iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VI **Month of publication:** June 2026

DOI: https://doi.org/10.22214/ijraset.2026.83687

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VI June 2026- Available at www.ijraset.com

Advanced Heart Attack Risk Prediction Using Stacked Hybrid Machine Learning

M. Vijaya Kumar[1], D. V. V. Manikanta Raju[2], A. D. M. Praveen[3], P. Baladithya[4], V. Sai Dinesh Kumar[5]
[1]*Assistant Professor,* [2,3,4,5]*Student, Department of CSE-Data Science, St. Ann's College of Engineering & Technology, Chirala, India*

Abstract: ***Heart disease remains one of the leading causes of mortality worldwide, making early prediction essential for effective prevention and timely treatment. This paper presents an advanced machine learning-based system for predicting heart attack risk using a stacked hybrid ensemble approach. The proposed system integrates multiple machine learning algorithms, including Random Forest, Decision Tree, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression, Gradient Boosting, and XGBoost.***
These models are combined using a stacking classifier with Logistic Regression as the meta-learner to achieve higher accuracy and reliability.
The system analyzes various patient health parameters such as age, cholesterol levels, blood pressure, heart rate, and other clinical factors from the Cleveland Heart Disease Dataset. Data preprocessing techniques including data cleaning, feature scaling using StandardScaler, feature selection, and stratified train-test splitting are applied to improve robustness. Experimental results demonstrate that the stacked hybrid model achieves a prediction accuracy of approximately 89–90%, outperforming individual base classifiers. The proposed solution offers a non-invasive, cost-effective, and accurate method for early detection of heart disease, thereby supporting healthcare professionals in making informed clinical decisions.
Keywords: Heart Disease Prediction, Machine Learning, Stacked Hybrid Model, Ensemble Learning, Random Forest, XGBoost, Support Vector Machine, Logistic Regression, Cleveland Heart Disease Dataset

I. INTRODUCTION

In today's rapidly advancing healthcare environment, the early detection and prevention of life-threatening diseases have become increasingly important. Among these diseases, heart disease continues to be one of the leading causes of death worldwide, accounting for a significant proportion of global mortality. Identifying the risk of a heart attack at an early stage can greatly reduce mortality rates and enhance patient survival and quality of life. However, conventional diagnostic approaches largely depend on manual evaluation by medical professionals, which can be time-consuming, less efficient, and sometimes prone to human error.

With the emergence of Artificial Intelligence (AI) and Machine Learning (ML), the healthcare sector is shifting towards more intelligent and data-driven systems [1]. Machine learning algorithms have the capability to process and analyze large volumes of medical data to uncover complex patterns and relationships that are often difficult to detect using traditional methods [2]. These technologies enable faster, more accurate, and consistent predictions, thereby supporting healthcare professionals in making more informed and timely decisions.

This paper presents an Advanced Heart Attack Risk Prediction System using a stacked hybrid machine learning approach. The system incorporates multiple machine learning algorithms, including Random Forest, Decision Tree, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression, XGBoost, and Gradient Boosting. These individual models are integrated using a stacking technique, where their outputs are combined through a meta-learner (Logistic Regression) to enhance overall prediction performance and accuracy [3]. The system evaluates various medical parameters such as age, blood pressure, cholesterol levels, heart rate, and other significant health indicators from the Cleveland Heart Disease Dataset [4].

Advanced techniques such as feature scaling using StandardScaler, hyperparameter tuning, cross-validation, and feature engineering are applied to improve model performance and reliability. Compared to traditional single-model approaches, the stacked hybrid model offers superior accuracy, better generalization across datasets, and increased dependability, making it suitable for deployment in hospitals, clinics, and personal health monitoring platforms [5].

II. LITERATURE SURVEY

A. Machine Learning for Heart Disease Prediction

Machine learning-based heart disease prediction systems play a vital role in enabling early detection of cardiovascular risks by efficiently analyzing patient data [2]. The key steps involved in such systems include data collection from reliable sources, data preprocessing for cleaning and handling missing values, feature selection for identifying important attributes, model training using classification algorithms, and prediction using ensemble techniques for improved accuracy [6].

Logistic Regression has been widely used for binary classification in medical prediction tasks, providing a simple yet effective baseline [7]. Decision Tree and Random Forest models have demonstrated improved accuracy through rule-based and ensemble approaches, respectively, though Random Forest incurs higher computational costs [8]. Support Vector Machine (SVM) has been shown to be effective for high-dimensional datasets, identifying optimal hyperplanes to separate data classes, though it requires careful parameter tuning [9]. Advanced boosting algorithms such as XGBoost and Gradient Boosting have been increasingly adopted for their ability to improve model performance and handle complex patterns efficiently [3].

B. Challenges in Heart Attack Prediction

Despite significant advancements, several challenges persist in heart attack risk prediction. Data quality issues arising from missing or inconsistent records can significantly reduce prediction accuracy, necessitating robust preprocessing techniques [4]. Imbalanced datasets, where the distribution between healthy and diseased patient records is unequal, can bias model predictions and require techniques such as SMOTE for resampling [10]. Feature selection complexity arises because not all features contribute equally to prediction outcomes, requiring algorithmic approaches to identify the most relevant attributes. Overfitting, where a model performs well on training data but poorly on unseen data, remains a concern that is addressed through cross-validation and regularization techniques [5].

C. Hybrid and Ensemble Models

Recent research has demonstrated that hybrid models combining multiple machine learning algorithms through techniques such as voting and stacking consistently outperform individual classifiers [3], [5]. These ensemble methods leverage the strengths of individual algorithms while mitigating their respective limitations, resulting in higher accuracy, better generalization, and reduced overfitting. Studies have shown that stacking classifiers with a meta-learner can achieve superior prediction performance compared to standalone models, making them particularly suitable for critical applications such as cardiovascular disease prediction [6], [7].

TABLE I
COMPARISON OF MACHINE LEARNING MODELS FOR HEART DISEASE PREDICTION

Model	Advantages	Limitations
Logistic Regression	Simple, fast, easy to interpret	Less accurate for complex data
Decision Tree	Easy to understand, rule-based	Prone to overfitting
Random Forest	High accuracy, reduces overfitting	More computational cost
KNN	Simple, works well with small datasets	Slow for large datasets
SVM	Effective for high-dimensional data	Requires proper parameter tuning
XGBoost	Improved performance and speed	Complex hyperparameter tuning

III. PROPOSED METHODOLOGY

A. Proposed System Overview

The proposed system is an Advanced Heart Attack Risk Prediction System developed using a stacked hybrid machine learning approach. Unlike traditional diagnostic methods that depend on manual evaluation and basic statistical techniques, this system offers an automated, data-driven, and more accurate prediction mechanism. The system evaluates patient medical data including age, blood pressure, cholesterol levels, heart rate, and other important health parameters [1], [4].

Multiple machine learning algorithms, including Random Forest, Decision Tree, K-Nearest Neighbors (KNN), Support Vector Machine (SVM), Logistic Regression, XGBoost, and Gradient Boosting, are utilized to assess the risk of a heart attack.

These individual models are integrated using a stacking technique, where the predictions generated by each model are passed to a meta-model (Logistic Regression) that produces the final output. This hybrid approach enhances prediction accuracy, and advanced techniques such as feature scaling, cross-validation, and hyperparameter tuning are applied to improve performance [3], [5].

B. System Workflow

The proposed system operates through a structured pipeline consisting of five sequential stages. In the first stage, patient medical data is collected from the Cleveland Heart Disease Dataset and subjected to preprocessing techniques including handling missing values, normalization using StandardScaler, and data cleaning. The second stage involves feature selection, where important features such as age, cholesterol level, blood pressure, and heart rate are identified using feature engineering and selection techniques to improve model performance and reduce complexity.

In the third stage, multiple machine learning models are trained independently using an 80:20 stratified train-test split with cross-validation and hyperparameter tuning. The fourth stage implements the stacking hybrid model, where predictions from all base models (Random Forest, XGBoost, SVM, and Gradient Boosting) are combined and passed to the meta-model (Logistic Regression) for final prediction generation. In the fifth stage, the system classifies the patient as either high-risk or low-risk based on the final output, thereby supporting further medical decision-making [6].

TABLE II
COMPARISON WITH TRADITIONAL DIAGNOSIS METHODS

Feature	**Traditional Methods**	**Proposed ML System**
Prediction Speed	Slow (Manual analysis)	Fast (Automated)
Accuracy	Moderate	High (~89–90%)
Human Dependency	High	Low
Data Handling	Limited	Handles large datasets
Early Detection	Difficult	Efficient
Reliability	Less consistent	More reliable

IV. SYSTEM DESIGN

A. System Architecture

The design of the Heart Attack Risk Prediction System is structured to ensure accurate prediction, efficient data processing, and reliable decision-making. The system follows a modular, multi-layered architecture where different modules interact to ensure accurate and efficient prediction. The architecture integrates machine learning models, data preprocessing techniques, and a stacked hybrid learning approach to deliver a scalable prediction process.

The architecture consists of five key components. The Data Acquisition Module collects patient medical data from datasets or user input, including parameters such as age, blood pressure, cholesterol, and heart rate. The Data Preprocessing Module cleans and prepares raw data by handling missing or inconsistent values and applying normalization and scaling techniques. The Feature Selection Module identifies the most important features influencing heart disease using feature selection and engineering techniques. The Machine Learning and Stacking Module applies multiple algorithms (Random Forest, XGBoost, Gradient Boosting, SVM, and Logistic Regression) independently and combines their outputs through a stacking classifier. Finally, the Prediction and Output Module analyzes the final stacking model output and classifies the result into high-risk or low-risk categories [3], [8].

B. Dataset Description

The dataset used in this study is the Cleveland Heart Disease Dataset, which is widely used for heart disease prediction tasks [4]. The dataset consists of multiple medical attributes related to patients that are essential for predicting heart attack risk. Each record corresponds to an individual patient with specific health-related features. The dataset contains 14 attributes including age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, ST depression (oldpeak), slope of the ST segment, number of major vessels, thalassemia type, and the target variable indicating presence (1) or absence (0) of heart disease.

TABLE III
DATASET ATTRIBUTES DESCRIPTION

Attribute	Description
Age	Age of the patient
Sex	Gender of the patient
Chest Pain Type	Type of chest pain experienced
Resting BP	Blood pressure at rest
Cholesterol	Serum cholesterol level
Fasting Blood Sugar	Blood sugar level after fasting
Resting ECG	Electrocardiographic results
Max Heart Rate	Maximum heart rate achieved
Exercise Angina	Exercise-induced chest pain
Oldpeak	ST depression value
Slope	Slope of ST segment
Major Vessels	Number of major vessels colored by fluoroscopy
Thalassemia	Blood disorder type
Target	Heart disease presence (0 = No, 1 = Yes)

C. Data Preprocessing

The dataset undergoes several preprocessing steps to improve data quality and ensure accurate predictions. Data cleaning removes missing, noisy, and inconsistent values along with duplicate records. Data transformation converts categorical data into numerical format suitable for machine learning algorithms. Normalization using StandardScaler is applied to scale features and bring them to a common range. Feature selection identifies the most relevant features for prediction and removes irrelevant or redundant attributes, improving model performance and reducing computational complexity. The dataset is divided using an 80:20 stratified train-test split to ensure balanced representation of classes in both training and testing sets [4], [10].

V. IMPLEMENTATION

A. Development Environment

The system is developed using Python 3.8+ and utilizes several essential libraries and frameworks. Pandas and NumPy are used for data handling and numerical computations. Scikit-learn provides implementations of machine learning models including Random Forest, Decision Tree, KNN, SVM, Logistic Regression, and the Stacking Classifier. XGBoost is used for gradient boosting in the hybrid ensemble model. Matplotlib and Seaborn are employed for data visualization and performance analysis. Joblib is used for serializing and loading trained models. The user interface is developed using the Streamlit framework, providing a web-based interactive platform for prediction.

B. Model Training and Stacking

The implementation follows a modular pipeline. The data input module collects patient details and ensures correct formatting. The preprocessing module handles missing values, applies StandardScaler normalization, and transforms categorical data. The feature selection module removes irrelevant features using engineering techniques. In the model training module, base models (Random Forest, XGBoost, SVM, and Gradient Boosting) are trained independently on the processed dataset. The stacking module then combines outputs from all base models using a StackingClassifier with Logistic Regression as the final estimator. Each base model is trained using cross-validation, and the meta-model processes the combined predictions to generate the final output classifying patients as high-risk or low-risk [3], [5].

C. User Interface

The prediction interface is implemented using Streamlit, providing a web-based application where users can enter patient medical details such as age, blood pressure, cholesterol, chest pain type, and other clinical parameters. The system processes the input data through the trained stacking model and displays the prediction result indicating whether the patient is at high or low risk of heart disease, along with a confidence score and probability distribution.

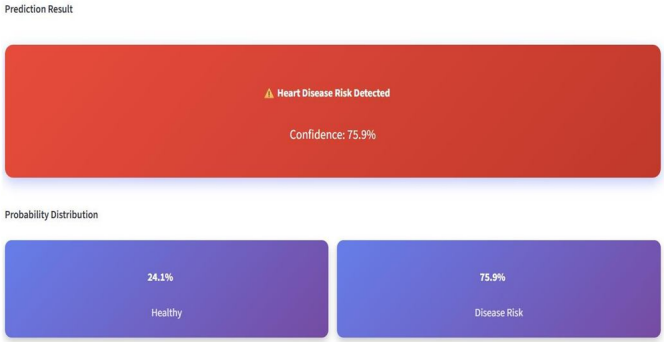


Figure 1. showing the Prediction Result for a sample Output.

VI. RESULTS AND DISCUSSION

A. Experimental Results

The Heart Attack Risk Prediction System was evaluated using the Cleveland Heart Disease Dataset with an 80:20 stratified train-test split. The stacked hybrid model, combining Random Forest, XGBoost, SVM, and Gradient Boosting with Logistic Regression as the meta-learner, achieved an overall prediction accuracy of approximately 89–90%. This performance surpassed that of the individual base classifiers, demonstrating the effectiveness of the stacking approach in improving prediction reliability [5], [6].

Evaluation metrics including precision, recall, and F1-score confirmed that the model performs well in identifying both high-risk and low-risk patients. The stacking model effectively reduced the prediction errors inherent in individual models by leveraging the complementary strengths of diverse algorithms. Performance testing demonstrated that the system generates predictions within a few seconds, ensuring practical applicability in clinical settings.

B. Test Case Validation

The system was validated through multiple test case scenarios. In Test Case 1, patient data representative of a high-risk profile was provided as input. The stacked hybrid model correctly classified the patient as high-risk with a confidence score of approximately 75.9%, with the probability distribution indicating a higher likelihood of disease presence. In Test Case 2, patient data representative of a low-risk profile was entered. The system correctly identified the patient as low-risk with a confidence score of approximately 62.6%, with the probability distribution showing a higher likelihood of the patient being healthy.

TABLE IV
TEST CASE SCENARIOS AND OUTCOMES

Test ID	Scenario	Expected Output	Status
TC-01	Valid patient data input	Accepted without errors	Pass
TC-02	Data preprocessing	Missing values handled	Pass
TC-03	Model training	Trained successfully	Pass
TC-04	Prediction (high-risk data)	High Risk (~75.9%)	Pass
TC-05	Prediction (low-risk data)	Low Risk (~62.6%)	Pass
TC-06	Stacking vs individual	Stacking outperforms	Pass
TC-07	Invalid input handling	Handled gracefully	Pass

All test scenarios passed successfully, confirming the robustness and reliability of the system. The comprehensive testing validated that the system effectively handles valid and invalid inputs, performs accurate predictions using the stacked hybrid model, and generates results within acceptable response times.

VII. CONCLUSION

This paper presented an Advanced Heart Attack Risk Prediction System using a stacked hybrid machine learning approach. The system integrates multiple machine learning algorithms, including Random Forest, Decision Tree, KNN, SVM, Logistic Regression, XGBoost, and Gradient Boosting, combined through a stacking classifier with Logistic Regression as the meta-learner. By leveraging the complementary strengths of diverse algorithms, the stacked hybrid model achieved a prediction accuracy of approximately 89%, outperforming individual base classifiers. The system demonstrated effective data preprocessing, feature selection, and model training using the Cleveland Heart Disease Dataset. A user-friendly web interface was developed using Streamlit, enabling users to input patient data and obtain rapid, accurate prediction results. The system provides a non-invasive, cost-effective, and reliable method for early detection of heart disease, supporting healthcare professionals in making informed clinical decisions. However, certain limitations exist. The accuracy of the system depends on the quality and size of the dataset used for training. A limited number of health parameters are considered, and the system may not generalize equally well to unseen datasets from different populations. Additionally, the system is not currently integrated with real-time hospital information systems.

VIII. FUTURE SCOPE

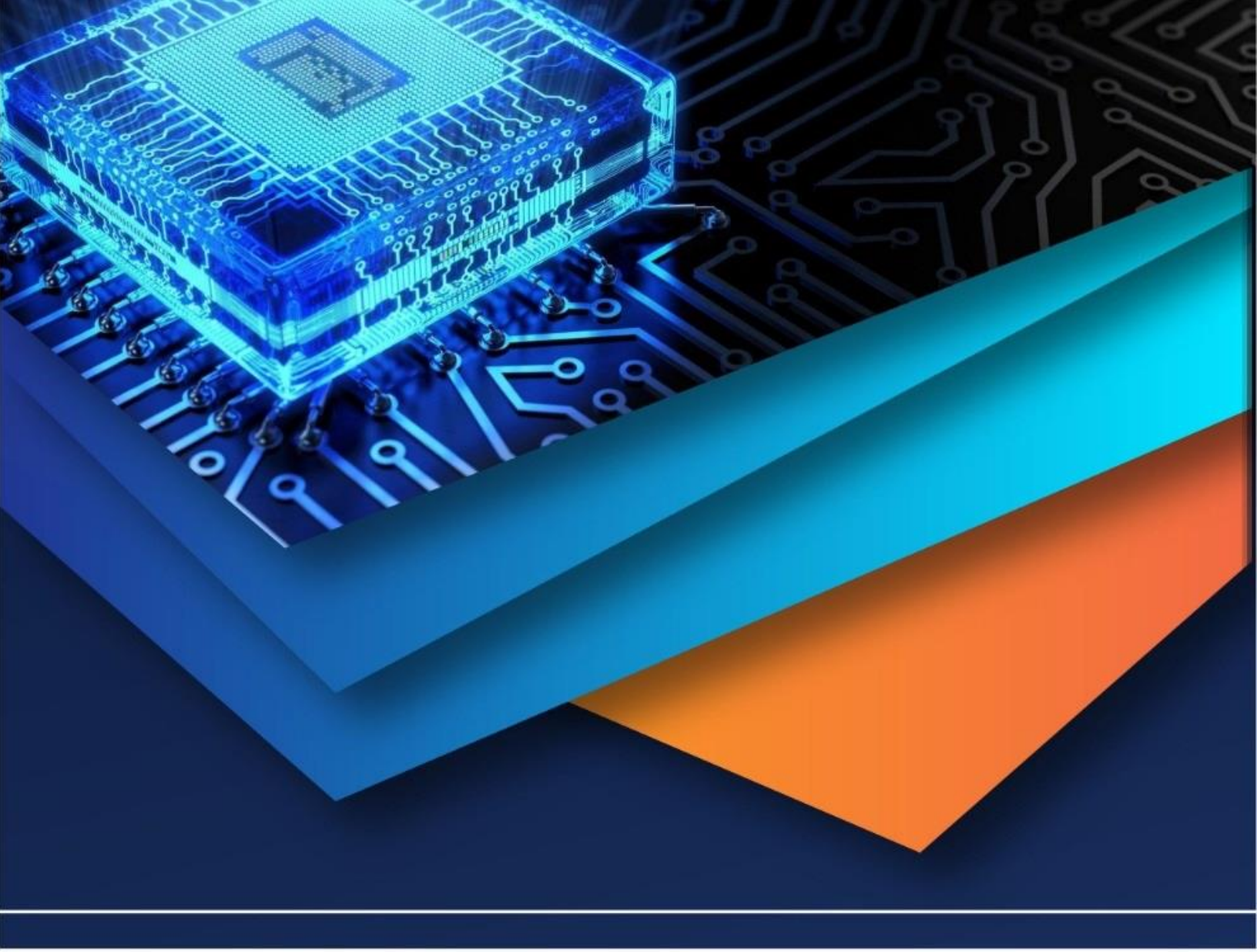
Future improvements can focus on enhancing accuracy, scalability, and real-world usability. The system can be improved by incorporating larger and more diverse medical datasets from multiple sources and populations. Advanced deep learning techniques, such as convolutional neural networks and recurrent neural networks, can be integrated for potentially better performance on complex feature interactions. Real-time integration with hospital databases and Electronic Health Record (EHR) systems can enable seamless clinical deployment. Integration with wearable devices such as smartwatches can enable continuous health monitoring and early warning capabilities. Cloud-based deployment and mobile application development can provide remote access and ease of use for both patients and healthcare providers. The system can also be extended to predict other cardiovascular and chronic diseases such as diabetes and stroke. Ensuring data privacy, security, and compliance with healthcare regulations will be critical for real-world deployment [8].

IX. ACKNOWLEDGMENT

The authors express sincere gratitude to Mr. M. Vijaya Kumar for his valuable guidance and supervision throughout this project. The authors also thank the Head of the Department, Dr. K. Subba Rao, the Principal, Dr. K. Jagadeesh Babu, and the management of St. Ann’s College of Engineering & Technology, Chirala, for providing the necessary support and infrastructure.

REFERENCES

[1] World Health Organization (WHO), "Cardiovascular Diseases (CVDs)," Fact Sheet. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
[2] S. Raheja and N. Ray, "Detection of heart disease using machine learning," in Proc. Int. Conf. on Artificial-Business Analytics, Quantum and Machine Learning, Singapore: Springer Nature, 2023, pp. 1–8.
[3] P. Sharma, R. Gupta, and A. Kaur, "Hybrid BiLSTM-GRU model for coronary heart disease prediction using randomized search cross-validation," Journal of Healthcare Engineering, vol. 2023, pp. 1–12, Apr. 2023.
[4] P. Balakrishnan and R. Kumar, "IoT-enabled cardiovascular risk prediction using recurrent convolutional neural networks and fuzzy C-means," IEEE Trans. on Industrial Informatics, vol. 19, no. 7, pp. 2345–2354, Jul. 2023.
[5] B. Nandy, A. Dey, and D. Goswami, "Swarm-ANN: A swarm intelligence-based artificial neural network for heart disease prediction," Applied Soft Computing, vol. 110, pp. 107677, Oct. 2021.
[6] R. Elsedimy, S. Ibrahim, and M. Abdelghany, "Quantum-behaved particle swarm optimization-support vector machine model for cardiovascular disease prediction," Int. J. of Computational Intelligence Systems, vol. 16, no. 4, pp. 239–254, Apr. 2023.
[7] X. Cai, J. Li, and Y. Wang, "Independent validation of AI cardiovascular risk models: A comprehensive review and development of independent validation score (IVS)," Journal of Medical Systems, vol. 48, no. 1, pp. 12–28, Jan. 2024.
[8] M. M. Islam, T. Nasrin, and A. Uddin, "Real-time cardiovascular disease prediction system using IoT and machine learning," Journal of Healthcare Informatics Research, vol. 7, no. 3, pp. 285–302, Sep. 2023.
[9] A. Hossain, M. Miah, and M. H. Kabir, "Feature selection in random forest models for accurate heart disease prediction," Computers in Biology and Medicine, vol. 153, pp. 106415, Aug. 2023.
[10] E. K. Dritsas and M. Trigka, "Ensemble machine learning for heart disease prediction with SMOTE: Addressing class imbalance in medical data," Medical Informatics and Decision Making, vol. 23, no. 5, pp. 89–103, Nov. 2023.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY