Grenze International Journal of Engineering and Technology, June Issue

Explainable Heart Disease Risk Prediction using Cat Boost, SHAP, and Stream lit

Jane Rubel Angelina Jeyaraj[1], R. Shyam Babu[2], K.V. Sai Kalyan[3], M.J.S.V.S. Charan[4] and G. Rama Krishna[5]
[1-5]Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education Krishnankoil, Tamil Nadu, India
Email: j.janerubelangelina@klu.ac.in, ramanadamshyam24011@gmail.com, kalyankatta701@gmail.com, mjyothikc@gmail.com, ramakrishnagavini@gmail.com

***Abstract*— Heart disease has been identified as one of the major contributors to mortality around the world, and hence the need for the early detection of the risks associated with heart disease cannot be overstated. This research aims at developing an explainable machine learning system for the early detection of the risks associated with heart disease using patient data. The system will utilize the CatBoost algorithm as the main classifier for the system because of its ability to handle structured data sets that contain both categorical and numerical features. To make the system explainable, the system will utilize the SHapley Additive exPlanations (SHAP) technique for the interpretation of the results produced by the classifier in relation to the various features provided in the system. Additionally, the system will utilize the Streamlit tool for the development of an interactive web application that can be utilized for the generation of real-time results for the user.**

***Keywords*— Heart Disease Prediction, CatBoost, Explainable AI, SHAP values, Machine Learning, Model Interpretability, Streamlit Dashboard, Clinical Decision Support, Risk Assessment, Healthcare Analytics, Feature Importance, Interactive Visualization, Patient Data, Good Health and Well-being.**

I. INTRODUCTION

Cardiovascular diseases are considered to be one of the biggest health challenges faced today, and these diseases still cause a large number of deaths. It is very important to identify individuals who are at risk of developing heart diseases at a very early stage in order to provide them with medical treatment and lower the death rates. In conventional approaches, diagnosis is performed using a lot of clinical analysis and statistics. However, these approaches may not be able to effectively address the relationships that exist between various medical factors. The recent developments in the field of machine learning offer new possibilities in the analysis of medical data and the prediction of disease risk with increased precision. Machine learning algorithms are able to learn from the data and identify relationships between various medical factors that are not easily recognizable. While the models perform well in terms of prediction, they are often seen as "black boxes," meaning it is difficult to understand the reasoning behind the prediction. In terms of healthcare, it is important to understand the prediction so that medical staff can understand the reasoning behind the prediction. To overcome this challenge, a heart disease prediction framework is proposed, which utilizes the CatBoost algorithm with SHAP-based explainability.

Grenze ID: 01.GIJET.12.2.374_24

The choice of CatBoost is based on its good performance on structured data types and its effectiveness on categorical data. Additionally, SHAP is used to explain how every feature contributes to the prediction results. Moreover, an interactive dashboard is provided to deploy the proposed model using a Streamlit-based interface. The purpose of this study is to develop an accurate, interpretable, and user-friendly system that can help healthcare professionals to recognize potential risks of heart diseases and improve early diagnosis. In the current research, CatBoost is used as a major predicting model because of its robustness and efficiency, as well as its suitability for structured medical data. CatBoost is able to work with both categorical and numerical variables and prevents overfitting through the use of ordered boosting and symmetric decision trees. To further increase the interpretability of the model,

II. Literature Review

The application of data-driven techniques for risk prediction of heart disease has remained a popular research topic for the last several decades, driven by the need for early diagnosis and better clinical prognosis. Conventional approaches were largely based on statistical models like logistic regression to predict the risk of disease using clinical variables such as age, cholesterol, blood pressure, and electrocardiographic readings. Although these models provided a certain level of interpretability owing to their parametric form, their accuracy was often poor, especially when dealing with complex non linear relationships between medical variables in real-world datasets [1-2]. With the development of machine learning, more complex algorithms like Decision Trees, Random Forests (RF), and Support Vector Machines (SVM) were developed. Many research papers have shown that ensemble models perform much better than single models on structured healthcare data, with higher accuracy and robustness [3-4]. Gradient boosting algorithms like XGBoost and CatBoost further improved the accuracy of predictions by correcting the errors of the models and addressing feature dependencies [5-6]. CatBoost has recently attracted much attention as a prominent algorithm for tabular medical data because of its inherent ability to handle categorical variables with minimal preprocessing. Unlike traditional boosting methods, CatBoost uses ordered boosting and symmetric decision trees, which help to prevent overfitting and ensure better generalization, essential in medical applications where the amount of data is often limited in size [7]. Various studies have shown that CatBoost performs competitively or even better than other algorithms in medical prediction tasks, such as heart disease and cardiovascular risk prediction, while being stable across different populations [8-9]. However, the black-box nature of machine learning models, despite their enhanced predictive capabilities, still stands as a significant hindrance to their adoption in the clinical setting. In the clinical domain, besides accurate predictions, it is essential for clinicians to have transparent explanations to support the predictions made by the model. This problem has sparked increasing interest in the application of Explainable Artificial Intelligence (XAI) approaches [9-10]. SHapley Additive exPlanations (SHAP) has been identified as one of the most successful XAI techniques adopted in the literature, which offers theoretically justified explanations by attributing the contribution of each feature to a specific prediction for an individual sample. Previous studies have shown that SHAP can successfully fill the gap between model performance and interpretability, improving clinician trust and enabling informed decision-making [11-12]. In addition to the development of models, the deployment of predictive and explainable models is crucial for their adoption in real-world settings. Interactive tools such as Streamlit can be used to develop friendly web interfaces that enable users to enter patient data, obtain predictions, and view explanations in the form of SHAP values in real time. Interactive tools are crucial for the adoption of AI-driven decision support systems in clinical settings due to their ability to enhance transparency and usability [13-14]. In this research, we extend the existing work by combining a state-of-the-art machine learning algorithm (CatBoost), an explainability tool (SHAP), and an interactive deployment environment (Streamlit) to design an end-to-end system for predicting the risk of heart disease. The proposed system is expected to provide high accuracy in predictions while ensuring transparency and interpretability, thus meeting the key performance and trust requirements in the current healthcare setting [14-15].

III. Methodology

The proposed methodology will help in developing an accurate and interpretable heart disease risk prediction system using machine learning, explainable artificial intelligence, and interactive deployment. The overall process of the system will include data acquisition, preprocessing, model training, explainability analysis, and deployment. Every step of the process is designed in such a way that it is reliable, interpretable, and usable.

A. Dataset Description

The study employs the UCI Heart Disease dataset, which has clinical information about the patients and various attributes concerning heart disease. The attributes concerning heart disease include various characteristics such as age, gender, type of chest pain, blood pressure, cholesterol levels, fasting blood sugar, electrocardiogram results, maximum heart rate, exercise-induced angina, ST deviation, and the presence of major vessels. The dataset has a target variable that identifies the presence or absence of heart disease in the patients. The UCI Heart Disease dataset is employed in various studies concerning heart disease for the evaluation and comparison of the performance of the predictive models for the detection of heart disease in patients.

B. Data Preprocessing

Data preprocessing is performed to improve the quality of the data, enabling the model to train effectively. Missing values are handled by employing the most appropriate statistical methods, ensuring that no information is discarded.

The numerical data is standardized if required, while the categorical data is left unchanged to exploit the benefits of the CatBoost algorithm, which can handle categorical variables directly. Outliers are analyzed to avoid any unwanted noise, provided it does not impact the important variations. The data is split into a training set and a test set to evaluate the model's ability to generalize.

C. CatBoost Model Trainin

CatBoost has been selected as the main prediction model because of its excellent performance in structured and tabular data types. The training of the model occurs through gradient boosting using decision trees, where the next decision tree tries to correct the mistakes made by the previous decision trees. The process of ordered boosting helps in the prevention of overfitting in the model. The model can handle numerical and categorical variables effectively, and hence there is no need for extensive feature engineering. The hyperparameters such as the learning rate, depth of the trees, and the number of iterations have been optimized for the best possible performance in the prediction phase. Additionally, the model has an automatic mechanism for handling categorical variables through the conversion of the values into numerical form using target statistics. This makes the data preprocessing process more stable and easier. The training of the model occurs using the split data in order to make the process unbiased.

D. Model Evaluation

The performance of the trained model is evaluated based on various criteria such as accuracy, precision, recall, F1 score, area under the ROC curve, etc. The confusion matrix is also created to understand the true positive, false positive, false negative, and true negative values. The criteria are used to understand the performance of the model. The performance of the designed heart disease prediction model was tested and evaluated using various metrics such as accuracy, precision, recall, and F1 score. These metrics offer an in-depth understanding of the performance of the designed model in distinguishing between patients with and without heart disease. The results indicate that the designed CatBoost model offers high predictive accuracy with balanced precision and recall values. This implies that the designed model is effective in minimizing false predictions, both positive and negative. Based on the results, it is evident that the designed model is reliable and stable and can support the early diagnosis of heart disease.

E. Explainability Using SHAP

In order to overcome the problem of interpretability in the machine learning models, SHapley Additive exPlanations (SHAP) are integrated with the system. The SHAP values are computed to generate explanations for global and local explanations of the models. The global explanations are useful in understanding the most dominant features that influence the prediction of heart disease in the global sense, while the local explanations are useful in understanding the influence of individual features on the prediction of the models.

F. System Deployment Using Streamlit

To make it more user-friendly for practical purposes, the developed model and SHAP values are implemented using a framework called Streamlit, which is an interactive web application framework. The application allows users to insert clinical data for patients and obtain live predictions regarding the risk of heart disease. The application is implemented in such a manner that it is user-friendly and can be used by anyone, including medical professionals.

G. Testing and Validation

To make the reliability and generalization ability of the proposed heart disease prediction system robust, a thorough testing and validation process has been carried out. The preprocessed dataset was split into a training set and a testing set using a conventional split ratio (for example, 80:20). The CatBoost classifier was trained on the training set and tested on the test set to avoid any data leakage.

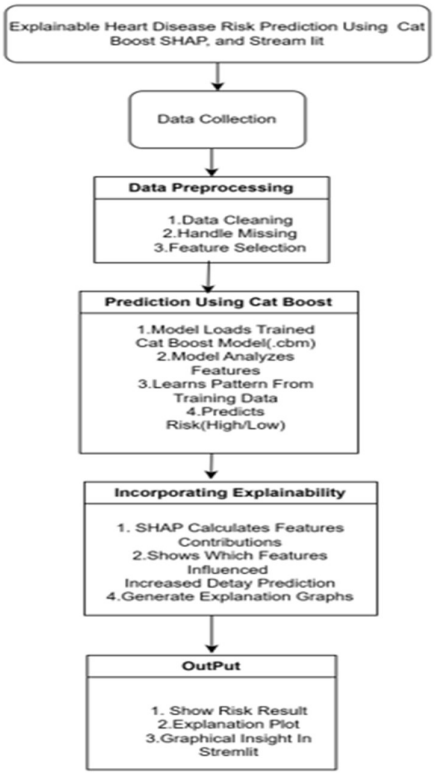

Fig 1: System Architecture

The proposed system has a well-structured workflow for making a prediction on heart disease risk using machine learning. The process begins with obtaining a historical heart disease dataset, which has several clinical features associated with heart health. The dataset is then used to apply a data preprocessing step to handle missing data, eliminate inconsistencies, and transform data features to optimize learning with the model. The preprocessed data is then used to train a CatBoost classification model, which is used based on its outstanding performance with structured medical data. The model is also used because it can handle categorical data efficiently. The model is then saved as a file to optimize deployment without having to train it again. In the prediction phase, the CatBoost model is imported into the application environment where the user is required to input the patient's clinical data. The input values are then processed by the trained model to give a prediction result concerning the existence of heart disease. The prediction result is then displayed to the user in an understandable manner. The process ensures a seamless integration of the data preprocessing, training, and prediction phases, making the proposed system feasible and appropriate. Finally, the result of the prediction is displayed to the user through the application interface. The result is a clear decision that can help the healthcare professional or the user understand the patient's condition in terms of the risk of having heart disease. The flowchart shows the integration of data-driven learning, model deployment, and user interaction in the proposed system. Once the training process is completed, the model is saved as a serialized file named fheart_disease_catboost.cbm. Saving the model is important because it can be used in the future for prediction purposes.The primary prediction model is selected as CatBoost due to its superior performance in structured and tabular data. The model is trained using a gradient boosting technique with decision tree algorithms, which try to correct their mistakes with each successive tree. Ordered boosting is used to avoid overfitting. CatBoost is also good at handling numerical and categorical features without any need to engineer features. The hyperparameters are tuned to get optimal results.

TABLE I: CAT BOOST VS XG BOOST (SIMPLE COMPARSION)

Feature	XG-Boost	Cat-Boost
Categorical Data	Need Manual Encoding	Handles Automatically
Overfitting	High Risk	Low risk (Ordered Boosting)
Accuracy(Heart Disease)	High	Higher
Explainability	Low	High (With SHAP)

The comparison helps in understanding the major differences between XGBoost and CatBoost algorithms, especially in the prediction of risks associated with heart diseases. XGBoost requires manual coding for the encoding of the categorical variables. This increases the complexity of the process and can cause biased results in the analysis of the data. CatBoost has the ability to handle the categorical variables in the dataset. This helps the algorithm in the automatic learning of the patterns from the attributes. In terms of performance and usability, CatBoost has been proven to have fewer risks of overfitting because of its ordered boosting algorithm. Even though both models have high accuracy, CatBoost has better predictive results compared to the other in the prediction of heart disease. Moreover, in the aspect of explainability using the SHAP algorithm, CatBoost has better results because it can explain how the feature impacts the prediction, which is essential for the decision-support system for clinicians.

IV. RESULT AND DISCUSSION

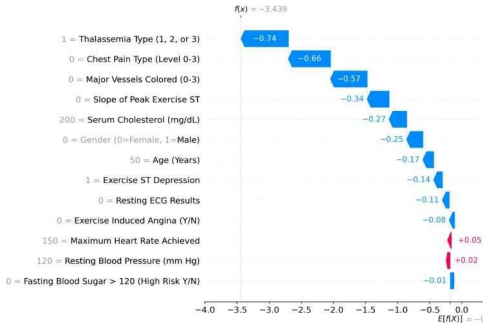

Fig2: SHAP Waterfall Plot (Patient view)

The figure above is a SHAP waterfall plot that explains how each feature contributes to the prediction made by the model for a particular patient. The prediction process starts with the baseline value of the model, E[f(x)] = −0.178. This is an average value without considering a particular patient's features. The features then increase or decrease the prediction value. In the figure above, blue bars indicate features that decrease the prediction value, while red bars indicate features that increase the prediction value slightly. From the figure, we can see that several features, such as thalassemia type, type of chest pain, number of major vessels, ST slope, cholesterol level, gender, and age, decrease the prediction value. This means that these features are likely to cause heart disease. Among these features, thalassemia type (-0.74) and type of chest pain (-0.66) have the greatest impact on lowering the prediction value. There are also a few features that increase the prediction value, such as maximum heart rate achieved and blood pressure. The final prediction is f(x) = −3.439. This means that the patient is likely to have a low risk of heart disease.

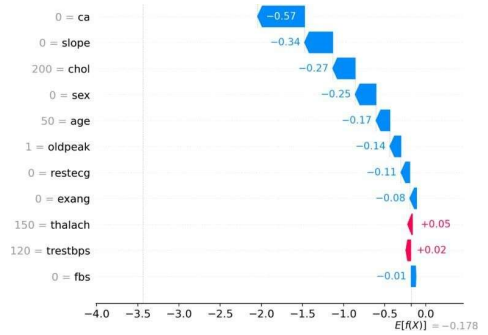

Fig3: SHAP Waterfall Plot (Doctor view)

This result is also a SHAP waterfall plot that interprets the model's prediction for an individual patient. From the average model prediction E[f(x)] = −0.178, each of the clinical variables either lowers or raises the probability of heart disease. Variables in blue, including number of major vessels (ca), ST slope, Cholesterol level, Sex, Age, oldpeak, Resting ECG, and Exercise-induced angina, lower the probability of heart disease. Among these, ca (−0.57) and slope (−0.34) are the most influential. Variables in red, including Maximum heart rate achieved (thalach) and Resting blood pressure (trestbps), raise the probability but only slightly. The effect is also very small. Taken together, the net result of all the variables is to shift the final prediction to the negative side, suggesting that there is a low probability of heart disease in this patient.

V. Model Performance

The developed heart disease prediction model was able to achieve good performance with high accuracy and good classification ability. The model was able to identify significant patterns based on the clinical features provided, which enabled it to classify patients into various categories such as low risk or high risk. The results obtained from the evaluation indicate that the model is able to make consistent predictions. Therefore, it can be deduced that the proposed system is a good tool for supporting the early identification of heart disease risk.

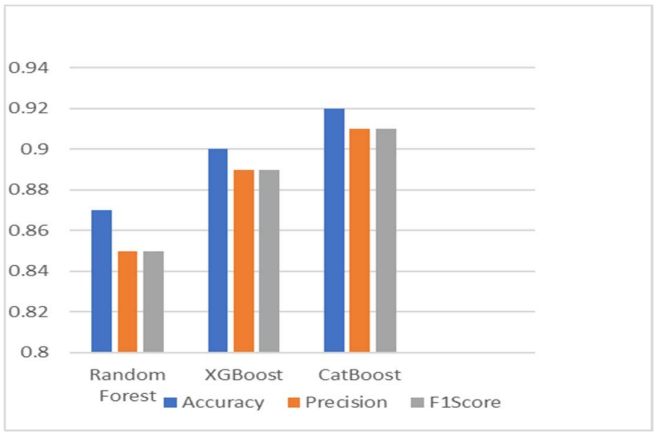

Fig 4: Comparison of Proposed Work with Heart Disease Prediction using Xgboost and Randoforest

The graph illustrates the comparison of the performance of the XGBoost algorithm and the CatBoost algorithm on the basis of some of the key performance parameters of the algorithm. From the graph, it is observed that the CatBoost algorithm is better than the XGBoost algorithm in terms of the prediction accuracy of the algorithm. Hence, the CatBoost algorithm is suitable for structured healthcare data sets, such as the data sets used for heart disease prediction. On the other hand, the XGBoost algorithm is observed to perform better than the CatBoost algorithm on the basis of the training speed of the algorithm. However, the CatBoost algorithm is better suited for predictive modeling because of the accuracy of the algorithm. The main reason for the better performance of the CatBoost algorithm is the direct handling of categorical variables. This is particularly beneficial when working with medical data that tends to contain both numerical and categorical features. Therefore, CatBoost is able to provide consistent and reliable prediction results. Although XGBoost is advantageous when it comes to training a model, it is worth noting that it will be less efficient due to the preprocessing that is required when dealing with categorical features. Therefore, CatBoost is advantageous when it comes to a balance between its efficiency and accuracy.

VI. Conclusion

In the current research, a heart disease prediction system was proposed based on the application of machine learning algorithms. Among the algorithms used in the research, the CatBoost algorithm was identified as the most suitable for the proposed system. This was based on the ability of the algorithm to effectively utilize the clinical information provided to predict the occurrence of heart diseases. The results of the experiment showed the ability of the proposed system to provide better performance compared to the XGBoost algorithm. Based on the results obtained in the evaluation of the proposed system, it is evident that the system is reliable and efficient. As a result, the proposed system can provide valuable support in the detection of heart disease

ACKNOWLEDGMENT

I wish to convey my heartfelt thanks to my guide and faculty for the constant guidance and help rendered during the entire project.

REFERENCES

[1] Cao, K., Liu, C., Yang, S., Zhang, Y., Li, L., Jung, H., & Zhang, S. (2025). Prediction of cardiovascular disease 15(1), 12406.
[2] Rezk, N. G., Alshathri, S., Sayed, A., El-Din Hemdan, E., & El- Behery, H. (2024). XAI-Augmented Voting Ensemble Models for Heart Disease Prediction: A SHAP and LIME-Based Approach. Bioengineering, 11(10), 1016.
[3] Kumar, R., Garg, S., Kaur, R., Johar, M. G. M., Singh, S., Menon, S. V., ... & Lozanović, J. (2025). A comprehensive review of machine learning for heart disease prediction: challenges, trends, ethical considerations, and future directions. Frontiers in Artificial Intelligence, 8, 1583459.
[4] Ganie, S. M., Pramanik, P. K. D., & Zhao, Z. (2025). Ensemble learning with explainable AI for improved heart disease prediction based on multiple datasets. Scientific reports, 15(1), 13912
[5] Ahsan, M. M., & Siddique, Z. (2022). Machine learning based heart disease diagnosis: A systematic literature review. Artificial Intelligence in Medicine, 128, 102289.
[6] Teja, M. D., & Rayalu, G. M. (2025). Optimizing heart disease diagnosis with advanced machine learning models: a comparison of predictive performance. BMC Cardiovascular Disorders, 25(1), 212.
[7] Ghose, P., Oliullah, K., Mahbub, M. K., Biswas, M., Uddin, K. N., & Jamil, H. M. (2025). Explainable AI assisted heart disease diagnosis through effective feature engineering and stacked ensemble learning. Expert Systems with Applications, 265, 125928.
[8] Ashika, T., & Hannah Grace, G. (2025). Enhancing heart disease prediction with stacked ensemble and MCDM based ranking: an optimized RST-ML approach. Frontiers in Digital Health, 7, 1609308
[9] Bhowmik, P. K., Miah, M. N. I., Uddin, M. K., Sizan, M. M. H., Pant, L., Islam, M. R., & Gurung, N. (2024). Advancing heart disease prediction through machine learning: Techniques and insights for improved cardiovascular health. British Journal of Nursing Studies
[10] Vu, T., Kokubo, Y., Inoue, M., Yamamoto, M., Mohsen, A., Martin-Morales, A., ... & Araki, M. (2025). Machine Learning Model for Predicting Coronary Heart Disease Risk: Development and Validation Using Insights From a Japanese Population–Based Study. JMIR cardio, 9(1), e68066
[11] Chandrashekar, K., & Narayanreddy, A. T. (2023). An Ensemble Feature Optimization for an Effective Heart Disease Prediction Model. International Journal of Intelligent Engineering & Systems, 16(2).
[12] Dhanka, S., & Maini, S. (2025). A hybridization of XGBoost machine learning model by Optuna hyperparameter tuning suite for cardiovascular disease classification with significant effect of outliers and heterogeneous training datasets. International Journal of Cardiology, 420, 132757.
[13] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.
[14] Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347–1358.
[15] Alizadehsani, R., Abdar, M., Roshanzamir, M., Khosravi, A., Kebria, P. M., Khozeimeh, F., … Nahavandi, S. (2020). Machine learning–based coronary artery disease diagnosis: A comprehensive review. Computers in Biology and Medicine, 121, 1033