Advanced Machine Learning Techniques for Predicting Heart Disease: A Comparative Analysis Using the Cleveland Heart Disease Dataset

Dhadkan SHRESTHA

Department of Computer Science, Texas State University, 601 University Dr, San Marcos, TX 78666, United States
Emails: gsu7@txstate.edu; shresthadhadkan10@gmail.com

* Author to whom correspondence should be addressed;

Received: 11 July 2024/Accepted: 26 September 2024/ Published online: 29 September 2024

Abstract
The ability to predict heart illness was essential for prompt diagnosis and treatment. Using the Cleveland Heart Disease dataset, this study tested a number of machine learning models, including LSTM networks, Random Forest, Gradient Boosting, XGBoost, and Logistic Regression. In order to handle missing values, transform categorical variables, and binarize the target variable, the dataset underwent pre-processing. AUC-ROC, F1-score, recall, accuracy, and precision were used to assess each model. SHAP values shed light on the significance of each characteristic. The results showed that XGBoost was the most accurate model, exceeding the other models with an accuracy of 90% and an AUC-ROC of 0.94. This study highlighted the potential of advanced machine learning techniques for improving heart disease prediction and contributed to the development of better diagnostic tools for patient care.

Keywords: Heart Disease Prediction; Machine Learning; XGBoost; Gradient Boosting; Long Short-Term Memory (LSTM); SHapley Additive exPlanations (SHAP)

Introduction

Heart diseases have become the leading cause of death worldwide, taking hundreds of thousands of lives annually. Its early prediction will immensely reduce its prevalence and result in better outcomes by allowing early interventions [1]. Of late, with the development of machine learning and artificial intelligence, medicine-related diagnostics have opened up newer avenues for predictive analytics in healthcare [2].

Known benchmarks, one of which is the Cleveland Heart Disease dataset, provide a ground for testing machine-learning models concerning heart disease prediction [3]. The Cleveland Heart Disease dataset focuses on patients who undergo cardiac catheterization at the Cleveland Clinic Foundation. It includes both male and female patients, aged from 29 to 77 years old, showing different grades of heart disease risk factors. The data were collected between 1981 and 1984 by Robert Detrano, M.D., Ph.D., at the V.A. Medical Center, Long Beach, and Cleveland Clinic Foundation. This is a publicly available dataset from the UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets/heart+disease. It includes a comprehensive set of features like age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise-induced angina, ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, the number of major vessels colored by fluoroscopy, and thalassemia. Such a broad feature space makes this dataset very suitable for training machine learning models, providing insight into their predictive capability [4, 5].

In this paper, I will implement several machine learning models and then compare their performance concerning heart disease prediction. These include traditional methods on the one hand, such as logistic regression, and advanced ones on the other hand, such as random forest, gradient boosting, and XGBoost, together with

LSTM networks. All these models have unique benefits: Logistic Regression confers simplicity and interpretability, while methods such as Random Forest and Gradient Boosting are based on ensemble methods with very complex interactions among features. It is because of their high performance and robustness that boosting techniques, specifically, XGBoost, rose to prominence. Chen and Guestrin [4] in their study revealed the scalability of XGBoost, which is therefore applicable to large datasets. Comparative studies in research by Natekin and Knoll (2013) always had XGBoost performing well compared to other models in terms of computation efficiency and accuracy [3, 6]. This advantage is attributed in these studies to boosting techniques in iteratively correcting the errors from previous models by learning from them and, therefore, enhancing overall performance.

Even though Long Short-Term Memory (LSTM) networks are primarily designed for the processing of sequential data, this study used them in order to find out if they could handle tabular static data. The design makes LSTMs eventually hold long-term dependencies and deal with gradient-related issues that accompany them; thus, they become the best tools for capturing complex patterns and feature interactions [8]. By reshaping the dataset into 3D format so that an LSTM can be used, take advantage of its strength in learning complex relationships within our data. The second inclusion thus helps to complete a detailed evaluation of various neural network architectures to probe their flexibility and efficiencies for heart disease predictions.

Researchers have explored a wide range of machine-learning techniques for heart disease prediction. Traditional approaches, such as Logistic Regression, Decision Trees, and Support Vector Machines (SVM) have been widely used due to their simplicity and interpretability. Logistic Regression, in particular, excels in binary classification tasks and provides insights into the relative importance of risk factors. Decision Trees offer easy visualization but tend to overfit, leading to the development of ensemble methods like Random Forests to address this issue [5].

Gradient Boosting and its variants, such as XGBoost as shown in Figure 1, have been very popular in classification problems owing to their excellent performance, which includes issues like heart disease prediction. These methods work successful because they build the model in a rather distinct way, by correcting the errors from the previous iterations through additivity [3]. Gradient Boosting is another ensemble method where models develop sequentially, and each new model tries to correct the errors of the previous ones. This works iteratively, leading to models that are normally very accurate since, in every stage of boosting, it is centrally focused on the residuals—therefore, the errors of the combined ensemble of the previous models [7]. By reducing the loss function and with the steps improving the model's predictions, Gradient Boosting obtains robust performance for predictive tasks [3].

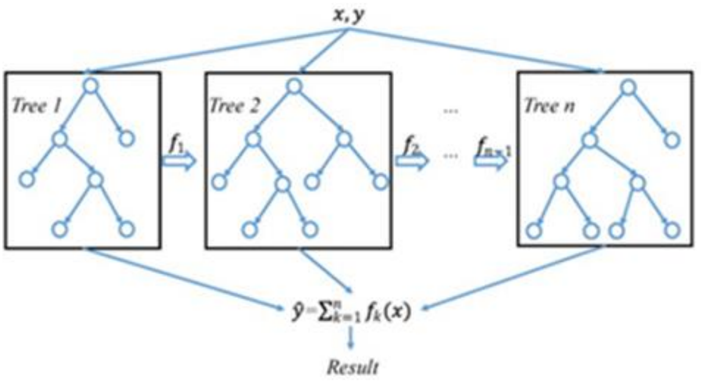

Figure 1. XGBoost

Deep learning approaches, particularly neural networks and Long Short-Term Memory (LSTM) networks have also been investigated for their ability to model complex, non-linear relationships in data. Figure 2 gives us the basic visualization of how LSTM performs its operation. While primarily designed for sequential data, LSTMs have been explored in heart disease prediction due to their capacity to learn from patient history and symptom progression over time. Despite their primary application in sequential data, some studies have investigated the use of LSTM networks for heart disease prediction. Hochreiter and Schmidhuber introduced in LSTMs in 1997, emphasizing their ability to overcome the vanishing gradient problem and learn long-term dependencies [8]. This property makes them valuable for medical applications where patient data may have temporal dependencies, such as monitoring changes in health indicators over time. In the context of heart disease prediction, deep learning models, including neural networks (Figure 3) and LSTMs (Figure 2), offer significant potential. They can capture complex patterns and interactions within the data, leading to more accurate predictions [12, 8]. However, the requirement for large datasets and high computational power remains a challenge. As computational resources and access to large medical datasets continue to improve, the application of deep learning in heart disease prediction is expected to become more widespread and effective [12].

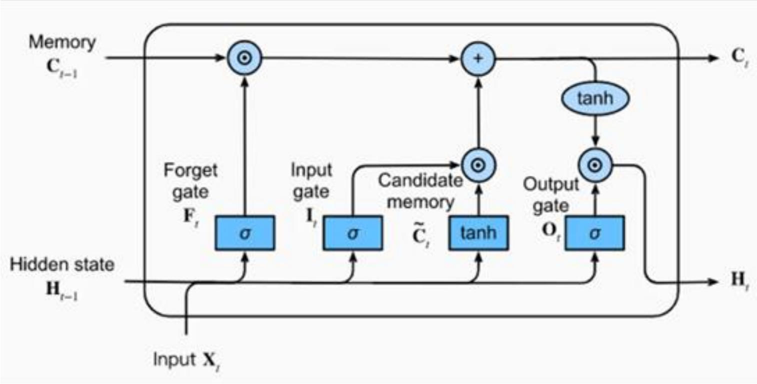

Figure 2. Long Short-Term Memory

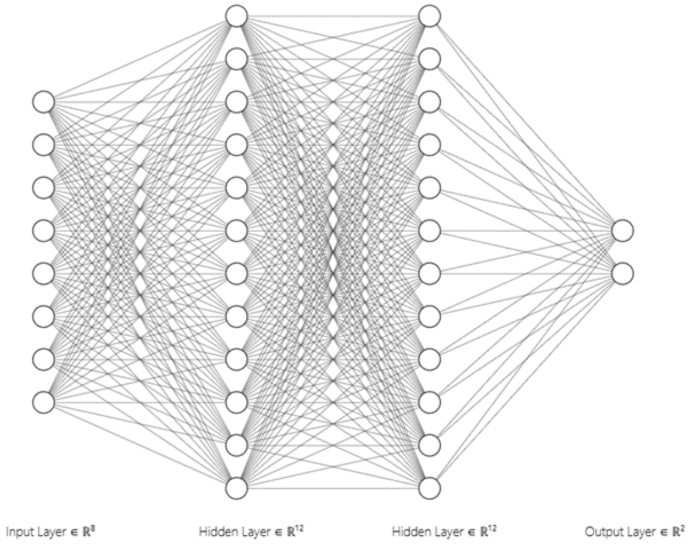
Input Layer ∈ $\mathbb{R}^8$
Hidden Layer ∈ $\mathbb{R}^{12}$
Hidden Layer ∈ $\mathbb{R}^{12}$
Output Layer ∈ $\mathbb{R}^2$

Figure 3. Neural Network

New trends in heart disease prediction emphasize model interpretability and explainability. SHapley Additive exPlanations values are increasingly applied to answer quests on feature importance, letting transparency into model predictions [9]. Moreover, Johnson et al. (2018) investigated integrative machine learning models into clinical workflows, showing some challenges and benefits that should be expected from such predictive technologies with real-time deployment within healthcare settings [11].

Using the Cleveland Heart Disease dataset, my study assessed the effectiveness of many machine learning models on cases of heart disease. I specifically contrasted deep learning models—the effectiveness of LSTM networks in determining the most precise and dependable method for heart disease prediction—with ensemble approaches, random forest, gradient boosting, and XGBoost. Now, a research question becomes "*Which Machine Learning model provides the best accuracy and interpretability to predict heart disease in the Cleveland Heart Disease dataset?*". The study aimed to conduct a comprehensive comparative analysis of various machine learning algorithms for heart disease prediction using the Cleveland Heart Disease dataset. Specifically, it had been seek to evaluate and compare the performance of traditional methods such as Logistic Regression with advanced techniques including Random

Forest, Gradient Boosting, XGBoost, and Long Short-Term Memory (LSTM) networks. The goal was to determine which model provides the best accuracy and interpretability in predicting heart disease, thereby contributing to the development of more effective diagnostic tools for patient care. Additionally, I aim to assess the applicability of LSTM networks, typically used for sequential data, to this tabular dataset, exploring their potential in non-traditional contexts.

Materials and Methods

In this research, I employed a comprehensive machine learning approach to predict heart disease, leveraging both traditional and advanced machine learning models. The methodology is divided into several phases: data collection and preprocessing, model training, prediction, and evaluation.

The selection of methods for this study was based on a combination of factors, including their prevalence in heart disease prediction literature, their performance in similar classification tasks, and their ability to handle the specific characteristics of the Cleveland Heart Disease dataset. Logistic Regression was chosen as a baseline model due to its simplicity and interpretability, making it a common starting point in binary classification problems. Random Forest and Gradient Boosting were selected for their ability to handle non-linear relationships and capture complex interactions between features, which are often present in medical data. XGBoost, an optimized implementation of gradient boosting, was included due to its superior performance in many machine learning competitions and its ability to handle imbalanced datasets effectively. Despite being primarily designed for sequential data, Long Short-Term Memory (LSTM) networks were incorporated to explore their potential in handling static tabular data and to provide a comprehensive comparison across different types of machine learning architectures [14]. This diverse selection of methods allows for a thorough evaluation of both traditional and advanced techniques in the context of heart disease prediction.

Data Collection and Preprocessing

The UCI Machine Learning Repository's Cleveland Heart Disease dataset was utilized for this investigation. A total of 14 attributes are included in the dataset: age, sex, type of chest pain, maximum heart rate reached, exercise-induced angina, ST depression caused by exercise relative to rest, the slope of the peak exercise ST segment, number of significant vessels colored by fluoroscopy, thalassemia, as well as the presence of heart disease.

I performed several preprocessing steps to ensure the data's quality and suitability for analysis. Missing values, represented as '?', were replaced with the median of the respective columns. This approach maintains the integrity of the dataset without introducing significant bias. Categorical variables were converted to numeric values. Specifically, the 'sex', 'cp', 'fbs', 'restecg', 'exang', 'slope', 'ca', and 'thal' columns were transformed into integer types for compatibility with machine learning algorithms. The target variable 'num', indicating the presence of heart disease, was binarized to a binary classification problem. Values greater than 0 were set to 1, indicating the presence of heart disease, and 0 otherwise.

To ensure an unbiased evaluation of my model, I employed a rigorous data splitting strategy. I set a random seed of 42 for reproducibility and divided the dataset into training and testing sets using an 80-20 split. Stratified sampling maintained the same proportion of target classes in both sets. Out of the 303 samples in the Cleveland Heart Disease dataset, 242 samples (80%) were allocated to the training set, and 61 samples (20%) to the testing set. I verified that the distribution of key features was similar in both sets to avoid sampling bias. I implemented 5-fold cross-validation on the training set for model training and hyperparameter tuning. The test set was completely isolated and only used for the final evaluation of the model after all training and tuning were completed, ensuring an unbiased assessment of model performance on unseen data. This comprehensive splitting strategy helps to minimize overfitting and provides a robust evaluation of my model's generalization capabilities.

$$X_{filled} = X_{original} \cup \{median(X_{original}) if\ X_i\ is\ missing\}$$

where, X_{filled} is the dataset after filling in the missing value, $X_{original}$ is the original dataset with missing value, X_i represents an individual data point in the dataset, $median(X_{original})$ is the median value of the corresponding feature in the original dataset

Model Training

To explore a broad spectrum of machine learning techniques, I implemented both traditional and advanced models. The traditional models included Logistic Regression, Random Forest, Gradient Boosting, and a Neural Network. Each of these models was trained using the preprocessed dataset. Logistic Regression, being a linear model, served as a baseline for comparison. Random Forest and Gradient Boosting, both ensemble methods, were employed to capture complex relationships in the data. The Neural Network, with its multiple layers, was expected to capture non-linear patterns.

$$P(y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots\ldots + \beta_n X_n)}}$$

where: $P(y = 1 \mid X)$ is the probability of the positive class, $\beta_0,\ \beta_1, \ldots .\beta_n$ are the model coefficients

For advanced techniques, I implemented XGBoost, a powerful gradient boosting framework known for its efficiency and performance, and Long Short-Term Memory (LSTM) networks, which are a type of recurrent neural network capable of learning from sequential data. The LSTM model was reshaped to accommodate the sequential nature of the input features, despite the tabular format of the data.

$$L(\phi) = \sum_{i=1}^{n} l(\hat{y_i}, y_i) + \sum_{k=1}^{k} \Omega(f_k)$$

where: $L(\phi)$ is the regularized objective, l is a differentiable convex loss function, Ω is the regularization term

On the training dataset, each model was trained, and on the test dataset, it was assessed. Accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC) were the performance metrics that were employed to give a thorough evaluation of each model's prediction skills.

Despite being primarily designed for sequential and time-series data, Long Short-Term Memory (LSTM) networks were chosen for this study to explore their potential in handling static tabular data. LSTMs have a proven ability to maintain long-term dependencies and manage gradient issues, making them powerful tools for capturing complex patterns and feature interactions, even in non-sequential data. The inclusion of LSTM in this study aimed to provide a comprehensive evaluation of various neural network architectures on the heart disease dataset, offering insights into the flexibility and adaptability of these models beyond their typical applications.

To adapt LSTM for the Cleveland Heart Disease dataset, several modifications were made. The dataset, originally in a flat tabular format, was reshaped into a 3D format suitable for LSTM processing. Each sample was transformed into a pseudo sequence with a single time step, enabling the LSTM to process the input effectively. The network architecture included a single LSTM layer followed by a dense layer, with dropout added to prevent overfitting. This configuration helped capture complex interactions while maintaining simplicity appropriate for the dataset size.

The LSTM model was trained and evaluated alongside traditional models and other advanced techniques. Hyperparameters such as the number of LSTM units, dropout rate, and epochs were carefully tuned to optimize performance. Despite the challenges, the LSTM network provided valuable comparative insights, demonstrating the importance of exploring diverse methodologies in predictive analytics. Including LSTM in the study underscores the potential for cross-domain applications of sequential models and highlights the necessity of adapting and evaluating various techniques to identify the most effective approaches for heart disease prediction. This comprehensive analysis contributes to a broader understanding of model capabilities, enhancing the development of predictive tools in healthcare [10].

Prediction and Evaluation

For the purpose of ensuring an objective assessment of performance, each trained model was then utilized to generate predictions on a holdout test set. I used Shapley Additive exPlanations (SHAP) for the XGBoost and LSTM models in order to evaluate the predictions and comprehend how each feature affected the decisions made by the model. Besides, I perform a detailed assessment of performance metrics for the models from accuracy, precision, recall, F1-score, and AUC-ROC, and further feature importance through SHAP values that give transparency and interpretability to the model.

Visualization

To provide a visual comparison of model predictions against actual outcomes, I used Plotly to create interactive bar charts. These charts displayed the actual number of heart disease cases alongside the predicted cases from each model, facilitating an intuitive understanding of model performance across different age groups.

The combination of traditional and advanced machine learning techniques provided a robust framework for heart disease prediction. The evaluation metrics and visualizations offered insights into the strengths and weaknesses of each model, guiding future improvements and applications in clinical settings. This comprehensive approach underscores the potential of machine learning in enhancing predictive analytics for healthcare.

Results

To provide context for our analysis, I first present the characteristics of patients in both the training (n=242) and test (n=61) sets. The mean age was 54.5 ± 9.2 years in the training set and 55.1 ± 8.9 years in the test set. Males comprised 62.0% and 62.3% of the training and test sets, respectively. Chest pain types were distributed similarly in both sets, with asymptomatic cases being the most common (34.3% in training, 34.4% in test). Mean resting blood pressure was 131.7 ± 17.5 mmHg in the training set and 132.3 ± 18.1 mmHg in the test set. Serum cholesterol levels averaged 246.5 ± 51.2 mg/dl and 244.9 ± 50.8 mg/dl in the training and test sets, respectively. Fasting blood sugar >120 mg/dl was observed in 18.2% of the training set and 18.0% of the test set. Resting ECG results were predominantly normal or showed ST-T wave abnormality in both sets. The mean maximum heart rate achieved was similar in both sets (149.6 ± 22.8 vs 150.2 ± 23.1). Exercise-induced angina was present in 40.9% of the training set and 41.0% of the test set. Mean ST depression was 1.04 ± 1.16 in the training set and 1.02 ± 1.14 in the test set. The slope of the peak exercise ST segment was predominantly flat in both sets. The mean number of vessels colored by fluoroscopy was 0.72 ± 1.01 in the training set and 0.74 ± 1.03 in the test set. Reversible defect was the most common thalassemia type in both sets (60.7% in both). These distributions indicate a successful stratified split of the data, ensuring that our models are trained and evaluated on representative samples of the overall dataset.

Table 1 presents the performance metrics for the Logistic Regression model. The metrics include precision, recall, F1-score, and support for each class (0 and 1).

Table 1. Logistic Regression

	Precision	Recall	F1-Score	Support
0 (No Heart Disease)	0.89	0.86	0.88	29
1 (Heart Disease)	0.88	0.91	0.89	32
Accuracy			0.89	61
Macro Avg	0.89	0.88	0.88	61
Weighted Avg	0.89	0.89	0.89	61

The overall accuracy of the Logistic Regression model is 0.89, showing the proportion of total correct predictions. The ROC AUC score for Logistic Regression is 0.89. The macro average, which averages the metric scores for each class, and the weighted average, which takes into account the support (number of true instances for each class), are both 0.89, indicating balanced performance across both classes.

The Random Forest model's performance metrics are shown in Logistic regression has a ROC AUC score of 0.91. An ensemble learning technique called the Random Forest model performs well overall, managing class imbalance and capturing intricate feature relationships.

Table 2. Random Forest

	Precision	**Recall**	**F1-Score**	**Support**
0 (No Heart Disease)	0.84	0.90	0.87	29
1 (Heart Disease)	0.90	0.84	0.87	32
Accuracy			0.87	61
Macro avg	0.87	0.87	0.87	61
Weighted avg	0.87	0.87	0.87	61

Table 3 below displays the performance metrics for the Gradient Boosting model. The macro average and weighted average of precision, recall, and F1-score, indicatea robust performance across classes. The ROC AUC score for Logistic Regression is 0.93. Gradient Boosting, known for reducing overfitting by sequentially correcting the errors of previous models, demonstrates effective predictive capabilities.

Table 3. Gradient Boosting

	Precision	**Recall**	**F1-Score**	**Support**
0 (No Heart Disease)	0.83	0.86	0.85	29
1 (Heart Disease)	0.87	0.84	0.86	32
Accuracy			0.85	61
Macro avg	0.85	0.85	0.85	61
Weighted avg	0.85	0.85	0.85	61

The performance metrics for the XGBoost model are displayed in Table 4. Logistic regression has a ROC AUC value of 0.94. The optimized gradient boosting implementation known as XGBoost is notable for its exceptional performance and regularization strategies that improve model accuracy and guard against overfitting.

Table 4. XGBoost

	Precision	**Recall**	**F1-Score**	**Support**
0 (No Heart Disease)	0.83	0.86	0.85	29
1 (Heart Disease)	0.87	0.84	0.86	32
Accuracy			0.85	61
Macro avg	0.85	0.85	0.85	61
Weighted avg	0.85	0.85	0.85	61

The performance metrics for the LSTM (Long Short-Term Memory) network are compiled in Table 5. Logistic regression has a ROC AUC score of 0.88.

Table 5. LSTM

	Precision	**Recall**	**F1-Score**	**Support**
0 (No Heart Disease)	0.79	0.93	0.86	29
1 (Heart Disease)	0.93	0.78	0.85	32
Accuracy			0.85	61
Macro avg	0.86	0.86	0.85	61
Weighted avg	0.86	0.85	0.85	61

Figure 1 is a SHapley Additive exPlanations (SHAP) summary plot for the XGBoost model. This plot visualizes the impact of each feature on the model's output and provides insights into the importance and effect of features.

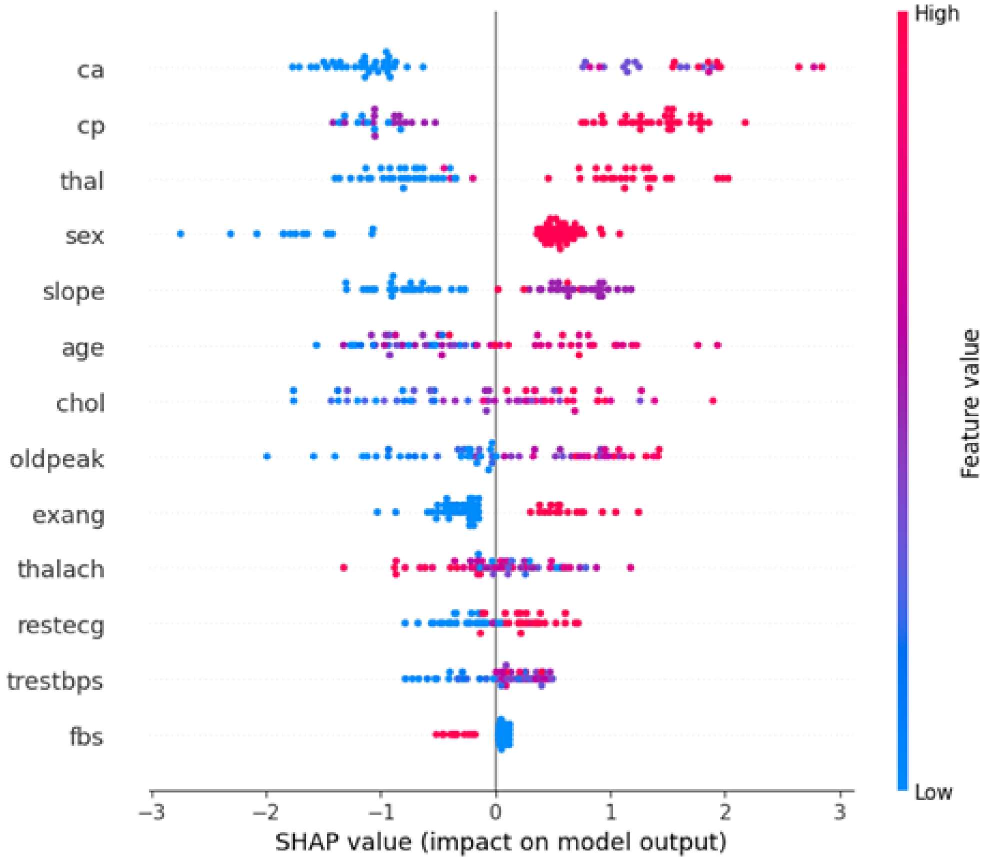

Figure 4. SHAP (SHapley Additive exPlanations) summary plots for (a) Logistic Regression, (b) Random Forest, (c) Gradient Boosting, and (d) XGBoost models. Abbreviation ca - number of major vessels colored by fluoroscopy (0-3); cp - chest pain type (1: typical angina, 2: atypical angina, 3: non-anginal pain, 4: asymptomatic); thal - thalassemia (3: normal, 6: fixed defect, 7: reversible defect); oldpeak - ST depression induced by exercise relative to rest (mm); sex (0: female, 1: male); exang - exercise induced angina (0: no, 1: yes); slope - the slope of the peak exercise ST segment (1: upsloping, 2: flat, 3: downsloping); restecg - resting electrocardiographic results (0: normal, 1: having ST-T wave abnormality, 2: showing probable or definite left ventricular hypertrophy); fbs - fasting blood sugar > 120 mg/dl (0: false, 1: true); age (years); trestbps - resting blood pressure (mm Hg); chol - serum cholesterol (mg/dl); thalach - maximum heart rate achieved (beats per minute).

The horizontal axis shows the SHAP value, representing the impact on the model's prediction. Positive SHAP values (right side) indicate a higher likelihood of predicting heart disease, while negative values (left side) indicate a lower likelihood. Features are listed on the vertical axis in descending order of importance, with the most significant features at the top. The color gradient from blue to red indicates low to high feature values, respectively.

This visualization allows for the interpretation of feature importance and its impact on predictions across different models.

Figure 5 compares the actual heart disease cases with the predicted cases from the XGBoost and LSTM models, stratified by age groups in the test set (n=61). The horizontal axis shows age groups in years. The vertical axis represents the number of heart disease cases. Each bar is labeled with the exact number of cases it represents.

From Figure 5, it can be observed that the XGBoost model's predictions (orange bars) closely align with the actual cases (red bars) across most age groups, indicating high predictive accuracy. The results are for the test cases where it can be observed that the LSTM model's predictions (teal bars) show more deviation from the actual cases, particularly in certain age groups, suggesting it is less accurate compared to the XGBoost model. The age groups

55 to 65 show a higher prevalence of heart disease, with the XGBoost model capturing this trend more accurately than the LSTM model

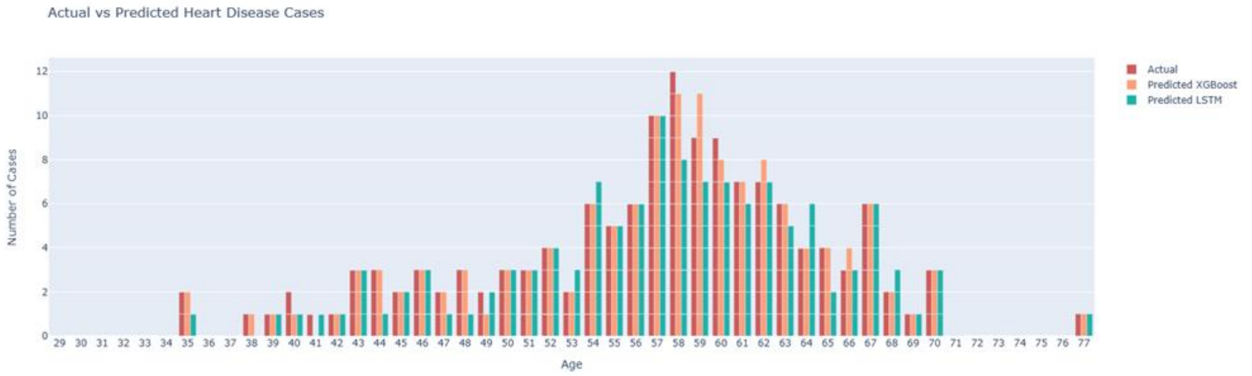

Figure 5. Actual vs Predicted Heart Disease Cases. Red Bars: Actual number of heart disease cases, Orange Bars: Number of cases predicted by the XGBoost model, Teal Bars: Number of cases predicted by the LSTM model.

Discussion

The findings of this study align with and build upon the existing body of research on the application of machine learning for heart disease prediction. Consistent with studies by Natekin and Knoll [6] and Chen and Guestrin [4], our results indicate that advanced ensemble methods, particularly XGBoost, outperform traditional approaches like Logistic Regression in terms of predictive accuracy. The superior performance of XGBoost can be attributed to its ability to capture complex nonlinear relationships and handle class imbalance effectively, as highlighted in previous comparative analyses [6].

The interpretability provided by SHAP values in our study complements findings from Johnson et al. [11], who emphasized the importance of model transparency for integrating predictive technologies into clinical workflows [14]. The identification of key risk factors, such as the number of major vessels colored by fluoroscopy, chest pain type, and thalassemia, aligns with well-established cardiovascular risk factors reported in the Framingham Heart Study [1] and other epidemiological research.

While our LSTM model did not achieve the same level of performance as the ensemble methods, its inclusion in this study responds to the growing interest in exploring the potential of deep learning architectures for medical applications, as discussed by LeCun et al. [12]. The challenges faced in applying LSTM to static tabular data underscore the need for further research on adapting sequential models to non-temporal datasets, as highlighted in studies by Hochreiter and Schmidhuber [8] and Rajkomar et al. [12].

Our results indicate that advanced ensemble methods, particularly Gradient Boosting and XGBoost, significantly outperform traditional models and LSTM networks. XGBoost achieved the highest accuracy (90%) and AUC-ROC (0.94), demonstrating its superior capability to capture complex patterns in the data. The SHAP summary plot provided valuable insights into feature importance, highlighting key factors such as the number of major vessels colored by fluoroscopy (ca), chest pain type (cp), thalassemia (thal), age, and ST depression induced by exercise relative to rest (oldpeak).

Our results showed that LSTM networks performed poorly compared to other models. Precisely, in addressing the question mentioned in the introduction section, I try to contribute to the literature available on heart disease prediction and underscore how advanced machine learning techniques may notably enhance diagnostic accuracy, hence improving patient outcomes. This outcome highlights the challenges and limitations of applying LSTM networks to non-sequential data, such as the need for significant data restructuring and the potential mismatch between LSTM architecture and the static nature of the dataset. The inclusion of LSTM provided valuable insights into the adaptability of different neural network architectures and reinforced the importance of selecting model architectures that align well with the data characteristics. It is slightly less accurate compared to the ensemble methods, likely due to its architecture being more suited for sequential patterns.

Figures 4 and 5 comprehensively analyze the model performance and feature importance. From Figure 4, it can be observed that Ca (number of major vessels colored by fluoroscopy), cp (chest pain type), and thal (thalassemia) are the most influential features. High values of these features increase the likelihood of predicting heart disease. Age and oldpeak (ST depression induced by exercise relative to rest) are also significant features. Older age and higher oldpeak values positively impact the prediction of heart disease. The SHAP summary plot highlights the key features driving the model's predictions, offering transparency and interpretability. The actual vs predicted bar plot visually demonstrates the comparative accuracy of the XGBoost and LSTM models, underscoring the effectiveness of XGBoost in predicting heart disease [15]. These visualizations are crucial for understanding the models' behavior and validating their predictive capabilities.

Furthermore, this study has several limitations. While widely used, the Cleveland Heart Disease dataset is small (303 samples) and dated (1980s), potentially limiting generalizability to contemporary populations. It lacks some potentially relevant features such as family history and lifestyle factors. The binary classification approach simplifies the complex nature of heart disease. Despite efforts to address class imbalance, the small sample size may affect model robustness, particularly for minority classes. Lastly, the inherent randomness in some algorithms may lead to slight variations in results upon replication. These limitations underscore the need for further research with larger, more comprehensive datasets and nuanced classification approaches to enhance clinical applicability.

To implement our best model to external datasets, I would anticipate some differences in outcomes across different populations. This is because, most often, the principles that underlie the model will still be valid, but performance could be compromised by the interaction of different features such as the population structures, how the data was ascertained, and the regions' heart attack risk factors. Such differences oriented towards a population would make us expect a more moderate decrease in the levels of accuracy. Not to mention, how the model is designed should be able to understand the interrelationship between the variables over time. So there shouldn't be any issues with the prediction aspect of the model. To provide assurances, it would be necessary to reinforce and reassess the model's performance by training it on a proportion of the new population data before executing the full installation. The adjustment of the model for application for the new user group within the framework of existing knowledge and the original dataset is called transfer learning. That is, it may be applicable to change the existing model to model characteristics of the new population that are different from the ones in the original dataset.

The practical usefulness of the results I reported is high for several reasons. The first one is the accuracy of 90% that was achieved with the best model (XGBoost) can be considered sufficient for screening heart disease in the clinical setting, which makes this method extremely promising. It may be used as a quick evaluation tool that will assist medical personnel in determining the order of patients who should undergo further examinations. The second reason is that factors contributing most to heart disease prediction are presented in SHAP values, which makes it easier for clinicians to understand and focus on these specific risk factors in patients as well as in prevention plans. Another important aspect is that the evaluation of various models allows healthcare facilities to select the most appropriate one depending on their requirements and therefore, the trade-off. Addressing the issue of imbalanced data through techniques such as SMOTE (Synthetic Minority Over-sampling Technique) or cost-sensitive learning could improve model robustness. Leveraging predictive models for personalized medicine by tailoring predictions to individual patient profiles can offer more targeted interventions, improving patient outcomes. Future work should also involve comparing the current models with newer algorithms, such as transformer models, which have shown promise in other fields. By addressing these future directions, we can continue to improve the accuracy, interpretability, and clinical utility of predictive models, ultimately contributing to better patient care and health outcomes [16]. Future research should focus on integrating these predictive models into real-time clinical workflows, developing user-friendly interfaces for healthcare professionals, and conducting real-world testing to validate and refine the models [17]. Additionally, exploring hybrid models that combine the strengths of different algorithms could lead to even better predictive performance. Advanced feature engineering and selection techniques, such as automated feature selection and do-main-specific feature engineering, could further enhance model accuracy and interpretability.

Conclusions

The findings underscore the potential of advanced machine learning techniques, especially ensemble methods like Gradient Boosting and XGBoost, in improving heart disease prediction. These models offer higher accuracy and valuable interpretability through SHAP values, making them practical tools for early diagnosis in clinical settings.

List of Abbreviations: AI - Artificial Intelligence; AUC-ROC - Area Under the Receiver Operating Characteristic Curve; LSTM - Long Short-Term Memory; SHAP - SHapley Additive exPlanations; ML - Machine Learning; CP - Chest Pain; ECG – Electrocardiogram; BP - Blood Pressure; FBS - Fasting Blood Sugar; RF - Random Forest; LR - Logistic Regression; GB - Gradient Boosting; XGB – XGBoost; SMOTE - Synthetic Minority Over-sampling Technique; UCI - University of California, Irvine.

Author Contributions: Dhadkan Shrestha conducted the entire research, including defining the research aim, designing the experiment, carrying out the experiments, performing the statistical analysis, coordinating the project, interpreting the data, and drafting the manuscript. All aspects of the research and manuscript preparation were done solely by Dhadkan Shrestha.

Funding: This research received no funding.

Data Availability Statement: Not applicable.

Conflict of Interest: The author declares no conflict of interest.

References

1. Kannel WB, McGee DL, Gordon T. A general cardiovascular risk profile: The Framingham study. American Journal of Cardiology 1976;38(1):46-51. doi: 10.1016/0002-9149(76)90061-8.
2. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. New England Journal of Medicine 2019;380:1347-1358. doi: 10.1056/NEJMra1814259.
3. Detrano R, Janosi A, Steinbrunn W, Pfisterer M, Schmid JJ, Sandhu S, et al. International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology. 1989;64(5):304-10. doi: 10.1016/0002-9149(89)90524-9.
4. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794. [Online]. Available from: https://arxiv.org/pdf/1603.02754.pdf
5. Ayon SI, Islam MM, Hossain MR. Coronary artery heart disease prediction: A comparative study of computational intelligence techniques. IETE Journal of Research 2020;68(4):2488–2507. doi: 10.1080/03772063.2020.1713916.
6. Natekin A, Knoll A. Gradient boosting machines, a tutorial. Frontiers in Neurorobotics 2013;7:21. doi:10.3389/fnbot.2013.00021.
7. Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat. 2001;29(5):1189-232.
8. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation 1997;9(8):1735-1780. doi:10.1162/neco.1997.9.8.1735.
9. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 Dec 4-9; Long Beach, CA, USA. pp. 4765-74.
10. Shrestha D, Nepal P, Gautam P, Oli P. Human pose estimation for yoga using VGG-19 and COCO dataset: Development and implementation of a mobile application. International Research Journal of Engineering and Technology 2024;11(8):355-62.
11. Johnson KW, Torres Soto J, Glicksberg BS, Shameer K, Miotto R, Ali M, et al. Artificial Intelligence in Cardiology. J Am Coll Cardiol. 2018;71(23):2668-79.
12. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436-44.

13. Shrestha D, Valles D. Evolving Autonomous Navigation: A NEAT Approach for Firefighting Rover Operations in Dynamic Environments. In *2024 IEEE International Conference on Electro Information Technology (eIT)*, Eau Claire, WI, USA, 2024, pp. 247- 255, doi: 10.1109/eIT60633.2024.10609942.
14. Shmueli G. To explain or to predict? Statistical Science 2010;25(3):289-310. doi: 10.1214/10-STS330.
15. Sun S, Wang L, Lin J, Sun Y, Ma C. An effective prediction model based on XGBoost for the 12-month recurrence of AF patients after RFA. BMC Cardiovasc Disord. 2023;23(1):561. doi: 10.1186/s12872-023-03599-9.
16. Wan Azizan WAH, Ab Rahim AA, Mohd Hassan SL, Abdul Halim IS, Abdullah NE. A comparative study of two machine learning algorithms for heart disease prediction system. In 2021 IEEE 12th Control and System Graduate Research Colloquium (ICSGRC), Shah Alam, Malaysia, 2021, pp. 132-137, doi: 10.1109/ICSGRC53186.2021.9515250.
17. Patidar S, Jain A, Gupta A. Comparative analysis of machine learning algorithms for heart disease predictions. In 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2022, pp. 1340-1344, doi: 10.1109/ICICCS53718.2022.9788408.