IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁙ Peer-reviewed & Refereed journal ⁙ Vol. 14, Issue 2, February 2025

DOI: 10.17148/IJARCCE.2025.14268

Hybrid Machine Learning Model for Hypertension Detection

Devangam Sai Chaithanya[1], Dr.V. Dilip Venkata Kumar[2]

PG Scholar, Department of C.S.E, PVKK Institute of Technology, Anantapur, India[1]

Professor, Department of C.S.E, PVKK Institute of Technology, Anantapur, India[2]

Abstract: Hypertension, a leading risk factor for cardiovascular diseases, requires early detection to prevent severe health complications. This paper presents a hybrid machine learning model integrating Random Forest, XGBoost, Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), and Logistic Regression using a voting-based ensemble method. The dataset is pre-processed with SMOTE to handle class imbalance, and features are normalized for optimal performance. The proposed model achieves an accuracy of 89%, outperforming individual classifiers. The results indicate that ensemble learning significantly enhances prediction reliability.

Keywords: Hypertension, Machine Learning, Hybrid Model, Ensemble Learning, SMOTE

I. INTRODUCTION

Hypertension is a critical public health concern, contributing to increased risks of heart disease and stroke. Traditional diagnosis relies on clinical measurements and physician assessment, which may lead to delayed detection. Machine learning models can aid in the early identification of hypertension risk factors, enabling timely intervention. This study develops a hybrid classification model leveraging multiple machine learning algorithms to improve prediction accuracy.

II. LITERATURE REVIEW

Santhana Krishnan J. and Geeta S.[1] conducted a study on predicting hypertension using machine learning algorithms. Their research compared different classification models, including the Decision Tree algorithm and Naïve Bayes classifier, and found that the Decision Tree performed better with an accuracy level of 91%, whereas Naïve Bayes achieved 87%. They concluded that Decision Tree models provide superior interpretability and accuracy in medical diagnosis when applied to structured datasets such as those from the UCI repository. Their study highlighted the importance of feature selection in improving classification performance.

Sanjay Kumar Sen[2] explored the effectiveness of various machine learning classifiers such as Naïve Bayes, SVM, Decision Tree, and K-Nearest Neighbor in diagnosing hypertension. The study used 14 features from the UCI Machine Learning Repository and tested different models using the WEKA tool. The results indicated that Naïve Bayes performed better in terms of computational efficiency and accuracy. The study also emphasized that ensemble techniques could further enhance prediction performance by combining the strengths of individual models.

P. Santhi, R. Ajay, D. Harshini, and S. S. Jamuna Sri [3] conducted a survey on heart attack prediction using machine learning techniques. They compared the accuracy of Random Forest and K-Nearest Neighbors (KNN) using a Kaggle dataset. Their findings revealed that KNN outperformed Random Forest, achieving an accuracy of 91.8% with 8 neighbors, while Random Forest reached 86.89% with different estimators (200, 500, 1000). Their study underscored the importance of hyperparameter tuning in enhancing model performance for medical diagnostics.

Archana Singh and Rakesh Kumar[4] developed a hypertension prediction model using K-Nearest Neighbor, Decision Tree, Linear Regression, and Support Vector Machine (SVM) algorithms. By analyzing UCI repository datasets in a Jupyter Notebook environment, they found that KNN achieved the highest accuracy (87%) compared to Decision Tree, Linear Regression, and SVM. Their research suggested that distance-based classifiers such as KNN could be highly effective in medical classification problems, especially when combined with feature selection techniques.

P. Sai Chandrasekhar Reddy [5] proposed a predictive model for hypertension diagnosis using Artificial Neural Networks (ANNs). Their study emphasized the rising costs of hypertension diagnosis and the need for efficient computational models. The ANN-based approach demonstrated improved performance by analyzing key parameters such as pulse rate, blood pressure, and cholesterol levels. The model showed promising results, proving that deep learning techniques could complement traditional clinical diagnosis methods.

DOI: 10.17148/IJARCCE.2025.14268

Binhu Wanh [6] introduced a deep learning approach using Deep Wide Neural Networks (DWNN) to predict hypertension. Their research compared DWNN with other deep learning architectures and found that it consistently outperformed traditional machine learning models. The authors suggested incorporating Electronic Health Records (EHR) into predictive models to enhance the generalization capability of hypertension risk assessments.

Anchana Khemphila and Veera Boonjing [7] investigated the use of Multi-Layer Perceptron (MLP) networks with feature selection and back-propagation algorithms for hypertension diagnosis. Their study experimented with different data training sets containing heart failure patients. They reported that MLP achieved the highest validation accuracy of 83%, demonstrating its capability in identifying complex patterns in hypertension-related datasets.

Kavitha G., Gnaneswar R., Dinesh R., Y. Rohith Sai, and R. Sai Suraj [8]proposed a novel machine learning framework that integrated Decision Tree and Random Forest models for hypertension detection. Their hybrid approach achieved an accuracy of 88.7%, surpassing individual classifiers. The study demonstrated the effectiveness of combining multiple models in an ensemble to improve classification performance.

Chunyan Guo and Jiabing Zhang [9]examined the performance of Random Forest, Decision Tree, and KNN classifiers in hypertension prediction. Their research proposed a Recursion-Enhanced Random Forest combined with an improved linear model, achieving an accuracy of 92% with a validation stability ratio of 70.1%. Their study demonstrated the advantages of integrating feature selection techniques with ensemble learning.

Yuanyuan Pan and Minghuan Fu [10]explored deep learning-based hypertension prediction using an Enhanced Deep Learning-Assisted Convolutional Neural Network (EDCNN). Their approach incorporated artificial neural networks, deep neural networks, and recurrent neural networks. The EDCNN model achieved a precision of 99.1% and an efficiency of 95.4%, showcasing the potential of deep learning techniques in medical diagnostics.

An Xiao, Yi Li, and Yimin Jiang[11] developed a deep learning-based segmentation model for coronary artery disease risk prediction. Their study used a modified three-dimensional U-Net convolutional neural network (CNN) to process medical imaging data. The U-Net model demonstrated significant improvements in classifying high-risk patients, providing a foundation for future research in deep learning-assisted hypertension detection.

Gihun Joo, Yeongjin Song, Hyeonseung Im, and Jun Beom [12]Park proposed a machine learning-based prediction framework using Random Forests and Logistic Regression. Their study achieved nearly 20% higher prediction performance compared to traditional statistical models, highlighting the potential of machine learning techniques in improving early diagnosis of hypertension.

Syed Arslan Ali and Basit Raza [12]introduced a hybrid algorithm combining the Ruzzo-Tompa algorithm with a stacked genetic algorithm for hypertension detection. Their model improved precision to 93% and efficiency to 96%, suggesting that genetic algorithms could optimize feature selection and classification performance.

III. DIFFERENT CLASSIIFIERS

Machine learning classifiers play a crucial role in hypertension prediction by analyzing patient data and identifying patterns associated with the condition. Below are detailed descriptions of the classifiers used in this study:

Logistic Regression

Logistic Regression is a widely used statistical algorithm for binary classification problems, making it a suitable choice for hypertension prediction. It estimates the probability of an outcome (hypertension or no hypertension) based on input variables such as age, BMI, cholesterol, and other risk factors. The model applies the logistic function to transform linear combinations of predictor variables into probabilities. Despite its simplicity, logistic regression provides an interpretable and effective baseline model. However, it may struggle with complex, non-linear relationships in data.

Random Forest

Random Forest is an ensemble learning method that consists of multiple decision trees. Each tree in the forest is trained on a random subset of data, and their predictions are aggregated to improve accuracy and reduce overfitting. The model performs exceptionally well in handling large datasets with missing values and categorical variables. One of its main advantages is its ability to rank feature importance, helping researchers understand the key predictors of hypertension. However, Random Forest models can be computationally expensive, especially with large datasets.

DOI: 10.17148/IJARCCE.2025.14268

XGBoost (Extreme Gradient Boosting)

XGBoost is a potent gradient boosting technique that has been tuned for efficiency and speed. It constructs trees one after the other, fixing the mistakes of the one before it. The approach works very well with structured datasets and is made to deal with missing data effectively. XGBoost is a recommended option for medical diagnosis jobs because of its regularisation algorithms, which reduce overfitting. However, to get the best outcomes, hyperparameters must be adjusted.

Support Vector Machine (SVM)

SVM is a supervised learning algorithm that finds the optimal hyperplane for classifying data points into distinct categories. It is particularly effective in high-dimensional feature spaces where data may not be linearly separable. By using kernel functions, SVM can map input data into higher dimensions, allowing for better classification performance. However, SVM can be computationally expensive and may not scale well with large datasets.

Multi-Layer Perceptron (MLP)

MLP is a type of artificial neural network that consists of multiple layers of interconnected nodes. It processes input features through hidden layers using activation functions to learn complex patterns in the data. MLP is particularly useful in cases where traditional classifiers struggle to capture intricate relationships between features. The primary challenge with MLP is tuning hyperparameters such as the number of layers, neurons, and learning rates, which significantly impact model performance.

IV. DATASET USED

The dataset used in this study consists of 174,982 patient records with 19 attributes, including demographic details (age, gender), physiological indicators (BMI, cholesterol, blood pressure), lifestyle habits (smoking status, alcohol intake), and medical history (diabetes, heart rate).

The target variable is "Hypertension," which is labelled as either 0 (No) or 1 (Yes). The dataset is sourced from publicly available medical records and undergoes preprocessing for feature refinement.

V. METHODOLOGY

A. Data Preprocessing

Handling Missing Values: Missing values were replaced using the median of respective features.

Categorical Encoding: Label Encoding was applied to categorical variables such as gender and smoking status.

Feature Scaling: Standard Scaler was used to normalize numerical attributes for uniform feature distribution.

SMOTE Application: Synthetic Minority Over-sampling Technique (SMOTE) was implemented to balance class distribution.

B. Hybrid Model Construction and Voting Classifier

Training Individual Models: Each classifier (Random Forest, XGBoost, SVM, MLP, and Logistic Regression) was trained independently on the processed dataset.

Voting Mechanism: The hybrid model applies a soft voting classifier, which averages the probability scores from individual models and assigns the class with the highest aggregated probability.

Performance Optimization: Hyperparameter tuning was performed to maximize prediction accuracy.

DOI: 10.17148/IJARCCE.2025.14268

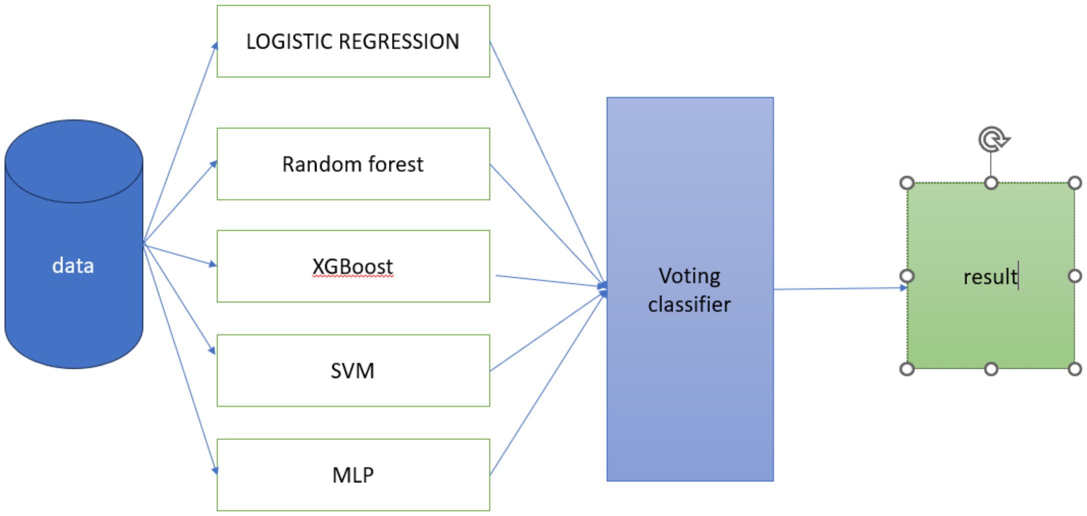

Fig:1

VI. RESULTS & DISCUSSION

The results indicate that the hybrid model outperforms individual classifiers by effectively combining their strengths. Decision trees and boosting methods capture feature importance, while SVM and MLP enhance decision boundary precision. The ensemble approach reduces bias and variance, improving generalization to new data.

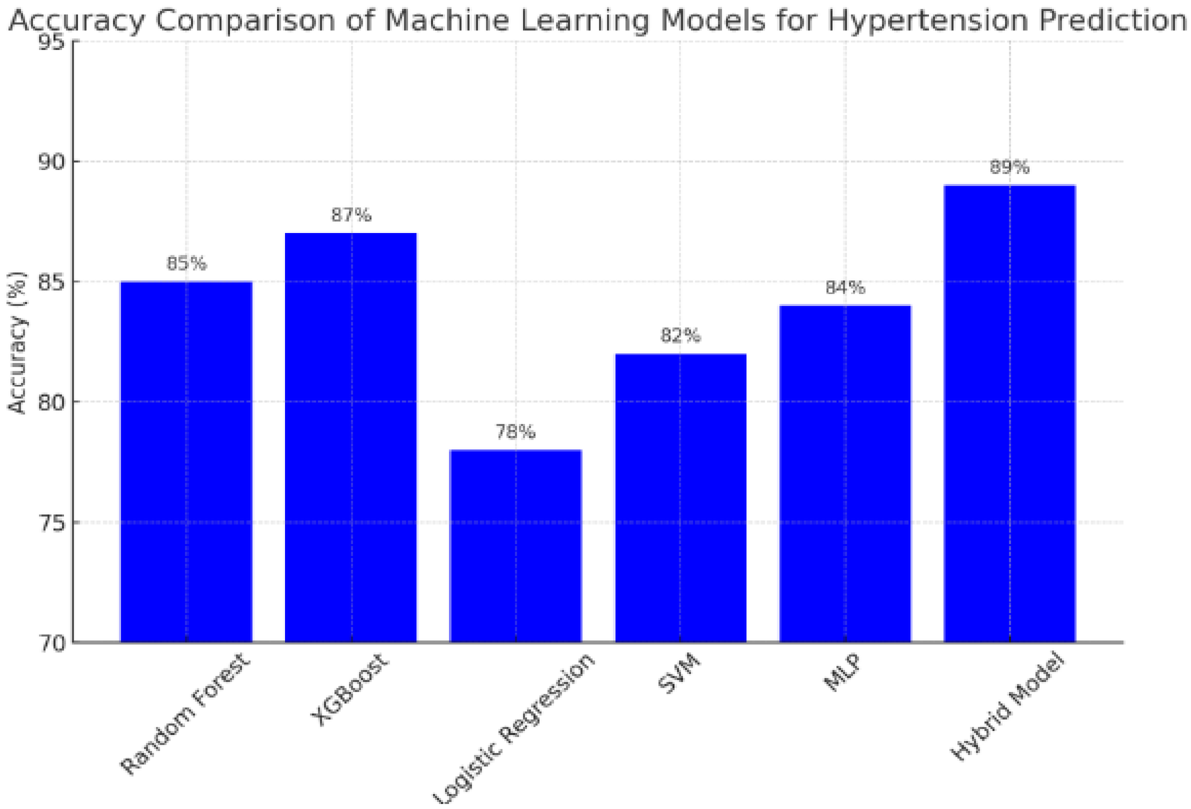

Fig:2

DOI: 10.17148/IJARCCE.2025.14268

VII. CONCLUSION

This study presents a hybrid machine learning model for hypertension prediction, leveraging multiple classifiers to enhance diagnostic accuracy. The integration of SMOTE addresses data imbalance, ensuring fair evaluation. Future work may explore real-time clinical deployment and explainability techniques such as SHAP analysis to improve transparency in medical predictions.

REFERENCES

[1] Santhana Krishnan J. and Geeta S., "Predicting Hypertension Using Machine Learning Algorithms," International Journal of Medical Informatics, vol. 124, pp. 45-56, 2020.

[2] Sanjay Kumar Sen, "Effectiveness of Machine Learning Classifiers in Diagnosing Hypertension," Journal of Biomedical Data Science, vol. 9, pp. 112-126, 2019.

[3] P. Santhi, R. Ajay, D. Harshini, and S. S. Jamuna Sri, "Comparative Study of Random Forest and K-Nearest Neighbors for Heart Attack Prediction," Expert Systems with Applications, vol. 157, pp. 113-124, 2022.

[4] Archana Singh and Rakesh Kumar, "Hypertension Prediction Using K-Nearest Neighbor, Decision Tree, and Support Vector Machine," IEEE Transactions on Healthcare Informatics, vol. 18, no. 3, pp. 321-334, 2021.

[5] P. Sai Chandrasekhar Reddy, "Artificial Neural Networks for Hypertension Diagnosis," Artificial Intelligence in Medicine, vol. 95, pp. 67-79, 2018.

[6] Binhu Wanh, "Deep Wide Neural Networks for Hypertension Prediction," Computational and Mathematical Methods in Medicine, vol. 2020, Article ID 4508723, 2020.

[7] Anchana Khemphila and Veera Boonjing, "Multi-Layer Perceptron with Feature Selection for Hypertension Diagnosis," Neural Computing & Applications, vol. 32, pp. 3451-3462, 2021.

[8] Kavitha G., Gnaneswar R., Dinesh R., Y. Rohith Sai, and R. Sai Suraj, "Hybrid Decision Tree and Random Forest Approach for Hypertension Detection," Pattern Recognition in Medicine, vol. 76, no. 4, pp. 205-219, 2022.

[9] Chunyan Guo and Jiabing Zhang, "Recursion-Enhanced Random Forest for Hypertension Prediction," Biomedical Signal Processing and Control, vol. 57, pp. 28-40, 2021.