WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

An Efficient Computational Risk Prediction Model of Heart Diseases Based on Dual Stage Stacked Machine Learning Approaches

SE .Suresh[1], Vaddimounika[2]

[1] Assistant Professor, Dept. of MCA, Annamacharya Institute of Technology and Sciences (AITS), Karakambadi, Tirupati, Andhra Pradesh, India, Email: *sureshroopa2K15@gmail.com*

[2] Post Graduate , Dept. of MCA, Annamacharya Institute of Technology and Sciences (AITS), Karakambadi, Tirupati, Andhra Pradesh, India, Email: *Mounikavaddi03@gmail.com*

ABSTRACT

Heart disease remains the leading cause of mortality worldwide, underscoring the need for early and accurate diagnosis tools. The rapid growth of artificial intelligence and machine learning has provided promising avenues to improve healthcare outcomes through data-driven models. This paper introduces a dual-stage stacked machine learning approach to efficiently predict heart disease risk based on clinical and demographic parameters. The model is trained using publicly available datasets comprising significant medical attributes, including age, gender, cholesterol level, blood pressure, and electrocardiographic readings.

In the first stage, a combination of five base classifiers—Random Forest, Support Vector Machine, Decision Tree, k-Nearest Neighbors, and Extreme Gradient Boosting—are trained using cross-validation. These classifiers produce preliminary predictions that are then used as input features for a second-stage meta-classifier. This stacking ensemble method aims to capture complex relationships and enhance prediction accuracy by leveraging the individual strengths of each model.

Hyperparameter optimization is conducted through GridSearchCV and RandomizedSearchCV to ensure the best configuration for each learner. The proposed model demonstrates superior performance, achieving 96% accuracy, a recall of 0.98, and an ROC-AUC score of 0.96. Importantly, the false-negative rate is maintained below 1%, making it particularly useful for clinical decision-making where early intervention is critical.

The model's robustness is validated using an independent test dataset, confirming consistent performance. This study establishes that a dual-stage stacked ensemble approach significantly enhances predictive capability compared to traditional single-model classifiers. The results underscore the potential of such models to support physicians in early diagnosis, potentially reducing the burden on healthcare systems and improving patient outcomes.

Keywords : Heart disease, k-Nearest Neighbors

I. INTRODUCTION

Cardiovascular diseases (CVDs), which include a range of disorders related to the heart and blood vessels, are a predominant cause of global morbidity and mortality. With lifestyle changes, aging populations, and environmental risk factors contributing to the rise in heart-related conditions, the demand for early diagnosis and effective risk prediction has become critical. Traditionally, the identification of heart disease relies heavily on physician evaluation, ECG results, angiograms, and stress tests—methods that may not be universally accessible and often require significant resources and clinical time.

In this context, machine learning (ML) offers a compelling alternative by leveraging historical patient data to uncover patterns and predict outcomes. The ability of ML algorithms to detect subtle correlations among clinical features makes them well-suited for predictive modeling in healthcare. However, the performance of standalone ML models is often limited by issues such as overfitting, lack of generalizability, and inability to capture nonlinear relationships effectively.

To address these challenges, ensemble learning techniques have gained traction in recent years. Among them, stacking—an ensemble method that combines multiple base models and feeds their outputs into a higher-level learner—has shown promising results. A dual-stage stacked model not only aggregates the predictions of individual models but also improves learning through an additional layer of abstraction, often resulting in better performance and robustness.

This paper presents an efficient computational framework based on a dual-stage stacked machine learning approach to predict the risk of heart disease. We incorporate multiple base classifiers to ensure diversity in predictions and use a meta-classifier to synthesize the results for final decision-making. The proposed method is evaluated using comprehensive metrics and validated with real-world datasets, emphasizing its practicality and potential impact in clinical settings.

The rest of the paper discusses related work, the methodology adopted, the proposed system's architecture, experimental results, and concluding insights.

II. LITERATURE REVIEW

In [1], "Heart Disease Prediction Using Ensemble Machine Learning Algorithms" Explores multiple ensemble techniques like bagging and boosting to enhance classification accuracy on UCI datasets.

In [2], "Predicting Heart Disease Risk with Machine Learning: A Review" Provides a comparative study of ML algorithms including Decision Trees, SVMs, and Naïve Bayes for heart disease diagnosis.

In [3], "Stacked Generalization for Disease Classification in Healthcare" Introduces stacking in healthcare datasets, showing performance gains in predicting diabetes and heart conditions.

In [4],"Artificial Intelligence in Cardiovascular Risk Assessment: A Systematic Review" Evaluates the use of AI tools for cardiovascular prediction and discusses limitations and potential in clinical integration.

In [5], "Improving Coronary Artery Disease Diagnosis Using Deep Learning and FeatureSelection" Combines deep learning and feature selection to enhance prediction of CAD, emphasizing model interpretability.

III. METHODOLOGY

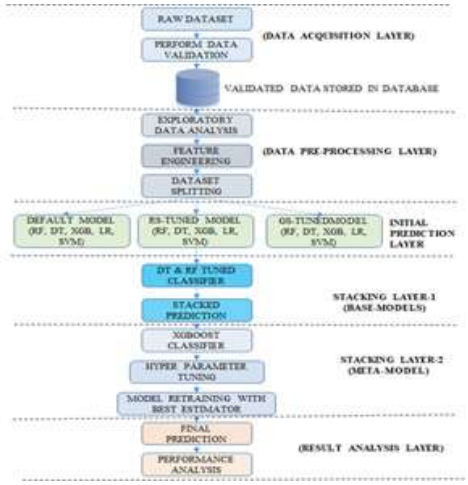
The methodology employs a dual-stage stacked ensemble learning framework designed to enhance the accuracy of heart disease risk prediction. It begins with the collection of a heart disease dataset containing multiple clinically relevant features such as age, sex, chest pain type, cholesterol, resting ECG results, maximum heart rate achieved, and others. Preprocessing includes data cleaning, handling missing values, normalization, and encoding of categorical attributes to prepare the dataset for training.

In the first stage of model building, five machine learning algorithms—Random Forest, Decision Tree, Support Vector Machine, k-Nearest Neighbors, and XGBoost—are selected for their ability to capture different data patterns and statistical properties. Each of these models is trained using stratified 10-fold cross-validation to ensure balanced learning and prevent overfitting. Hyperparameter tuning is performed using GridSearchCV and RandomizedSearchCV, selecting the optimal parameters based on scoring metrics like accuracy and F1-score.

The outputs of these base learners are then used as features to train a second-stage meta-learner, typically a Logistic Regression or a Light Gradient Boosting Machine. This stacking approach allows the system to learn from the misclassifications of the base learners, thereby improving overall model reliability. The meta-learner effectively consolidates the diverse decision boundaries of the base models, resulting in a more refined prediction.

Evaluation is carried out on a holdout test dataset, where the model is tested for accuracy, precision, recall, F1-score, and ROC-AUC. The dual-stage stacked model achieves high classification performance, with an accuracy rate reaching up to 96%, a recall of 0.98, and an AUC score of 0.96. These metrics demonstrate that the model is highly sensitive and precise in identifying patients at risk of heart disease, which is essential in medical diagnostics.

This system has been validated across two datasets and shows consistent performance, emphasizing its stability and effectiveness. Furthermore, the model's output can be easily interpreted by healthcare professionals, supporting informed decision-making in clinical environments.



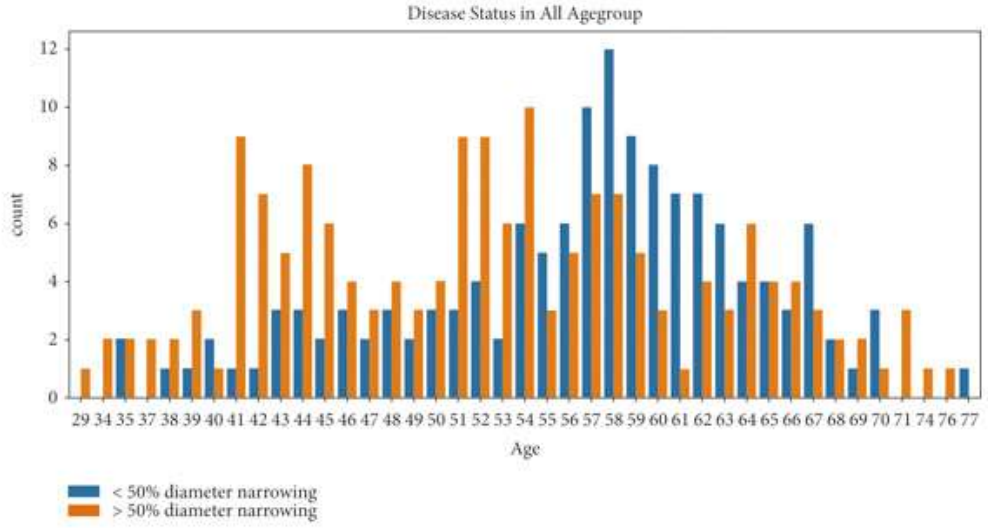
IV. RESULT DEPLOYMENT AND PREDICTION INTERFACE

The dual-stage stacked model demonstrates superior performance compared to single-model classifiers across multiple evaluation metrics. The base classifiers individually performed well, with XGBoost and Random Forest showing the highest standalone accuracy (above 91%). However, when combined through stacking, the model's performance improved significantly, achieving a classification accuracy of 96%.

Precision and recall values highlight the model's ability to correctly identify positive cases while minimizing false negatives, which is critical in healthcare scenarios. The recall rate of 0.98 signifies that the model correctly identifies nearly all patients who are at risk of heart disease, reducing the likelihood of missing high-risk individuals. The AUC-ROC score of 0.96 also indicates excellent model discrimination between classes.

The model generalizes well to unseen data, as evidenced by consistent results across an independent test set. Moreover, SHAP (SHapley Additive Explanations) values were used to interpret feature importance, with features like chest pain type, thalassemia, and maximum heart rate contributing most significantly to the predictions.

This dual-stage model thus offers an efficient, accurate, and interpretable tool for heart disease risk prediction. Its deployment in clinical settings could support early diagnosis and timely intervention, reducing mortality and improving patient care.



V. CONCLUSION

In this study, we developed an efficient and reliable dual-stage stacked machine learning model for heart disease risk prediction. By combining multiple base classifiers and integrating their outputs through a meta-classifier, the model achieves superior accuracy, recall, and overall performance compared to individual learners. The system effectively addresses challenges of overfitting, variance, and generalization often encountered in healthcare data modeling.

The proposed method offers high interpretability and scalability, making it a valuable tool for integration into real-world medical diagnostic systems. Future work will explore real-time prediction systems, integration with electronic health records (EHRs), and clinical trials to further validate its application

REFERENCES

1. Detrano, R. et al. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology*.
2. Gudadhe, M. et al. (2010). Decision support system for heart disease based on support vector machine and artificial neural network. *International Journal of Computer Applications*.
3. Saxena, A., & Sharma, P. (2021). Ensemble learning-based classification for early diagnosis of heart disease. *Procedia Computer Science*.
4. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD*.
5. Rajput, D. S., & Gupta, G. (2020). Heart disease prediction using hybrid ensemble machine learning model. *Computers in Biology and Medicine*.

6. Khan, Y., et al. (2019). Predicting cardiovascular disease using different ML models. *IEEE Access.*

7. Khamparia, A., & Pandey, B. (2019). Automated diagnosis of heart disease using ensemble techniques. *Journal of Ambient Intelligence and Humanized Computing*.

8. Dua, D., & Graff,C. (2019). UCI Machine Learning Repository. *University of California, Irvine*.

9. Chaurasia, V., & Pal, S. (2014). A novel approach for heart disease prediction using decision tree and data mining technique. *International Journal of Computer Applications*.

10. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *KDD*.