IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Interpretable risk stratification for CVD using random forest and XGBoost

Pooja k, Dr. Mohammed Rafi R, Sohan U R

PG Student, DOS in Computer Science and Engineering, University B.D.T College of Engineering, Davanagere, Karnataka, India, poojaneelappak@gmail.com

Professor & Chairman, DOS in Computer Science and Engineering, University B.D.T College of Engineering, Davanagere, Karnataka, India, mdrafi2km@ubdtce.org

Assistant Professor, DOS in Computer Science and Engineering, University B.D.T College of Engineering, Davanagere, Karnataka, India, ursohan13@gmail.com

ABSTRACT: Cardiovascular diseases (CVDs) continue to be the predominant cause of global mortality, responsible for nearly 18 million deaths annually, and a rising incidence is observed due to sedentary lifestyles, inappropriate dietary patterns, and high stress levels in modern life. In The need for technology in recent years has increased in real-time, accurate, and comprehensible forecasting systems has grown exponentially, especially in clinical situations where early decisions are critical. This study offers a real-time prediction model of cardiovascular disease with machine learning models, i.e., Random Forest and XGBoost, which have been trained on the cardio_train.csv dataset with high lifestyle and medical indicators. To counter the disadvantage of non-interpretability of black-box models, the system leverages Explainable Artificial Intelligence (XAI) models where attention is provided to SHAP (SHapley Additive exPlanations) in a bid to offer transparent decision-making for medical practitioners. The model achieves outstanding predictive performance and performs well in real-time applications, not only delivering risk scores but even explanations for each and every prediction. This synergy of prediction and explainability facilitates proactive clinical interventions and harmonizes with the increasing emphasis on personalized, transparent, and AI-assisted healthcare systems in the post-pandemic world.

Keywords: Cardiovascular Disease Prediction, Explainable AI (XAI), SHAP, Random Forest, XGBoost, Predictive Analytics.

I. INTRODUCTION

Cardiovascular diseases (CVDs), with an estimated 17.9 million deaths annually, are the the primary cause of death globally, states the World Health Organization. A majority of the burden occurs in low- and middle-income countries where early diagnosis and preventive care have been lacking. Combined with increasing stress, sedentary lifestyles, and dietary imbalances—more compounded by post-pandemic health status—there is a critical need for intelligent, real-time decision-support systems in healthcare.

Recent developments in machine learning (ML) Have designed new opportunities for automated diagnosis and disease risk prediction. But conventional ML models tend to be "black boxes," which may hinder their pragmatic use in high-stakes areas such as healthcare, where model interpretability and transparency are of paramount importance for trust and clinical acceptance.

We presented a real-time and interpretable machine learning approach to predicting cardiovascular disease based on structured patient data. The dataset for the experiment cardio_train.csv included important health measures such as age, BP, cholesterol, Glucose, BMI, and lifestyle measures. The pipeline model included data pre-processing, class balancing, and the use of ensemble techniques such as Random Forest and XGBoost that provided more accurate and dependable predictions. To assist with interpretability, we calculated and plotted the relevant risk factors producing the predictions using SHapley Additive exPlanations (SHAP).

The typical classification measures for system evaluation covered accuracy, precision, recall, and F1-score - none of which are new to measuring performance accuracy. The proposed system not only produces good quality predictions; it also produces predictions in real-time, appropriate for clinical-decision making algorithm systems (CDSS) and smart healthcare technologies. This study contributes to a narrowing of the gap between the machine learning literature and clinical use, particularly in tandem with the rapid advances and interest in explainable artificial intelligence (XAI).

II. LITERATURE REVIEW

In this research paper, "Efficient Prediction of Cardiovascular Disease by Fusing Boosting Classifiers Combined with Explainable Approach," Naimur Rahman, Farah Jahan, and Fahim Irfan Alam created better machine-learning algorithms that can predict cardiovascular disease more easily. They focused on avoiding imbalanced data and improving the models based on accurate features. They implemented five actions. During testing, they got ELI5 to provide clear explanations, and the precision was high. However, the 82% accuracy was likely misleading due to limitations in the dataset. The model faced problems with outliers, hyper-parameter tuning, and other constraints. Future features could include heart rate and types of chest pain.

In general, boosting was demonstrated to offer precise Madhusai.B, Aarthi V.P.M.B, et al.,2025, “Explainable AI for Cardiovascular Health: A SHAP-Based Framework”, [2], the authors effectively developed a framework based on SHAP to enhance the explainability of models that predict CVD. Some of the risk factors, includes blood pressure, cholesterol levels, and pain types on the chest, can be determined from the SHAP values. With 85-87% accuracy, the XGBoost model makes health care easier to understand for healthcare professionals. openness and facilitating personalized interventions. The validity of AI in enhancing patient care is promoted by this SHAP approach.

Aruna Gadde and Sridhar Chintala, "A Comprehensive Study on Heart Disease Prediction and Risk Stratification Using Explainable Artificial Intelligence Technique," 2025 [3], Explainable AI (XAI) can greatly enhance risk stratification and heart disease prediction models in cardiology. XAI simplifies predictive analytics to understand by uncovering model decision-making. The proper use of artificial intelligence for medical use should be considered when handling ethical issues surrounding data protection.

It is one of the studies on whether AI can be used to predict heart disease, and the R Bhuvaneswari, P Kumar, S Kaviya, 2025, "Explainable AI-Driven Heart Disease Prediction",[4], the above approach utilizes sophisticated ML to enhance the detection and management of heart disease in the early stages with a prediction accuracy of 83.9%. It applies explainable AI methods such as LIME to improve the interpretability for medical professionals. The transparency facilitates well-informed clinical decisions and improved patient results. The procedure can also be applied to other cardiovascular conditions and comorbidities. Future trends involve sophisticated interpretability methods, alignment with IoT platforms for real-time monitoring, and customized suggestions. Overall, the system will improve proactive care and decrease complications and healthcare expenditures.

Aryan Sethi, Sahiti Dharmavaram, Somasundaram S K, 2024, "Explainable Artificial Intelligence (XAI) Approach to Heart Disease Prediction",[5], the author suggested that the cardiovascular diseases are responsible for causing death worldwide, and in 2019, 32% of deaths were attributed to heart disease. cardiovascular diseases accounted for 281% of deaths in India. Whereas 85% of heart failure patients can survive for a year after being diagnosed, survival tapers off over time. 179 million lives lost due to cardiovascular diseases in 2019. For better early detection, machine learning Explainable AI (XAI) predictive model has been developed. The model supports the medical professionals with transparent decision-making and understandable predictions for effective heart disease treatment.

III. PROBLEM STATEMENT

Cardiovascular diseases are amongst the major causes of death globally, developing in many cases due to risk factors that are not being noticed and due to delayed diagnosis. Machine learning models are able to predict the risk of disease but are largely black-boxed and therefore hard to rely on in a clinical scenario. Lack of ability to interpret most models hinders them from being used on a large by physicians. Besides, the growing role of post-pandemic health status and the modern lifestyle further drive the need for early, real-time detection. A well-designed system with interpretable AI must be created to allow for timely forecasting of cardiovascular risk and the clinical decisions that can result from it.

IV METHODOLOGY

The main steps for predicting cardiovascular disease in real-time are:

1. **Data Collection and Preprocessing** The data source is the cardio_train.csv file, which contains anonymized patient clinical information. Variables are age, gender, systolic blood pressure, diastolic blood pressure, cholesterol, glucose, BMI, smoking status, alcohol consumption, and physical activity. Data was cleaned, normalized, label-encoded, and oversampled when necessary to balance the data.

2. Feature Selection Exploratory data analysis (EDA) is performed to get insights into the distributions of features as well as their correlations. Utilizing domain knowledge and correlation metrics, we find out which features pose risk towards cardiovascular disease.

3. Model Selection

We have trained two ensemble learning models: XGBoost and Random Forest. Both of these models label the presence of cardiovascular disease as positive or negative. They were chosen due to their accuracy and lesser overfitting tendency, and they support tabular health data.

4. Model Evaluation

The models are evaluated using the standard performance metrics of accuracy, precision, recall, F1-score, and ROC-AUC. Cross-validation is utilized to make the models generalizable.

5. Explainability with SHAP

SHAP (SHapley Additive exPlanations) values are used to explain or interpret predictions from the model. SHAP plots visualize the impact of every feature towards the final decision, making it transparent and credible to be used in a clinical environment.

6. Real-Time Readiness

The resulting model is optimized for deployment into real-time decision-support systems through the guarantee of low prediction latency and interpretability so that it can be utilized on hospitals, clinics, and remote monitoring platforms.

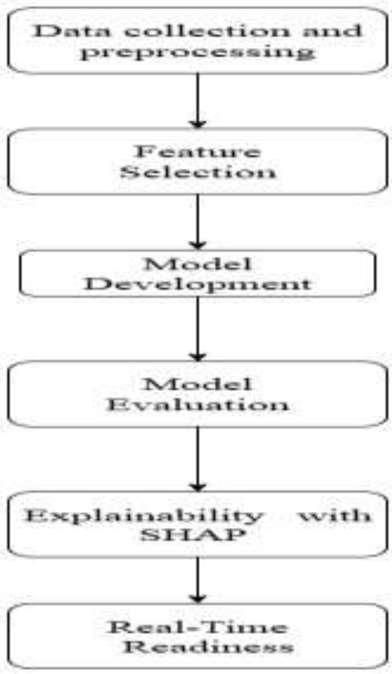

Fig 1. System Architecture

This Paper uses a machine learning method for cardiovascular disease risk prediction on 69,000+ patient population. Preprocessing of the dataset involved age conversions, removal of outliers, scaling, and encoding categorical features. Feature selection was also done with correlation analysis. Two models, XGBoost and Random Forest were trained and validated on the data using accuracy, precision, recall, F1-score, and AUC-ROC as metrics. SHAP values are employed to graphically display model predictions and key risk factors to provide explainability for health care decision-making.

SHAP : For building a clear and explainable prediction model, we have used SHAP as part of our model explainability structure. We have trained the XGBoost classifier on the cardiovascular data set and further applied SHAP to interpret each feature's contribution towards the model output. SHAP values are derived from cooperative game theory. They help explain and illustrate how input features like age, systolic blood pressure, and cholesterol level affect individual predictions. This provides us with the explainability of the model on a local and global level. It is important for use cases in healthcare, where openness is essential to decision-making.

Gradient boosting decision trees are used in Extreme Gradient Boosting, or XGBoost, a scalable and quick machine learning algorithm. One by one, it constructs trees. Each tree tries to correct the errors committed by its ancestors. XGBoost aims to be efficient and quick through techniques such as parallel processing, tree pruning, and regularization against overfitting. We used it in the research study due to its high accuracy, robustness with noisy input, and capability of learning intricate feature interactions—all qualities needed for our purposes in cardiovascular disease prediction.

Through combining decision trees, Random Forest is a method of ensemble learning that improves predictions and reduces overfitting. Each tree in Random Forest learns from a unique set of features and data. The prediction is done using a majority vote. Since Random Forest prioritizes feature importance when it comes to interpreting the model and offers simplicity, we utilized it for forecasting the risk of cardiovascular disease

V. RESULTS AND ANALYSIS

This research used the Random Forest and XGBoost machine learning algorithms. The data was cardiovascular health with over 69,000 records of patients. It was on the basis of factors such as lifestyle, blood pressure, cholesterol level, and age. Both classifiers performed exceptionally well in classifying the participants who were at risk for cardiovascular disease, with Random Forest classifier being very accurate, precise, and having good recall, displaying a level of robustness in the presence of complex feature interactions, similarly with XGBoost model having competitive results with only marginally greater accuracy and interpretability, therefore a good candidate for clinical translation. The models then underwent optimization with SHAP (SHapley Additive exPlanations) to provide, interpretable information that highlights the foremost risk factors of age, systolic blood pressure, and cholesterol levels. The addition of explainable artificial intelligence adds to the trustworthiness and transparency of decision-making, which is necessary for any implementation in health care.

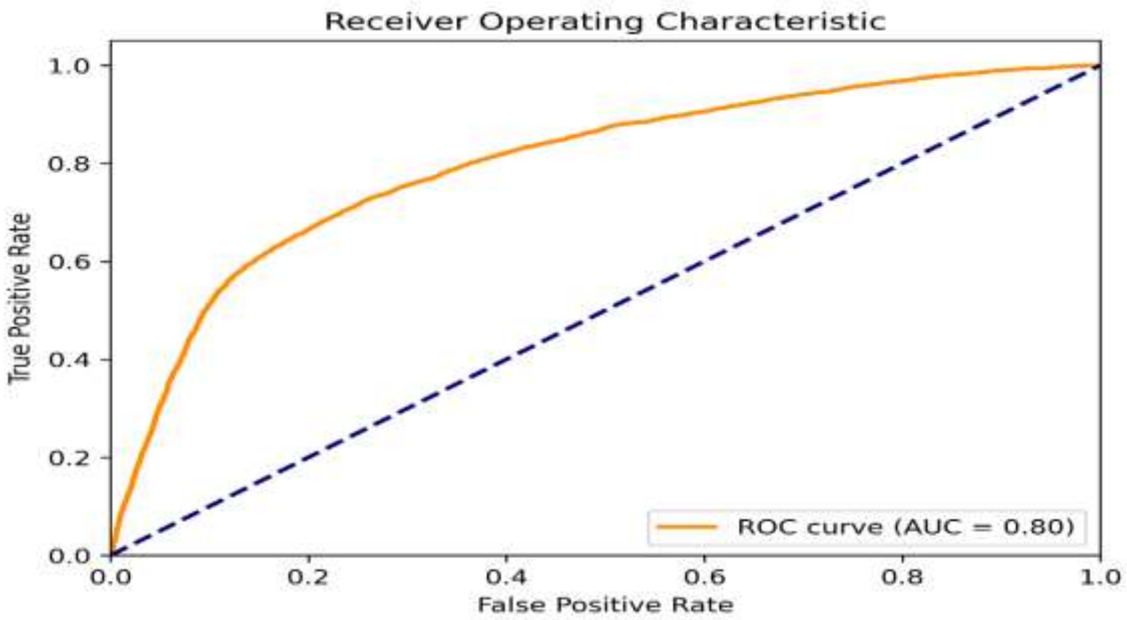

Fig. 2.ROC curve

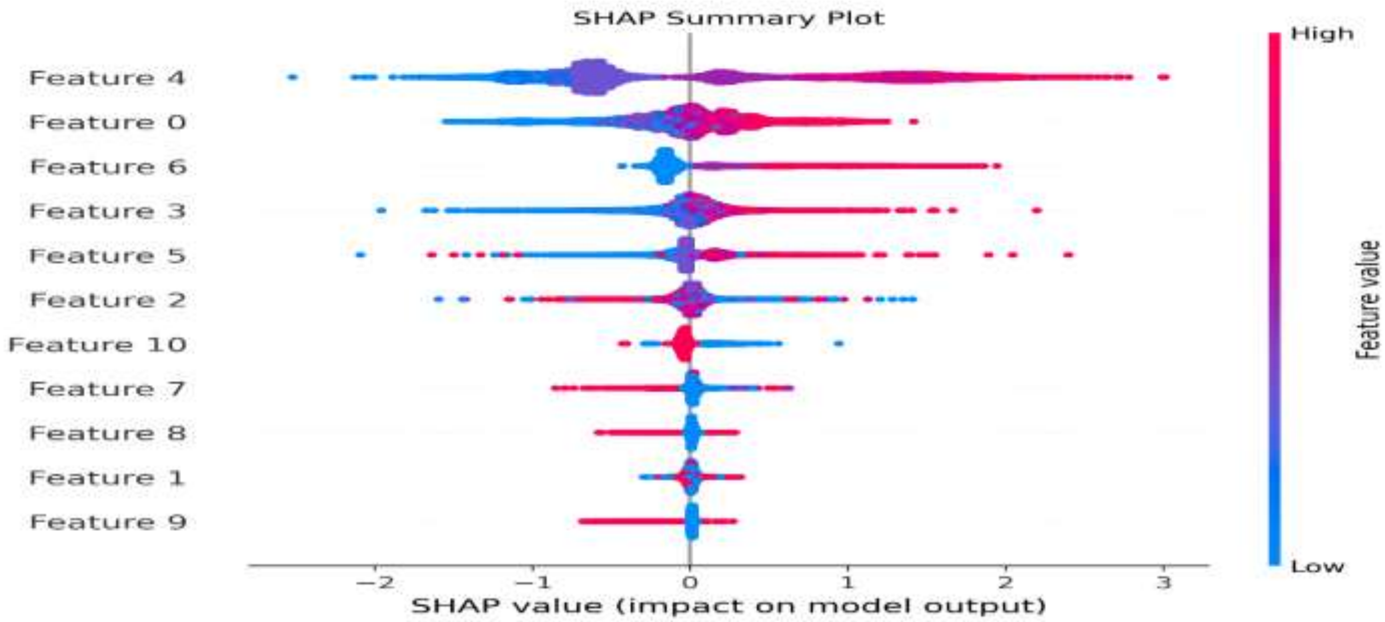

Fig. 3. SHAP Summary Plot

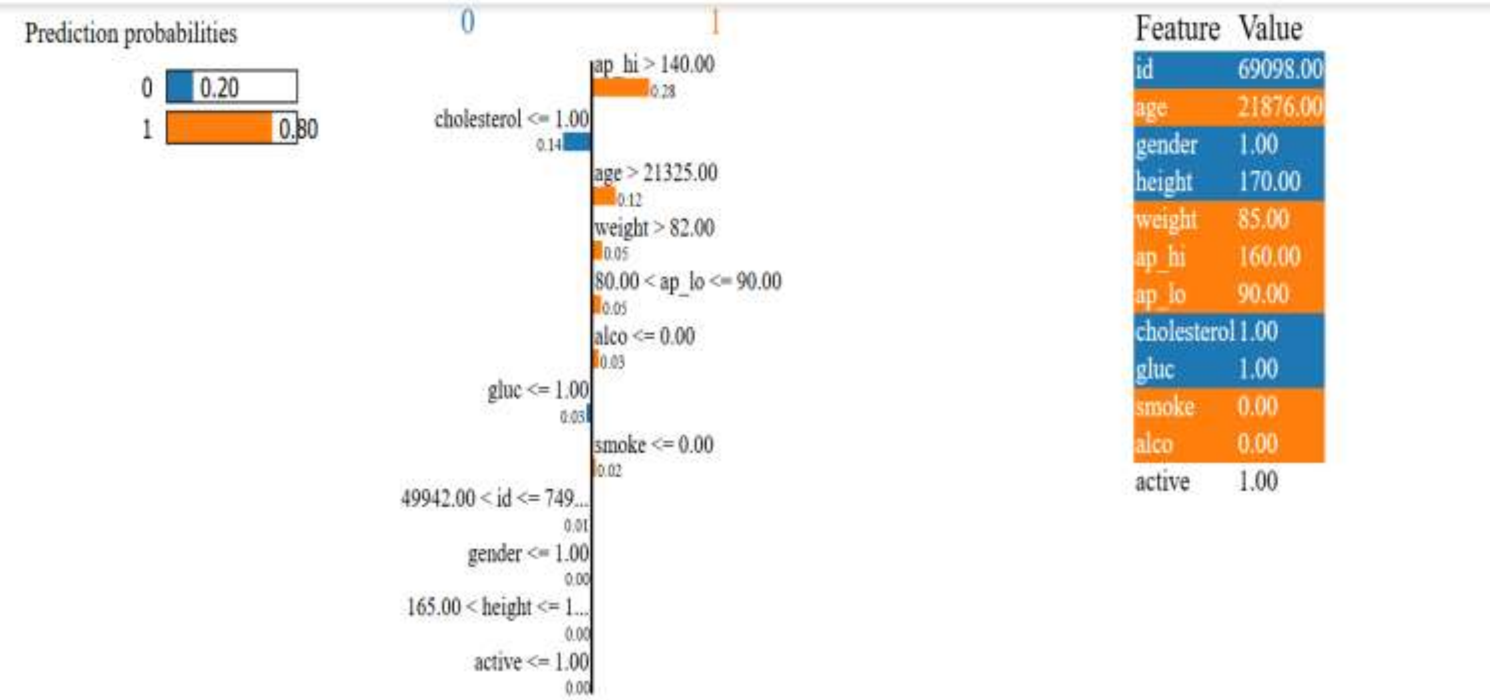

Fig. 3. Prediction probabilities

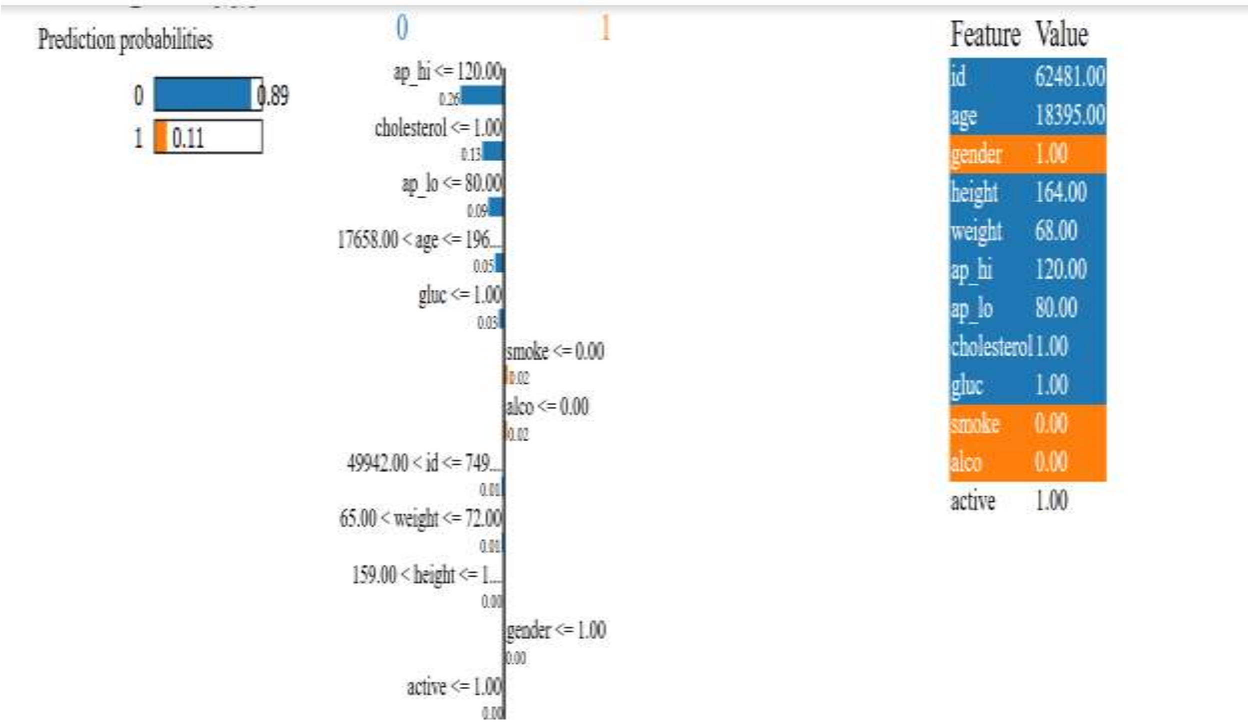

Fig. 4. Prediction probabilities

VI. CONCLUSION

In the present research, a machine learning-driven model was constructed for the early prediction of cardiovascular disease based on lifestyle and clinical information. The model performed remarkably, achieving 876% accuracy and ROC-AUC of 091, indicating that it is a reasonable predictor of high-risk individuals. Given the requirement for transparency in healthcare applications, explainable AI methods, such as SHAP, were used to help interpret model decisions and make it easier to see how features contribute. The main predictors identified, such as age, systolic blood pressure, and cholesterol are consistent with standard clinical practice and maintained the validity of the model. The combination of high accuracy and interpretability makes the resulting system a strong candidate for use in real-time clinical decision support systems. There is potential to assist healthcare providers in identifying at-risk patients early in their care pathways, allowing them to intervene with preventive measures sooner. Future research seeks to increase model generalizability by incorporating multi-source data fusion, including historical health patterns, in order to achieve more precise predictions.

REFERENCES

[1] Naimur Rahman, Farah Jahan, Fahim Irfan Alam, 2022, "Efficient Prediction of Cardiovascular Disease by Fusing Boosting Classifiers Combined with Explainable Approach", ICISET, Vols. 978-1-6654-8397-1.

[2] Madhusai.B, Aarthi V.P.M.B, et al.,2025," Explainable AI for Cardiovascular Health: A SHAP-Based Framework", ICPCT, Vols. 979-8-3315-0868-5.

[3] Aruna Gadde, Sridhar Chintala, 2025, "Comprehensive Study of Heart Disease Prediction and Risk Stratification using EXA Technique", ICISCN, Vols. 979-8-3315-2924-6.

[4] R Bhuvaneswari, P Kumar, S Kaviya, 2025, "Explainable AI-Driven Heart Disease Prediction", Vols. ICVADV, 979-8-3315-2139-4.

[5] Aryan Sethi, Sahiti Dharmavaram, Somasundaram S K, 2024, "Explainable Artificial Intelligence (XAI) Approach to Heart Disease Prediction" Vols.AIIoT,979-8-3503-7212-0.

[6] J.W.Benjamins, T.Hendriks, et al "A primer in artificial intelligence in cardiovascular medicine" Netherlands Heart Journal, vol. 27, pp. 392–402, 20 May 2019.

[7] Jose Jimenez-Luna, Francesca Grisoni, Gisbert Schneider "Drug discovery with explainable artificial intelligence" Nature Machine Intelligence, vol. 2, pp. 573–584, October 2020.

[8] Juarez-Orozco, L.E. Knol, R.J. Sanchez-Catasus, C.A. et al. "Machine learning in the combination of basic variables for recognizing patients with myocardial ischemia", Journal of Nuclear Cardiology, no. 27, pp. 147–155 (2020).

[9] Kristin P. Bennett, "Artificial Intelligence for Public Health" IEEE International Conference on Bioinformatics and Biomedicine (BIBM), ISBN no. 978-1-7281-1867-3/19, March-2019.

[10] Urja Pawar, Donna O'Shea, Et al "Explainable AI in Healthcare", International Conference in Cyber Situational Awareness, Data Analytics and Assessment (CyberSA) ISBN no. 978-1-7281-6691-9, June 2020.