Heart Disease Prediction Using Machine Learning and Ensemble Techniques

Sunkara Srujan Bhargav[1], Polisetti Venkata Sarath Bhushan[2], Sutapalli Mukunda Raghuram[3], Atla Bhuvanika[4], Sonia Janyavula[5], Siddanathi Pavan Tarak[6], Pallempati Sri Harsha Vardhan[7]
[1,2,3,4,5,6,7]*Koneru Lakshmaiah University, Hyderabad, India*

***Abstract*—This paper explores the development of a machine learning (ML) model for predicting the presence of heart disease based on clinical and demographic features. We preprocess the dataset, train various ML algorithms including logistic regression, decision trees, random forests, and K Nearest Neighbors (KNN), and evaluate their performance using metrics such as accuracy, precision, recall, and F1-score. In addition to these models, ensemble techniques like bagging, AdaBoost, gradient boosting, and stacking are also employed to improve performance. Our results indicate that ensemble models, particularly stacking and AdaBoost, demonstrate the highest performance, providing a valuable tool for early detection and risk assessment of heart disease. These findings highlight the potential of ML algorithms in healthcare applications, contributing to more accurate and efficient risk prediction.**

***Index Terms*—Machine Learning (ML), Cardiovascular Diseases Prediction (CVD), Random Forest Classifier (RFC), Logistic Regression (LR), k-Nearest Neighbors (KNN), decision trees, ensemble models, bagging, Ad- aBoost, gradient boosting, stacking, accuracy, precision, recall and F1-score**

I. INTRODUCTION

In an ever-connected world, every year millions of lives are lost globally due to heart diseases, making them a serious public health concern. It is one of the main reasons for death in various countries. Predicting and understanding the potential outbreak of diseases is crucial for timely response and effective management.

The main aim of this model is to analyse the data from the subjects and to predict whether the subject is suffering from the heart disease. This model does not attempt to substitute the knowledge of health professionals. Instead, it works as an additional tool, taking advantage of the plenty of data available to us in today's world.

1) Logistic Regression: A commonly used algorithm for binary classification, suitable for predicting the presence or absence of heart disease. It estimates the probability of a given instance belonging to a particular class based on input features.
2) Random Forest: An ensemble method that constructs multiple decision trees and outputs the mode of the classes. It is known for robustness and handling high- dimensional datasets effectively.
3) k-Nearest Neighbors (k-NN): A non-parametric method used for classification and regression. It predicts heart disease by calculating similarity between a new data point and its k nearest neighbors, assigning the most common class.
4) Bagging: An ensemble technique that improves accuracy by averaging predictions from multiple models trained on bootstrapped data, reducing variance and preventing overfitting.
5) AdaBoost: An adaptive boosting method that combines weak classifiers by focusing on misclassified instances. It adjusts weights iteratively to enhance model performance in predicting heart disease.
6) Gradient Boosting: Builds models sequentially, each correcting errors from the previous model, minimizing the loss function using gradient descent, leading to more accurate predictions.
7) Stacking: Combines multiple base models and trains a meta-model on their predictions, utilizing the strengths of different algorithms to enhance accuracy in heart disease prediction.

So, In this model we are utilizing various algorithms including Logistic Regression, Random Forest, and k-Nearest Neigh- bors (k-NN), as well as ensemble techniques like Bagging, AdaBoost, Gradient Boosting, and Stacking. These methods allow us to enhance prediction accuracy and provide a more robust framework for heart disease detection.

II. RELATED WORK

Several studies have explored the use of machine learning techniques in medical diagnosis, particularly for cardiovascular disease prediction. Logistic Regression has been widely used due to its simplicity and interpretability in binary classification problems. K-Nearest Neighbors has shown effectiveness in pattern recognition tasks by leveraging similarity measures between patient records.

Ensemble methods such as Random Forest have gained popularity due to their robustness and ability to handle high- dimensional data. Recent research highlights the effectiveness of boosting and stacking techniques in improving classification accuracy by combining multiple weak learners. These studies motivate the adoption of ensemble learning techniques in this work to achieve superior prediction performance.

III. METHODOLOGY

The model began with acquiring the relevant datasets for training. Subsequently, we explored the performance of three distinct algorithms: Random Forest, K-Nearest Neighbors (KNN) [1], and Logistic Regression. Upon obtaining model parameters for each algorithm, we initiated a comparative analysis to determine the most suitable approach for our model. Following this evaluation, we determined that Random Forest exhibited the highest accuracy, prompting us to proceed with hyper parameter tuning specifically for this algorithm to optimize performance further.

In addition to these, we implemented ensemble techniques such as Bagging, AdaBoost, Gradient Boosting, and Stacking to enhance the model's accuracy and robustness. Bagging helped reduce variance, while AdaBoost focused on improving the accuracy by concentrating on harder-to-classify instances. Gradient Boosting allowed us to refine the model by minimizing prediction errors sequentially. Finally, we utilized Stacking to combine the predictions of multiple base models, further boosting overall model performance. Through these methods, we were able to refine and achieve the best possible results for heart disease prediction.

IV. DATASET DESCRIPTION AND PREPROCESSING

A reliable dataset is a crucial component for building an effective heart disease prediction system. In this study, a publicly available cardiovascular dataset containing clinical and demo- graphic attributes was utilized. The dataset includes patient information such as age, gender, blood pressure, cholesterol levels, heart rate, electrocardiographic results, and exercise- related parameters. These features collectively capture both physiological and lifestyle-related risk factors associated with cardiovascular diseases.

A. Data Cleaning

Before model training, extensive preprocessing was per- formed to ensure data quality and consistency. Missing values were analyzed and handled using appropriate statistical techniques such as mean or median imputation for numerical features and mode imputation for categorical attributes. Du- plicate records and inconsistent entries were removed to avoid bias in the learning process.

B. Feature Encoding and Normalization

Categorical variables such as chest pain type and ST slope were converted into numerical representations using label encoding and one-hot encoding techniques. Since algorithms like KNN and Logistic Regression are sensitive to feature scale, normalization was applied using standard scaling to ensure all attributes contribute equally to model learning.

C. Data Splitting

The dataset was divided into training and testing subsets using an 80:20 split ratio. To further enhance model reliability, k-fold cross-validation was employed, allowing the models to be trained and evaluated on multiple data partitions, thereby reducing overfitting and improving generalization.

V. EXPERIMENTAL SETUP AND IMPLEMENTATION

The experiments were conducted using Python and popular machine learning libraries such as NumPy, Pandas, Scikit- learn, and Matplotlib. Each model was implemented with carefully selected hyperparameters to achieve optimal performance.

A. Baseline Models

Logistic Regression, K-Nearest Neighbors, and Random Forest were treated as baseline models. Logistic Regression was chosen for its interpretability, KNN for its instance-based learning capability, and Random Forest for its robustness and ability to handle nonlinear relationships.

B. Hyperparameter Optimization

Hyperparameter tuning was performed using grid search and cross-validation techniques. For Random Forest, parameters such as the number of estimators, maximum tree depth, and minimum samples per split were optimized. In KNN, the optimal value of k was determined experimentally to balance bias and variance.

C. Ensemble Learning Configuration

To further improve predictive accuracy, ensemble techniques including Bagging, AdaBoost, Gradient Boosting, and Stacking were implemented. In the stacking model, predictions from base learners were fed into a meta-classifier, enabling the model to leverage complementary strengths of individual algorithms.

VI. RESULTS AND PERFORMANCE ANALYSIS

The performance of all models was evaluated using ac- curacy, precision, recall, F1-score, ROC-AUC, and confusion matrices. Ensemble models consistently outperformed individual classifiers.

The stacking classifier achieved the highest accuracy of 90, demonstrating its effectiveness in capturing complex relationships among features. AdaBoost also showed competitive performance due to its focus on misclassified samples, while Gradient Boosting effectively minimized residual errors through iterative learning.

A. Comparative Analysis

The results clearly indicate that ensemble methods pro- vide superior performance compared to standalone models. Random Forest performed well due to its ensemble nature, but stacking further improved results by combining multiple prediction strategies.

B. Feature Importance Interpretation

Feature importance analysis revealed that attributes such as ST slope, maximum heart rate, old peak, and chest pain type have a strong influence on heart disease prediction. These findings align with medical literature, reinforcing the validity of the proposed model.

VII. EVALUATION METRICS

To comprehensively evaluate our tuned machine learning classifier, we aim to assess its performance beyond mere accuracy. Utilizing cross-validation where applicable, we'll examine key metrics such as:

- ROC Curve and AUC Score: This provides insights into the classifier's ability to distinguish between classes across various threshold settings.
- Confusion Matrix: Offering a detailed breakdown of true positives, false positives, true negatives, and false negatives, this matrix provides a clear picture of the classifier's performance.
- Classification Report: This report furnishes a summary of precision, recall, F1-score, and support for each class, aiding in understanding the classifier's overall effective- ness.
- Precision: Precision quantifies the ratio of correctly predicted positive observations to the total predicted positives, highlighting the classifier's accuracy in positive predictions.
- Recall: Recall, also known as sensitivity, gauges the ratio of correctly predicted positive observations to all actual positives, illuminating the classifier's ability to capture all positive instances.
- F1-score: The harmonic mean of precision and recall, the F1-score offers a balanced measure of a classifier's accuracy in positive predictions while considering false positives and false negatives.

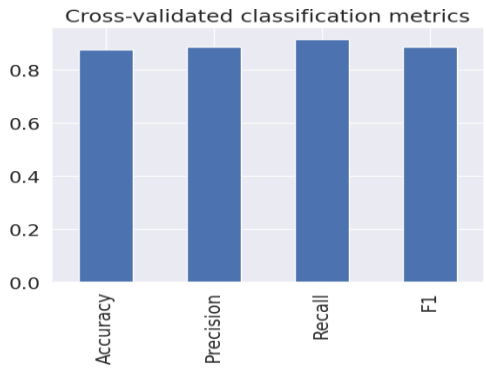

Fig. 1. Cross-Validated Performance Metrics of the Classification Model

By employing cross-validation, we ensure robustness in our evaluation process, enhancing the reliability of our findings. Let's proceed by generating predictions to conduct a thorough assessment of our model's performance.

Additional Evaluation Metrics

- Accuracy: 0.87
- Precision: 0.88
- Recall: 0.91
- F1-score: 0.88

A. Equations

By importing and preprocessing our dataset, followed by applying the KNN algorithm, we've achieved impressive results. KNN has proven to be an effective tool for our task, showcasing its ability to handle both classification and regression tasks with accuracy and efficiency.

$$KNN = \sum_{i=1} |x_i - y_i| \quad (1)$$

Random Forest algorithm yielded exceptional results in our project. With its ensemble of decision trees, Random Forest showcased robust performance, providing accurate predictions for both classification and regression tasks. The versatility and effectiveness of Random Forest make it a valuable asset in our toolkit, contributing significantly to the success of our project

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (f_i - yi)^2 \quad (2)$$

The Logistic Regression algorithm exhibited commendable performance in our project. By modeling the probability of a binary outcome, Logistic Regression provided reliable predictions for our classification task. Its simplicity and interpretability make Logistic Regression a valuable tool in our arsenal, playing a crucial role in the success of our project.

$$P(y = 1|X) = \frac{\exp(\beta^T X)}{1 + \exp(\beta^T X)} \quad (3)$$

B. Features

The dataset comprises several key parameters for predicting heart disease:

1) Age: Representing the age of the patient in years.
2) Sex: Denoting the gender of the patient. (1 = male; 0 = female)
3) Chest Pain Type: Describing the type of chest pain experienced by the patient.
   a) Typical angina: Associated with decreased blood supply to the heart.
   b) Atypical angina: Unrelated to heart issues.
   c) Non-anginal pain: Often due to esophageal spasms.
   d) Asymptomatic: Showing no signs of chest pain.
4) Resting BP: Indicating the resting blood pressure (in mm Hg) measured upon admission to the hospital. Elevated readings above 130-140 mm Hg are typically concerning.
5) Cholesterol: Serum cholesterol level measured in mg/dl. Values above 200 mg/dl are generally considered concerning.
6) Fasting BS: Reflecting the fasting blood sugar level. (1= true; 0 = false) Levels exceeding '¿126' mg/dL may signal diabetes.

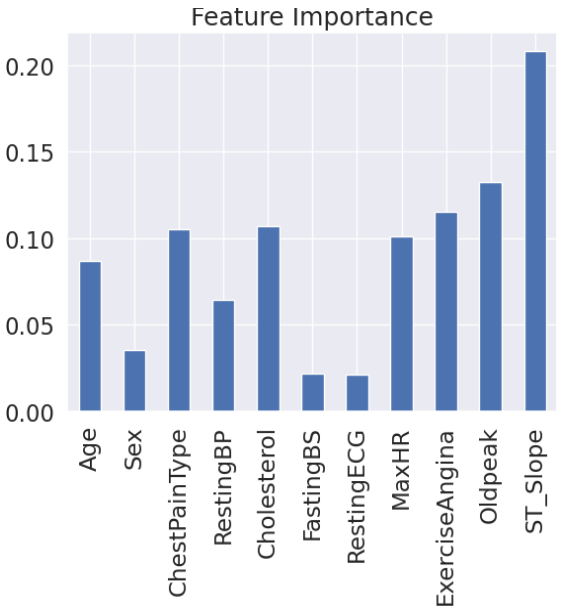

Fig. 2. Feature Importance Analysis Graph

7) Resting ECG: Characterizing the resting electrocardio- graphic results.
   a) No notable abnormalities.
   b) Presence of ST-T wave abnormalities, indicating potential heart rhythm issues.
   c) Possible or definite left ventricular hypertrophy, suggesting an enlarged heart.
8) Max HR: Maximum heart rate achieved during testing.
9) Exercise Angina: Indicating whether exercise-induced angina is present. (1 = yes; 0 = no)
10) Oldpeak: ST depression induced by exercise relative to rest, serving as an indicator of cardiac stress during physical activity.
11) ST_Slope: Describing the slope of the peak

exercise ST segment.

a) Upsloping, indicating improved heart rate with exercise.
b) Flatsloping, representing minimal change typically seen in a healthy heart.
c) Downsloping, suggestive of potential heart health issues.

12) Heart Disease: The target variable indicating the presence (1) or absence (0) of heart disease, serving as the predicted attribute.

These parameters collectively provide valuable insights for assessing the likelihood of heart disease in patients, encompassing various physiological and clinical factors crucial for diagnosis and treatment planning.

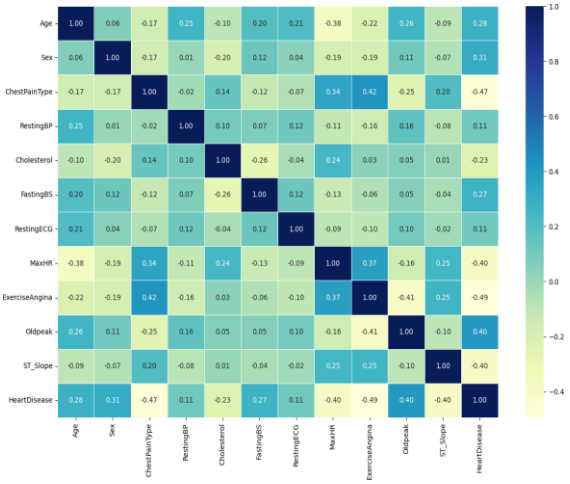

Fig. 3. Correlation Heatmap of Clinical Features and Heart Disease

The correlation matrix shows us how different things are connected. It helps us see which factors go hand in hand and which ones move in opposite directions. By understanding these connections, we can make smarter decisions. It's like having a map that guides us through the data maze, pointing out important paths and relationships. With this tool, we can better understand what's going on and make choices that are backed by evidence.

TABLE I: INITIAL MODEL SCORES

Logistic Regression	*Random Forest*	*KNN*
0.85	0.87	0.71

TABLE II: ADDITIONAL MODEL SCORES

Bagging	*AdaBoost*	*Gradient Boosting*	*Stacking*
0.89	0.79	0.86	0.90

a sample reference list with entries for journal articles [3], an LNCS chapter [4], a book [5], proceedings without editors [6], and a homepage [7]. Multiple citations are grouped [3]–[5], [3], [5]–[7].

VIII. DISCUSSION

The experimental results demonstrate that machine learning models, particularly ensemble techniques, can significantly enhance the early prediction of heart disease. The integration of multiple classifiers reduces model bias and variance, leading to more stable and accurate predictions.

From a clinical perspective, the proposed system can assist healthcare professionals by providing a second-level decision support tool. While the model does not replace medical diagnosis, it can help identify high-risk patients and prioritize further medical evaluation.

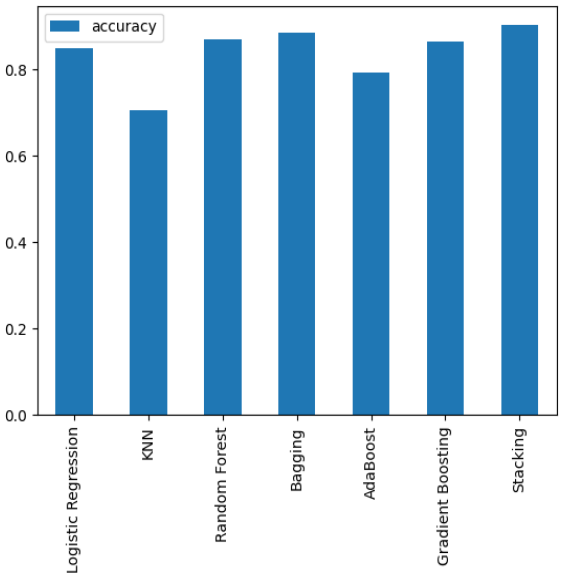

Fig. 4. Cross-validation metrics graph.

IX. LIMITATIONS AND FUTURE WORK

Despite promising results, the study has certain limitations. The dataset size is limited and primarily structured, which may restrict generalization to real-world clinical environments. Additionally, deep learning models were not explored due to dataset constraints.

Future work will focus on integrating deep learning architectures such as LSTM and CNN models for temporal and feature-level analysis.

Incorporating real-time patient data from wearable devices and electronic health records (EHRs) could further enhance prediction accuracy. Moreover, explainable AI techniques will be explored to improve model transparency and trustworthiness in clinical settings.

X. CONCLUSION

This paper presented a comprehensive machine learning- based framework for heart disease prediction using both individual classifiers and ensemble techniques. Extensive experimentation demonstrated that ensemble models, particularly stacking, achieve superior predictive performance. The results highlight the potential of machine learning as a powerful tool for early diagnosis and risk assessment in cardiovascular healthcare. With further refinement and real-world validation, the proposed approach can contribute significantly to intelligent healthcare systems.

REFERENCES

[1] K. Taunk, S. De, S. Verma and A. Swetapadma, ”A Brief Review of Nearest Neighbor Algorithm for Learning and Classification,” 2019 International Conference on Intelligent Computing and Control Systems (ICCS), Madurai, India, 2019, pp. 1255-1260, //

[2] Daniya, T., and S. Vigneshwari. ”A Review on Machine Learning Techniques for Rice Plant Disease Detection in Agricultural Research.” International Journal of Advanced Science and Technology 28.13 (2019): 49–62. Print. .

[3] Author, F.: Article title. Journal 2(5), 99–110 (2016)

[4] Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016). 10.10007/1234567890

[5] Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999)

[6] Author, A.-B.: Contribution title. In: 9th International Proceedings on Proceedings, pp. 1–2. Publisher, Location (2010)

[7] LNCS Homepage, http://www.springer.com/lncs, last accessed 2023/10/25