Enhancing Heart Disease Prediction and Analysis: An Efficient Voting Ensemble model

Sasank Talapaneni
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram
Andhra Pradesh, India.
2100030529cseh@gmail.com

Chaitanya Sai Kota
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram
Andhra Pradesh, India.
2100030085cseh@gmail.com

Narahari Yalagala
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram
Andhra Pradesh, India.
2100030712cseh@gmail.com

Rakesh Nunna
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram
Andhra Pradesh, India.
2100031947cseh@gmail.com

Radha Mothukuri
Associate Professor
Department of Computer Science and Engineering
Koneru Lakshmaiah Education Foundation
Vaddeswaram
Andhra Pradesh, India.
radha@kluniversity.in

2024 International Conference on Communication, Computer Sciences and Engineering (IC3SE) | 979-8-3503-6684-6/24/$31.00 ©2024 IEEE | DOI: 10.1109/IC3SE62002.2024.10593602

***Abstract* - The prediction of heart disease is of significant importance in the domain of healthcare, where accurate diagnosis can save lives. In this paper, we propose a proficient model for prediction of heart disease using a Voting Ensemble technique. We employ a integration of Logistic Regression, Decision Tree classifier, and Random Forest classifiers to create a robust predictive model. This ensemble method has shown encouraging results in terms of accuracy and performance. In our experimental analysis, we demonstrate the advantages of our method and offers a comprehensive assessment of its performance on a real-world heart disease dataset. The results highlight the need of ensemble techniques in improving diagnostic accuracy. Our Voting Ensemble model attained an outstanding accuracy of 98% on the test dataset, outperforming individual classifiers. This paper contributes to the advancement of forecasting models for predicting heart disease and provides insights for future investigations in the field of healthcare analytics.**

Keywords—heart disease, ensemble models, Voting ensemble technique, hard voting, soft voting.

I. INTRODUCTION

Heart diseases remains a global health challenge, standing as a leading cause of mortality worldwide. The sheer magnitude of its impact underscores the critical need for prompt and precise diagnosis. According to the reports of World Health Organization (WHO), cardiovascular diseases make up roughly 17.9 million fatalities annually, establishing it as the leading cause of death globally. The heart diseases are highly widespread in both developed and developing nations, transcending geographical and socio-economic-boundaries.

Several factors participate in the prevalence of heart diseases, including unhealthy lifestyle choices, such as sedentary habits, poor dietary practices, and tobacco use. Additionally, genetic predispositions, environmental factors, and underlying medical conditions contributes to the development of heart disease. The complex interplay of these factors makes predicting the likelihood of heart disease a challenging yet essential endeavour for healthcare professionals.

In this scenario, the algorithms of machine learning have gained prominence for their potential to enhance diagnostic accuracy. The domain of healthcare analytics has witnessed notable progress in recent years, propelled by the accessibility of extensive medical data and the development of sophisticated predictive models. These models leverage clinical and demographic features to offer valuable perspectives for early intervention and treatment planning.

This paperwork addresses the challenge of heart disease prediction. We propose the application of a Voting Ensemble technique, which harnesses the collective strengths of three powerful classifiers: Logistic Regression, Decision Tree classifier, and Random Forest classifiers. The main objective of our investigation is to develop a forecasting model that not only meets the requirements but surpasses the effectiveness of individual classifiers, contributing to more reliable diagnostics in the healthcare domain.

In the context of this paper, we introduce a novel approach that employs a Voting Ensemble technique. The voting ensemble method is a form of ensemble learning that integrates the forecasts from numerous base classifiers to make a final prediction. Specifically, we utilize Logistic Regression, Decision Trees, and Random Forest classifiers as our base models. These classifiers were selected for their distinct characteristics and complementary strengths our study not only emphasizes the potential of ensemble techniques in healthcare analytics but also imparts important perspectives into improving diagnostic accuracy in the sector of cardiovascular prediction. The outcomes derived from the work highlight the significance of combining diverse models to enhance prediction accuracy, setting the stage for further advancements in the field.

979-8-3503-6684-6/24/$31.00 ©2024 IEEE

II. Literature Survey

The Voting Ensemble technique operates based on the foundation of combining the collective wisdom of multiple classifiers[1] by utilizing Logistic Regression, Decision Tree classifier, and Random Forest classifier, to achieve the result with highest accuracy and resilience.

These classifiers can be diverse, each having its strengths and weaknesses. By aggregating their predictions, Voting Ensemble seeks to harness the advantages of individual algorithms while mitigating their limitations[2]. This methodology enables greater informed decisions by considering multiple viewpoints in the predictive process.[3]

Voting Ensembles can be categorized into two primary types:

1. ***Hard voting***: In the context of hard voting, the ultimate prediction is determined by taking the majority decision from the individual classifiers. This is particularly useful when combining classifiers with binary predictions.[4]

2. ***Soft voting***: Soft voting considers the probability scores provided by each classifier and selects the class with gives high average probability value. This offers a more fine-grained approach to prediction.[5][6]

TABLE I. Description Of Algorithms With Merits And De-Merits

Algorithm	Description	Advantages	Disadvantages
Logistic Regression	Logistic Regression is an easy to understand and interpretable linear classification algorithm. It calculates probability for a given data point that belongs to particular class. It is a well-suited model that is designed for binary classification purposes and can furnish probability scores for prediction confidence.[7]	Simplicity and interpretability	May not capture complex nonlinear relationships
Decision Trees	Decision Tree classifiers are non-linear classification algorithms that create a tree-like structure by partitioning the dataset into subgroups based on feature conditions. They are recognized for their simplicity and interpretability, making them valuable for visualizing and explaining classification rules.[8]	Ability to partition data into meaningful segments [9]	Prone to excessive model fitting
Random Forests classifier	Random Forests represent an ensemble approach that utilizes several decision trees to enhance predictive accuracy and mitigate overfitting. By aggregating the predictions from individual trees, Random Forests offers a more robust model with the ability to handle complex, high-dimensional data.[10]	Improved generalization through an ensemble approach	Complexity in managing multiple decision trees [11]
Voting Ensemble	Voting Ensemble is a approach that integrates the predictions from numerous individual classifiers to make a final prediction. It can use either hard voting (majority vote) or soft voting (average probability) to predict the conclusive class label. Voting Ensemble leverages the strengths of diverse algorithms, offering improved accuracy and model robustness.[12]	Enhanced accuracy and robustness through a combination of individual classifiers.[13]	Requires additional computational resources [14]

III. Proposed Methodology

The proposed methodology in this research paper introduces an ensemble model to enhance the prediction of heart disease. The methodology starts with the acquisition of a heart disease dataset, which serves as the foundation for subsequent data-driven decision-making. Following data ingestion, comprehensive preprocessing procedures, including feature standardization, are applied to ensure data quality and consistency.

In this framework, three individual classifiers, Logistic Regression, Decision Tree, and Random Forest, are incorporated into a Voting Classifier. This ensemble method harnesses the power of "hard" voting to amalgamate predictions from the constituent classifiers. The methodology employs a train-test split to facilitate robust model development and evaluation, thereby ensuring its adaptability to real-world healthcare scenarios. The accuracy of the ensemble model, serving as a key performance metric, is complemented by supplementary assessments like confusion matrix and a detailed classification report. This approach not only demonstrates the potential for ensemble learning in healthcare analytics but also contributes to more precise and reliable diagnostic solutions for prediction of heart disease.

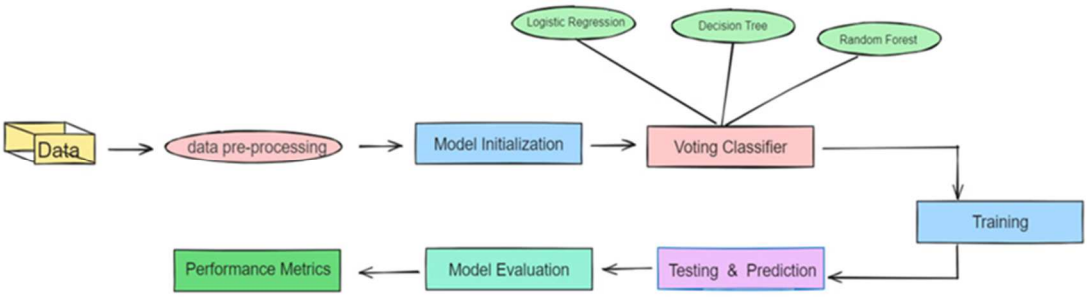

Figure 1. Flowchart of proposed methodology

A. ***Model Combination***: In the realm of model combination, our research leverages a fusion of three robust classifiers: Logistic Regression, Decision Tree classifier, and Random Forest classifier. Each classifier contributes its unique set of decision boundaries and parameter configurations, creating a diverse set of predictive models.

Numerically, let MLR, MDT and MRF represent the Logistic Regression, Decision Tree classifier, and Random Forest classifiers, respectively. The combination is achieved by aggregating the outputs of these models, forming an ensemble representation denoted as $M_{Ensemble}$. Mathematically, this can be expressed as:

$$M_{Ensemble} = 1/3x(M_{LR} + M_{DT} + M_{RF}) \quad (1)$$

B. ***Voting Mechanism***: The voting mechanism employed in this methodology involves a "hard" voting strategy. In this context, each classifier within the ensemble casts a singular vote for the predicted class. The final prediction is determined by the majority decision, where the class receiving the most individual votes is considered the ensemble prediction.

For instance, if M_{LR} determines the class A , M_{DT} identifies the class B, and M_{RF} identifies the class A , the ensemble prediction would favour class A due to the majority. Mathematically, the ensemble decision $D_{Ensemble}$ is given by:

$D_{Ensemble}$ = arg max (Votes for class A, Votes for class B, Votes for class C, ……..)

Algorithm 1: Proposed Voting Ensemble framework

Let D = { $d_1,d_2,d_3,……..,d_n$ } be the taken dataset.
M = { M_{LR},M_{DT},M_{RF} } be the machine learning classifiers.
X1 = 80% of dataset reserved for training , X1 ∈ D
X2 = 20% of dataset reserved for testing , X2 ∈ D
V = meta - level classifier
N = n(D)
for i = 1 to N do
  Train M(i) on X1
  Next i
End for
$M_{Ensemble}$ = 1/3 x (M_{LR} + M_{DT} + M_{RF})
for i = 1 to N do
  E(i) = $M_{Ensemble}$
  Next i
End for
E = E U V
RESULT = $D_{Ensemble}$ classifies X2

TABLE II. DESCRIPTION OF NOTATIONS USED IN ALGORITHM 1

Symbol	Description
h	Properties of the dataset taken
M	ML classifier methods
X1	Dataset used for Training
X2	Dataset used for testing
E	Ensemble model
V	Meta level classifier
M_{LR}	Represents Logistic Regression model
M_{DT}	Represents Decision Tree classifier model
M_{RF}	Represents Random Forest classifier model
$D_{Ensemble}$	Decision of the Ensemble model

IV. EXPERIMENTAL ANALYSIS

A. ***Data Source***: The dataset employed in this work for prediction of heart disease is sourced from an external dataset repository, such as Kaggle. It consists of a total of 1025 instances, each representing a patient's data entry. The dataset consists of 14 attributes, out of these 13 features were considered for testing, the final feature represents the outcome variable.

The dataset was categorized into two subsets in the ratio of 80:20 for testing and training respectively.

B. ***Description of the Dataset***:

TABLE III. PARAMETER DESCRIPTION OF DATASET

Attribute	Description	Type
age	Patient's Age	Integer
sex	Patient's Gender (Values - 0: female, 1: male)	Integer
cp	Type of the Chest pain (0-3)	Integer

trestbps	Blood Pressure during resting stage (mm Hg)	Integer
chol	Serum cholesterol level (mg/dl)	Integer
fbs	Results of blood sugar levels at fasting > 120 (0 represents false, 1 represents true)	Integer
restecg	Resting ECG results (0-2)	Integer
thalach	Value of heart rate attained in maximum during exercise	Integer
exang	Exercise-induced angina (0 represents no , 1 represents yes)	Integer
oldpeak	Exercise-induced Segment and T-wave depression relative to rest	Float
slope	Gradient of the Segment and T-wave at the peak of exercise (0-2)	Integer
ca	Count of major vessels visualized by fluoroscopy (0-3)	Integer
thal	Thallium heart scan results (0-3)	Integer
target	Existence of disease (Values - 0:no , 1:yes)	Integer

C. ***Performance Metrics***: During the assessment of the heart disease prediction using the Voting Ensemble technique, various performance measures are employed to gauge the model's accuracy and reliability. These measures offer a deeper understanding into the model's effectiveness in distinguishing between the existence and non-existence of heart disease. The subsequent performance indicators are calculated and presented in the table given below:

TABLE IV. PERFORMANCE METRICS RESULTS

Metric	Result (in %)
Weighted Area Under the Curve (AUC)	98.5436
Macro Area Under the Curve	98.5436
Micro Area Under the Curve	98.5436
Macro Average Precision	98.550
Micro Average Precision	98.550
Weighted Average Precision	98.550
Macro F1 Performance	98.5365
Micro F1 Performance	98.5365
Weighted F1	98.5365
Macro Recall	98.5436
Micro Recall	98.5436
Weighted Recall	98.5436
Accuracy	98.557

D. *Confusion Matrix*

TABLE V. CONFUSION MATRIX

	Predicted No-Disease	Predicted Disease
Actual No-Disease	102	0
Actual Disease	3	100

E. ***Differentiation with other Algorithms***: The dataset has been processed, trained, and assessed using various algorithms available. Among all these algorithms, the proposed voting ensemble technique stands out, demonstrating the highest accuracy and minimized Root Mean Square Error (RMSE) and Mean Square Error (MSE) values in our evaluations.

The Voting Ensemble's superior accuracy and lower error metrics indicate that the integration of diverse classifiers in the ensemble contributes to more robust and reliable predictions.

The proposed voting ensemble approach leverages the capabilities of individual models, mitigating their weaknesses and providing a more thorough and precise predictive outcome.

TABLE VI. COMPARISON OF ACCURACY WITH OTHER ALGORITHMS

Name of Algorithm	Accuracy (in %)
Voting Ensemble	98.5
Stack Ensemble	88.2
Logistic Regression	76.4
Random Forest classifier	81.2
Decision tree classifier	79.1

TABLE VII. ERROR PERFORMANCE METRICS

Name of Algorithm	Value of RMSE	Value of MSE
Voting Ensemble	0.1209	0.01463
Stack Ensemble	0.382	0.1459
Random Forest classifier	0.425	0.1806
Logistic Regression	0.452	0.2043
Decision Tree classifier	0.449	0.2016

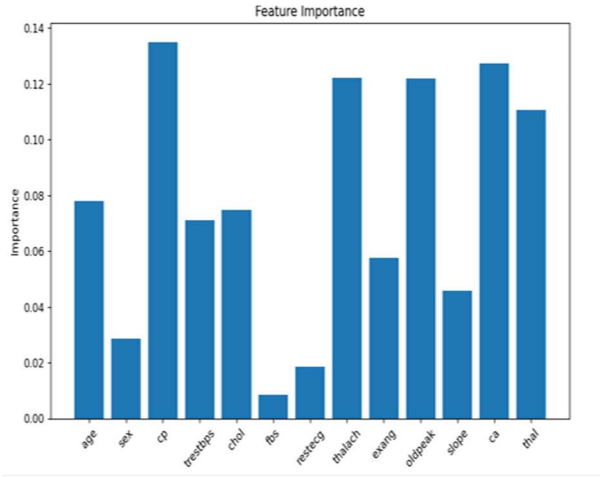

Figure 2. Bar plot of Feature Importance

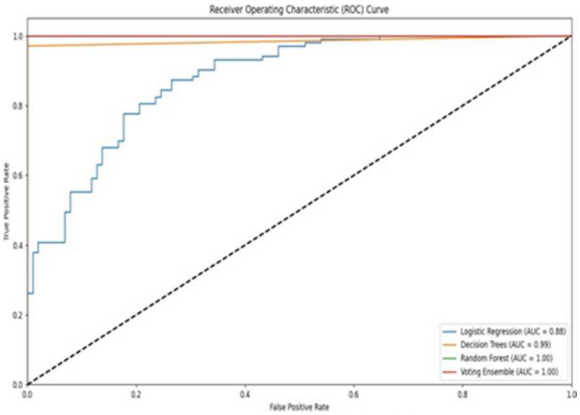

Figure 3. ROC curve of algorithms

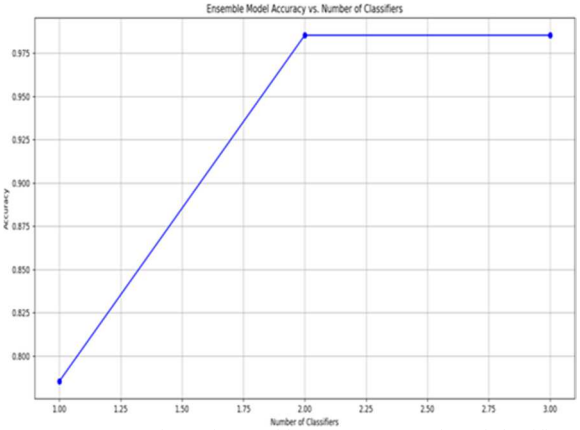

Figure 4. Comparison of Ensemble accuracy to number of classifiers

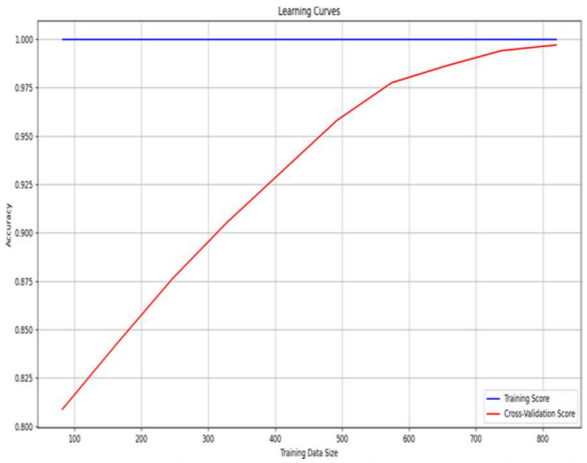

Figure 5. Correlation between size of training dataset and model accuracy

V. Conclusion

In conclusion, our study demonstrates the efficacy of a Voting Ensemble technique for heart disease prediction. By combining Logistic Regression classifier, Decision Tree classifier, and Random Forest classifier, we have achieved a higher level of accuracy and durability in diagnostic predictions. The ensemble model outperforms individual classifiers, underlining the potential of ensemble techniques in healthcare analytics. This research contributes to the ongoing efforts to improve diagnostic accuracy in the realm of predicting heart disease. Future work implementation in this area may explore additional ensemble methods and consider larger datasets for even more reliable predictions.

References

[1] Puneet, Deepika, P. Singh, R. Bansal and S. Sharma," Coronary Heart Disease Prediction Using Voting Classifier Ensemble Learning," 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), 2021, pp. 181-185, doi: 10.1109/ICAC3N53548.2021.9725705.

[2] S. Asif, Y. Wenhui, Y. Tao, S. Jinhai and H. Jin,"An Ensemble Machine Learning Method for the Prediction of Heart Disease," 2021 4th Inter- national Conference on Artificial Intelligence and Big Data (ICAIBD), 2021, pp. 98-103, doi: 10.1109/ICAIBD51990.2021.9459010.

[3] Premkumar Duraisamy, Yuvaraj Natarajan, EbinN L, Jawahar Raja P, "Efficient Way of Heart Disease Prediction and Analysis using different Ensemble Algorithm: A Comparative Study", 2022 6th International Conference on Electronics, Communication and Aerospace Technology, pp.1425-1429, 2022.

[4] I. B. Mijoya, S. Khurana, N. Gupta and K. Gupta, "Malware Detection in Mobile Devices Using Hard Voting Ensemble Technique," *2023 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, Greater Noida, India, 2023,pp.116-121,doi: 10.1109/ICCCIS60361.2023.10425513.

[5] D. K. Behera, S. Dash, A. K. Behera and C. S. K. Dash, "Extreme Gradient Boosting and Soft VotingEnsemble Classifier for Diabetes Prediction," 2021 19th OITS International Conference on Information Technology (OCIT), Bhubaneswar, India, 2021, pp. 191-195, doi: 10.1109/OCIT53463.2021.00046.

[6] K. Sen and B. Verma, "Heart Disease Prediction Using a Soft Voting Ensemble of Gradient Boosting Models, RandomForest, and Gaussian Naive Bayes," *2023 4th International Conference for Emerging Technology (INCET)*, Belgaum, India, 2023, pp. 1-7, doi: 10.1109/INCET57972.2023.10170399.

[7] R. Bhuvana, S. Maheshwari and S. Sasikala, "Predict the Heart Disease Using a Logistic Regression Classifier Algorithm," *2023 12th International Conference on System Modeling & Advancement in Research Trends (SMART)*, Moradabad, India, 2023, pp. 649-652, doi: 10.1109/SMART59791.2023.10428486.

[8] G. S. Reddy Thummala and R. Baskar,"Prediction of Heart Disease using Decision Tree in Comparison with KNN to Improve Accuracy," 2022 International Conference on Innovative Computing,Intelligent Communication and Smart Electrical Systems (ICSES), Chennai, India, 2022, pp. 1-5, doi: 10.1109/ICSES55317.2022.9914044.

[9] K. Mittal, K. S. Gill, D. Upadhyay, V. Singh and S. Aluvala, "Applying Machine Learning for Autism Risk Evaluation Using a Decision Tree Classification Technique," *2024 2nd International Conference on Computer, Communication and Control (IC4)*, Indore, India, 2024, pp. 1-6, doi: 10.1109/IC457434.2024.10486622.

[10] G. S. Reddy Thummala, R. Baskar and R. S, "Prediction of Heart Disease using Random Forest in Comparison with Logistic Regression to Measure Accuracy," *2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI)*, Chennai, India, 2023, pp.1-5, doi: 10.1109/ACCAI58221.2023.10199851.

[11] S. Naveen, S. K. Ravindran, S. G and S. N. Ameen, "Effective Heart disease prediction framework using Random Forest and Logistic regression," *2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN)*, Vellore, India, 2023, pp. 1-6, doi: 10.1109/ViTECoN58111.2023.10157078.

[12] S. N. Soualihou, J. -D. Kim, C. Zonyfar, H. Lee, T. Lee and J. -B. Lee, "An ensemble learning-based machine learning with voting mechanism for chronic disease prediction," *2024 International Conference on Artificial Intelligence, Computer, Data Sciences and Applications (ACDSA)*, Victoria, Seychelles, 2024, pp. 1-6, doi: 10.1109/ACDSA59508.2024.10467844.

[13] N. V. R and V. C. S. S., "Lung Cancer Malignancy detection Using Voting Ensemble Classifier," *2023 2nd International Conference on Computational Systems and Communication (ICCSC)*, Thiruvananthapuram, India, 2023, pp. 1-5, doi: 10.1109/ICCSC56913.2023.10142984.

[14] A. Batool and Y. -C. Byun, "Toward Improving Breast Cancer Classification Using an Adaptive Voting Ensemble Learning Algorithm," in *IEEE Access*, vol. 12, pp. 12869-12882, 2024, doi: 10.1109/ACCESS.2024.3356602.