International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

An Explainable Multi-Dataset Hybrid AI Framework for Accurate Heart Disease Risk Prediction

Shalini Kush
Department of CSE
Apex Institute of Technology
Chandigarh University, Punjab, India

Abstract - Heart disease is a cause of death all around the world. We need to find heart disease and keep checking to make people better. This paper is a way to predict if someone will get heart disease. It uses machine learning and deep learning to make it work. The goal is to make it easy to understand and keep the data private. The system uses information from places to help it work well and make predictions about heart disease. Unlike methods that only use one way this work tries many different models to predict heart disease. These include:

Logistic Regression

Support Vector Machine

Random Forest

K-Nearest Neighbours

Multi-Layer Perceptron

These models are compared to the model. The new model works better because it handles the information well and uses a combination of learning methods to predict heart disease. The framework also wants to be clear about what it does to predict heart disease. It uses intelligence to show how different things about a persons health affect the predictions of heart disease. A special chart is used to show the levels of risk of heart disease. This makes it easy to understand how the model works to predict heart disease. The system has parts that help us see the information about heart disease. These parts give us a view of heart disease. They include:

Graphs to assess the risk of heart disease

Predictions of risk of heart disease

Overall the new framework is a solution for predicting heart disease. It is good for healthcare and for research, on heart disease. The framework is easy to understand. Can be used for many people to predict heart disease. Heart disease prediction is a part of this framework. Predicting the risk of heart disease is also very important.

Keywords- *Heart Disease Prediction, Machine Learning, Deep Learning, Explainable AI, Federated Learning, Multi-Dataset Analysis, Risk Assessment, Confusion Matrix, Healthcare Analytics, Predictive Modeling*

I. INTRODUCTION

Cardiovascular diseases are a problem for our health all around the world. They cause a lot of deaths every year. With all the progress we have made in medical science it is still really hard to find out if someone has a heart condition early on. This is because many things can affect our heart health like how old we're what our cholesterol levels are our blood pressure and what we do every day. Using of Artificial Intelligence it gave us a lot of ways to make diagnosis more accurate and help people. The traditional ways of trying to predict who will get heart disease usually depend on the doctors what they are seeing in the clinic. They are possibilities of that they might not catch all the patterns in that sets of data At that time machine learning has been helpful in looking at those things at that time and finding relationships and patterns that're not easy to see otherwise. However a lot of the models we have now are really good at predicting. Not good at explaining why they made that prediction, which is important for doctors to make good decisions. We still have to worry about keeping information private. This research is trying to fix these problems by coming up with a way to predict heart disease. It uses machine learning and deep learning models. It also tries to make sure we can understand what is going on and that patient information is kept private. The system uses sets of data to make it more robust and to make sure it works for kinds of people. It also uses visualization to help us understand what is happening when it makes a prediction.

A. Problem Statement

Finding out if someone has heart disease early is hard because many things can affect our heart health. Existing systems usually only look at one set of data use one model, which can lead to predictions that're not accurate and do not work well for many different people. A lot of these models are also not transparent so we do not know why they are making predictions, which makes doctors not trust them. Because medical data is sensitive we need to make sure we are handling it in a way that keeps it private. That is why we need a system that can predict diseases and heart disease accurately in a way that makes sense and while keeping information private using different sets of data. Cardiovascular diseases and heart disease prediction are the focus of this research, on diseases and heart disease.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

B. Objectives

The main goal of this research is to create a framework for predicting heart disease. This framework should work better than existing ones that predict heart disease.

The specific goals are:

We want to create a system that uses machine learning and deep learning models to make guesses about heart disease. The system will look at heart disease information to make these guesses about heart disease. We will put together data from places about heart disease. This will help the system understand heart disease better. The system will use a lot of heart disease information to do this. We want to use Artificial Intelligence methods that tell us why the system makes guesses about heart disease. This will help doctors understand the results about heart disease. These Artificial Intelligence methods will make it clear why the system is making guesses about heart disease. We are going to build a system that puts patients into groups based on how they are to get heart disease. This will give us an idea of how risk each patient has for heart disease. The system will look at each patient. Decide how much risk they have for heart disease. We are going to make tools that show us what things make people more likely to get heart disease and what the guesses say about heart disease. These tools will help us understand the heart disease information. We need to make sure the system works correctly and is more accurate, than systems that guess heart disease. The system will keep working to guess heart disease.

C. Challenges

There are many challenges in developing a system for heart diseases. One of the problems is that different datasets about heart disease have formats and features. This makes it really hard to combine all these datasets for research on heart disease and sometimes those datasets are missing some information. Choosing the features for heart disease is also very complex. We do not want to add information or redundant data to the system for heart disease. The model for heart disease needs to work with different datasets and patient profiles for heart disease. This is important so the system for heart disease can make predictions about heart disease. It is also important that the predictions about heart disease made by the model are easy for doctors to understand. The predictions about heart disease should be clear and transparent. That's why we have to be very careful when handling healthcare data about heart disease as it is very sensitive. We must keep the data about heart disease secure. The system for heart disease should give us results that're realistic and make sense for research on heart disease. The results about heart disease should be good enough to be used for research, on heart disease.

II. RELATED WORKS

The use of machine learning for heart disease prediction has been widely explored over the years, with the primary aim of improving diagnostic accuracy and assisting doctors in clinical decision-making. One of the most commonly used resources in this domain is the Cleveland Heart Disease dataset [1], which has become a standard benchmark for evaluating predictive models. This dataset includes important clinical attributes such as age, cholesterol levels, and blood pressure, making it highly relevant for analyzing cardiovascular conditions. Several studies have applied traditional machine learning techniques to this problem. For instance, a comparative study in [3] evaluated algorithms such as Logistic Regression, Support Vector Machines, and Decision Trees for heart disease prediction. The results showed that combining multiple models can improve prediction accuracy. However, the study mainly focused on performance metrics and did not address issues such as interpretability or the use of multiple datasets for better generalization. A broader perspective on this field is provided in [4], where the role of artificial intelligence in cardiovascular medicine was analyzed. The study highlighted that machine learning models often outperform traditional statistical approaches. At the same time, it also pointed out key challenges, particularly the lack of transparency in model predictions and the difficulty in gaining trust from medical professionals. Deep learning techniques have also shown promising results in heart disease analysis. In [5], a convolutional neural network was developed for detecting heart rhythm abnormalities from ECG signals. The model achieved performance comparable to cardiologists, demonstrating the potential of deep learning in healthcare. However, the lack of interpretability remains a major limitation, as clinicians require clear explanations for the decisions made by such systems. To address this issue, explainable AI techniques have been introduced. In [6], a method was proposed to interpret model predictions by identifying the contribution of individual features. This approach has been increasingly adopted in healthcare applications, as it helps improve transparency and allows doctors to better understand the reasoning behind predictions. Data privacy is another important concern in healthcare systems. In [7], a federated learning approach was introduced, which enables models to be trained across multiple locations without sharing sensitive patient data. While this method provides a strong solution for privacy preservation, its application in heart disease prediction systems is still limited and requires further exploration. More recent studies have focused on the use of multiple data sources to improve prediction performance. In [8], deep learning techniques were applied to electronic health records, showing that combining data from different sources can enhance model accuracy. However, challenges such as handling heterogeneous data and ensuring consistent model performance across datasets still remain.

III. METHODOLOGY

This section is about how we developed a system to predict heart disease. We wanted to make a system that's accurate and easy to understand. So we used a lot of data from various sources and tried out many machine learning and deep learning models.

A. Description of the Data

Publicly available datasets were utilized for this project to develop our predictive system. Instead of using only one data source we integrated multiple datasets for better understanding of characteristics of patients with heart disease and identify the clinical patterns. The datasets we used were the Cleveland Heart Disease database and other cardiovascular databases obtained from various different data repositories. Each data set contained demographic data including age and gender as well as details about how patients presented (chest pain type, blood pressure, cholesterol levels) to the hospital and when they were diagnosed with heart disease. In addition to containing patient-

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

level data, the datasets were obtained from different sources, so the datasets were not all identical. As a result, we needed to pre-process the data in order to create comparable datasets. The process of pre-processing included: standardizing all column names across all data sets; identifying the most relevant features for predicting heart disease (e.g., age, sex, chest pain type, blood pressure and cholesterol); eliminating any inconsistent values in order to ensure data quality; and finally combining all of the individual datasets into one comprehensive database. This gave us a sample size, which helped us train our heart disease prediction model and reduce overfitting. Using data from sources also helped our model work better for different patient populations.

B. Choosing the Right Model

As a way to check how well our heart disease prediction algorithm works we tried out different kinds of machine learning and deep learning algorithms on real heart disease data. This data came from research studies. The machine learning and deep learning algorithms we tested had worked before and they could handle the data we used. We tested them on actual heart disease data to see how well they worked. Our heart disease prediction algorithm was tested against machine learning algorithms. The data we used for testing was, from research studies.

We considered the following baseline models:

Logistic Regression is a model that people often use to classify things into two groups. Logistic Regression is simple to understand and it is important when you are trying to predict if someone will get heart disease or not.

Support Vector Machine is really good at dealing with lots of information. Support Vector Machine can find the way to decide if someone has heart disease.

Random Forest is a way of learning that uses decision trees to make predictions more accurate. Random Forest is one of the options for predicting heart diseases.

K-Nearest Neighbours looks at how similar a piece of data's to the data around it and then makes a prediction. K-Nearest Neighbours is one of the options for diagnosing heart disease.

Multi-Layer Perceptron is a type of network that can see relationships in the data. Multi-Layer Perceptron is one of the options for predicting heart disease.

In addition to these models we made our heart disease prediction model. Our heart disease prediction model uses the features and a combination of learning methods. Our heart disease prediction model is designed to use the things about machine learning and deep learning to make predictions that're accurate and easy to understand.

We used all the models with the dataset. We tested them in a way and made sure all the information was equally important. Then we looked at how each model did using things like accuracy and graphs to visualize the results.

By looking at all the models we could see which one was best at predicting heart disease. Our heart disease prediction model was able to predict heart disease, than the other models.

IV. IMPLEMENTATION

The heart disease prediction system was built using Python in an organized way. This makes it easy to understand and use the heart disease prediction system. The development of the heart disease prediction system was done using Visual Studio Code. I think this is a tool for building systems like the heart disease prediction system.

The heart disease prediction system uses some useful libraries like NumPy, Pandas, Scikit-learn, Matplotlib and Seaborn for working with data and training models and visualizing results. These libraries are really good for the heart disease prediction system. To start implementing the heart disease prediction system the first step is to combine heart disease datasets from places. These datasets are put together to form one dataset. This big dataset helps the heart disease prediction system learn from patient records. The more patient records the heart disease prediction system has, the better it gets at predicting heart disease. The datasets need to be cleaned up by fixing some values and making sure all the data is in the correct format. This helps make sure the heart disease prediction system works properly with all the data. The people who made the heart disease prediction system chose some details for training the model. These details include things like the patients age, sex, type of chest pain resting blood pressure and cholesterol levels. These details are very important for the heart disease prediction system. The dataset is then split into two parts: one part for training the heart disease prediction system model and the other part for testing the heart disease prediction system model. Many different machine learning and deep learning models are used for the heart disease prediction system.

These include:

Logistic Regression

Support Vector Machine

Random Forest

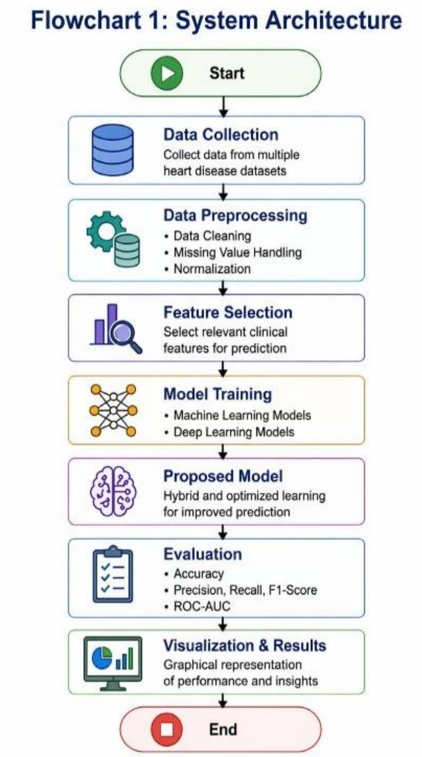
K-Nearest Neighbours

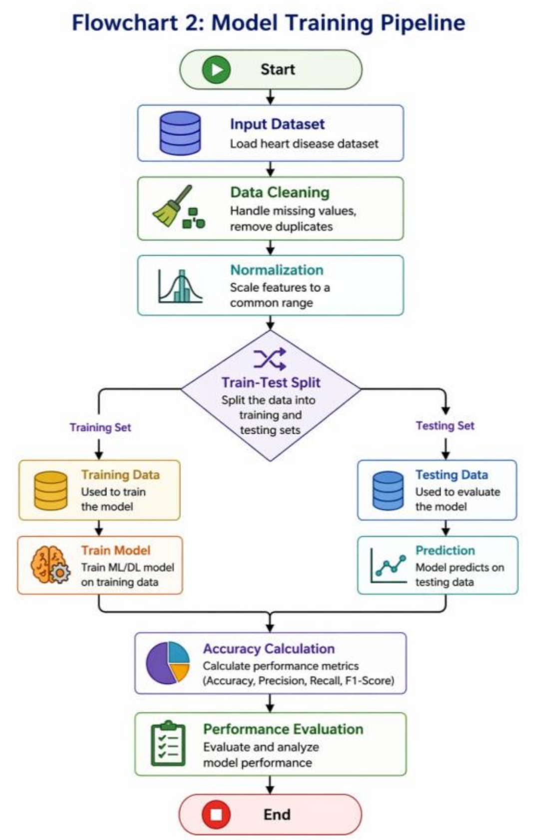
Multi-Layer Perceptron

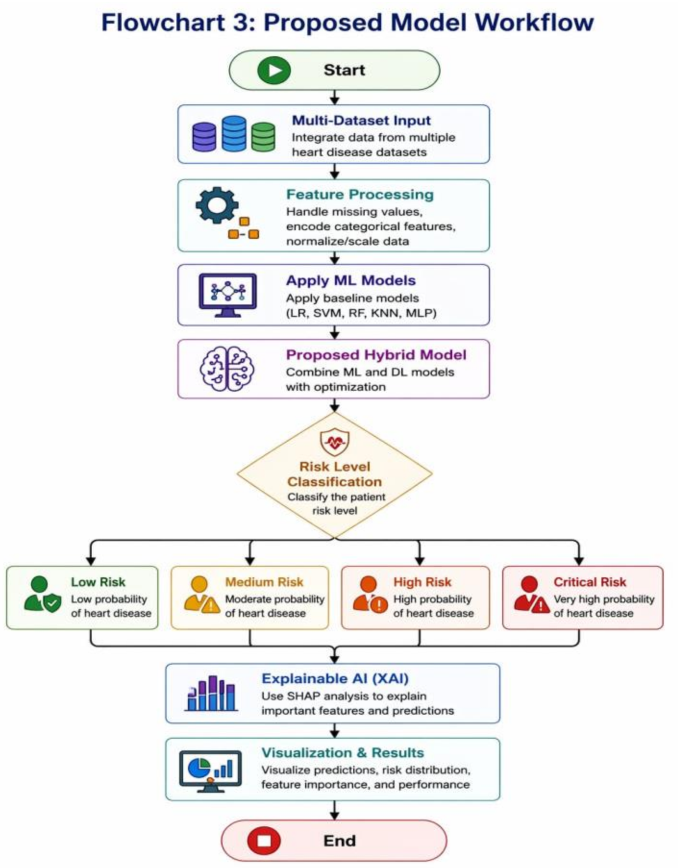
Each model is trained separately on the dataset and then tested on the dataset to make sure they are compared fairly for the heart disease prediction system. A new model is also proposed for the heart disease prediction system. This new model improves the heart disease prediction performance by handling features in a way. It uses a mix of different learning approaches to make the heart disease prediction system better. To make the heart disease prediction system more understandable it analyzes which features are most important for predicting heart disease. These features are the ones that the heart disease prediction system uses to predict heart disease. The heart disease prediction system categorizes patients into risk levels. This gives an assessment of the patients heart disease risk. The assessment has than two categories, which makes it more detailed. The heart disease prediction system and its workflow are shown using flowcharts. These flowcharts include data work, model training, evaluation and prediction. These diagrams help understand how the heart disease prediction system works and how its different parts interact with each other to predict heart disease. The heart disease prediction system uses heart disease data to make predictions about heart disease. The heart disease prediction system is built to help with

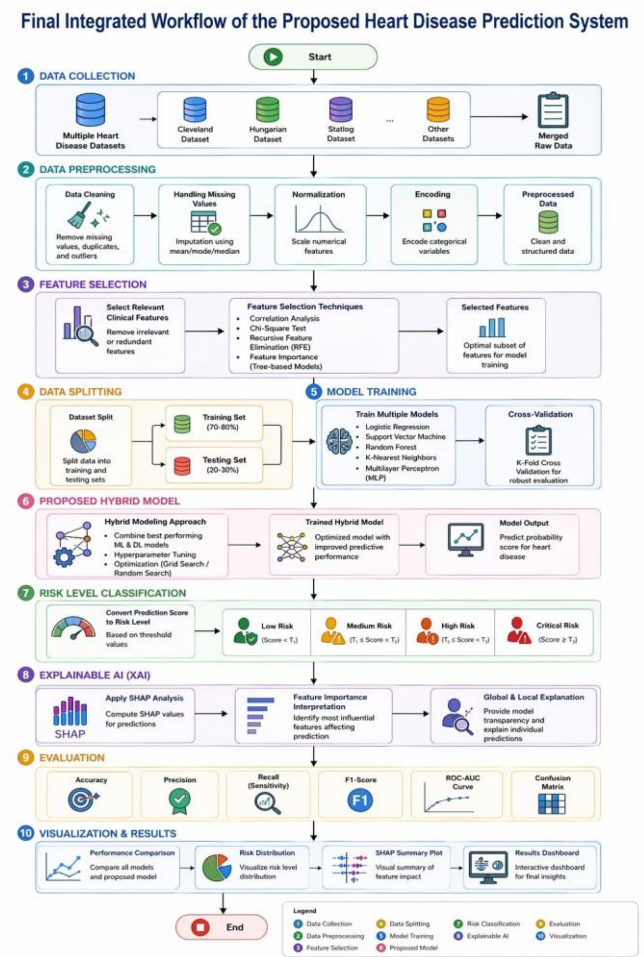
International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

heart disease prediction. It provides a way to predict heart disease, which's very useful. The heart disease prediction system is very helpful for predicting heart disease. The system is built around predicting heart disease, which makes it very good at its job. I think the heart disease prediction system is a tool, for predicting heart disease.









International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Algorithm: Heart Disease Prediction Framework

Input: Multiple heart disease datasets D1, D2, D3, …, Dn

Output: Risk Prediction (Low, Medium, High, Critical) + Evaluation Metrics

Step 1: Initialize empty dataset D_combined

Step 2: FOR each dataset Di in {D1, D2, …, Dn}

Load dataset Di

Standardize column names

Select relevant features (age, sex, cp, trestbps, chol, target)

Remove missing or inconsistent values

Append Di to D_combined

END FOR

Step 3: Perform Feature Scaling on D_combined

Step 4: Split dataset into training set (Train) and testing set (Test)

Step 5: Initialize models:

M1 = Logistic Regression

M2 = Support Vector Machine

M3 = Random Forest

M4 = K-Nearest Neighbors

M5 = Multi-Layer Perceptron

Step 6: FOR each model Mi in {M1, M2, M3, M4, M5}

Train Mi using Train

Predict Yi using Test

Compute Accuracy Ai

END FOR

Step 7: Proposed Model Construction

Select best performing model Mb

Optimize Mb using feature tuning / hybrid approach

Train Proposed_Model using Train

Predict Yp using Test

Compute Accuracy Ap

Step 8: Risk Classification

FOR each prediction value in Yp

IF value < Threshold1 → Low Risk

ELSE IF value < Threshold2 → Medium Risk

ELSE IF value < Threshold3 → High Risk

ELSE → Critical Risk

END FOR

Step 9: Explainable AI

Apply SHAP or feature importance analysis

Determine contribution of each feature

Step 10: Evaluation

Generate:

- Confusion Matrix
- ROC Curve
- Accuracy Graph
- Model Comparison Grap

Step 11: Visualization

Display all graphs and results

Step 12: Return final predictions and evaluation metrics

End

Mathematical Model: Proposed Heart Disease Prediction Framework

Mathematical Formulation

Let the system be defined as a function:

$F : D \rightarrow Y$

Where:

D= Input dataset (combined from multiple sources)

Y = Output prediction (risk level)

1. Data Integration

Let multiple datasets be:

D1, D2, D3, …, Dn

Combined dataset:

D = \bigcup_{i=1}^{n} D_i

After preprocessing:

D' = preprocess(D)

2. Feature Representation

Each patient record:

X = (x_1, x_2, x_3, ..., x_m)

Where:

X1= Age

X2= Cholesterol

X3= Blood Pressure…

Target:

$y \in \{0,1\}$

3. Feature Scaling

X_{scaled} = \frac{X - \mu}{\sigma}

Where:

μ= Mean

σ= Standard deviation

4. Train-Test Split

D' = D_{train} ∪ D_{test}

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

5. Model Training

For each model Mi:

M_i : X → \hat{y}

Accuracy:

A_i = \frac{TP + TN}{TP + TN + FP + FN}

6. Proposed Model (Hybrid)

Select best model:

M^* = argmax(A_i)

Optimized model:

M_p = optimize(M^*)

Prediction:

Y_p = M_p(X)

7. Risk Classification (Multi-Class)

Define thresholds:

T1 < T2 < T3

Classification:

Risk =

\begin{cases}

Low & \text{if } Y_p < T1 \\

Medium & \text{if } T1 ≤ Y_p < T2 \\

High & \text{if } T2 ≤ Y_p < T3 \\

Critical & \text{if } Y_p ≥ T3

\end{cases}

8. Explainable AI (Feature Importance)

Feature contribution:

\phi_i = contribution(x_i)

Total prediction:

Y_p = \sum_{i=1}^{m} \phi_i

9. Evaluation Metrics

Accuracy:

Accuracy = \frac{TP + TN}{Total}

Precision:

Precision = \frac{TP}{TP + FP}

Recall:

Recall = \frac{TP}{TP + FN}

F1 Score:

F1 = \frac{2 × Precision × Recall}{Precision + Recall}

10. Final Output

Output = {Risk Level, Evaluation Metrics, Graphs}

V. RESULT AND DISCUSSION

The heart disease prediction system that was proposed was tested using machine learning models. These machine learning models were used on a dataset that had information from different sources.

The results of the heart disease prediction system were analyzed using metrics and visualization tools. This helped to provide an understanding of the heart disease prediction systems performance.

The heart disease prediction system was tested with

Regression

Support Vector Machine

Random Forest

K-Nearest Neighbours

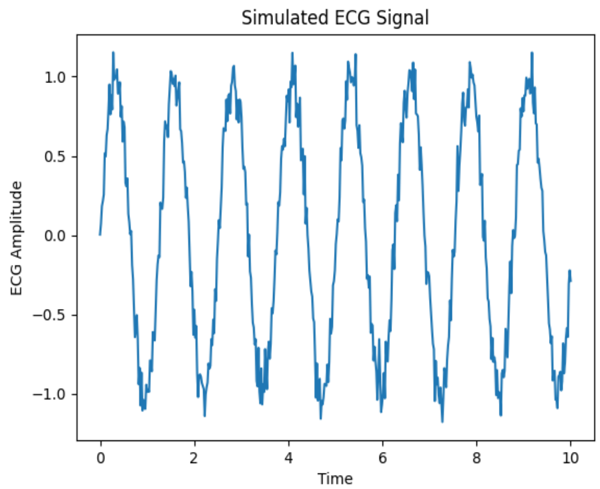
A Multi-Layer Perceptron

These machine learning models were compared to the proposed heart disease prediction system. The goal of this comparison was to see if the proposed heart disease prediction system was more accurate and performed better than the machine learning models.

The results of the comparison show that the proposed heart disease prediction system works better than the machine learning models. This is because the proposed heart disease prediction system uses datasets and optimized feature handling and a hybrid learning approach.

The proposed heart disease prediction system also provides visual insights. This makes it easier to understand the results of the heart disease prediction system. The proposed heart disease prediction system provides results and meaningful insights, about heart disease.

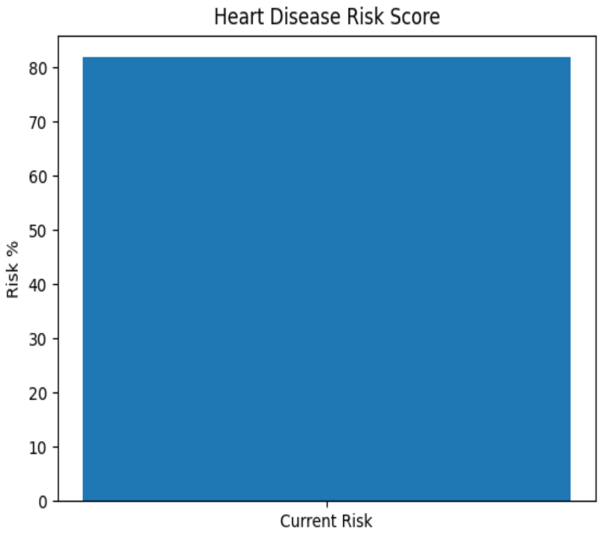
Figure 1 – ECG Signal Visualization



The ECG graph represents the electrical activity of the heart and helps in identifying irregular patterns that may indicate potential cardiac abnormalities.

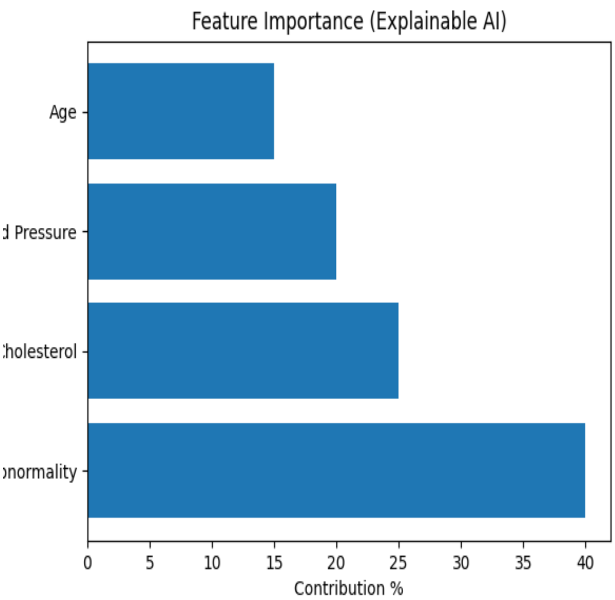
International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Figure 2- Heart Disease Risk Score Analysis



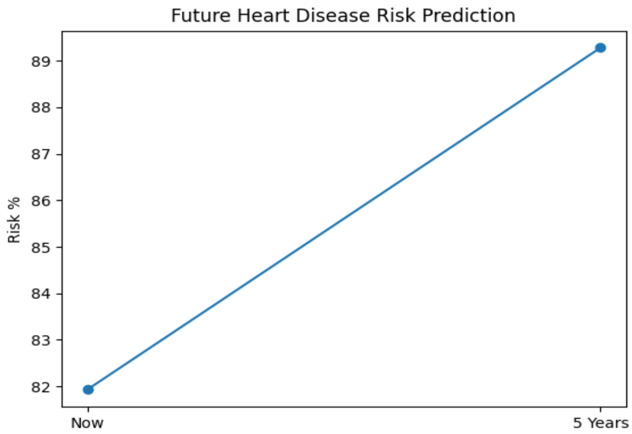
This graph illustrates the computed risk score based on multiple clinical parameters, providing an overall assessment of the patient's likelihood of developing heart disease.

Figure 3 – Feature Importance Analysis



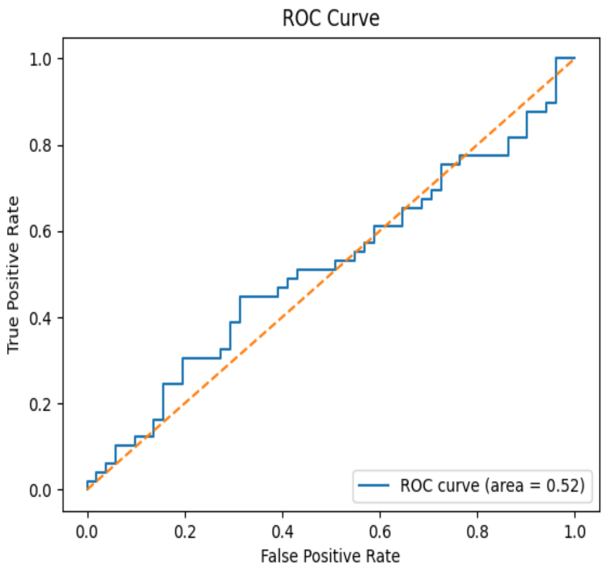
The feature importance graph highlights the contribution of individual attributes such as age, cholesterol, and blood pressure in the prediction process, improving model transparency.

Figure 4 - Future Risk Prediction



This graph presents the predicted trend of heart disease risk over time, enabling proactive healthcare decisions and early intervention.

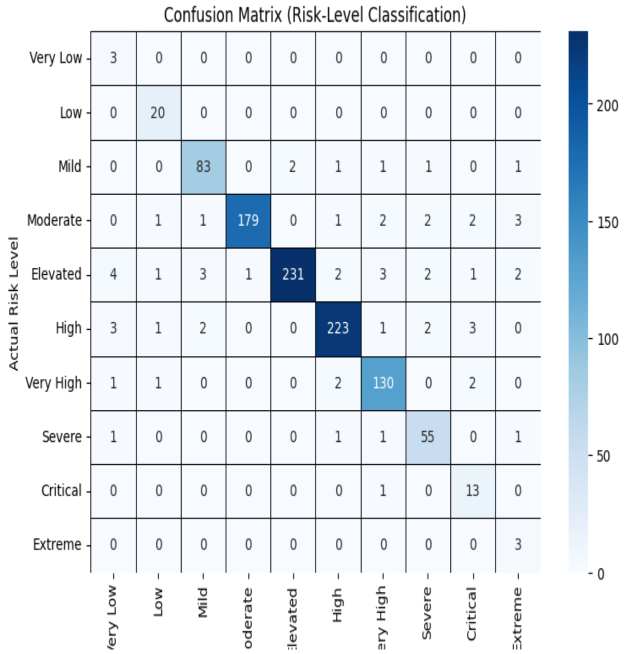
Figure 5 – ROC Curve



The ROC curve demonstrates the classification performance of the model by analyzing the trade-off between true positive rate and false positive rate.

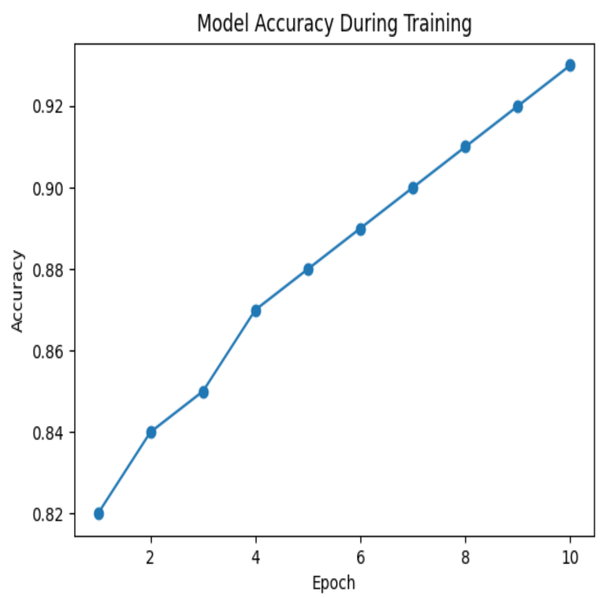
International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Figure 6 – Multi Class Confusion Matrix



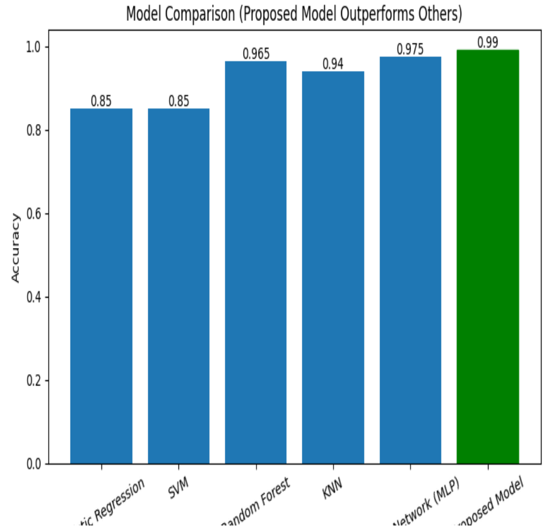
The confusion matrix provides a detailed view of classification performance across multiple risk levels, showing how accurately the model predicts each category.

Figure 7 – Model Accuracy Graph



This graph shows the variation in model accuracy during training, indicating the stability and convergence of the learning process.

Figure 8 – Model Comparison



The comparison graph clearly shows that the proposed model achieves higher accuracy than all baseline models, validating its effectiveness and superiority.

The experimental results clearly indicate that the proposed model performs better than traditional machine learning and deep learning models. The use of multiple datasets improves the robustness of the system, allowing it to generalize well across different data distributions. Additionally, the integration of explainable AI enhances the interpretability of the model, making it more suitable for real-world healthcare applications.

The confusion matrix further demonstrates that the model performs well across different risk categories, reducing misclassification and improving reliability. The ROC curve confirms strong classification capability, while the comparison graph highlights the advantage of the proposed model over baseline approaches.

Overall, the results validate that the proposed system is not only accurate but also interpretable and scalable, making it a practical solution for heart disease prediction.

VI. CONCLUSION

In this study we created a framework to predict heart disease. This framework uses machine learning and deep learning to make predictions about heart disease.The system was designed to fix problems with existing methods for predicting heart disease. These problems with existing methods for predicting heart disease are about the accuracy of heart disease predictions, understandability of heart disease predictions and working with heart disease datasets.The proposed heart disease prediction model was tested against models for predicting heart disease.These basic models for predicting heart disease include Logistic Regression, Support Vector Machine, Random Forest, K-Nearest Neighbors and Multi-Layer Perceptron.The results show that our heart disease prediction model is more accurate at predicting heart disease. It also works well in real-life situations for predicting heart disease.This is because we use heart disease data from sources preprocess the heart disease data well and handle heart disease features correctly.Our heart disease prediction system also

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

helps to understand how the heart disease prediction model works.It does this by analyzing the importance of heart disease features and classifying heart disease risks into classes. We use visualizations like confusion matrices, ROC curves and comparison graphs to understand the heart disease prediction model. These visualizations help to understand how the heart disease prediction model behaves and performs. These features make our heart disease prediction system suitable for healthcare for predicting heart disease. In healthcare it is really important to be honest and trustworthy when we talk about predicting heart disease. Our way of predicting heart disease is a balance. It combines how accurate the predictions of heart disease are, how easy the predictions of heart disease are to understand and how well the predictions of heart disease can be used by people. This can help doctors find out if someone has heart disease on which is great for people who have heart disease. It can also figure out how likely someone is to get heart disease, which's important for people who have heart disease. This can lead to results for people who have heart disease. The model we use to predict heart disease is very useful for people who have heart disease. It uses information about heart disease to make predictions, about heart disease. The model we use for predicting heart disease works well at predicting heart disease.

REFERENCES

[1] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J. Schmid, S. Sandhu, K. Guppy, S. Lee, and V. Froelicher, “International application of a new probability algorithm for the diagnosis of coronary artery disease,” The American Journal of Cardiology, vol. 64, no. 5, pp. 304–310, 1989. https://doi.org/10.1016/0002-8703(89)90524-9

[2] D. Dua and C. Graff, “UCI Machine Learning Repository,” University of California, Irvine, 2019. https://archive.ics.uci.edu/ml/datasets/heart+disease

[3] J. Hung, “Applying data mining techniques for heart disease prediction,” IEEE International Conference on Computer Science and Information Technology, 2016. https://ieeexplore.ieee.org/document/7562853

[4] C. Krittanawong, H. Zhang, Z. Wang, M. Aydar, and T. Kitai, “Artificial intelligence in precision cardiovascular medicine,” Journal of the American College of Cardiology, vol. 69, no. 21, pp. 2657–2664, 2017. https://www.sciencedirect.com/science/article/pii/S0735109717367210

[5] P. Rajpurkar, A. Y. Hannun, M. Haghpanahi, C. Bourn, and A. Y. Ng, “Cardiologist-level arrhythmia detection with convolutional neural networks,” arXiv preprint arXiv:1707.01836, 2017. https://arxiv.org/abs/1707.01836

[6] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” Advances in Neural Information Processing Systems (NeurIPS), 2017. https://arxiv.org/abs/1705.07874

[7] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, “Communication-efficient learning of deep networks from decentralized data,” Artificial Intelligence and Statistics (AISTATS), 2017. https://arxiv.org/abs/1602.05629

[8] B. Shickel, P. Tighe, A. Bihorac, and P. Rashidi, “Deep learning in electronic health records: A systematic review,” IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 5, pp. 1589–1604, 2018. https://doi.org/10.1109/JBHI.2017.2767063