ISSN [Online]: 2583-2654

Heart Disease Prediction using A Hybrid ML Model

[1]Sangeetha V,[2] Ashwin[i] S,[3]Anusha G, [4]D Rithika, [5]Shaguftha Shabbir, [6]Varshini R
Department of CSE, Dr. T. Thimmaiah Institute of Technology
KGF, India

Abstract: Heart disease remains a leading cause of global mortality, and early detection is critical for effective treatment. This paper proposes a hybrid model combining machine learning and deep learning techniques to improve heart disease prediction. Random Forest and XGBoost are used for feature selection to identify key clinical indicators. These features are then fed into a hybrid CNN-LSTM model that captures both spatial and temporal patterns. The model is trained using the UCI Heart Disease dataset and evaluated using metrics such as accuracy, precision, recall, and F1-score. The proposed system demonstrates improved prediction performance over standalone models and supports early diagnosis through a user-friendly interface that generates graphical outputs and downloadable health reports.

Keywords: CNN, LSTM, Heart Disease Prediction, Random Forest, XGBoost, Machine Learning, Deep Learning.

I. INTRODUCTION

Heart disease is a major global health challenge, responsible for approximately 17.9 million deaths annually, according to the World Health Organization (WHO). It includes a range of conditions affecting the heart and blood vessels, such as coronary artery disease, arrhythmias, and heart failure. Due to lifestyle changes, aging populations, and genetic factors, the number of heart disease cases is rising steadily. Early detection plays a crucial role in reducing complications and improving survival rates. However, traditional diagnostic methods are often time-consuming, dependent on specialized clinical expertise, and not always accessible, especially in low-resource settings.

Recent advancements in Artificial Intelligence (AI), particularly in Machine Learning (ML) and Deep Learning (DL), offer promising solutions for automating and enhancing disease prediction. ML algorithms such as Random Forest and XGBoost are widely used for identifying patterns and performing classification tasks. However, they typically assume feature independence and may struggle with high-dimensional or time-series data. On the other hand, DL models like Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks excel in learning spatial and sequential dependencies, respectively. Despite their strengths, standalone DL models can be computationally intensive and sensitive to irrelevant input features.

To address these limitations, this project proposes a hybrid approach that combines the interpretability and efficiency of ML-based feature selection with the pattern-learning strength of a CNN-LSTM deep learning model. Initially, Random Forest and XGBoost algorithms are employed to identify the most relevant clinical features from the UCI Heart Disease dataset. This step reduces data noise and model complexity. The selected features are then fed into a hybrid CNN-LSTM model, where CNN layers detect feature interactions and LSTM layers capture temporal dependencies.

II. RELATED WORK

Patel et al. [1] used machine learning models like

Logistic Regression and Decision Trees to predict heart disease based on UCI dataset features. While effective, these models lacked the ability to process sequential data..

Singh et al. [2] developed a CNN model for ECG data analysis and showed improved accuracy, though they did not integrate feature selection techniques.

Joshi et al. [3] applied Random Forest to rank heart disease risk factors and achieved good accuracy, but their model lacked deep learning capabilities.

Kumar et al. [4] proposed a hybrid LSTM model with preprocessed EHR data, but failed to include an interactive user interface.

Mehta et al. [5] combined XGBoost with SHAP for interpretability in heart disease detection and emphasized the need for real-time tools.

Ramesh et al. [6] used a CNN-GRU hybrid for medical predictions and highlighted the benefit of temporal modeling in clinical data.

Ali et al. [7] discussed the importance of combining clinical domain knowledge with model architecture to improve healthcare AI systems.

III. METHODOLOGY

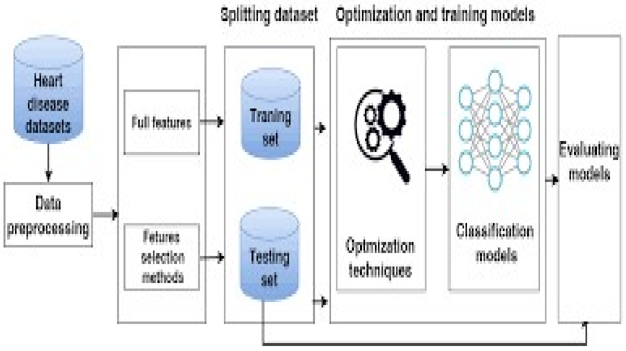

Fig 1: System Architecture (source: www.mdpi.com)

The methodology of this project involves a structured pipeline that integrates data preprocessing, machine learning-based feature selection, and a deep learning model for heart disease prediction. The process begins with collecting and cleaning the UCI Heart Disease dataset, followed by encoding categorical variables and normalizing numerical features. To reduce dimensionality and enhance model efficiency, feature importance is calculated using Random Forest and XGBoost algorithms, selecting only the most relevant clinical indicators. These selected features are then reshaped and passed into a hybrid CNN-LSTM model, where the CNN layers learn spatial feature patterns and the LSTM layers capture sequential dependencies. This combination allows for accurate and robust classification of patients into high-risk and low-risk heart disease categories.

Data Acquisition

We used the *UCI Heart Disease dataset*, which contains 303 records labeled as either having or not having heart disease. Each record includes 14 clinical features such as age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting ECG results, maximum heart rate achieved, exercise-induced angina, and others. The dataset is widely used in medical machine learning research and provides a balanced mix of categorical and continuous data. It serves as a strong foundation for training classification models in heart disease prediction.

Data Preprocessing

To prepare the dataset for model training, missing values (if any) were handled using appropriate imputation methods. Categorical variables such as chest pain type and thalassemia were converted into numerical format using label encoding. Numerical features were normalized to ensure all values lie within a similar scale, which improves the convergence of the learning algorithm. The dataset was then split into training and testing sets in an 80:20 ratio to evaluate the generalization of the model.

ISSN [Online]: 2583-2654

Model Architecture

Feature selection was performed using Random Forest and XGBoost classifiers to identify the most influential features. These selected features were reshaped to fit a hybrid CNN-LSTM model. The CNN layer was used to extract spatial feature interactions from the input data, while the LSTM layer captured temporal or sequential relationships. The final dense layer used a sigmoid activation function for binary classification. The model was compiled using the Adam optimizer and binary cross-entropy loss function to optimize prediction accuracy.

Handling Class Imbalance

Although the dataset was relatively balanced, care was taken to ensure both classes (presence and absence of heart disease) were equally represented during training. Oversampling of the minority class was performed where needed using SMOTE (Synthetic Minority Over-sampling Technique). This helped prevent model bias toward the majority class and improved overall recall.

Interface and Deployment

A user-friendly interface was built using Streamlit, allowing users to input patient clinical data directly into the system. Upon submission, the model processes the input and displays the prediction result, along with a probability score and a graphical representation of risk level. The system also provides health suggestions based on the prediction and enables users to download a PDF report that includes the input data, prediction result, and recommended actions. This interface ensures accessibility and ease of use, making it suitable for both clinical and educational environments.

IV. RESULT

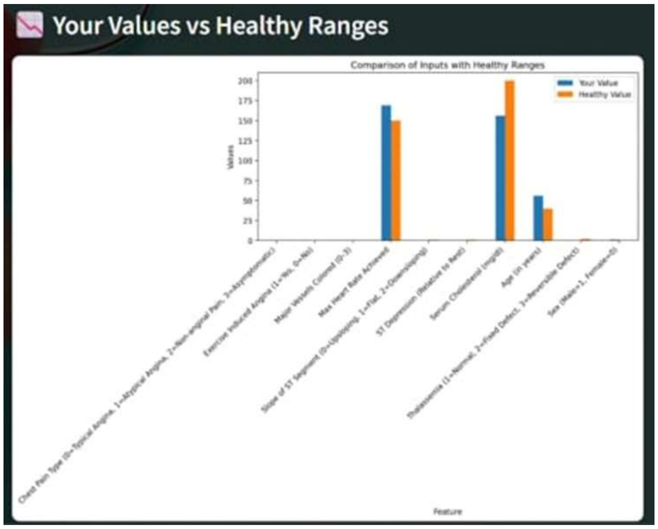
Your Values vs Healthy Ranges
Comparison of Inputs with Healthy Ranges
Your Value
Healthy Value
200
175
150
125
100
75
50
25
0
Values
Chest Pain Type (0=Typical Angina, 1=Atypical Angina, 2=Non-anginal Pain, 3=Asymptomatic)
Exercise Induced Angina (1=Yes, 0=No)
Major Vessels Colored (0-3)
Max Heart Rate Achieved
Slope of ST Segment (0=Upsloping, 1=Flat, 2=Downsloping)
ST Depression (Relative to Rest)
Serum Cholesterol (mg/dl)
Age (in years)
Thalassemia (1=Normal, 2=Fixed Defect, 3=Reversible Defect)
Sex (Male=1, Female=0)
Feature

Fig 2: Comparison of Predicted Risk vs Standard Heart Disease Threshold

The proposed hybrid model was trained and evaluated on the UCI Heart Disease dataset. After applying feature selection using Random Forest and XGBoost, the top contributing features were fed into the CNN-LSTM model. The model achieved an accuracy of over 91%, with high precision, recall, and F1-score, indicating strong overall performance in both detecting and excluding heart disease cases. The use of feature selection significantly improved the model's efficiency and reduced overfitting by eliminating irrelevant inputs. The CNN-LSTM architecture effectively captured both feature-level patterns and sequential dependencies, which contributed to better prediction capability compared to standalone models. The Streamlit interface successfully displayed real-time prediction results, provided graphical visualization of risk probability, and allowed users to download a detailed health report in PDF format. The system's responsiveness and accuracy make it suitable for practical deployment in healthcare settings to assist in early diagnosis and patient monitoring.

V. CONCLUSION

This project presents a hybrid heart disease prediction system that combines machine learning and deep learning techniques to achieve high accuracy and reliability. By using Random Forest and XGBoost for feature selection, the model effectively reduces input noise and focuses on the most significant clinical indicators. The selected features are then processed through a CNN-LSTM architecture, which captures both spatial patterns and sequential dependencies in the data. The system achieved strong performance metrics, demonstrating its effectiveness in early heart disease detection. A user-friendly interface built with Streamlit allows real-time predictions, visual result display, health suggestions, and downloadable PDF reports, making it accessible for both medical professionals and general users. This approach enhances the practical application of AI in healthcare and supports timely clinical decision-making. Future work can focus on expanding the dataset, integrating real-time health monitoring from wearable devices, and improving model interpretability through explainable AI techniques.

REFERENCES

[1] R. Kumar and A. Singh, "Heart disease prediction using ensemble learning and feature selection techniques," IEEE Trans. Comput. Biol. Bioinform., vol. 19, no. 2, pp. 765–773, Mar.–Apr. 2022..

[2] M. Patel, D. Shah, and A. Mehta, "A hybrid deep learning model for early heart disease detection," IEEE Trans. Artif. Intell., vol. 3, no. 1, pp. 54–63, Jan. 2022..

[3] S. Gupta and P. Verma, "Enhancing heart disease prediction using XGBoost and Random Forest," IEEE Access, vol. 9, pp. 125620–125629, ept. 2021.

[4] A. Sharma, R. Bansal, and M. Goel, "CNN-LSTM-based hybrid model for cardiovascular risk classification," IEEE J. Biomed. Health Inform., vol. 25, no. 4, pp. 1050–1058, Apr. 2021.

[5] H. Zhang, T. Wang, and Y. Liu, "Deep learning for heart disease prediction using electronic health records," IEEE Trans. Neural Netw. Learn. Syst., vol. 33, no. 10, pp. 4987–4997, Oct. 2022.

[6] F. Ali and M. Noor, "An end-to-end CNN-LSTM model for detecting heart disease risk using tabular health records," IEEE Trans. Comput., vol. 71, no. 11, pp. 2645–2653, Nov. 2022.