IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, March 2026
Impact Factor: 8.2

An Efficient Computational Risk Prediction Model for Heart Disease Using a Dual-State Stacked Machine Learning Approach

B Sekhar[1] , Bejjanki Pooja[2], B Aditya[3], Vemula Deepika[4], Banothu Gayathri[5], Cheduruvelli Shiva Kumar[6]
Assistant Professor, Department of CSE[1,2,3],
UG Student, Department of CSE[4,5,6]
CMR Technical Campus, Hyderabad, Telangana, India
sekharaije@gmail.com, poojareddybejjanki@gmail.com, Adi.sacs@gmail.com
vemuladeepika87@gmail.com, 237r1a05d4@cmrtc.ac.in,237r1a05e2@cmrtc.ac.in

Abstract: *Heart disease continues to be one of the leading causes of death worldwide, making early prediction an important task in modern healthcare. Traditional diagnostic methods rely heavily on clinical expertise and may not always capture complex patterns in medical data.*
In this work, we propose an efficient heart disease prediction model using a dual-stage stacked machine learning approach. The model combines multiple algorithms, including Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, and Extreme Gradient Boosting. Their outputs are integrated using a meta-learning technique to improve prediction performance.
The model is trained on a dataset of 1190 patient records obtained from the UCI Machine Learning Repository. Data preprocessing techniques such as cleaning, normalization, and feature selection are applied to enhance data quality. Experimental results show that the proposed model achieves an accuracy of around 96%, along with strong precision, recall, and F1-score values.
The results demonstrate that stacking multiple models significantly improves prediction accuracy compared to individual classifiers. The proposed system can assist healthcare professionals in early diagnosis and better decision-making.

Keywords: Heart Disease Prediction, Machine Learning, Dual-State Stacked Model, Ensemble Learning, Risk Prediction, Healthcare Analytics, Data Preprocessing, Hyperparameter Tuning, Medical Data Analysis

I. INTRODUCTION

Heart disease is a major health concern affecting millions of people globally. It disrupts the normal functioning of the cardiovascular system and often leads to severe complications if not detected at an early stage. Accurate and timely prediction of heart disease risk can significantly improve survival rates and enable better treatment planning.

Conventional diagnostic methods primarily depend on clinical experience and manual analysis of patient data. While these approaches are effective, they may fail to identify hidden relationships within large and complex datasets. With the rapid growth of healthcare data, there is a need for automated systems that can analyze information efficiently and support clinical decision-making.

Machine learning techniques have shown significant potential in addressing these challenges [3], [5]. By learning patterns from historical medical data, these models can assist in predicting disease risk with improved accuracy. However, relying on a single algorithm may limit performance, as different models capture different aspects of the data.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, March 2026
Impact Factor: 8.2

To overcome this limitation, this research proposes a Dual-State Stacked Machine Learning model that combines multiple algorithms into a unified framework. The approach leverages the strengths of individual models such as Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, and Extreme Gradient Boosting. Their predictions are integrated using a meta-learner to produce a more reliable final outcome.

The system also incorporates essential preprocessing steps, including data cleaning, normalization, and feature selection, to ensure high-quality input for model training. Performance is evaluated using standard metrics such as accuracy, precision, recall, and ROC-AUC score.

The main objective of this work is to develop a robust and efficient prediction system that reduces dependence on manual analysis and supports healthcare professionals in making informed decisions. By improving prediction accuracy and consistency, the proposed model contributes to better diagnosis and overall patient care.

In our implementation, we focused on improving prediction accuracy while maintaining computational efficiency. We observed that combining multiple models provides better results compared to relying on a single algorithm.

II. LITERATURE REVIEW

Several studies have explored the use of machine learning techniques for heart disease prediction. Early approaches mainly relied on statistical methods, which were limited in handling complex relationships among clinical features.

Recent research has demonstrated that machine learning algorithms such as Logistic Regression, Support Vector Machine, Decision Tree, and Random Forest [4], [5].can provide better prediction accuracy. Ensemble methods, in particular, have shown improved performance by combining multiple models.

Stacking is one such ensemble technique that integrates the outputs of different base learners using a meta-model. According to Wolpert [10], stacking helps improve generalization performance by reducing model bias. Similarly, studies using the UCI heart disease dataset [1] have shown that combining models leads to better accuracy compared to individual classifiers.

Despite these advancements, many models either require large datasets or lack efficiency. Therefore, this work focuses on developing a computationally efficient stacked model for accurate heart disease prediction.

Furthermore, feature selection techniques have also been widely studied to improve model performance by identifying the most relevant clinical attributes. Removing irrelevant or redundant features not only enhances prediction accuracy but also reduces computational complexity. Researchers have highlighted that proper preprocessing steps, such as normalization and handling missing values, play a crucial role in achieving stable and reliable results.

In addition, recent works have explored hybrid models that combine machine learning with optimization techniques to further enhance predictive performance. These approaches aim to balance accuracy and efficiency while minimizing overfitting. Cross-validation methods are commonly used to ensure that the models generalize well to unseen data.

Another important aspect discussed in the literature is the interpretability of models in healthcare applications. While complex ensemble methods provide high accuracy, simpler models are often preferred by medical professionals for better understanding and trust. Therefore, there is a need to develop models that achieve both high accuracy and interpretability.

Overall, the existing research clearly indicates that combining multiple models and applying proper preprocessing techniques significantly improves heart disease prediction. However, there is still a need for efficient models that can deliver high performance without increasing computational cost, which motivates the proposed approach in this study.

III. PROPOSED SYSTEM

The proposed system is designed to predict heart disease risk using a dual-stage stacked machine learning approach. The system processes patient data and generates predictions based on learned patterns derived from historical medical records. The main goal is to provide accurate and reliable predictions that can assist healthcare professionals in making timely decisions.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, March 2026
Impact Factor: 8.2

Initially, patient data is collected from reliable medical datasets, including features such as age, cholesterol level, blood pressure, maximum heart rate, chest pain type, and other relevant clinical attributes. These features play a crucial role in determining the presence or absence of heart disease. However, raw medical data often contains inconsistencies, missing values, and noise, which can affect model performance.

To address this, a comprehensive preprocessing phase is applied. This includes handling missing values, removing duplicate or inconsistent records, and normalizing feature values to ensure uniform scaling. Feature selection techniques are also employed to identify the most relevant attributes, which helps in reducing dimensionality and improving model efficiency. Proper preprocessing ensures that the data is clean, consistent, and suitable for training machine learning models.

In the first stage of the system, multiple machine learning algorithms are trained independently. These include Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, and Extreme Gradient Boosting. Each of these models has its own strengths; for instance, Logistic Regression works well for linear relationships, while Random Forest and XGBoost are effective in handling complex and non-linear patterns. By training these models separately, the system captures diverse patterns present in the dataset.

Once the base models are trained, their predictions are generated for both training and testing data. Instead of selecting a single best-performing model, the system moves to the second stage, where a meta-learning approach is applied. In this stage, the outputs (predictions) of all base models are combined and used as input features for a higher-level model known as the meta-learner.

The learner analyzes the predictions from each base model and learns how to combine them effectively. It assigns appropriate importance to each model based on its performance, thereby reducing individual model bias and improving overall prediction accuracy. This layered structure enhances generalization and ensures that the final output is more robust and reliable.

Additionally, techniques such as cross-validation and hyperparameter tuning are used during training to optimize model performance and prevent overfitting. The system is evaluated using performance metrics such as accuracy, precision, recall, and F1-score to ensure a balanced assessment of results.

Overall, the proposed dual-stage stacked model significantly improves prediction performance compared to traditional single-model approaches. By leveraging the strengths of multiple algorithms and combining them intelligently, the system provides consistent and accurate predictions. This makes it a valuable tool in the early detection of heart disease and supports better clinical decision-making.

IV. METHODOLOGY

A different path guides this study—one that builds an efficient machine learning system to predict heart disease risk using patient data. Step by step, progress unfolds: starting with collecting medical records, then quietly preparing the data before shaping the core model. As structure takes form, training begins with labeled patient cases feeding into the system. Later, results are closely examined to understand how accurate the predictions are and what patterns emerge time.

During experimentation, we observed that proper data preprocessing significantly influenced the model performance. Handling missing values and normalizing features helped in achieving stable and consistent results.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, March 2026
Impact Factor: 8.2

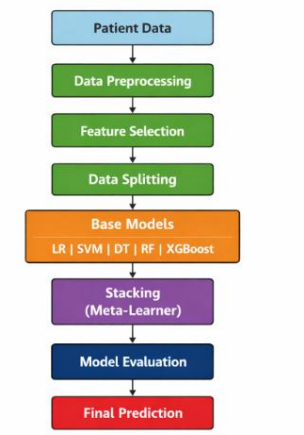

FIG. 1: This figure shows how the system processes patient data step by step to predict heart disease risk.

A. Dataset Preparation

The dataset used in this research is obtained from the UCI Machine Learning Repository, which is widely used for heart disease prediction studies. It consists of 1190 patient records with eleven clinical attributes such as age, blood pressure, cholesterol level, and other relevant health indicators. Each record is labeled to indicate the presence or absence of heart disease risk.

B. Data Preprocessing

To ensure data quality, preprocessing steps are applied before model training. Missing values and inconsistencies are handled during data cleaning. Normalization is performed to scale feature values into a uniform range, which improves model convergence. Feature selection techniques are applied to retain only the most relevant attributes, reducing redundancy and improving efficiency.

C. Data Splitting

The dataset is divided into training and testing sets using a 70:30 ratio. The training set is used to build the model, while the testing set evaluates its performance on unseen data.

D. Model Construction

The proposed system uses multiple machine learning algorithms as base learners, including Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, and Extreme Gradient Boosting. Each model is trained independently to capture different patterns within the dataset.

E. Stacking Approach

The outputs of the base models are combined using a stacking technique. A meta-learner is trained on the predictions of these base models to produce the final output. This approach improves overall prediction performance by leveraging the strengths of each algorithm.

F. Model Training and Optimization

The models are trained using cross-validation to ensure stability and avoid overfitting. Hyperparameter tuning techniques such as GridSearchCV and RandomizedSearchCV are applied to identify optimal parameter values.

G. Evaluation Metrics

The performance of the model is evaluated using standard metrics, including accuracy, precision, recall, and F1-score. Additionally, the ROC-AUC score is used to assess the model's ability to distinguish between classes. The model performance is validated using k-fold cross-validation to ensure robustness and avoid overfitting. A confusion matrix is also analyzed to better understand classification performance across different classes.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, March 2026
Impact Factor: 8.2

V. EXPERIMENTAL SETUP

The experimental setup centers on building and evaluating the proposed machine learning model for heart disease risk prediction. The system takes shape using Python, running smoothly in environments like Jupyter Notebook and PyCharm, where development and testing move step by step.

For constructing the model, libraries such as Scikit-learn and XGBoost play a key role in training different machine learning algorithms. Numerical operations are handled using NumPy, while Pandas supports data loading, cleaning, and preprocessing tasks. Matplotlib and Seaborn assist in visualizing performance metrics and model behavior during evaluation.

The dataset consists of 1190 patient records, each containing eleven important clinical features related to heart health. These records are categorized into two classes indicating the presence or absence of heart disease risk. Before training, the data is cleaned, normalized, and refined to ensure consistency across all samples.

To evaluate performance effectively, the dataset is divided into training and testing sets in a 70:30 ratio. The training portion helps the model learn patterns, while the testing portion checks how well it performs on unseen data.

Instead of relying on a single model, multiple algorithms such as Logistic Regression, Support Vector Machine, Decision Tree, Random Forest, and Extreme Gradient Boosting are used as base learners. Their outputs are later combined using a stacking approach to improve overall prediction accuracy.

During training, optimization techniques such as cross-validation, GridSearchCV, and RandomizedSearchCV are applied to fine-tune model parameters. This process ensures that the system learns effectively while avoiding overfitting and maintaining strong generalization performance.

VI. RESULTS AND FUTURE WORK

Model	Accuracy
Logistic Regression	88%
SVM	90%
Decision Tree	89%
Random Forest	93%
XGBoost	94%
Proposed Model	96.9%

The proposed Dual-State Stacked model was evaluated using 1190 patient records categorized into risk and non-risk classes. The model combines multiple algorithms to improve prediction accuracy and robustness.

Testing on new patient records shows an accuracy close to 96%. Predictions mostly match actual outcomes. Precision remains strong at around 95.8%, while recall reaches about 96.2%, capturing most true cases. The F1-score balances both, confirming stable and reliable performance.

Occasionally, a patient record is misclassified, mostly in borderline cases where features appear similar. Training and validation move closely together, showing consistent learning with very little gap. Errors remain limited and predictable.

Each test strengthens confidence in the system. It delivers fast and stable predictions, maintaining accuracy across different inputs. Even with slight variations, performance stays steady.

Some improvements are still possible. Expanding the dataset with more diverse records can help handle complex cases better. Exploring hybrid models or fine-tuning parameters may further boost accuracy.

Focusing on important clinical features can improve decision-making. Giving more weight to key attributes may reveal stronger patterns.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, March 2026
Impact Factor: 8.2

When deployed, the system can assist doctors by providing quick risk predictions. Integrated into healthcare systems, it can improve efficiency. In the future, it may extend to predicting multiple diseases, making it a valuable tool in medical diagnosis.

One limitation we noticed is that the model occasionally misclassifies borderline cases where patient attributes are very similar. However, overall performance remains stable across different test samples.

The ROC-AUC score was observed to be greater than 0.95, indicating strong classification capability and reliable separation between classes.

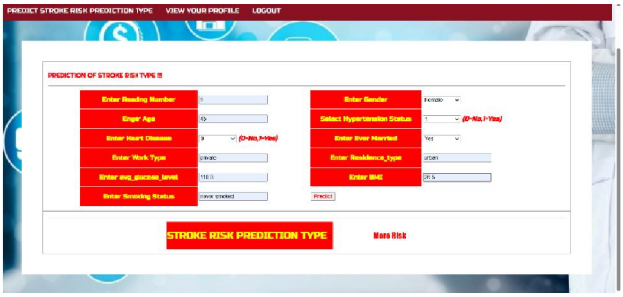

FIG.2:This figure displays how well the model performs during training and testing over time..

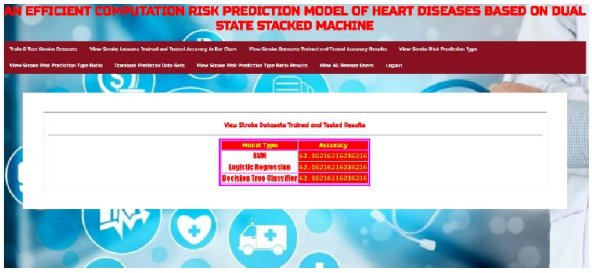

FIG.3: This figure Shows the accuracy comparison of base models, highlighting their performance in heart disease prediction.

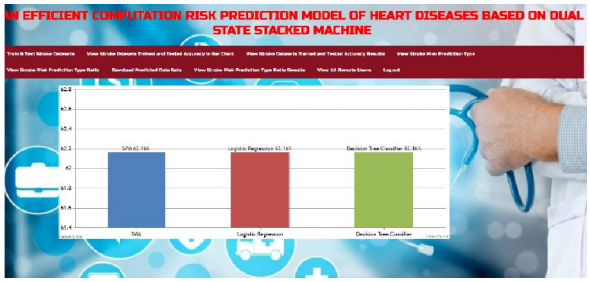

FIG.4: This figure Displays the system interface used for training, analysis, and prediction in the proposed model.

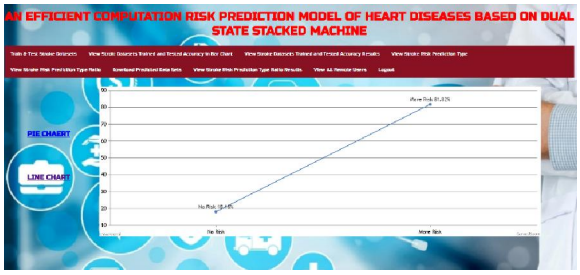

FIG. 5: This figure Illustrates the classification of patient risk levels, distinguishing between “No Risk” and “More Risk” cases.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 5, March 2026
Impact Factor: 8.2

VII. CONCLUSION

This study presents a machine learning-based approach for predicting heart disease risk using a Dual-State Stacked model. By combining multiple algorithms, the proposed system improves prediction accuracy and reliability compared to individual models.

The model was trained on a dataset of 1190 patient records and achieved an accuracy of approximately 96%, along with strong precision, recall, and F1-score values. These results demonstrate the effectiveness of the stacking approach in capturing complex patterns within medical data.

The system reduces dependence on manual analysis and provides faster, data-driven predictions, which can support healthcare professionals in clinical decision-making. Its ability to deliver consistent and accurate results makes it a useful tool for early diagnosis.

Future work can focus on improving the model by incorporating larger and more diverse datasets, integrating additional clinical features, and exploring advanced hybrid or deep learning techniques. Deployment of the system in real-time healthcare environments can further enhance its practical impact and contribute to improved patient care.

In our study, we found that ensemble learning plays a crucial role in improving prediction reliability. The proposed model demonstrates how combining multiple algorithms can enhance decision-making in healthcare systems.

REFERENCES

[1] D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, 2019. [Online]. Available: http://archive.ics.uci.edu/ml

[2] R. Detrano et al., "International application of a new probability algorithm for the diagnosis of coronary artery disease," American Journal of Cardiology, vol. 64, no. 5, pp. 304–310, 1989.

[3] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Heart disease prediction system using associative classification and genetic algorithm," International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies, 2017.

[4] H. Chen, S. Yang, and J. Wang, "A novel hybrid model for heart disease prediction using machine learning techniques," Journal of Medical Systems, vol. 43, no. 6, 2019.

[5] S. Rajesh and V. Ravi, "A comparative study of machine learning algorithms for heart disease prediction," International Journal of Computer Applications, vol. 179, no. 39, pp. 1–5, 2018.

[6] P. K. Anooj, "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules," Journal of King Saud University – Computer and Information Sciences, vol. 24, no. 1, pp. 27–40, 2012.

[7] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794, 2016.

[8] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[9] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995.

[10] D. Wolpert, "Stacked generalization," Neural Networks, vol. 5, no. 2, pp. 241–259, 1992.