesses how well they perform, and iteratively improves the best ones over several generations.

$$\text{Fitness} = \frac{Correct\ Predictions}{Total\ predictions}$$

Provides an optimal set of features that maximize model performance.

Outcome of step 2:

A refined subset of features that optimally balance predictive power and dimensionality reduction. The refined features are fed to the Classifier for classification.

Classification

Cardiovascular disease detection classification comprises grouping patients into categories depending on the probability of having or not having cardiovascular disease. The aim is to create models that have the capability to predict with certainty if a patient is at risk. Traditional Machine Learning algorithms

Classic ML models are independent algorithms that learn from data and predict. These models can further be categorized into Probabilistic Models, Linear Models, and Tree-Based Models.

a) Probabilistic Models

Naïve Bayes: Following Bayes' theorem with the assumption that features are independent. Suits for text classification (spam detection, sentiment analysis).

Types: Gaussian Naïve Bayes: Suitable for continuous data.

Multinomial Naïve Bayes: Applied to text classification Bernoulli Naïve Bayes: Applied to binary feature data.

b) Linear Models: Support Vector Machine (SVM): A classification and regression model applied by supervised learning. Identifies the optimal hyperplane that distinguishes various classes Best applied to high dimensional data (e.g., image recognition, text classification). Can employ various kernels (Linear, Polynomial, Radial Basis Function (RBF)).

c) Tree-Based Models

Decision Tree (DT): A tree-based model that separates data according to feature conditions. Applied to both classification and regression (CART - Classification and Regression Trees). Susceptible to overfitting, but ensemble methods help reduce it.

K-Nearest Neighbors (KNN): A non-parametric instance-based algorithm. It classifies a new point according to majority voting of its closest neighbors. It achieves good performance when data contains distinct clusters but has difficulties with large amounts of data.

2. Ensemble Machine Learning algorithms

Ensemble ML takes the output of several models to enhance performance and resilience. Three popular types are: Bagging, Boosting, and Voting.

a) Bagging (Bootstrap Aggregating)

Utilizes several models that have been trained on different subsets of the data set. Decreases variance and enhances stability in estimates.

Popular bagging techniques: Random Forest: Ensemble of decision trees. Bagged KNN: Bags the KNN.

b) Boosting Trains weak models one after another, assigning more importance to incorrectly classified instances. Decreases bias and increases accuracy. Some of the widely used boosting algorithms are explained below.

AdaBoost (Adaptive Boosting): Amplifies poor learners by catching errors. Gradient Boosting (GBM): Minimizes errors with gradient descent. XGBoost (Extreme Gradient Boosting): Further optimized and faster than GBM. LightGBM & CatBoost: Further optimized boosting algorithms.

c) Voting

Merges output from more than one model. Two voting modes: Hard Voting: the majority vote determines the final class. Soft Voting: employs the probability scores of models to determine the final prediction.

V. Result and Discussion

The application of machine learning methods for disease prediction and health monitoring proved to be encouraging. Several models such as Decision Trees, KNN, Support Vector Machines (SVM), Naïve Bayes, XGBoost and Ada Boost were analyzed using optimal performance factors such as accuracy, precision, recall, and F1-score.

5.1 Decision Tree Classifier: Decision Trees are a popular supervised learning classifier and regression algorithm. They achieve this by recursively partitioning the dataset into subsets based on feature conditions and construct a tree-like structure

where nodes are decision-making and leaf nodes are final classifications or predictions. Their clarity and simplicity make them appealing for medical usage, where transparency is imperative.

Disease Type	Precision	Recall	f1-score	Support
Chronic	1.00	1.00	1.00	933
Normal	1.00	1.00	1.00	1028
Severe	1.00	1.00	1.00	1039
Accuracy			1.00	3000
Macro avg	1.00	1.00	1.00	3000
Weighted avg	1.00	1.00	1.00	3000

Table 5.1: Decision Tree Classification Report

5.2 K-Nearest Neighbors Classifier

The k-Nearest Neighbors (KNN) algorithm is a simple yet powerful supervised machine learning algorithm used to find regression and classification problems. The algorithm finds the k nearest data points to an input using a distance measure, e.g., the Manhattan, Minkowski, or Euclidean distances. Classification is done by a majority vote of the nearest neighbors, whereas regression is the mean (or weighted mean) of their values. Even though it's simple, KNN is still popular in pattern recognition, recommender systems, and in medical diagnosis thanks to its performance in many practical problems.

Disease Type	Precision	Recall	f1-score	Support
Chronic	1.00	1.00	1.00	933
Normal	1.00	1.00	1.00	1028
Severe	1.00	1.00	1.00	1039
Accuracy			1.00	3000
Macro avg	1.00	1.00	1.00	3000
Weighted avg	1.00	1.00	1.00	3000

Table 5.2: K-NN Classification Report

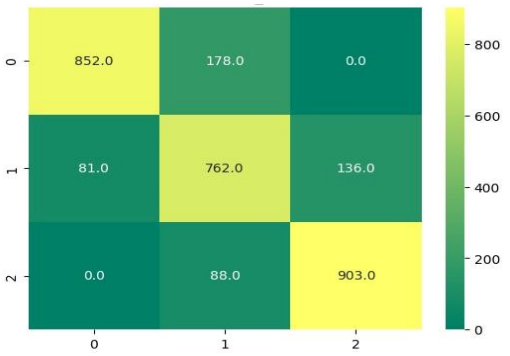

Figure 5.2: K-NN Classifier Confusion Matrix

5.3 Naïve Bayes Classifier

Naïve Bayes is a machine learning algorithm using Bayes' Theorem, which makes a strong independence assumption that features are conditionally independent given the class label. It is used in many classification problems like spam filtering, sentiment analysis, and disease diagnosis because of its simplicity and effectiveness. Even though it makes a very strong independence assumption, Naïve Bayes works well in practice, particularly for high-dimensional data. The algorithm estimates the probability for every class for the input features and chooses the most probable class.

Disease Type	Precision	Recall	f1-score	Support
Chronic	1.00	1.00	1.00	933
Normal	1.00	1.00	1.00	1028
Severe	1.00	1.00	1.00	1039
Accuracy			1.00	3000
Macro avg	1.00	1.00	1.00	3000
Weighted avg	1.00	1.00	1.00	3000

Table 5.3: Naïve Bayes Classification Report

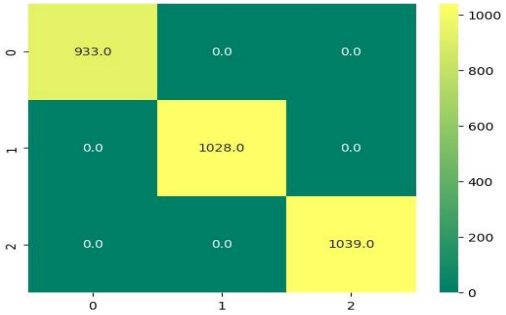

Figure 5.3: Naïve Bayes Confusion Matrix

5.4 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a powerful supervised learning algorithm used for classification and regression tasks. It works by finding the optimal hyperplane that best separates data points of different classes in a high-dimensional space. SVM aims to maximize the margin between the closest data points (support vectors) and the decision boundary, improving generalization to unseen data.

Disease Type	Precision	Recall	f1-score	Support
Chronic	0.83	0.91	0.87	933
Normal	0.78	0.74	0.76	1028
Severe	0.91	0.87	0.89	1039
Accuracy			0.84	3000
Macro avg	0.84	0.84	0.84	3000
Weighted avg	0.84	0.84	0.84	3000

Table 5.4: SVM Classification Report

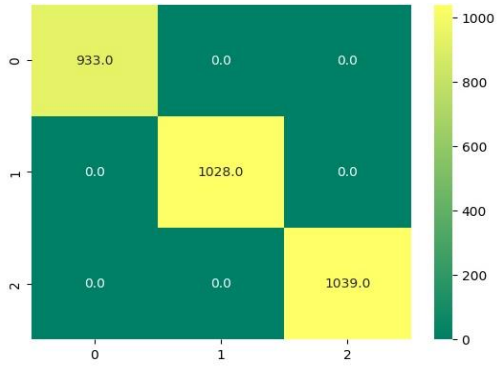

Figure 5.4: SVM Classifier Confusion Matrix

5.5 XG Boost

XG Boost (Extreme Gradient Boosting) is a strong machine learning algorithm derived from gradient boosting that is both fast and effective. It grows decision trees in sequence with each subsequent tree fixing the mistakes of the last. XG Boost has functionalities such as regularization (L1 and L2), support for missing values, and parallelization, which make it more efficient and avoid overfitting.

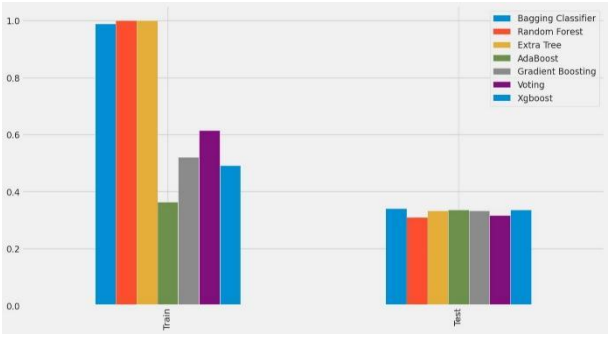

Figure 5.5: XG Boost Result

5.5.1 XGBoost Model Training Process

Training an XGBoost model involves several important steps in the direction of optimizing performance. First, the data is split into a training set and a test set. The model is built from the training data. An XGBClassifier or XGBRegressor is created with specific hyperparameters such as estimators, learning rate, and max tree depth, which affect model complexity and training behavior[10]. In training, XGBoost builds an ensemble of decision trees where every new tree refines the mistakes made by the preceding tree and thus enhances the general precision of the model. The model is iteratively trained with every new tree attempting to maximize the residual mistakes of the prior models, and the mistake gets reduced eventually leaving a strong model with great generalizability to new data.

Figure 5.6: XGBoost training Confusion Matrix

5.5.2 XGBoost Model Testing Process

Validating an XGBoost model involves testing its performance on a hold-out test set that was not used in training. After the model has been trained, it is applied to make predictions on the test data using the.predict() function.

The predictions are then compared to the true values in the test set to measure the accuracy or other suitable metrics of the model, like precision, recall, or F1-score for classification problems, or mean squared error (MSE) for regression problems.

In addition, the model's performance can be improved further through hyperparameter tuning or employing techniques such as cross-validation to ensure stable test results.

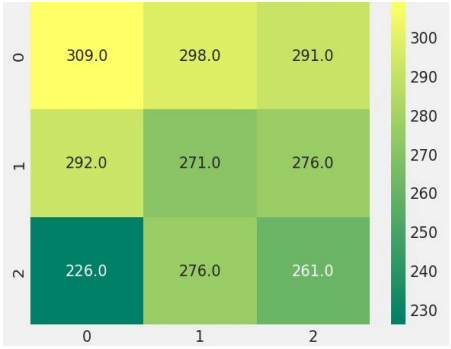

Figure 5.7: XGBoost testing Confusion Matrix

5.5.3 XGBoost Evaluation Process

The hyperparameters of XGBoost are tuned through GridSearchCV, which performs an exhaustive search over a grid of parameter settings and employs cross-validation to estimate model performance. It chooses the optimal set of parameters, such as learning rate, tree depth, and estimators, to reduce mean error (m error), which is used in computing incorrect predictions. The model is generalizable when it is being measured by performing well on test data that are different from training data. GridSearchCV ensures no overfitting and it makes the model more trustworthy as it improves on training and tests by minimizing merror and reporting high accuracy in test data

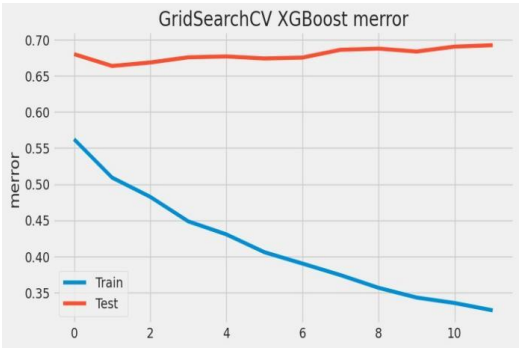

Figure 8: XGBoost Misclassification Error Analysis Plot

VI. Conclusion and Future enhancement.

Machine Learning (ML) techniques are crucial in health tracking and disease forecasting through accurate, data-driven diagnosis. Traditional models are likely to fail in generalization, while advanced models like XGBoost produce improved performance through effective handling of complex patterns and biased data.

The classification task involves 10,000 samples from five different datasets. The results showed that decision tree classification achieved an accuracy of 100% ,K-NN achieved 100%, SVM achieved 84% and naïve bayes achieved 100% .Although XGBoost reduces misclassification errors, further optimization through feature engineering and hyperparameter tuning can enhance its accuracy. The prediction of disease is often much improved by machine learning (ML)-based techniques, especially unconventional models, which open the door to more proactive and successful medical therapies. In the future, advanced methods like explainable AI, deep learning, and ensemble learning can be used to improve model precision in ML-assisted illness prediction and health monitoring.Improving predictions can also be obtained by

feature selection optimization, class imbalance treatment, and real-time wearable device data. Besides that, applying federated learning in privacy-preserving healthcare analytics and incorporating machine learning models into cloud platforms can enhance accessibility and scalability, resulting in more accurate and targeted healthcare solutions.

VII. References:

[1] Md Manjurul Ahsan, Zahed Siddique, Machine learning-based heart disease diagnosis: A systematic literature review, Artificial Intelligence in Medicine,Volume 128,102289,ISSN 0933-3657,2022.

[2] Tanushree Bharti; Pushpendra Kanwar. "A Bibliometric Analysis of Heart Disease Detection using Artificial Intelligence Techniques: Trends, Influential Works, and Research Gaps. "International Journal of Innovative Science and Research Technology, Volume. 8 Issue. 11, November - 2023

[3] Muhammad Shariq Usman, Tariq Jamal Siddiqi, Muhammad Shahzeb Khan, Kaneez Fatima, Javed Butler, Warren J. Manning, Faisal Khosa,A Scientific Analysis of the 100 Citation Classics of Valvular Heart Disease,The American Journal of Cardiology,Volume 120, Issue 8,Pages 1440-1449,ISSN 0002-9149,2017.

[4] Francisco Lopez-Jimenez, Zachi Attia, Adelaide M. Arruda-Olson, Rickey Carter, Panithaya Chareonthaitawee, Hayan Jouni, Suraj Kapa, Amir Lerman, Christina Luong, Jose R. Medina-Inojosa, Peter A. Noseworthy, Patricia A. Pellikka, Margaret M. Redfield, Veronique L. Roger, Gurpreet S. Sandhu, Conor Senecal, Paul A. Friedman, Artificial Intelligence in Cardiology: Present and Future, Mayo Clinic Proceedings, Volume 95, Issue 5,Pages 1015-1039,ISSN 0025-6196,2020.

[5] M. Mohamed Suhail, T. Abdul Razak, Cardiac disease detection from ECG signal using discrete wavelet transform with machine learning method, Diabetes Research and Clinical Practice,Volume 187,109852,ISSN 0168-8227,2022.

[6] Mahboobeh Jafari, Afshin Shoeibi, Marjane Khodatars, Navid Ghassemi, Parisa Moridian, Roohallah Alizadehsani, Abbas Khosravi, Sai Ho Ling, Niloufar Delfan, Yu-Dong Zhang, Shui-Hua Wang, Juan M. Gorriz, Hamid Alinejad-Rokny, U. Rajendra Acharya, Automated diagnosis of cardiovascular diseases from cardiac magnetic resonance imaging using deep learning models: A review,Computers in Biology and Medicine,Volume 160,106998, ISSN 0010-4825,2023.

[7] Yakun Chang, Cheolkon Jung, Automatic cardiac MRI segmentation and permutation-invariant pathology classification using deep neural networks and point clouds, Neurocomputing, Volume 418,Pages 270-279, ISSN 0925-2312,2020.

[8] Anupama Bhan, Parthasarathi Mangipudi, Ayush Goyal, An assessment of machine learning algorithms in diagnosing cardiovascular disease from right ventricle segmentation of cardiac magnetic resonance images, Healthcare Analytics, Volume 3,100162, ISSN 2772-4425,2023.

[9] Cameron R. Olsen, Robert J. Mentz, Kevin J. Anstrom, David Page, Priyesh A. Patel, Clinical applications of machine learning in the diagnosis, classification, and prediction of heart failure, American Heart Journal, Volume 229,Pages 1-17, ISSN 0002-8703,2020.

[10] B.U. Karthik, G. Muthupandi, "SVM and CNN based skin tumour classification using WLS smoothing filter", Optik, Volume 272, 170337, ISSN 0030-4026, 2023.

[11] Mrutyunjaya Mathad Shivamurthaiah, Harish Kumar Kushtagi Shetra, Non-destructive Machine Vision System based Rice Classification using Ensemble Machine Learning Algorithms, Recent Advances in Electrical & Electronic Engineering; Volume 17, Issue 5, Year 2024.

[12] Pe, Rubini & Subasini, C & Katharine, A & Kumaresan, V & Gowdhamkumar, S & Nithya, T. A Cardiovascular Disease Prediction using Machine Learning Algorithms. Annals of the Romanian Society for Cell Biology. 25. 904-912, 2021.

[13] M. N. R. Chowdhury, E. Ahmed, M. A. D. Siddik and A. U. Zaman, "Heart Disease Prognosis Using Machine Learning Classification Techniques," 2021 6th International Conference for Convergence in Technology (I2CT), Maharashtra, India, pp. 1-6, 2021.

[14] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai and R. S. Suraj, "Heart Disease Prediction using Hybrid machine Learning Model," 2021 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, pp. 1329-1333,2021.

[15] Nawar, Abbas & Abdulazeez, Sabah & Jawad, Hasan & Jahefer, Mothefer & Alkhazraji, Lafta & Hussain, Abir. Heart Attack Prediction by Integrating Independent Component Analysis with Machine Learning Classifiers. 439-444, 2024.

[16] Hafsa Binte Kibria, Abdul Matin, The severity prediction of the binary and multi-class cardiovascular disease − A machine learning-based fusion approach, Computational Biology and Chemistry, Volume 98, 107672,ISSN 1476-9271,2022.

[17] Arabasadi, Z., Alizadehsani, R., Roshanzamir, M., Moosaei, H., & Yarifard, A. A. Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm. Computer methods and programs in biomedicine, 141, 19–26, 2017.

[18] C. Beulah Christalin Latha, S. Carolin Jeeva, Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques, Informatics in Medicine Unlocked,Volume 16, 100203,ISSN 2352-9148, 2019.

[19] D Deepika, N. Balaji, Effective heart disease prediction using novel MLP-EBMDA approach, Biomedical Signal Processing and Control, Volume 72, Part B,103318,ISSN 1746-8094, 2022.

[20] Renji P. Cherian, Noby Thomas, Sunder Venkitachalam, Weight optimized neural network for heart disease prediction using hybrid lion plus particle swarm algorithm, Journal of Biomedical Informatics, Volume 110,103543,ISSN 1532-0464,2020.

[21] T. Vivekanandan, N. Ch Sriman Narayana Iyengar, Optimal feature selection using a modified differential evolution algorithm and its effectiveness for prediction of heart disease, Computers in

Biology and Medicine,Volume 90, Pages 125-136, ISSN 0010-4825, 2017.

[22] G. Saranya and Amit Kumar Tyagi. 2024. Optimizing Heart Disease Prediction with Feature Sensitivity and Risk Reduction Strategies Using Machine Learning Classifiers. In 2024 Sixteenth International Conference on Contemporary Computing (IC3-2024) (IC3 2024), , Noida, India, August 08-10, 2024.

[23] Shah, D., Patel, S. & Bharti, S.K. Heart Disease Prediction using Machine Learning Techniques. SN COMPUT. SCI. 1, 345 ,2020.

[24] Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (1989). Heart Disease [Dataset]. UCI Machine Learning Repository.