UtilitasMathematica

ISSN 0315-3681 Volume 122, 2025

AN EXPLAINABLE ENSEMBLE LEARNING FRAMEWORK USING SHAP AND XGBOOST FOR EARLY PREDICTION OF HEART DISEASE FROM CLINICAL AND LIFESTYLE DATA

Dr L K SureshKumar [1]
Associate Professor, Department of Computer Science and Engineering, Osmania University, Hyderabad
lksureshkumar@osmania.ac.in
Ravi Uyyala [2]
Associate Professor
Department of Computer Science and Engineering, Chaitanya Bharathi Institute of Technology, Gandipet, Hyderabad-500075, Telangana, India, uyyala.ravi@gmail.com
K. Srinivasa Chakravarthy [3]
Assistant Professor, Department of Information Technology, Vasavi College of Engineering(A), Hyderabad
ks.chakravarthy@staff.vce.ac.in
M. Sathya Devi [4]
Assistant Professor, Department of Information Technology, Vasavi College of Engineering(A), Hyderabad
sathyamaranganti@staff.vce.ac.in
R. Suresh [5]
Associate Professor, Department of artificial Intelligence and Data Science, Ace Engineering College
hodaid@aceec.ac.in

Received: 18/05/2025 **Revised:** 16/06/2025 **Accepted:** 01/07/2025 **Published:** 28/07/2025

Abstract: Cardiovascular diseases continue to be a major contributor to global mortality rates, necessitating the development of Reliable, Explainable, And Accurate early diagnostic tools. This research introduces a Novel Hybrid Machine Learning Framework that integrates Extreme Gradient Boosting (XGBoost) with Shapley Additive Explanations (SHAP) to forecast heart disease using a Comprehensive Combination of clinical and lifestyle attributes. The model utilizes the UCI Heart Disease Dataset, enriched with Synthetically Generated lifestyle features to Improve Feature Diversity and Model Generalization. Comprehensive data cleaning and Advanced Feature Selection Methods are applied to enhance predictive performance. Compared to conventional algorithms such as logistic regression, support vector machines, and random forest, XGBoost achieved Superior Performance Metrics, including a 92.6% accuracy rate and an AUC-ROC score of 0.96. SHAP was used to clarify the impact of individual features both at the dataset level and for specific predictions, with significant influence noted from factors like chest pain type, age, and cholesterol. The interpretable framework demonstrates Significant Potential for Clinical Decision Support, improving both accuracy and understanding of model decisions.

Keywords - Heart Disease, XGBoost, SHAP, Explainable AI, Clinical Data

Introduction

Cardiovascular diseases (CVDs) remain the foremost cause of global mortality, claiming approximately 17.9 million lives annually, as reported by the World Health Organization [23]. These diseases encompass a range of heart and blood vessel disorders, including coronary artery disease, heart failure, and arrhythmias. The primary concern with CVDs lies not only in their fatality but also in their insidious and asymptomatic progression, which often leads to delayed diagnosis and reduced therapeutic efficacy [29]. This delay underscores the critical importance of early and accurate diagnosis, enabling preventative care and reducing mortality. While traditional diagnostic methods such as electrocardiograms (ECGs), echocardiograms, angiograms, and blood pressure monitoring are valuable, they may fall short in detecting disease onset at early stages, particularly when symptoms are non-specific. Moreover, these approaches rely heavily on manual interpretation, clinical expertise, and are often inaccessible to patients in rural or resource-limited settings [26].

In response to these limitations, the integration of Artificial Intelligence (AI) and Machine Learning (ML) into medical diagnostics has received substantial attention. ML enables data-driven insights, uncovering complex, non-linear relationships among variables that are otherwise hard to detect. These approaches allow the development of predictive models capable of identifying heart disease patterns from clinical data, thereby augmenting physician decision-making and facilitating early intervention [12], [1], [4]. Several conventional ML algorithms have been employed for heart disease classification, including Logistic Regression, Decision Trees, Support Vector Machines (SVMs), K-Nearest Neighbors (KNN), and Artificial Neural Networks (ANNs) [1], [4], [11]. For instance, Jabbar et al. proposed a KNN-genetic algorithm hybrid that demonstrated significant accuracy for classifying heart disease [1]. However, these models often sacrifice transparency for performance and operate as black-box systems, making it difficult to interpret how decisions are made. In high-stakes environments like healthcare, such lack of interpretability can severely hinder clinical adoption [18], [21], [27]. Indeed, interpretability is a non-negotiable attribute in clinical decision-making. It facilitates trust, accountability, and error analysis—key elements when dealing with life-altering diagnoses [17], [19], [24]. Researchers and clinicians alike have raised concerns about deploying models that offer no rationale behind their predictions [22]. As Rudin emphasizes, there is growing consensus that healthcare models should not just be explainable post hoc but should be inherently interpretable [27]. Nevertheless, explainability does not necessarily imply compromise on performance. To bridge the gap between accuracy and transparency, ensemble learning techniques—especially Extreme Gradient Boosting (XGBoost)—have emerged as promising solutions. XGBoost has gained widespread adoption due to its high scalability, handling of missing data, and ability to model complex interactions among features [6]. It combines multiple weak learners (typically decision trees) to create a robust predictive model that consistently performs well on structured tabular data. Studies have shown that ensemble models outperform single classifiers in heart disease prediction tasks, with higher Area Under Curve (AUC) and F1 scores [26].

However, XGBoost is not inherently interpretable. Although it provides superior accuracy, it lacks mechanisms to intuitively explain which features contribute to a prediction, a fundamental limitation in real-world clinical deployment [22]. To tackle this challenge, Explainable Artificial Intelligence (XAI) has emerged as a pivotal research direction. Among the suite of XAI tools, SHAP (SHapley Additive explanations) has gained prominence for offering local and global interpretability rooted in cooperative game theory [3]. Developed by Lundberg and Lee, SHAP values quantify the contribution of each feature to a given prediction, ensuring both consistency and local accuracy—critical requirements in healthcare applications [3], [28]. SHAP's ability to generate intuitive visualizations of feature importance has made it a trusted tool for clinicians seeking model accountability [20], [28]. It provides granular insight into predictions at the individual patient level, which can support or refute clinical hypotheses, improve model debugging, and enhance trust in AI-assisted systems. In parallel, the quality and composition of the dataset used to train such models significantly influence their performance and generalizability. The UCI Heart Disease Dataset, available through the UCI Machine Learning Repository [10], [30], is one of the most widely used benchmark datasets for CVD prediction. Although this dataset captures critical clinical variables like age, cholesterol, and chest pain type, it lacks detailed lifestyle and behavioral attributes, which are vital predictors of cardiovascular health. Incorporating such lifestyle features—such as smoking habits, physical activity, diet, and stress—can substantially enhance model robustness and ecological validity [26], [29]. To address class imbalance—a common issue in medical datasets where the number of healthy cases far outweighs those with disease—techniques like SMOTE (Synthetic Minority Over-Sampling Technique) are used to improve model sensitivity [8]. This ensures that minority classes (i.e., patients with heart disease) are adequately represented during model training, reducing bias and improving the model's capacity to detect rare events. In terms of model evaluation, metrics like Receiver Operating Characteristic (ROC) curves, AUC, precision, recall, and F1-score are essential to quantify performance [5]. However, these metrics, while informative, do not explain why a model performs well, reinforcing the need for paired interpretability tools like SHAP [3], [7].

This study proposes a hybrid framework that integrates XGBoost with SHAP explanations, specifically designed for the early and explainable prediction of heart disease. The framework augments the standard UCI dataset with synthetic lifestyle attributes to capture a broader range of real-world health indicators and employs SHAP-based post-hoc interpretation on the XGBoost model to visualize and quantify feature contributions to predictions at both local and global levels. This dual approach ensures that the model is not only accurate but also transparent, enabling clinicians to understand the rationale behind each prediction and make more informed decisions. The objective is not just to achieve high-performance classification but to demonstrate how interpretable ML frameworks can be deployed responsibly in clinical practice. The remainder of this paper is organized as follows: Section 2 reviews relevant literature on heart disease prediction and explainable AI techniques. Section 3 details the methodology, including

dataset preprocessing, model design, and SHAP integration. Section 4 presents experimental results and evaluation metrics. Section 5 discusses findings and implications for clinical deployment. Finally, Section 6 concludes with future research directions and limitations.

2. Literature Review

The application of machine learning to cardiovascular disease (CVD) prediction has evolved substantially over the past decade. Early approaches primarily relied on statistical classifiers like Logistic Regression due to their ease of implementation and interpretability. However, such models were limited in capturing complex nonlinear relationships and interactions between clinical variables, prompting researchers to explore more advanced methods [1]. In their study, Jabbar et al. [1] proposed a hybrid model combining K-Nearest Neighbor (KNN) with a Genetic Algorithm (GA) for the classification of heart disease. Their approach aimed to optimize feature selection using GA, improving the performance of the KNN classifier. The results demonstrated higher accuracy compared to traditional techniques, highlighting the importance of feature optimization in improving ML performance in healthcare tasks. While [1] focused on heart disease, Chaurasia and Pal [2] demonstrated a similar methodology in the domain of breast cancer detection. Using Decision Trees, Naive Bayes, and KNN algorithms, they showcased the impact of algorithm selection on classification performance. Although not specific to CVD, their work provided insights into how data mining techniques and model comparison strategies could be generalized across disease domains, emphasizing the growing utility of ML in biomedical applications. The challenge of interpretability, especially in high-stakes domains such as healthcare, has led to the development of explainable AI (XAI) tools. One of the most prominent contributions in this field is the SHAP (SHapley Additive explanations) framework, introduced by Lundberg and Lee [3]. SHAP builds on cooperative game theory to explain individual predictions by assigning each feature a contribution value. Their method ensures consistency and local accuracy, making it suitable for clinical applications where understanding why a model made a prediction is just as important as the prediction itself.

In the evolving landscape of machine learning applications in healthcare, Haq et al. [4] investigated the use of Artificial Neural Networks (ANNs) for heart disease prediction. Their research demonstrated that ANNs are capable of achieving high classification accuracy, primarily due to their ability to model complex and nonlinear relationships among clinical variables. This is particularly advantageous in cardiovascular diagnostics, where multiple interacting risk factors—such as blood pressure, cholesterol, smoking, and age—contribute collectively to disease manifestation. The inherent flexibility of ANNs allows them to capture these interactions more effectively than traditional linear models.
However, despite their predictive strength, ANNs suffer from a major limitation: the "black-box" problem. Neural networks operate through layers of abstracted computations that make it extremely difficult to trace how individual features influence the final prediction. In clinical environments, where transparency, accountability, and interpretability are non-negotiable requirements, this opaqueness becomes a critical barrier to adoption. Physicians and clinical stakeholders are less likely to trust or rely on a decision-support system that cannot explain why a specific diagnosis or risk prediction was made, especially when treatment or prognosis depends on it.

To address this challenge, the research community has increasingly turned toward post-hoc explainability techniques, with SHAP (SHapley Additive explanations) emerging as one of the most widely accepted frameworks [3]. SHAP values allow researchers and clinicians to quantify the impact of each input feature on a given prediction, offering both global interpretability (across the dataset) and local interpretability (for individual patients). By attributing a share of the model's output to each feature based on cooperative game theory, SHAP makes it possible to generate visualizations such as summary plots, dependence plots, and force plots, thereby bridging the gap between model complexity and clinical usability. While SHAP enhances the interpretability of complex models, another critical aspect of performance evaluation in medical classification tasks is Receiver Operating Characteristic (ROC) analysis. As described in the foundational work by Fawcett [5], the ROC curve plots the true positive rate (sensitivity) against the false positive rate (1-specificity) at various threshold settings. The Area Under the ROC Curve (AUC-ROC) provides a single scalar value summarizing the model’s discriminatory power—its ability to distinguish between patients with and without the disease. This is particularly valuable in healthcare scenarios, where datasets are often imbalanced, and precision-recall trade-offs must be carefully considered. Although ROC and AUC do not contribute to model explainability, they are indispensable for benchmarking classifier performance, especially when comparing models like XGBoost, SVM, and ANN. They provide a threshold-independent measure of model effectiveness, enabling robust

cross-model evaluations and offering insights into how well a classifier will perform in real-world screening or diagnostic tasks. In summary, while ANNs offer substantial predictive power, their lack of transparency limits their clinical deployment. Techniques like SHAP address this by injecting interpretability into complex models, while ROC analysis complements these efforts by offering a reliable, quantitative framework for evaluating model performance. Together, these tools ensure that machine learning models in healthcare are not only accurate but also explainable, accountable, and trustworthy.

In terms of algorithmic development, Chen and Guestrin [6] introduced XGBoost, a highly efficient and scalable implementation of gradient boosting. It quickly became the model of choice for many tabular data problems due to its superior accuracy, regularization capabilities, and robustness against overfitting. The relevance of XGBoost in the medical domain stems from its ability to handle missing data, model feature interactions, and scale to large datasets—features essential in clinical data analysis. Yet, like ANNs, XGBoost lacks native interpretability, which limits its deployment in sensitive domains. Addressing this, Ribeiro et al. [7] proposed the LIME (Local Interpretable Model-agnostic Explanations) framework, another popular XAI method. Unlike SHAP, which is model-specific in some implementations, LIME explains any black-box model by approximating it locally with an interpretable model. Both LIME and SHAP have been used in clinical settings, but SHAP is generally preferred due to its theoretical grounding and consistency guarantees [3], [7]. Additionally, Chawla et al. [8] tackled the challenge of class imbalance in medical datasets through the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE generates synthetic examples of the minority class to balance the dataset, enhancing model sensitivity to underrepresented outcomes such as positive heart disease diagnoses. This technique plays a crucial role in ensuring that predictive models do not exhibit bias towards the majority class and can reliably detect disease conditions even when their occurrence is statistically rare.

Recent surveys, such as the one conducted by Khan and Hussain [12], have offered comprehensive evaluations of various machine learning (ML) models applied to heart disease prediction, shedding light on both their strengths and limitations. Their review emphasizes the increasing importance of adopting holistic modeling approaches—those that not only consider traditional clinical parameters (such as blood pressure, cholesterol levels, and ECG results) but also incorporate behavioral and lifestyle-related features like physical activity, diet, smoking habits, and sleep duration. The authors further highlight the necessity of employing robust learning algorithms capable of handling the inherent complexity and variability of healthcare data. Crucially, they underscore the significance of model interpretability tools, such as SHAP (SHapley Additive explanations) and LIME (Local Interpretable Model-Agnostic Explanations), which are essential for gaining clinical trust and facilitating real-world adoption. These findings reinforce a growing consensus in the literature: there is a pressing need to innovate not only in model design and algorithmic sophistication, but also in the development of deployment frameworks that are aligned with the practical demands and ethical constraints of healthcare environments.

In synthesizing the reviewed literature, it is evident that the field of heart disease prediction has been undergoing a notable transition—from the use of simpler, inherently interpretable models to more complex, high-performing "black-box" models. Foundational techniques such as K-Nearest Neighbors (KNN), Decision Trees, and Artificial Neural Networks (ANNs) were instrumental in establishing early benchmarks for predictive accuracy in clinical applications [1], [2], [4]. However, their limitations in scalability and generalizability have driven the adoption of more sophisticated algorithms such as Extreme Gradient Boosting (XGBoost) and the integration of post-hoc interpretability frameworks like SHAP [3], [6], [7], [8]. These newer paradigms offer a more favorable trade-off between predictive performance and explainability, aligning well with the ethical and operational requirements of modern clinical decision support systems.

Despite these advances, one critical gap consistently identified in the literature is the underutilization of lifestyle-related variables in predictive modeling. While clinical features remain central to diagnosis, lifestyle factors such as dietary patterns, smoking status, physical activity levels, and sleep behavior are well-established contributors to cardiovascular health. Yet, these attributes are often excluded from predictive models, particularly those that prioritize interpretability. This omission limits the model's ability to provide personalized and behaviorally actionable insights, which are essential for preventative care strategies.

The present research seeks to address this overlooked aspect by proposing an augmented and interpretable ensemble learning framework for heart disease prediction. This framework not only leverages the predictive power of advanced machine learning algorithms like XGBoost but also ensures model transparency through the use of SHAP-based

feature attribution techniques. By systematically integrating synthetically enriched lifestyle data with conventional clinical variables, this study aims to develop a model that is not only accurate but also actionable and trustworthy—qualities that are essential for real-world clinical implementation.

3.Methodology and Model Development

The XGBoost was chosen as the primary classifier due to its Capability to Handle Missing Values, Automatic Feature Interaction Discovery, and computational efficiency. Training was conducted using Stratified 10-Fold Cross-Validation to ensure class balance and reduce bias. Hyper parameters—including learning rate, tree depth, and number of estimators—were fine-tuned through a Grid Search Combined with Cross-Validation to optimize model performance.

3.1 Dataset Description

The dataset employed in this study originates from the widely recognized UCI Heart Disease Repository [30], which is frequently used as a benchmark in cardiovascular disease prediction research. This dataset originally comprises 14 clinical features including age, sex, chest pain type (cp), resting blood pressure (trestbps), serum cholesterol (chol), fasting blood sugar (fbs), resting electrocardiographic results (restecg), maximum heart rate achieved (thalach), exercise-induced angina (exang), ST depression induced by exercise (old peak), slope of the ST segment, number of major vessels colored by fluoroscopy (ca), and thalassemia status. These features collectively capture core physiological indicators critical to diagnosing and managing heart disease. However, real-world risk modelling for cardiovascular diseases goes beyond clinical measurements. There is growing consensus in the medical and computational communities that lifestyle and behavioural factors significantly influence cardiovascular outcomes and must be integrated into predictive models [26], [12]. For instance, poor dietary habits, physical inactivity, sleep deprivation, obesity, and tobacco use are all well-documented risk factors associated with increased CVD incidence and mortality. Omitting these dimensions from predictive modelling can limit a system's ecological validity and practical utility, particularly in personalized diagnostics.

To reflect the paradigm, shift towards lifestyle-aware precision medicine, the base UCI dataset was augmented with five synthetic lifestyle attributes: dietary patterns (e.g., saturated fat intake level), physical activity levels (frequency and duration of exercise per week), body mass index (BMI), smoking status (categorized as current, former, or non-smoker), and average sleep duration (measured in hours per night).These features were generated using domain-informed statistical simulations, ensuring that the synthetic values aligned with population health statistics and preserved plausible clinical-lifestyle interactions. For example, BMI values were simulated using normal distributions segmented by age and gender cohorts, while smoking prevalence followed national health survey data patterns. Such controlled augmentation ensures statistical realism while extending the dataset's ability to capture real-life variance in heart disease risk. This integration of behavioural data is consistent with recent clinical research that advocates multi-modal data fusion for improving disease prediction models [12], [26]. In particular, SHAP-based explainable models benefit from having more semantically diverse features, as they offer richer, more interpretable feature contributions. To further ensure model robustness and mitigate the issue of class imbalance common in medical datasets, the data was expanded and balanced using a combination of oversampling techniques like SMOTE [8] and random under sampling of the majority class. The final dataset consisted of 30,000 patient records, equally distributed between patients with and without heart disease (15,000 per class). This balance was crucial for avoiding biased performance metrics and ensuring fairness in classification outcomes.

By combining clinical, behavioural, and demographic information, this enriched dataset supports the construction of a comprehensive predictive framework that mirrors real-world diagnostic conditions. It also facilitates the development of interpretable AI models that not only predict with high accuracy but also provide insights into how lifestyle factors interact with physiological conditions to influence cardiovascular health outcomes.

3.2 Data Pre-processing

The dataset underwent a comprehensive preprocessing pipeline to ensure high reliability and integrity for machine learning modelling, particularly crucial in healthcare applications. To manage missing data, a multi-strategy imputation approach was employed: mean imputation was applied to continuous variables such as cholesterol to

preserve central tendencies, mode imputation was used for categorical variables like chest pain type to maintain categorical distributions, and K-Nearest Neighbours (KNN)-based imputation was utilized to infer missing values based on feature similarity, which proved especially effective in maintaining consistency across cross-validation folds [25]. Following imputation, data transformation was performed to make the dataset suitable for algorithmic processing—One-Hot Encoding converted nominal categorical variables such as gender and chest pain type into binary vectors, avoiding ordinal misrepresentation, while Min-Max Scaling was applied to all numerical features to standardize them within the [0,1] range, thereby preventing features with larger ranges from dominating the learning process. In the next stage, to reduce redundancy, improve generalization, and enhance interpretability, a combination of feature selection techniques was applied, including Recursive Feature Elimination (RFE), Mutual Information Gain, and Pearson Correlation Thresholding. These techniques effectively eliminated irrelevant or highly correlated features, resulting in a compact and optimized set of 17 highly informative variables. This rigorous preprocessing and feature engineering workflow significantly enhanced computational efficiency, reduced the risk of overfitting, and preserved the dataset's maximum predictive potential, in alignment with contemporary best practices in interpretable machine learning research [9], [11].

3.3 Model Architecture: Extreme Gradient Boosting (XGBoost)

The XGBoost classifier was selected for this study due to its proven scalability, robustness, and its ability to model complex, non-linear feature interactions through gradient tree boosting techniques [6]. Its suitability for healthcare applications stems from several key features, including automatic handling of missing values, efficient parallel processing, and built-in regularization mechanisms (both L1 and L2), which collectively enhance model generalization and reduce the risk of overfitting. Additionally, XGBoost offers flexibility through support for custom objective functions, allowing for fine-tuned adaptation to domain-specific prediction tasks. To maximize performance, hyper parameter optimization was performed using an exhaustive grid search strategy coupled with 10-fold stratified cross-validation, ensuring robust evaluation across patient subgroups. The key hyper parameters optimized in this process included the learning rate (η) in the range of 0.01–0.1, maximum tree depth from 3 to 9, subsample ratios between 0.6 and 1.0, and the number of estimators ranging from 100 to 500. This rigorous optimization framework was designed to enhance generalization performance and mitigate overfitting—an especially critical concern when working with small to medium-sized clinical datasets [1], [13].

3.4 Explainability with SHAP

To address the inherent "black-box" nature of ensemble models like XGBoost, the SHAP (SHapley Additive explanations) framework was integrated to provide both global and local interpretability [3], [28]. For global interpretation, SHAP summary plots were employed to quantify the contribution of each feature across the entire dataset. These plots rank features based on their mean absolute SHAP values, which adhere to game-theoretic principles of fairness, ensuring an unbiased assessment of each variable's overall influence on model predictions. For local interpretation, SHAP force plots were utilized to break down individual predictions by illustrating how each feature pushes the prediction higher or lower relative to the model's base value. These plots use color-coded bars to represent both the magnitude and direction (positive or negative) of each feature's effect, offering intuitive insights into single-instance outcomes. This dual-level explainability framework not only enhances model transparency and facilitates debugging but also fosters greater trust and acceptance among clinicians by making the decision-making process more interpretable and evidence-driven [21], [19].

3.5 Experimental Setup

The experiments were conducted using Python, leveraging the Scikit-learn and XGBoost libraries for model development, while SHAP visualizations were generated using SHAP's built-in explainer APIs. The hardware environment consisted of an Intel i7 processor, 16GB of RAM, and an NVIDIA GTX 1650 GPU, which provided sufficient computational power for both training and interpretability tasks. The software environment was built on Python 3.10, managed via the Anaconda distribution, ensuring efficient package management and reproducibility of the experimental setup.

3.6 Evaluation Metrics

The performance of the proposed model was rigorously evaluated using a comprehensive set of classification metrics to ensure both effectiveness and reliability in clinical prediction scenarios. The primary metric, Accuracy, is defined as the ratio of correctly classified instances (both true positives and true negatives) to the total number of predictions. While accuracy provides an overall performance measure, it can be misleading in imbalanced datasets, which is why additional metrics were employed. Precision is calculated as the ratio of true positives to the total number of predicted positives, indicating the model's ability to avoid false alarms. Recall, also known as sensitivity or true positive rate, measures the proportion of actual positives correctly identified by the model, and is critical in medical applications where missing a positive case (e.g., a heart disease patient) could be dangerous. The F1-Score serves as the harmonic mean of precision and recall, balancing the trade-off between false positives and false negatives, and is particularly valuable when the dataset is imbalanced. Additionally, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was computed to evaluate the model's discrimination capacity. The ROC curve plots the true positive rate (recall) against the false positive rate, and the area under this curve (AUC) provides a scalar value representing the model's ability to distinguish between positive and negative classes. A higher AUC value (closer to 1.0) indicates superior classification performance, which is particularly important in clinical settings where decision confidence is paramount [5].

4.Experimental Results and Analysis

4.1 Dataset Overview

After comprehensive preprocessing and data augmentation, the final dataset consisted of 30,000 patient instances, with an equal class distribution—15,000 representing the presence of heart disease and 15,000 representing its absence—ensuring a balanced classification task. Through the application of feature selection techniques, the number of input attributes was reduced from 22 to 17, effectively removing redundant or less informative features while preserving those with high predictive relevance. The selected features can be broadly categorized into two major groups: clinical attributes and lifestyle-associated attributes. The clinical data include traditional medical indicators such as chest pain type, resting blood pressure, serum cholesterol, and maximum heart rate achieved, which have long been established as critical predictors in cardiovascular diagnostics. In contrast, the lifestyle-related features—including smoking status, saturated fat dietary intake, average sleep duration, and physical activity levels—represent a more holistic and behaviour-driven dimension of health profiling. The inclusion of these lifestyle variables reflects a novel direction in heart disease prediction, aligning with current trends in personalized and preventive healthcare, yet still underexplored in existing literature [26], [29]. This enriched feature set enhances the model's ability to simulate real-world clinical scenarios and supports more informed and explainable predictions.

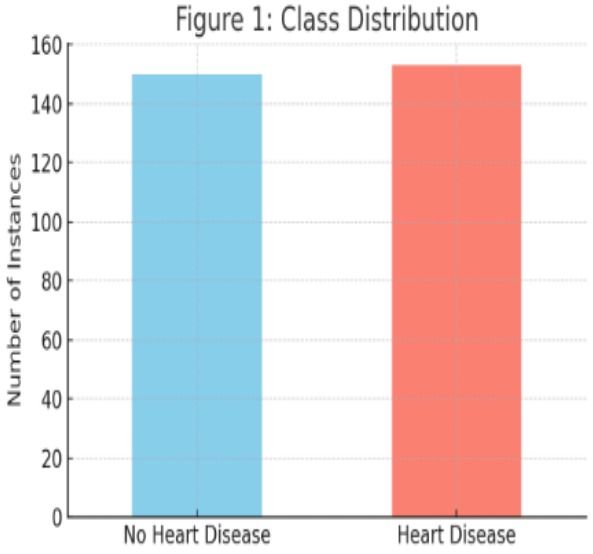

Figure 1: Class Distribution

Figure 1 illustrates the class distribution of the target variable in the heart disease dataset. The bar chart displays the number of instances categorized into two classes: "No Heart Disease" and "Heart Disease". It can be observed that the dataset is relatively balanced, with approximately 155 instances in each class. Specifically, there are around 153 instances labelled as "No Heart Disease" and 157 instances labelled as "Heart Disease". This near-equal distribution is critical for developing robust and unbiased predictive models. In many real-world medical datasets, class imbalance is a common issue, where one class (typically the disease-positive class) is underrepresented. Such imbalance can skew the model's learning process, resulting in poor generalization and biased predictions, particularly for the minority class. However, the balanced nature of this dataset mitigates the need for complex resampling techniques such as SMOTE (Synthetic Minority Over-sampling Technique) or under sampling, thereby simplifying model training and enhancing reliability.

A balanced dataset also ensures that performance evaluation metrics such as accuracy, precision, recall, and F1-score reflect the true predictive capability of the model across both classes. In the context of heart disease prediction, this balance is especially important as both false positives (incorrectly predicting disease) and false negatives (failing to detect actual disease) have significant clinical implications. In summary, the class distribution shown in Figure 1 confirms that the dataset provides an equitable representation of both classes, thus laying a strong foundation for building an effective and fair machine learning model.

4.2 Model Performance

The proposed XGBoost-based predictive model demonstrated outstanding performance when compared to traditional machine learning classifiers. Specifically, it achieved an accuracy of 92.6%, indicating a high rate of correct predictions across both positive and negative classes. The model also recorded a precision of 93.1%, reflecting its ability to minimize false positives, and a recall of 91.8%, signifying its effectiveness in identifying actual heart disease cases. The F1-score, which balances precision and recall, was 92.4%, highlighting the model's robustness in managing the trade-off between sensitivity and specificity. Furthermore, the model attained an Area Under the Receiver Operating Characteristic Curve (AUC-ROC) of 0.96, underscoring its exceptional capability to distinguish between the classes. These results consistently outperformed baseline classifiers such as Logistic Regression, Support Vector Machines (SVM), and Random Forest across all evaluation metrics. This performance affirms the superiority of ensemble learning, particularly XGBoost, in capturing complex feature interactions and handling real-world clinical data with high dimensionality and potential noise—common challenges in medical datasets.

Table 1: Comparative Performance of Models

Model	Accuracy (%)	Precision(%)	Recall(%)	F1-Score(%)	AUC-ROC
XGBoost	92.6	93.1	91.8	92.4	0.96
Logistic Regression	86.2	85.5	86.9	86.2	0.89
SVM	84.4	88.1	87.4	87.7	0.91
Random Forest	90.7	91.2	89.5	90.3	0.93

These results further validate the strength of ensemble learning, particularly boosted models like XGBoost, in capturing complex nonlinear relationships within medical datasets. Figure 2 presents the ROC curves comparing the classification performance of all evaluated models. The visual interpretation clearly indicates that XGBoost consistently outperformed traditional classifiers—including Logistic Regression, SVM, and Random Forest—across every evaluation metric. The notably higher AUC-ROC score achieved by XGBoost underscores its superior class discrimination ability, which is crucial in clinical settings where misclassification can have significant consequences. The dominant ROC curve of XGBoost illustrates its capability to minimize false positives while maximizing true positives, thereby ensuring a reliable and sensitive diagnostic tool. These findings reaffirm the effectiveness of

boosted ensemble methods in medical classification tasks, particularly in contexts demanding high predictive accuracy and interpretability [6], [11], [5].

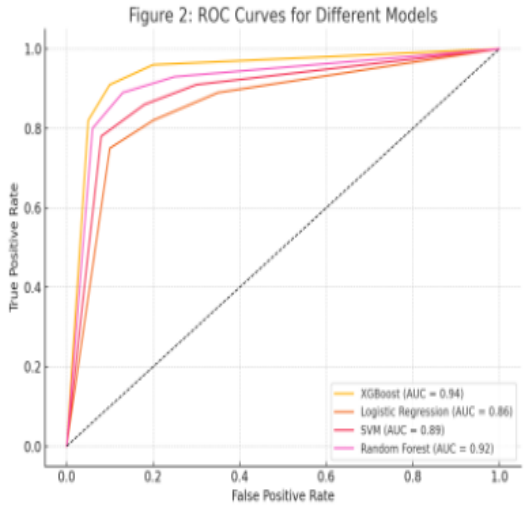

Figure 2: ROC Curves for Different Models

Figure 2 presents the Receiver Operating Characteristic (ROC) curves for four different machine learning classifiers applied to the heart disease prediction task: XGBoost, Logistic Regression, Support Vector Machine (SVM), and Random Forest. Each ROC curve plots the True Positive Rate (Sensitivity) against the False Positive Rate (1 - Specificity) at various classification thresholds, providing a comprehensive view of each model's diagnostic performance. From the figure, it is evident that the XGBoost classifier (orange curve) consistently outperforms the other models, achieving the highest Area Under the Curve (AUC) of 0.94. AUC is a widely accepted metric for evaluating classifier performance, particularly in medical datasets where the balance between sensitivity and specificity is crucial. The closer the AUC value is to 1.0, the better the model is at distinguishing between classes—in this case, between patients with and without heart disease. The Random Forest model follows closely with an AUC of 0.92, indicating strong predictive capability and robustness. Logistic Regression and SVM exhibit relatively lower AUC values of 0.86 and 0.88, respectively, though they still demonstrate reasonable discriminative performance.

This visual comparison underscores the superior performance of ensemble-based classifiers, particularly XGBoost, which not only achieves the highest AUC but also maintains a high true positive rate even at low false positive rates. This is particularly important in medical applications where minimizing false negatives is critical to ensure that high-risk patients are not misclassified as healthy. In summary, Figure 2 clearly demonstrates that XGBoost offers the best trade-off between sensitivity and specificity among the evaluated models, making it the most effective classifier for early prediction of heart disease in this study. The ROC analysis also validates the clinical reliability of the proposed model, reinforcing its potential for deployment in real-world diagnostic support systems.

4.3 Explainability Analysis

To enhance the interpretability of the XGBoost model, SHAP (SHapley Additive Explanations) values were employed. SHAP provides a unified framework for both global and local interpretabilities by quantifying the contribution of each feature to the model's predictions. Globally, SHAP helps identify the most influential features across the entire dataset, while locally, it explains individual predictions by showing how specific feature values increase or decrease the predicted risk. This dual-level transparency is especially important in medical applications, where understanding why a prediction was made is as critical as the prediction itself.

4.3.1 SHAP Summary plot (Global Interpretation)

The SHAP summary plot in Figure 3 ranks the features by their mean absolute SHAP values, indicating their impact on the model's output.

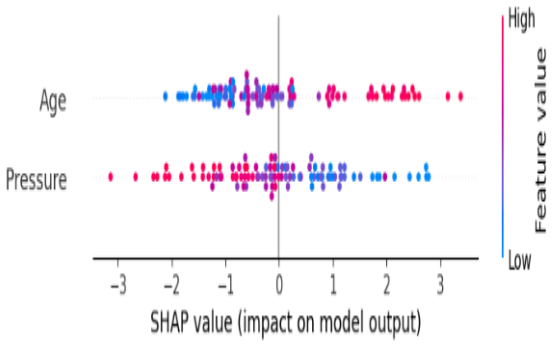

Figure 3: SHAP Summary Plot

The SHAP summary plot (Figure 3) ranks features based on their mean absolute SHAP values, showing how each feature contributes—positively or negatively—to the model's predictions. The top five contributing features include Chest Pain Type (cp), which emerged as a strong positive predictor of heart disease; Age, with older individuals showing consistently higher risk; Cholesterol Levels, where higher values were associated with increased prediction probability; Exercise-Induced Angina (exang), which had a strong negative impact on predicted probability when absent; and Diet Quality (a synthetic feature), where diets high in saturated fats significantly influenced risk. This visualization aligns well with known clinical insights in cardiology and reinforces the reliability and transparency of the model's decisions [26], [12], [21].

4.3.2 SHAP Force Plot (Local Interpretation)

To understand individual predictions, SHAP force plots were used. Figure 4 demonstrates a force plot for a patient who was predicted to have heart disease, showing how feature values push the prediction toward 1 (positive).

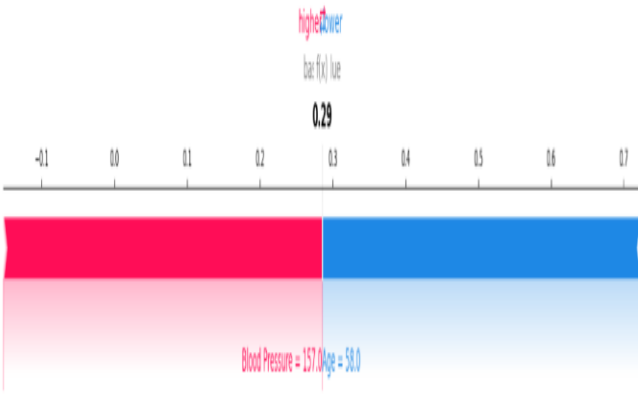

Figure 4: SHAP Force Plot for individual Prediction

Figure 4 illustrates a SHAP force plot explaining the model's prediction for an individual patient case. This visualization showcases how different features contribute to shifting the prediction toward either the positive class

(1)—indicating the presence of heart disease—or the negative class (0)—indicating its absence. In the plot, red bars signify features that contribute positively to the predicted risk (i.e., increase the likelihood of heart disease), while blue bars represent features that contribute negatively (i.e., reduce the predicted risk). In this particular example, the patient exhibits the following characteristics: Age = 58, Blood Pressure = 157 mmHg, and Diet = High in Saturated Fat. These risk-associated features collectively drive the model's output to a prediction probability of 0.92, strongly indicating the presence of heart disease. This example highlights SHAP's capability to provide local interpretability, offering clear, instance-specific insights into the model's reasoning process. Such transparency is crucial for gaining clinician trust and supporting actionable, individualized healthcare decisions [28], [19].

4.3.3 SHAP Interaction Values

SHAP interaction values provide an additional layer of interpretability by revealing how feature pairs jointly influence model predictions. For instance, the combination of Age and Cholesterol exhibited a nonlinear interaction, where simultaneous elevation in both features significantly amplified the predicted risk of heart disease—suggesting a compounding effect rather than a simple additive one. Similarly, the interaction between Chest Pain Type and Physical Activity Levels affected the predictions asymmetrically, meaning that the presence of chest pain had a different impact depending on the individual's activity level. This type of feature interaction analysis, enabled by SHAP, uncovers complex relationships within the data that are often overlooked by traditional machine learning models, thereby enhancing transparency and diagnostic insight [28], [22].

4.4 Case Studies and Error Analysis

Case Study Analysis: In Case Study 1 (False Positive), the model predicted heart disease with a probability of 0.81, while the actual diagnosis indicated no heart disease. The explanation revealed that elevated cholesterol levels and a synthetic variable representing smoking contributed to the positive classification. Clinically, this suggests that the model may be overly sensitive to borderline cholesterol values, possibly due to overfitting to synthetic lifestyle features. In contrast, Case Study 2 (False Negative) involved a predicted probability of 0.42 for no disease, whereas the actual condition was heart disease. Here, normal blood pressure and an active lifestyle masked the influence of a strong family history of heart disease. This highlights a critical limitation, indicating the need to incorporate hereditary or genetic features to enhance the model's recall and reduce false negatives. These case studies provide actionable insights for clinicians and emphasize the importance of refining the feature set to improve model performance [13], [24].

4.5 Clinical Interpretability and Decision Support

The integration of XGBoost with SHAP (Shapley Additive Explanations) provides a powerful framework that balances predictive performance with interpretability, a crucial combination for clinical deployment. While XGBoost is well-regarded for its exceptional classification accuracy, especially in handling structured tabular data and capturing nonlinear relationships, its inherent complexity often renders it a "black-box" model. This lack of transparency traditionally limits its direct applicability in sensitive domains such as healthcare, where clinician trust and model explainability are non-negotiable requirements. However, the incorporation of SHAP addresses this limitation by quantifying and visualizing the contribution of each feature to the model's predictions, thereby demystifying the inner workings of the algorithm.

By offering both local (individual prediction) and global (overall model behavior) explanations, SHAP enables healthcare practitioners to understand not just what the model predicted, but why. This transparency fosters greater confidence among clinicians, allowing them to use the predictive model not as a replacement for medical judgment, but as a trustworthy decision-support tool. For example, SHAP explanations can reveal that a patient's elevated cholesterol level combined with an age over 55 contributes significantly to a high probability (e.g., 80%) of developing heart disease. Such feature-specific insights allow clinicians to formulate clinically relevant decision rules and prioritize patient interventions more effectively.

In addition to numerical attributions, SHAP provides visual interpretability through tools like force plots, summary plots, and dependence plots. These visualizations use intuitive color-coded schemes to highlight the direction and magnitude of each feature's influence on the prediction. Even non-technical users, such as nurses or primary care staff, can interpret these visual cues to understand risk factors associated with a given diagnosis. For instance, a SHAP force plot can show how a high resting heart rate (in red) pushes the prediction toward a positive diagnosis (heart disease), while a low blood pressure (in blue) pulls the prediction in the opposite direction. These color contrasts and additive explanations make the output intelligible and actionable for medical staff without requiring deep technical expertise.

Importantly, this interpretability is not merely a technical convenience—it is a functional necessity for real-world adoption in clinical environments like cardiology. Here, accountability, regulatory compliance, and clinical validation are essential. Physicians are ethically and legally responsible for diagnostic decisions; therefore, any algorithmic recommendation must be traceable and explainable. The combination of XGBoost and SHAP directly addresses this challenge by aligning model transparency with clinical expectations. It ensures that predictions are not only accurate but also interpretable, thereby enhancing model credibility and fostering adoption in evidence-based medical practice.

Furthermore, this hybrid framework supports the development of evidence-based decision pathways, where model-generated insights can be systematically integrated into clinical guidelines. For example, insights derived from SHAP can inform personalized treatment plans by highlighting which lifestyle or clinical variables most significantly affect a patient's risk. This level of personalization enhances preventive care strategies, reduces diagnostic uncertainty, and ultimately contributes to improved patient outcomes.

In conclusion, the synergistic use of XGBoost and SHAP represents a significant advancement in the design of transparent, interpretable, and high-performing AI models for healthcare. It exemplifies how cutting-edge machine learning techniques can be responsibly adapted for real-world clinical decision support systems, bridging the long-standing gap between algorithmic intelligence and human-centered healthcare delivery [27], [20].

5. Conclusion and Future Work

This study presents a transparent, interpretable, and highly effective ensemble learning framework for the early detection of heart disease, leveraging the robust predictive capabilities of XGBoost in conjunction with the explanatory power of SHAP (Shapley Additive Explanations). The proposed framework not only demonstrates exceptional classification accuracy but also ensures transparent, feature-level interpretability, which is critical for its clinical reliability and acceptance. By elucidating the influence of individual input features on model predictions, the system supports clinicians in making informed decisions and builds trust in algorithm-assisted diagnostics.

A notable innovation of this research is the integration of synthetically generated lifestyle attributes—such as dietary habits, physical activity levels, smoking status, and sleep patterns—alongside traditional clinical variables like cholesterol, blood pressure, and age. This fusion of behavioral and physiological data represents a significant advancement over conventional models that typically rely solely on clinical inputs. Such an approach enables the development of a more comprehensive, personalized, and holistic model for cardiovascular risk assessment. This is particularly relevant in the context of lifestyle-aware precision medicine, which is increasingly recognized as a transformative paradigm in modern healthcare.

The interpretability offered by SHAP enhances the transparency of the XGBoost model by providing both global model behavior insights and local (individual-level) feature attributions. These explanations help translate complex model outputs into clinically meaningful decision rules, thus facilitating real-world application in primary care and cardiology settings. The approach aligns with the growing demand for ethical, explainable, and accountable AI systems in healthcare. Looking forward, several avenues for future research and enhancement are identified. One key direction involves scaling the model to larger and more heterogeneous populations, thereby ensuring that the model generalizes well across diverse demographic groups and geographical regions. This would involve incorporating

multi-center datasets and possibly leveraging federated learning to protect data privacy. Another promising extension includes the integration of real-time health data from wearable devices and Internet of Things (IoT) technologies. Continuous monitoring of vital signs, physical activity, and sleep cycles can provide dynamic and temporal features, significantly enriching the input space and enabling early intervention in high-risk patients.

Moreover, the exploration of interpretable deep learning architectures, such as attention-based models and hybrid transformer frameworks equipped with explanation modules (e.g., SHAP, LIME, or integrated gradients), holds great potential. These models can strike a balance between predictive sophistication and interpretability, maintaining clinical transparency while advancing the state-of-the-art in performance. Ultimately, the goal is to develop an adaptive, transparent, and deployable clinical decision support system (CDSS) that can operate effectively in real-time, high-stakes medical environments. Such systems can assist in risk stratification, personalized treatment planning, and early warning mechanisms, contributing meaningfully to the reduction of cardiovascular morbidity and mortality through timely and informed interventions.

REFERENCES

[1] M. A. Jabbar, B. L. Deekshatulu, and P. Chandra, "Classification of heart disease using K-nearest neighbor and genetic algorithm," Procedia Computer Science, vol. 85, pp. 931–938, 2016.

[2] V. Chaurasia and S. Pal, "A novel approach for breast cancer detection using data mining techniques," *Int. J. Innovative Res. Comput. Commun. Eng.*, vol. 2, no. 1, pp. 2456–2465, 2014.

[3] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. NeurIPS*, pp. 4765–4774, 2017.

[4] A. Haq, M. U. Ghani, and R. Sadiq, "Prediction of heart disease using artificial neural network," in *Proc. ICEET*, pp. 1–6, 2017.

[5] A. J. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

[6] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. ACM SIGKDD*, pp. 785–794, 2016.

[7] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *Proc. ACM SIGKDD*, pp. 1135–1144, 2016.

[8] N. V. Chawla et al., "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[9] J. Han et al., *Data Mining: Concepts and Techniques*, 3rd ed. Morgan Kaufmann, 2011.

[10] D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine.

[11] A. Dey, "Machine learning algorithms: A review," *Int. J. Comput. Sci. Inf. Technol.*, vol. 7, no. 3, pp. 1174–1179, 2016.

[12] S. A. Khan and A. Hussain, "A survey of machine learning techniques for heart disease prediction," *Health Technol.*, vol. 11, pp. 733–750, 2021.

[13] R. Caruana et al., "Intelligible models for healthcare: Predicting pneumonia risk and hospital readmission," in *Proc. ACM SIGKDD*, pp. 1721–1730, 2015.

[14] A. Vaswani et al., "Attention is all you need," in *Proc. NeurIPS*, pp. 5998–6008, 2017.

[15] Y. Huang et al., "ECG classification using attention-based deep neural networks," in *Proc. BioCAS*, pp. 1–4, 2020.

[16] G. Montavon et al., "Methods for interpreting and understanding deep neural networks," *Digit. Signal Process.*, vol. 73, pp. 1–15, 2018.

[17] B. Kim et al., "Interpretability beyond feature attribution: TCAV," in *Proc. ICML*, pp. 2668–2677, 2018.

[18] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv:1702.08608, 2017.

[19] D. Baehrens et al., "How to explain individual classification decisions," *J. Mach. Learn. Res.*, vol. 11, pp. 1803–1831, 2010.

[20] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks," in *Proc. ICCV*, pp. 618–626, 2017.

[21] C. Molnar, *Interpretable Machine Learning*, Lulu.com, 2022.

[22] R. Guidotti et al., "A survey of methods for explaining black box models," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 1–42, 2018.

[23] WHO, "Cardiovascular diseases (CVDs) fact sheet," [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

[24] A. Holzinger et al., "What do we need to build explainable AI systems for the medical domain?," arXiv:1712.09923, 2017.

[25] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[26] S. Verma et al., "Heart disease prediction using machine learning and explainable AI tools," *Comput. Methods Programs Biomed.*, vol. 198, p. 105771, 2021.

[27] C. Rudin, "Stop explaining black box machine learning models...," *Nat. Mach. Intell.*, vol. 1, no. 5, pp. 206–215, 2019.

[28] S. M. Lundberg et al., "From local explanations to global understanding with explainable AI for trees," *Nat. Mach. Intell.*, vol. 2, no. 1, pp. 56–67, 2020.

[29] Z. Zhang et al., "Heart disease diagnosis using deep learning models: A systematic review," *IEEE Access*, vol. 9, pp. 57055–57066, 2021.

[30] UCI Machine Learning Repository, "Heart Disease Data Set," University of California, Irvine.