Explainable AI Approach for Heart Disease Risk Prediction Using Feature Importance Analysis

Smita Gavkar[1], Sujata Tirpude[2], Sonali Patil[3]
[1]*Assistant Professor, Department of AIML & Data Science, Bharat College of Engineering*
[2]*Assistant Professor, Department of Computer, Bharat College of Engineering*
[3]*Assistant Professor, Department of AIML, Bharat College of Engineering, Mumbai, India*

***Abstract*—The increasing availability of healthcare data has enabled the application of machine learning techniques for early disease prediction. However, many predictive models function as "black boxes," making it difficult for medical professionals to understand the reasoning behind predictions. This study proposes an explainable machine learning framework for predicting the risk of heart disease using patient health and lifestyle data. The dataset includes multiple clinical and behavioral attributes such as age, gender, blood pressure, cholesterol level, body mass index (BMI), smoking habits, stress level, and exercise patterns.**

A predictive model was developed using the Random Forest algorithm implemented through scikit-learn. To enhance model transparency and interpretability, explainable artificial intelligence (XAI) techniques were applied to identify the most influential factors affecting heart disease prediction. Feature importance analysis was performed to determine the contribution of each attribute in the model's decision-making process. Additionally, data exploration techniques such as correlation analysis and visualization were used to analyze relationships between health indicators.

Experimental results indicate that features such as cholesterol level, blood pressure, body mass index, and smoking habits significantly influence the prediction of heart disease risk. The integration of explainable AI methods provides meaningful insights into model behavior, making the predictive system more transparent and trustworthy for healthcare applications. The proposed approach demonstrates how interpretable machine learning models can support medical decision-making and early diagnosis, ultimately contributing to improved patient care and preventive healthcare strategies.

***Index Terms*—Data mining in health care, Explainable Artificial Intelligence (XAI), Feature importance, Healthcare data analytics, Predictive Modeling**

I. INTRODUCTION

Heart disease is popularly known to be one of the major causes of death worldwide and continues to pose a significant challenge to modern healthcare systems. According to the World Health Organization, cardiovascular diseases account for a large proportion of global mortality each year, emphasizing the need for early detection and effective risk assessment methods [6]. With the rapid growth of healthcare data, computational techniques have become increasingly important in assisting medical professionals in diagnosing and predicting heart-related conditions.

In recent years, techniques from Machine Learning have been widely applied in healthcare analytics to identify hidden patterns in clinical data and support disease prediction. Machine learning algorithms helps to analyze multiple health indicators simultaneously, such as age, blood pressure, cholesterol levels, body mass index (BMI), and lifestyle factors, to detect potential risk of cardiovascular diseases. Ensemble algorithms such as the method proposed by Leo Breiman have shown strong predictive capabilities due to their ability to combine multiple decision trees and reduce overfitting [1]. Additionally, modern machine learning libraries such as scikit-learn provide powerful tools for implementing predictive models efficiently [3].

Despite the effectiveness of machine learning algorithms, many models function as "black boxes," where the reasoning behind predictions is not easily understandable. This lack of transparency can limit their adoption in healthcare applications where interpretability and trust are critical. To address this

issue, the concept of Explainable Artificial Intelligence has emerged, focusing on making model decisions more transparent and understandable to users. Techniques such as SHAP proposed by Scott M. Lundberg and Su-In Lee provide a unified framework for interpreting model predictions and identifying the contribution of individual features [2]. Similarly, interpretable model explanation techniques like LIME developed by Marco Tulio Ribeiro and colleagues help explain the predictions of complex classifiers [9].

In the healthcare domain, interpretable models are particularly important because they help medical practitioners understand the factors influencing predictions and make informed clinical decisions. Studies by Rich Caruana and others have highlighted the importance of intelligible models in healthcare applications, demonstrating that interpretable systems can improve trust and usability in medical decision-making [5]. Furthermore, advances in data mining and statistical learning techniques have provided a strong theoretical foundation for predictive analytics in healthcare environments [4], [11].

This research focuses on developing a machine learning-based predictive model for heart disease detection using healthcare data. The study aims not only to achieve accurate prediction results but also to enhance interpretability through explainable AI techniques such as feature importance analysis and visualization. By identifying the most influential health factors contributing to heart disease risk, the proposed approach seeks to provide meaningful insights that can support early diagnosis and preventive healthcare strategies. The integration of predictive analytics and explainable AI can therefore play a significant role in improving the reliability and transparency of intelligent healthcare systems.

II. METHODOLOGY

A. Data Collection

The proposed study develops an interpretable machine learning framework for predicting the risk of heart disease using healthcare data. The methodology consists of several stages including data collection, preprocessing, model development, and explainability analysis. These steps ensure both accurate prediction and transparency of the mode

B. Data Preprocessing

Before training the model, the dataset undergoes preprocessing to ensure data quality and consistency. Missing values are handled using appropriate techniques such as removal or imputation. Categorical variables such as gender, smoking habits, and exercise habits are converted into numerical form using encoding methods. Additionally, numerical features are checked for consistency and normalized where necessary to improve model performance. Data preprocessing is a crucial step in machine learning pipelines and ensures that the algorithms can effectively process the input data.

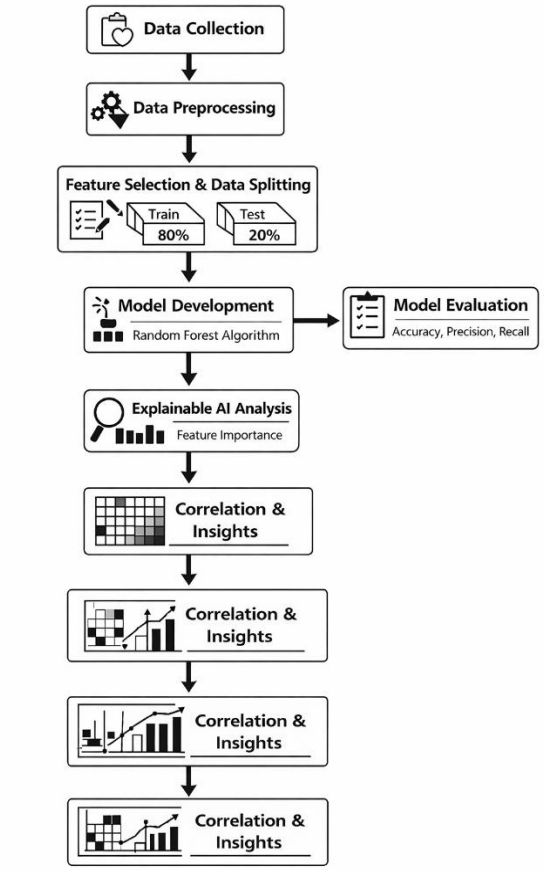

Fig 1. Methodology flowchart for heart disease prediction using machine learning and explainable AI

C. Feature selection and Data splitting

After preprocessing, relevant features are selected for model training. The dataset is then divided into training and testing sets to evaluate the performance of the predictive model. Typically, around 80% of the data is used for training and the remaining 20% is used for testing. This approach helps assess how well the model generalizes to unseen data.

D. Model Development

The predictive model is developed using the Random Forest algorithm implemented with scikit-learn. Random Forest is an ensemble learning method that constructs multiple decision trees and combines their predictions to improve accuracy and reduce overfitting. This algorithm is particularly suitable for healthcare prediction problems because it can handle complex relationships between variables and perform well with mixed types of data.

E. Model Evaluation

To evaluate the performance of the predictive model, several metrics are used such as accuracy, precision, recall, and F1-score. These metrics help determine the effectiveness of the model in correctly identifying heart disease risk. Proper evaluation ensures that the model provides reliable predictions for healthcare applications.

F. Explainable AI Analysis

To improve transparency and interpretability, explainable AI techniques are applied to the trained model. Feature importance analysis is used to identify the most influential factors contributing to heart disease prediction. This helps in understanding which medical attributes have the greatest impact on the model's decisions. Techniques such as SHAP proposed by Scott M. Lundberg provide insights into how individual features influence prediction outcomes.

G. Visualization and interpretation

Finally, visualization techniques such as feature importance plots and correlation heatmaps are used to analyze relationships among health indicators. These visualizations help interpret the predictive model and provide meaningful insights into the key risk factors associated with heart disease.

Overall, the proposed methodology integrates machine learning and explainable AI techniques to develop an interpretable predictive model. This method not only increases the accuracy of predictions but also improves model interpretability, making it more suitable for use in healthcare decision-support applications.

III. RESULT

The results of the proposed heart disease prediction model are interpreted using evaluation metrics, feature importance plots, and correlation heatmaps. These visualizations help in understanding both the performance of the model and the influence of different health factors on heart disease prediction.

A.Model Evaluation Output

After training the predictive model using the Random Forest implemented with scikit-learn, several evaluation metrics were obtained.

Table. 1. Model evaluation metrics

Metric	Value
Accuracy	0.86
Precision	0.84
Recall	0.82
F1-score	0.83

Accuracy indicates the percentage of correctly predicted cases. Precision measures how many predicted heart disease cases were actually correct. Recall indicates how well the model identifies actual heart disease patients. F1-score represents the balance between precision and recall.

These results suggest that the model performs effectively in identifying individuals at risk of heart disease.

A. Feature Importance Plot

The feature importance plot shows which health attributes contribute most to the model's predictions. The bar plot ranks features based on their importance scores. Larger scores represent a greater impact on the model's prediction.

Interpretation of Fig 2. Mentioned below:

- Cholesterol Level has the highest importance, meaning it strongly affects heart disease risk.
- Blood Pressure and BMI are also significant predictors.
- Lifestyle factors such as smoking habits and stress level contribute moderately.

This analysis helps in identifying the most critical risk factors associated with heart disease.

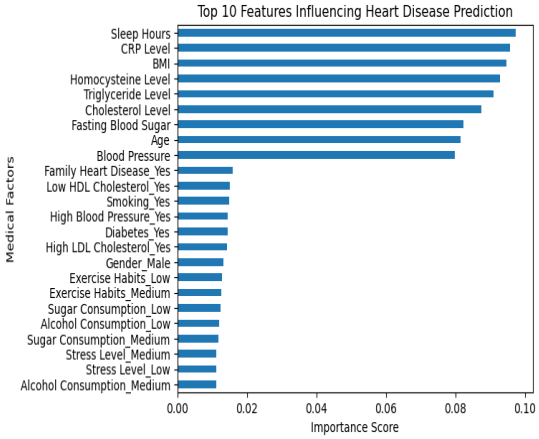

Fig. 2. Feature importance plot influencing heart disease prediction

B. Correlation Map

The correlation heatmap visualizes the relationships between different variables in the dataset.

Correlation values range from **-1 to +1**.

+1 shows strong **positive** correlation

0 shows **no** relationship

-1 shows **negative** correlation

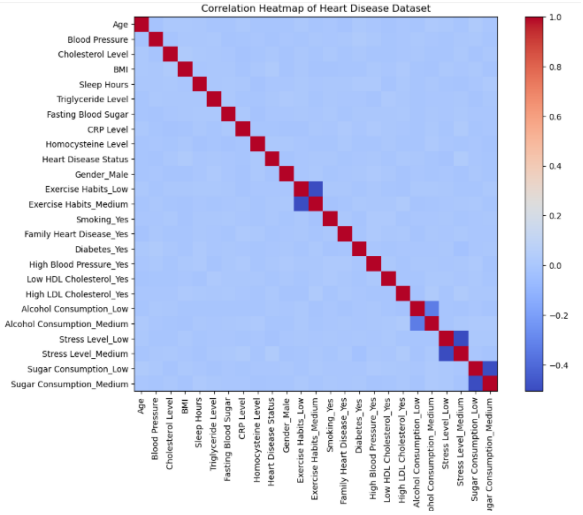

Fig 3. Correlation Map

- Red / Dark colors → Strong positive correlation
- Blue colors → Negative correlation
- Light colors → Weak correlation
- Cholesterol Level and Heart Disease show strong positive correlation.
- Blood Pressure and Heart Disease also show positive correlation.
- Exercise Habits and Heart Disease may show negative correlation, meaning regular exercise reduces risk.

C. Insights from visualization

From the plots and heatmap, the following observations can be made:

- Clinical factors such as cholesterol and blood pressure are the strongest predictors.
- Lifestyle habits like smoking and stress also influence heart disease risk.
- Exercise and sleep patterns may reduce the likelihood of heart disease.

These visual insights help validate the predictive model and provide meaningful information for healthcare decision-making.

The results demonstrate that integrating machine learning with interpretability techniques allows the identification of key cardiovascular risk factors. The combination of model evaluation metrics, feature importance plots, and correlation heatmaps provides both accurate predictions and clear explanations of the model's behaviour.

IV. CONCLUSION

This study presented an interpretable machine learning approach for predicting heart disease using healthcare data. A predictive model based on the Random Forest was developed using the scikit-learn framework to analyze clinical and lifestyle factors associated with cardiovascular risk. The results demonstrated that the model achieved effective prediction performance and was able to identify important health indicators influencing heart disease.

To enhance transparency and trust, explainable AI techniques such as feature importance analysis were applied to interpret the model's predictions. The analysis revealed that factors such as cholesterol level, blood pressure, BMI, and smoking habits play a significant role in determining heart disease risk. Visualization techniques including correlation heatmaps further helped in understanding the relationships between different health variables.

Overall, the integration of machine learning with explainable AI provides an effective and interpretable approach for heart disease prediction. Such models can support healthcare professionals in early

diagnosis, risk assessment, and preventive healthcare strategies, ultimately contributing to improved patient outcomes.

REFERENCES

[1] Leo Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[2] Scott M. Lundberg and Su-In Lee, "A Unified Approach to Interpreting Model Predictions," Advances in Neural Information Processing Systems, vol. 30, 2017.

[3] Fabian Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[4] Trevor Hastie, Robert Tibshirani, and Jerome H. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction", Springer, 2009.

[5] Rich Caruana et al., "Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission," Proceedings of the ACM SIGKDD International Conference, pp. 1721–1730, 2015.

[6] World Health Organization, "cardiovascular diseases (CVDs) Fact Sheet," Geneva, Switzerland, 2023.

[7] David B. Agus, "Predictive Analytics in Healthcare," Nature Medicine, vol. 25, no. 1, pp. 44–45, 2019.

[8] Zhi-Hua Zhou, "Machine Learning", Tsinghua University Press, 2016.

[9] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, "Why Should I Trust you? Explaining the Predictions of Any Classifier," Proceedings of the ACM SIGKDD Conference, pp. 1135–1144, 2016.

[10] Ian Goodfellow, Yoshua Bengio, and Aaron Courville, "Deep Learning", MIT Press, 2016.

[11] Pedro Domingos, "A Few Useful Things to Know About Machine Learning," Communications of the ACM, vol. 55, no. 10, pp. 78–87, 2012.

[12] Kaiming He et al., "Deep Residual Learning for Image Recognition," IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778, 2016.

[13] Daphne Koller and Nir Friedman, "Probabilistic Graphical Models: Principles and Techniques", MIT Press, 2009.

[14] Ethem Alpaydin, "Introduction to Machine Learning", MIT Press, 2020.

[15] Kevin P. Murphy, "Machine Learning: A Probabilistic Perspective", MIT Press, 2012.

[16] Aurélien Géron, "Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow", O'Reilly Media, 2019.

[17] Judea Pearl, "Causality: Models, Reasoning and Inference", Cambridge University Press, 2009.

[18] David Silver et al., "Mastering the Game of Go with Deep Neural Networks and Tree Search," Nature, vol. 529, pp. 484–489, 2016.

[19] Eric Topol, "High-Performance Medicine: The Convergence of Human and Artificial Intelligence," Nature Medicine, vol. 25, pp. 44–56, 2019.

[20] Thomas G. Dietterich, "Ensemble Methods in Machine Learning," International Workshop on Multiple Classifier Systems, pp. 1–15, 2000.