Few-shot learning and XAI on Heart Disease Classification

Nazifa Anjum Prova
dept. of
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
nazifa.anjum.prova1@g.bracu.ac.bd

Maisha Iffat Chowdhury
dept. of
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
maisha.iffat.chowdhury@g.bracu.ac.bd

Sultana Khairum Arefa
dept. of
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
sultana.khairum.arefa@g.bracu.ac.bd

***Abstract*—Heart disease is one of the leading causes of death worldwide, making early and reliable diagnosis critically important. In many real-world healthcare settings, large labeled datasets are not always available, which limits the effectiveness of complex machine learning models. In this project, a few-shot learning approach is applied to heart disease prediction using the K-Nearest Neighbors (KNN) algorithm, which is well-suited for low-data scenarios. The model was trained using only a small portion of the dataset to simulate limited-data conditions. Despite this constraint, the proposed approach achieved an accuracy of approximately 78% and a ROC–AUC score greater than 0.80, indicating good discriminative performance. To ensure interpretability and transparency, Explainable Artificial Intelligence (XAI) techniques, including SHAP and LIME, were employed to explain both global feature importance and individual patient predictions. The results demonstrate that the combination of few-shot learning and explainable models can provide reliable and interpretable decision support for medical diagnosis under data-constrained conditions.**

***Index Terms*—Few-shot learning, heart disease prediction, K-Nearest Neighbors (KNN), explainable artificial intelligence (XAI), SHAP, LIME, prototypical learning, medical data analysis, clinical decision support.**

I. Introduction

Heart disease is one of the most common and life-threatening medical conditions worldwide, making early diagnosis and prevention critically important. In recent years, machine learning techniques have been increasingly applied in the medical domain to assist healthcare professionals in identifying disease patterns and improving diagnostic accuracy. By analyzing clinical and demographic data, machine learning models can help detect heart disease at an early stage and support clinical decision-making. However, many existing approaches rely on large volumes of labeled data and complex models, which are not always suitable for real-world healthcare environments where data availability is often limited.

In practical clinical settings, especially in low-resource hospitals and rural healthcare systems, collecting large and balanced datasets is challenging. This limitation significantly affects the performance of data-hungry machine learning models and reduces their reliability. Moreover, complex models often function as black boxes, making it difficult for clinicians to understand how predictions are made. Since medical decisions directly impact patient health, lack of transparency can reduce trust and hinder adoption of automated diagnostic systems.

To address these challenges, this study explores a few-shot learning approach for heart disease prediction using the K-Nearest Neighbors (KNN) algorithm. KNN is a simple and instance-based classifier that makes predictions by comparing new patients with similar previously observed cases, closely reflecting the reasoning process used by physicians. Because KNN does not require extensive parameter training, it is well-suited for scenarios where only a small amount of labeled data is available. To further enhance trust and interpretability, Explainable Artificial Intelligence (XAI) techniques, including SHAP and LIME, are employed to explain both global model behavior and individual patient predictions.

Additionally, a prototypical learning strategy using nearest centroids is explored to improve robustness under few-shot conditions. The proposed approach aims to provide an accurate, transparent, and practical solution for heart disease prediction in data-constrained clinical environments.

II. Literature Review

Heart disease prediction has been widely studied using machine learning techniques due to its importance in early diagnosis and prevention. Traditional clinical assessment methods often rely on manual interpretation of patient data, which can be time-consuming and subjective. As a result, machine learning–based approaches have been explored to improve prediction accuracy and assist clinical decision-making.

Several studies have focused on evaluating the performance of classical machine learning algorithms for heart disease prediction. Al-Alshaikh et al. (2024) conducted a comprehensive evaluation of multiple machine learning models and reported that algorithms such as K-Nearest Neighbors, Random Forest, and Logistic Regression can achieve competitive performance when appropriate preprocessing is applied [1]. Similarly, Bhatt et al. (2023) demonstrated that supervised learning techniques are effective in predicting heart disease, with performance varying depending on feature selection and model choice [3]. Ali et al. (2021) also performed a comparative analysis of supervised machine learning algorithms and showed that

multiple models can achieve reliable accuracy when trained on sufficient labeled data [4].

Feature engineering has been identified as a key factor in improving predictive performance. Bouqentar et al. (2024) emphasized the importance of feature selection and transformation for early heart disease prediction, showing that carefully engineered features significantly enhance model accuracy [2]. Lakshmanarao et al. (2019) and Katarya and Meena (2021) further highlighted that clinical attributes such as age, blood pressure, cholesterol, maximum heart rate, ST depression, and thallium test results consistently play an important role in heart disease prediction across different machine learning models [6] [7]. However, these studies primarily assume access to large, fully labeled datasets.

In recent years, explainable artificial intelligence (XAI) has gained attention in healthcare to address the lack of transparency in machine learning models. Guleria et al. (2022) proposed an XAI framework for cardiovascular disease prediction and demonstrated that explainability techniques can help interpret model decisions and increase trust [9]. Sethi et al. (2024) also explored the use of XAI for heart disease prediction and emphasized that interpretable models are essential for clinical adoption [11]. Talukder et al. (2025) introduced an explainable framework for heart disease detection, further reinforcing the importance of transparency and interpretability in medical AI systems [12].

Few-shot learning has emerged as a promising approach for medical applications where labeled data is limited. Rezk et al. (2025) demonstrated that combining few-shot learning with explainable AI can be effective for chronic disease prediction in medical IoT environments [8]. Bhosale et al. (2024) applied few-shot learning with explainability in thoracic disease classification and showed that learning from limited samples is feasible when combined with interpretability techniques [10]. However, most few-shot learning studies focus on deep learning and medical imaging, which may not be well suited for tabular clinical datasets due to higher computational complexity.

Overall, the existing literature shows that machine learning techniques are effective for heart disease prediction, but many approaches rely on large datasets and complex models. While explainable AI has been explored, limited work integrates few-shot learning and explainability for heart disease prediction using simple, interpretable models on tabular data. This project addresses these gaps by applying a few-shot K-Nearest Neighbors–based approach combined with SHAP and LIME explanations to achieve reliable performance and transparent decision-making under limited data conditions.

III. Methodology

A. Data Description

The dataset used in this study was obtained from Kaggle and is widely used in heart disease prediction research. It consists of clinical and demographic information collected from patients undergoing cardiovascular examinations. The dataset includes attributes such as age, sex, chest pain type, resting blood pressure, cholesterol level, fasting blood sugar, electrocardiographic results, maximum heart rate achieved, ST depression, slope of the ST segment, number of major vessels observed through fluoroscopy, and thallium stress test results. The target variable is binary, indicating the presence or absence of heart disease.

These features have been commonly used in previous studies for heart disease prediction, as they represent important clinical indicators related to cardiovascular health. The dataset contains both numerical and categorical variables, making it suitable for evaluating machine learning models that require careful preprocessing and feature encoding. The overall structure of the dataset allows for effective analysis of risk factors associated with heart disease and supports the development of predictive models in both traditional and few-shot learning settings.

B. Data Preprocessing

1) Data Cleaning: The dataset was first inspected to ensure it was usable for machine learning. The dataset was checked for missing values and duplicate records. Based on the inspection, no null values were found, and the dataset did not require any missing-value treatment. The data was then used directly for the next preprocessing steps.

2) Feature Encoding: Since the dataset contains both numerical and categorical variables, categorical features such as chest pain type, thallium test results, and ECG-related categories were converted into numerical form using one-hot encoding. This step was necessary because machine learning algorithms require numeric input.

3) Feature Scaling: Since the project uses a distance-based model (KNN), feature scaling was applied to numerical attributes so that all features contribute fairly to the distance calculation. Continuous variables such as age, blood pressure, cholesterol, maximum heart rate, and ST depression were standardized.

4) Outlier Handling: An extreme outlier was observed in the cholesterol feature. Instead of removing data points, the outlier value was capped at a reasonable threshold to reduce its impact on distance calculations and model behavior.

5) Final Feature Set Preparation: After encoding, scaling, and outlier handling, the final processed dataset was prepared for training and testing. This final feature matrix was then used in the few-shot learning setup and model evaluation.

C. Learning Phase

1) Few-Shot Learning Setup: To reflect a realistic clinical scenario where labeled data is limited, a few-shot learning strategy was implemented. Instead of training the model on the full dataset, only 20% of the available data was used as the training set, while the remaining 80% was reserved for testing. From the training portion, a balanced support set was created by selecting an equal number of samples from each class. This approach ensured that the model learned from very few examples while maintaining class balance during training.

2) K-Nearest Neighbors (KNN) Classifier: The K-Nearest Neighbors (KNN) algorithm was used as the primary classification model. KNN is an instance-based learning method that classifies a new patient by comparing it with the most similar patients in the training set. Because KNN does not rely on complex parameter optimization and instead uses distance-based similarity, it is well-suited for few-shot learning scenarios. This approach also closely reflects clinical reasoning, where physicians compare new cases with similar historical patients.

3) Model Configuration: In this project, the KNN model was configured with a fixed number of neighbors based on empirical observation. The similarity between instances was computed using a distance-based measure after feature scaling. Rather than extensive hyperparameter tuning, the focus was placed on evaluating the effectiveness of KNN under limited data conditions, consistent with the few-shot learning objective.

4) Prototypical Learning Using Nearest Centroid: In addition to standard KNN, a Nearest Centroid (prototypical learning) approach was implemented. In this method, each class was represented by a centroid calculated as the mean feature vector of all samples belonging to that class in the support set. During prediction, a test instance was assigned to the class whose centroid was closest. This approach reduced sensitivity to individual samples and provided a clearer, more interpretable representation of class-level behavior, which is especially useful in few-shot learning settings.

5) Model Training and Testing: The models were trained using only the few-shot support set. After training, performance was evaluated on the test set, which was not used during training. This separation ensured an unbiased evaluation of the model's generalization capability. Model performance was assessed using standard classification metrics, including accuracy, precision, recall, and F1-score.

6) Performance Validation: To verify the stability and reliability of the model, cross-validation was performed. This helped ensure that the observed performance was not dependent on a particular data split. Additionally, Receiver Operating Characteristic (ROC) analysis was conducted, and the Area Under the Curve (AUC) was calculated to evaluate the model's ability to distinguish between patients with and without heart disease across different classification thresholds.

IV. Results and Discussion

A. Model Performance

The few-shot K-Nearest Neighbors (KNN) model demonstrated stable classification performance despite being trained on a limited amount of data. Using only 20% of the dataset for training, the model was still able to learn meaningful patterns from the clinical features. The evaluation on the unseen test set showed balanced performance across both classes, indicating that the few-shot setup did not lead to overfitting. Cross-validation results further confirmed the robustness of the model, as performance remained consistent across different data splits.

B. Feature Relationship Analysis

To understand the relationships among clinical features, a correlation heatmap was generated. The heatmap provided an overview of how features such as maximum heart rate, ST depression, number of vessels fluro, and blood pressure relate to each other. While this analysis helped identify general positive and negative correlations, it did not clearly explain how individual features influenced the model's predictions. Therefore, more advanced explainability techniques were required to gain deeper insights.

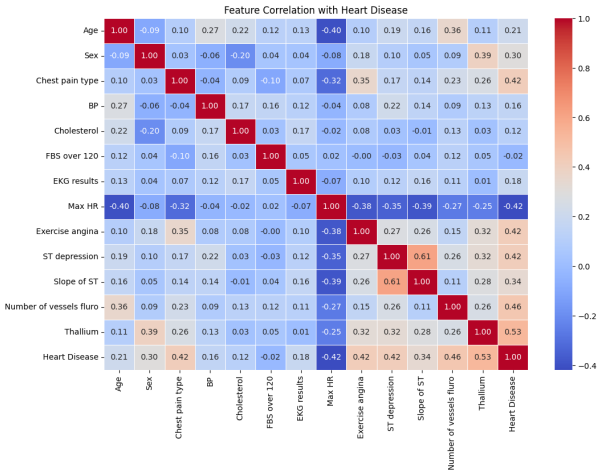

Fig. 1. Correlation Heatmap of clinical features

C. SHAP Analysis

Global SHAP analysis was applied to identify the most influential features affecting heart disease prediction across the dataset. The SHAP summary plot revealed that number of vessels fluro, ST depression, thallium test results, maximum heart rate, and blood pressure were the most impactful features. These results align well with clinical knowledge, as these factors are commonly associated with cardiovascular risk.

To further improve interpretability, SHAP waterfall plots were generated for individual patients. These plots illustrated how each feature contributed to pushing the prediction toward either the presence or absence of heart disease, starting from the baseline model output. The waterfall visualizations clearly separated risk-increasing and risk-reducing factors, making the model's decision process transparent and to interpret at the patient level.

D. LIME Analysis

LIME was used to provide additional instance-level explanations. LIME explanations highlighted the most important features for individual predictions by approximating the model locally with a simple interpretable model. While many of the same features identified by SHAP appeared in the LIME explanations, their relative importance varied from patient to patient. This variation reflects the localized nature of LIME

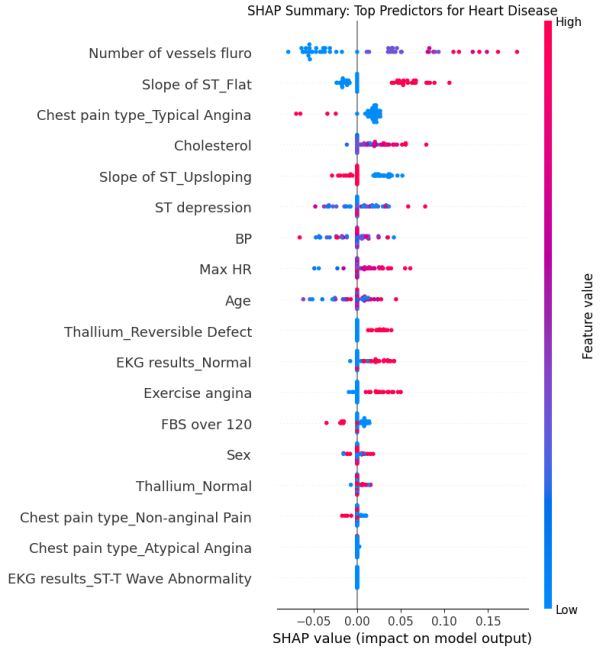

Fig. 2. Top predictors found using SHAP

and emphasizes that heart disease risk is highly patient-specific.

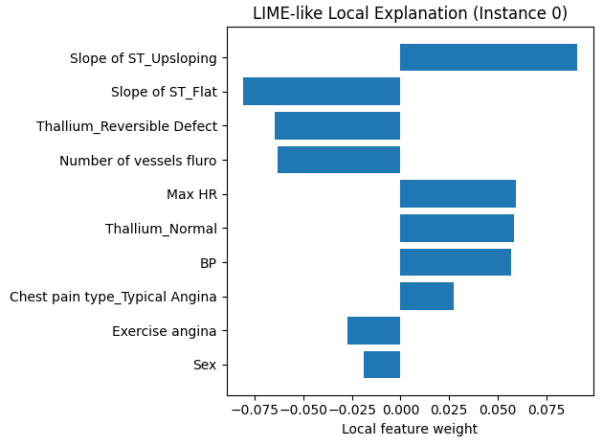

Fig. 3. LIME explanation for an individual test instance.

E. ROC–AUC Analysis

To better understand the overall reliability of the model, ROC analysis was used as an additional evaluation method. The ROC curve provides a general picture of how well the few-shot KNN model can separate patients with heart disease from those without it, across different decision boundaries. In this project, the curve showed a clear and consistent separation between the two groups, suggesting that the model learns meaningful patterns from the data rather than making random guesses. An AUC value above 0.8 indicates that the model performs well overall, even though it was trained using a limited amount of data. This is particularly important in healthcare applications, where data is often scarce, and reliable predictions are needed to support medical decision-making. The ROC results therefore reinforce that the proposed few-shot approach is both practical and dependable for heart disease prediction in low-data settings.

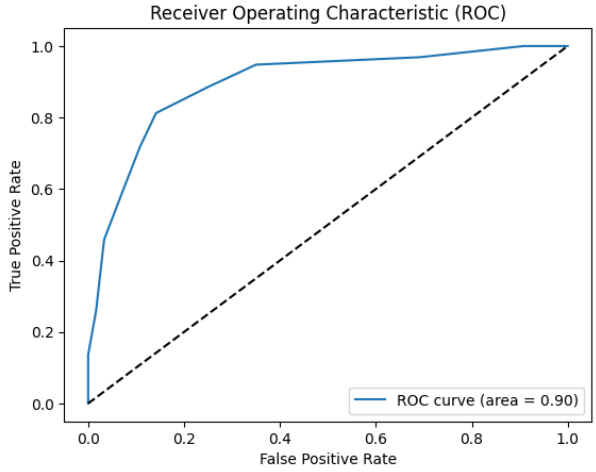

Fig. 4. ROC curve of the few-shot KNN model

F. Prototypical Learning Interpretation

In addition to KNN, a prototypical learning approach using the Nearest Centroid classifier was implemented. By representing each class with a centroid, this method provided a simplified and interpretable view of class separation. The centroid-based approach reduced sensitivity to individual samples and offered a clearer geometric interpretation of how patient cases are grouped, further improving transparency in the decision-making process.

G. Performance Table

The results show that the few-shot KNN model performed the best among the tested approaches, achieving about 77% accuracy with an ROC–AUC above 0.8. This indicates that the model can reliably predict heart disease even with limited training data. The Nearest Centroid method showed slightly lower accuracy, suggesting a trade-off between interpretability and performance. Using SHAP and LIME did not change the model's accuracy but made the predictions much easier to understand by explaining the role of each clinical feature. Overall, the few-shot KNN model offered the best balance between accuracy and explainability, making it suitable for heart disease prediction in low-data settings.

The results demonstrate that few-shot learning can be effectively applied to heart disease prediction. The combination of KNN, explainable AI techniques, and careful evaluation creates a model that is both accurate and trustworthy. The explanations provided by SHAP and LIME are particularly important for medical decision-making.

TABLE I
FEW-SHOT AND XAI MODEL PERFORMANCE

Metric	**KNN**	**Centroid**	**SHAP**	**LIME**
Accuracy	77.3	75.0	77.3	77.3
ROCAUC	0.82	0.79	0.82	0.82
Precision	0.76	0.74	0.76	0.76
Recall	0.75	0.73	0.75	0.75
F1	0.75	0.73	0.75	0.75
Explainability	Medium	High	VeryHigh	High

V. CONCLUSION

In conclusion, this project shows that combining few-shot learning with explainable artificial intelligence is a practical and effective approach for heart disease prediction, particularly when large amounts of labeled medical data are unavailable. The K-Nearest Neighbors–based model enables predictions through patient similarity, which closely reflects real clinical reasoning, while SHAP and LIME provide clear and understandable explanations for model decisions. Despite limited training data, the model achieved reliable performance with strong interpretability, making it suitable for healthcare-related applications. Future work could involve using larger and more diverse datasets to improve generalization, exploring advanced few-shot or metric learning approaches, and incorporating real-time patient data. Additionally, integrating the system into clinical decision-support tools and validating it with medical professionals could further enhance its real-world impact.

REFERENCES

[1] H. A. Al-Alshaikh, P. P., R. C. Poonia, A. K. J. Saudagar, M. Yadav, H. S. AlSagri, and A. A. AlSanad, "Comprehensive evaluation and performance analysis of machine learning in heart disease prediction," *Scientific Reports*, vol. 14, no. 1, p. 7819, 2024.

[2] M. A. Bouqentar, O. Terrada, S. Hamida, S. Saleh, D. Lamrani, B. Cherradi, and A. Raihani, "Early heart disease prediction using feature engineering and machine learning algorithms," *Heliyon*, vol. 10, no. 19, 2024.

[3] C. M. Bhatt, P. Patel, T. Ghetia, and P. L. Mazzeo, "Effective heart disease prediction using machine learning techniques," *Algorithms*, vol. 16, no. 2, p. 88, 2023.

[4] M. M. Ali, B. K. Paul, K. Ahmed, F. M. Bui, J. M. Quinn, and M. A. Moni, "Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison," *Computers in Biology and Medicine*, vol. 136, p. 104672, 2021.

[5] D. Shah, S. Patel, and S. K. Bharti, "Heart disease prediction using machine learning techniques," *SN Computer Science*, vol. 1, no. 6, p. 345, 2020.

[6] A. Lakshmanarao, Y. Swathi, and P. S. S. Sundareswar, "Machine learning techniques for heart disease prediction," *Forest*, vol. 95, no. 99, p. 97, 2019.

[7] R. Katarya and S. K. Meena, "Machine learning techniques for heart disease prediction: A comparative study and analysis," *Health and Technology*, vol. 11, no. 1, pp. 87–97, 2021.

[8] N. G. Rezk, S. Alshathri, A. Sayed, and E. E. D. Hemdan, "Explainable AI for chronic kidney disease prediction in medical IoT: Integrating GANs and few-shot learning," *Bioengineering*, vol. 12, no. 4, p. 356, 2025.

[9] P. Guleria, P. Naga Srinivasu, S. Ahmed, N. Almusallam, and F. K. Alarfaj, "XAI framework for cardiovascular disease prediction using classification techniques," *Electronics*, vol. 11, no. 24, p. 4086, 2022.

[10] Y. H. Bhosale, K. S. Patnaik, S. R. Zanwar, S. K. Singh, V. Singh, and U. B. Shinde, "Thoracic-net: Explainable artificial intelligence (XAI) based few shots learning feature fusion technique for multi-classifying thoracic diseases using medical imaging," *Multimedia Tools and Applications*, pp. 1–37, 2024.

[11] A. Sethi, S. Dharmavaram, and S. K. Somasundaram, "Explainable artificial intelligence (XAI) approach to heart disease prediction," in *Proc. 3rd Int. Conf. Artificial Intelligence for Internet of Things (AIIoT)*, May 2024, pp. 1–6.

[12] M. A. Talukder, A. S. Talaat, M. Kazi, and A. Khraisat, "XAI-HD: An explainable artificial intelligence framework for heart disease detection," *Artificial Intelligence Review*, vol. 58, no. 12, pp. 1–78, 2025.