computers

MDPI

Article

Application of Deep Learning for Heart Attack Prediction with Explainable Artificial Intelligence

Elias Dritsas * and Maria Trigka

Athena Research and Innovation Center, Industrial Systems Institute (ISI), 26504 Patras, Greece; trigka@isi.gr
* Correspondence: dritsas@isi.gr

Abstract: Heart disease remains a leading cause of mortality worldwide, and the timely and accurate prediction of heart attack is crucial yet challenging due to the complexity of the condition and the limitations of traditional diagnostic methods. These challenges include the need for resource-intensive diagnostics and the difficulty in interpreting complex predictive models in clinical settings. In this study, we apply and compare the performance of five well-known Deep Learning (DL) models, namely Multi-Layer Perceptron (MLP), Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), and a Hybrid model, to a heart attack prediction dataset. Each model was properly tuned and evaluated using accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC) as performance metrics. Additionally, by integrating an Explainable Artificial intelligence (XAI) technique, specifically Shapley Additive Explanations (SHAP), we enhance the interpretability of the predictions, making them actionable for healthcare professionals and thereby enhancing clinical applicability. The experimental results revealed that the Hybrid model prevailed, achieving the highest performance across all metrics. Specifically, the Hybrid model attained an accuracy of 91%, precision of 89%, recall of 90%, F1-score of 89%, and an AUC of 0.95. These results highlighted the Hybrid model's superior ability to predict heart attacks, attributed to its efficient handling of sequential data and long-term dependencies.

Keywords: heart attack; deep learning; XAI; SHAP; e-Health; prediction; data analysis

check for updates

Citation: Dritsas, E.; Trigka, M. Application of Deep Learning for Heart Attack Prediction with Explainable Artificial Intelligence. *Computers* **2024**, *13*, 244. https://doi.org/10.3390/computers13100244

Academic Editor: Paolo Bellavista

Received: 31 July 2024
Revised: 18 September 2024
Accepted: 23 September 2024
Published: 25 September 2024

1. Introduction

Heart disease remains the leading cause of mortality worldwide, with heart attacks being a predominant and severe manifestation of this disease. The timely and accurate prediction of heart attacks can significantly reduce mortality rates by enabling early intervention and appropriate medical care [1,2].

A heart attack, medically referred to as a myocardial infarction, occurs when the flow of blood to a part of the heart is blocked for a long enough time that part of the heart muscle is damaged or dies. This is most often caused by a buildup of fat, cholesterol, and other substances, which form a plaque in the coronary arteries that feed the heart (coronary artery disease). When a plaque in a heart artery breaks, a blood clot forms around the plaque. If the clot becomes large enough, it can mostly or completely block blood flow through a coronary artery [3,4].

If the blockage is not quickly resolved, the portion of the heart muscle fed by the artery begins to die. Healthy heart tissue is replaced with scar tissue, affecting the heart's ability to pump effectively and can lead to a lifetime of complications and possible heart failure. Symptoms of a heart attack can vary widely, from the classic sudden and intense chest pain to subtle symptoms that may include chest discomfort, palpitations, upper body pain, nausea, and shortness of breath. Immediate medical attention can be critical in restoring blood flow and minimizing the damage to heart tissue [5,6].

Advancements in medical treatment have improved the survival rates and quality of life for heart attack victims. Emergency interventions such as angioplasty and coronary artery bypass grafting can restore blood flow to the heart. Medications like thrombolytics can dissolve blood clots that obstruct coronary arteries. Post-heart attack treatments may involve cardiac rehabilitation, lifestyle adjustments, and medication regimes to manage risk factors and prevent subsequent attacks. Continuous patient education on recognizing symptoms and managing health effectively plays a crucial role in improving outcomes and preventing future incidents [7,8].

In recent years, the application of DL in medical diagnostics has shown tremendous potential, particularly in predicting critical conditions like heart attacks. However, one of the significant challenges in deploying these models in clinical practice is their interpretability. Healthcare professionals require not only accurate predictions but also a clear understanding of how these predictions are made to be trusted and acted on. XAI techniques have emerged as crucial tools in addressing this challenge by making the decision-making process of complex models transparent and interpretable [9–11]. Among these techniques, SHAP is prominent. SHAP values, rooted in cooperative game theory, provide a unified measure of feature importance, quantifying the impact of each feature on the model's output. By integrating SHAP into the prediction process, this study not only enhances the accuracy of heart attack predictions using advanced DL models but also ensures that these predictions are understandable and actionable for healthcare professionals [12,13].

The primary motivation for this work is to address the critical need for the accurate and timely prediction of heart attacks, which remain a leading cause of death globally. Traditional methods for diagnosing and predicting heart attacks, while effective, often require significant time and resources, and may not always provide the early warnings necessary to prevent adverse outcomes. By exploring the capabilities of DL models, we aim to harness the power of advanced computational techniques to improve predictive accuracy and efficiency. This study is driven by the potential to enhance clinical decision-making processes, reduce the burden on healthcare systems, and ultimately save lives by enabling earlier and more accurate detection of heart attack risks.

This work makes several significant contributions to the field of medical diagnostics and predictive analytics:

- We provide a comprehensive comparison of five prominent DL models, including MLP, CNN, RNN, LSTM, GRU, and a Hybrid model in the context of heart attack prediction. This analysis helps identify the strengths and weaknesses of each model, offering valuable insights into their suitability for medical applications.
- By assessing the models using a range of evaluation metrics, including accuracy, precision, recall, F1-score, and AUC, we ensure a thorough evaluation of their performance. This multifaceted approach provides a more holistic understanding of each model's effectiveness in predicting heart attacks.
- This study presents the key steps of the adopted methodology, including data preprocessing, optimal hyperparameters, and features importance measurement assuming an 80–20 train-test split, to ensure robust and reliable model evaluation.
- By incorporating SHAP, an XAI technique, the study provides transparent and interpretable insights into the decision-making processes of the DL models, making the predictions more understandable and trustworthy.
- The results of this study offer practical insights for healthcare professionals and researchers, highlighting which DL models may be most effective for heart attack prediction. These insights can guide future research and the development of predictive tools in clinical settings, contributing to improved patient outcomes.

This research article is structured as follows. In Section 2, we review related works pertinent to the subject. Section 3 details the dataset utilized and examines the methodology adopted. Section 4 presents the experimental results of the ML models and evaluates their performance. Finally, Section 5 summarizes our research findings and discusses potential future directions.

2. Related Work

In recent years, the application of ML and DL techniques in medical diagnostics has gained significant traction, particularly in predicting heart attacks. Numerous studies have explored various models and approaches to improve the accuracy and reliability of heart attack predictions, emphasizing the need for interpretable and clinically applicable solutions.

Firstly, the contribution of work [14] lies in the development of an automated heart disease prediction system using a Deep Neural Network (DNN) model. This proposed model addresses key anomalies in previous methods such as accuracy issues and the need for manual data pre-processing by implementing a fully automated data pre-processing stage. The model's effectiveness was tested on a dataset from the UCI repository, Cleveland, achieving a minimum accuracy of 87.64%, which outperforms other ML algorithms. By leveraging the complexity and feature extraction capabilities of DL, this method promises more reliable predictions and opens up opportunities for future enhancements in medical data analysis using DL techniques.

The proposed work [15] leverages CNN to enhance the accuracy of heart disease prediction using the Cleveland dataset. The CNN model is designed with a dense input layer of 128 neurons, and two hidden layers with specific pooling sizes, and uses the Adam optimizer to handle noise and sparse gradient issues. Experimental results demonstrate a high prediction accuracy of 75.2%, showing the effectiveness of CNN in this domain. This model significantly reduces the error metrics, making it a robust tool for the early detection of heart disease, potentially saving numerous lives by enabling timely medical intervention.

Moreover, work [16] integrates a collection of ML algorithms and DL techniques to predict heart disease with high accuracy. By utilizing algorithms such as a Random Forest (RF) classifier and a custom DL model, the study achieved significant accuracy improvements. The DL model demonstrated superior performance with an accuracy of 92.35%, precision of 90.84%, recall of 94.20%, and F1-score of 92.49% using oversampling to make the dataset balanced. A dataset stemming from the Behavioral Risk Factor Surveillance System (BRFSS) was utilized, keeping 19 variables that relate to lifestyle factors of a person that can contribute to being at risk with any form of cardiovascular disease.

A Deep CNN named CardioHelp effectively predicts the probability of cardiovascular disease [17]. This model outperforms existing methods by achieving a prediction accuracy of 94.78%, which is significantly higher than previous techniques. The CNN architecture leverages layers of convolution and pooling to handle temporal data modeling, enhancing the prediction at early stages. Experimental results demonstrate that CardioHelp not only improves the accuracy but also the overall performance in comparison to state-of-the-art methods, making it a robust tool for the early detection of heart disease.

Similarly, ref. [18] presents a hybrid model combining Support Vector Machine (SVM), K-Nearest Neighbors (K-NN), and Decision Trees (DT) to predict heart disease, achieving an accuracy of 95.5%. This model optimally integrates the strengths of each algorithm, thereby enhancing the prediction performance. A well-known heart disease dataset was used and carefully preprocessed to ensure reliability and validity. By implementing a layered approach, the model effectively minimizes prediction errors, providing a robust tool for clinical decision-making. The comparative analysis highlights its superiority over traditional models, demonstrating the significant potential for real-world medical applications.

Research work [19] employed a DNN with Talos hyper-parameter optimization (TO) to predict heart disease. The specific model achieved a classification accuracy of 90.78%, outperforming other algorithms such as K-NN with 90.16%, logistic regression with 85.25%, and SVM with 81.97%. This approach demonstrates the effectiveness of Talos optimization in enhancing prediction accuracy, making it a robust tool for the early detection of heart disease. The model's high accuracy can significantly aid in the timely intervention and treatment of patients.

Additionally, ref. [20] utilized a modified version of traditional Artificial Neural Networks (ANNs), the Deep Learning Modified Neural Network (DLMNN), enhanced with a Cuttlefish Optimization Algorithm (CFOA) for better accuracy. The Hungarian Heart Disease dataset from the UCI repository was selected as a benchmark for the model training. These data have the same features as Cleveland except for the attribute that captures the number of major vessels. The actual patient data collected through the IoT sensor network were utilized to recognize the disease's presence using the DLMNN, achieving an accuracy of 96.8%, along with high specificity 85%, a sensitivity of 98.13%, and F1-score of 98.3%), outperforming ANN.

The authors in [21] propose a DL-based prediction model and in particular ANN for cardiovascular disease prediction. The ANN model was designed with a 13-node input layer, a hidden layer of 4 nodes, and a final output layer, optimized using the Adam optimizer and binary cross-entropy loss function. The ANN reached an accuracy of 85.24%, outperforming (in terms of accuracy) traditional ML models such as SVM (81.97%) and DT (81.97%). The ANN's performance in terms of the rest metrics includes a precision, recall and F1-score of 0.85, demonstrating its robustness and effectiveness in predicting heart disease. Finally, ref. [22] introduces an efficient SMOTE-based DL model for heart attack prediction, achieving an accuracy of 96.1%. This model leverages the Synthetic Minority Oversampling Technique (SMOTE) to address data imbalance and employs an ANN without requiring extensive feature engineering. The ANN model showed superior performance with precision and recall values of 95.7% each, and an F1-score of 95.7%. This approach outperformed other machine learning models and existing systems, demonstrating high reliability and effectiveness in predicting heart attacks while minimizing computational costs.

Table 1 compares the proposed approach against related works, emphasizing the dataset, model and performance metric. The table highlights how the proposed approach stands out from existing models in a key metric AUC (which the other works did not assume) while also addressing the critical need for model transparency through the integration of SHAP. This comparison underscores the advantages of our method in both predictive power and clinical applicability, setting it apart from the models discussed in the related works. For a fair performance comparison between the prevailing models in each work and our study, the same features dataset, preprocessing, tuning, and training methods should have been applied. For example, the use of SMOTE highly favors the model's performance as demonstrated in [22], where the same dataset was used, but in our work, such a method was not applied so no straightforward comparisons can be made.

Table 1. Comparison of related works.

Ref.	Dataset	Number of Features	Model	Metric
[14]	Cleveland	14	DNN	Accuracy 87.64%
[15]	Cleveland	14	CNN	Accuracy 75.2%
[16]	BFRSS	19	DL and oversampling	Accuracy 92.35%, Precision 90.84%, Recall 94.20%, F1-score 92.49%
[17]	Cleveland	14	Deep CNN	Accuracy 94.78%
[18]	Cleveland	14	DL	Accuracy 94.2%, Recall 82.3%, Specificity 83.1%
[19]	Cleveland	14	DNN	Accuracy 90.78%
[20]	Hungarian	13	DLMNN	Accuracy 97%, Specificity 85%, Recall 98.13%, F1-score 98.3%
[21]	Cleveland	14	ANN	Accuracy 85.24%
[22]	Cleveland	14	SMOTE-based DL model	Accuracy 96.1%, F1-score 96.0%, Precision and Recall 96.1%
Proposed	Cleveland	14	Hybrid model (CNN, GRU) with SHAP	Accuracy 91%, Precision, 89%, Recall 90%, F1-score 89%, AUC 0.95

However, the distinct advantages over the abovementioned approaches are the following:

- By combining CNN and GRU, the model effectively captures both spatial and temporal features. Furthermore, the model's architecture is designed to handle complex patterns in sequential (medical) data, making it more robust against varied clinical scenarios (more details are provided in Section 3.3).
- The model is evaluated using multiple metrics—accuracy, precision, recall, F1-score, and AUC— ensuring a well-rounded assessment of its performance, unlike many studies that focus on a single metric.
- The integration of SHAP provides transparent insights into the model's decision-making process, addressing a critical need for interpretability in clinical settings, which most other models lack. By enhancing both performance and interpretability, the model is better suited for real-world clinical implementation, making it a more viable option for improving patient outcomes.

To summarize, these combined advantages make the proposed approach not only technically superior but also more aligned with the practical needs of healthcare professionals, setting it apart from other methods in the literature.

3. Materials and Methods

Initially, we emphasized dataset description and analysis. Then, we focused on the key methodological steps applied in this study: data preparation, model training, evaluation, and the application of the XAI technique. Each step was meticulously executed to ensure the reliability and robustness of the results.

3.1. Heart Attack Dataset Description and Analysis

The dataset [23] on which we relied comprises 14 attributes and 303 instances (Cleveland), including 6 categorical, 2 nominal (Yes/No, Male/Female), and 6 numerical attributes. A slightly modified version of this dataset was used by only excluding 6 rows with missing values; thus a dataset of 297 instances emerged. The obtained dataset's details are provided in Table 2.

The dataset used in this study was selected due to its balance of features and instances, which provides a manageable complexity for initial model development and comparison, while alternative dataset repositories [24–26] offer larger sample sizes (augmented datasets generated through instance duplication employing diverse machine learning classifiers) or different feature sets, the selected dataset was chosen to ensure a focused exploration of model performance with a moderate size feature set that includes critical clinical variables. This allowed for a detailed evaluation of model capabilities before scaling to larger, more complex datasets.

It includes subjects aged 29 to 77, with male patients represented by a gender value of Yes and female patients with a value of No. There are four types of chest pain recorded: typical (resulting from reduced blood flow due to narrowed coronary arteries), atypical, non-angina chest pain (arising from various causes and might not indicate heart disease), and asymptomatic angina which might not be related to heart disease as well. Additional attributes include RestBP (resting blood pressure), Chol (serum cholesterol level), FastBS (with values assigned as Yes if fasting blood sugar level is higher than 120 mg/dL and No if below), RestECG (resting electrocardiographic result), MaxHR (maximum heart rate achieved), ExIndAgina (exercise-induced angina, recorded as Yes if there is pain and No if not), STD (ST depression induced by exercise), slope (slope of the peak exercise ST segment), NMajV (number of major vessels colored by fluoroscopy), thal (a blood disorder called thalassemia), and heart attack (class attribute, with a value of No for normal and Yes for indicating the absence or presence of a heart attack).

In the following subsections, we separate the analysis of the features that capture the occurrence of heart attack into numerical and nominal. Concerning features with numerical values, we especially compare the two observed groups (Yes, No) providing a high-level graphical illustration of the minimum, 25% quartile, median (or 50% quartile), mean, 75% quartile and maximum information, also unveiling the two groups' data symmetry, skewness, variance, and outliers. As for the nominal features, their values (categorical data) prevalence per group are presented and properly discussed.

Table 2. Feature information in the dataset.

Attribute Name	Type	Description	Range of Values
Age	Numerical	Age of the person–years	29–77
Sex	Nominal	Gender of the person	Male, Female
ChestP	Categorical	Chest pain type (type of angina)	Typical, Atypical, Non-angina pain, Asymptomatic
RestBP	Numerical	Resting Blood Pressure—mm/Hg	94–200
Chol	Numerical	Serum cholesterol—mg/dL	126–564
FastBS	Nominal	Fasting Blood Sugar—mg/dL	No, Yes
RestECG	Categorical	Resting ECG Results	Normal, Hypertrophy, ST-T-wave abnormal
MaxHR	Numerical	Maximum Heart Rate Achieved—bpm	71–202
ExIndAng	Nominal	Exercise-Induced Angina	No, Yes
STDp	Numerical	ST depression induced by exercise relative to rest	0–6.20
Slope	Categorical	Slope of the Peak Exercise ST segment	Downsloping, Flat, Upsloping
NMajV	Numerical	Number of major vessels colored by fluoroscopy	0, 1, 2, 3
Thal	Categorical	Thalassemia disorder level	Normal, Fixed Defect, Reversible Defect
Heart Attack	Nominal	The target class variable indicates the presence or absence of a heart attack.	No, Yes

3.1.1. Heart Attack and Numerical Feature Distributions Using Box Plots

The first box plot from (Figure 1) concerns the feature "Age" for "No Heart Attack" (blue) and "Yes Heart Attack" (orange) classes. For the "No Heart Attack" class, the median age is 58 years, with an interquartile range (IQR) from 52 years to 62 years. The minimum age is 38 years, and the maximum age is 77 years, with a mean age of 56.70 years, as indicated by the cross mark. There is an outlier at 35 years, suggesting that a few individuals are significantly younger than the rest of the group. This indicates a relatively wide age range among those who did not experience a heart attack.

In contrast, the "Yes Heart Attack" class has a median age of 52 years, with an IQR from 44.50 years to 59 years. The ages range from 29 years to 76 years, with the mean age at 52.65 years, marked by the cross. This class does not have any significant outliers, indicating a more consistent age distribution compared to the "No Heart Attack" group. Overall, the data reveal that individuals without heart attacks tend to be older on average and have a wider age range compared to those who experienced heart attacks, who generally have a younger and more consistent age distribution.

The second box plot from (Figure 1) illustrates RestBP for the "No Heart Attack" and "Yes Heart Attack" classes. For the "No Heart Attack" class, the median RestBP value is 130 mm Hg, with an IQR from 120 mm Hg to 145 mm Hg. The minimum value is 100 mm Hg, and the maximum value is 180 mm Hg, with a mean RestBP of 134.42 mm Hg as indicated by the cross mark. There are outliers at 192 mm Hg and 200 mm Hg, suggesting a few individuals with significantly higher resting blood pressure values. This indicates a relatively wide spread of resting blood pressure values among those who did not experience a heart attack.

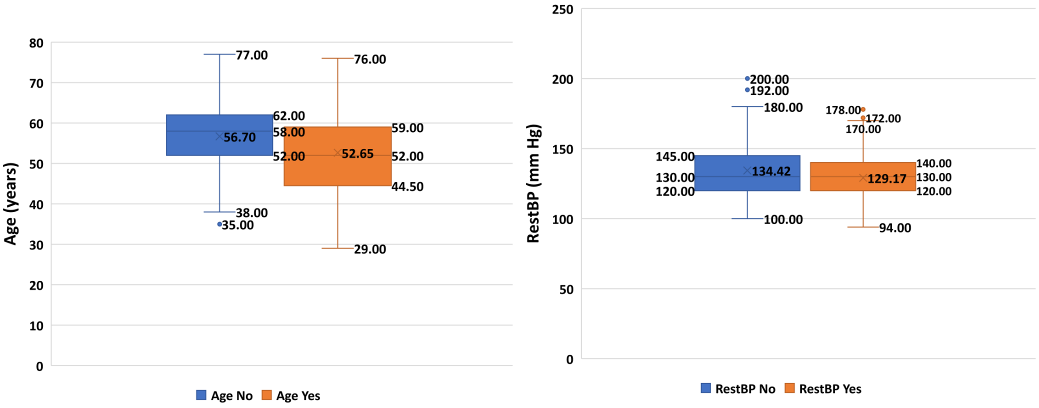

Figure 1. Box plot illustrating the age and resting blood pressure distribution within the dataset.

The "Yes Heart Attack" class has a median RestBP value of 130 mm Hg, with an IQR from 120 mm Hg to 140 mm Hg. The values range from 94 mm Hg to 170 mm Hg, with the mean RestBP at 129.17 mm Hg, marked by the cross. Outliers in this class are observed at 172 mm Hg and 178 mm Hg. This indicates that individuals in the heart attack group generally have similar resting blood pressure values on average, with fewer high outliers compared to the "No Heart Attack" group. Overall, the data reveal that individuals without heart attacks tend to have slightly higher and more variable resting blood pressure values compared to those who experienced heart attacks, who generally have a more consistent range of values with some extremely high outliers.

The first box plot from (Figure 2) shows "Cholesterol" levels for "No Heart Attack" (blue) and "Yes Heart Attack" classes. For the "No Heart Attack" class, the median cholesterol level is 249 mg/dL, with an IQR from 216.5 mg/dL to 283.5 mg/dL. The minimum value is 131 mg/dL, and the maximum value is 353 mg/dL, with a mean cholesterol level of 251.12 mg/dL, as indicated by the cross mark. There is an outlier at 407 mg/dL, suggesting a few individuals with significantly higher cholesterol values. This indicates a relatively widespread cholesterol level among those who did not experience a heart attack.

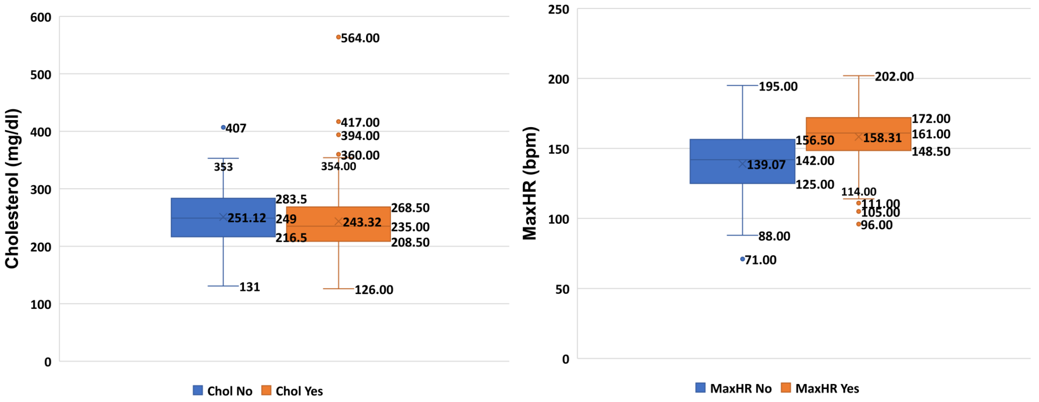

Figure 2. Box plot illustrating the cholesterol and maximum heart rate within the dataset.

The "Yes Heart Attack" class has a median cholesterol level of 235 mg/dL, with an IQR from 208.50 mg/dL to 268.50 mg/dL. The values range from 126 mg/dL to 354 mg/dL, with the mean cholesterol level at 243.32 mg/dL, marked by the cross. Outliers in this class are observed at 360 mg/dL, 394 mg/dL, 417 mg/dL, and a notably high value at 564 mg/dL. This indicates that individuals in the heart attack group generally have lower cholesterol levels on average, although there are a few with notably higher values. Overall, the data reveal that individuals without heart attacks tend to have higher and more consistent cholesterol values compared to those who experienced heart attacks, who generally have lower values with some extremely high outliers.

The second box plot from (Figure 2) represents the "MaxHR" for the "No Heart Attack" and "Yes Heart Attack" classes. For the "No Heart Attack" class, the median MaxHR value is 142 bpm, with an IQR from 125 bpm to 156.50 bpm. The minimum value is 88 bpm, and the maximum value is 195 bpm, with a mean MaxHR of 139.07 bpm as indicated by the cross mark. There is an outlier at 71 bpm, suggesting a few individuals with significantly lower MaxHR values. This indicates a relatively widespread of MaxHR values among those who did not experience a heart attack. Differently, the "Yes Heart Attack" class has a higher median MaxHR value of 161 bpm, with an IQR from 148.50 bpm to 172 bpm. The values range from 114 bpm to 202 bpm, with the mean MaxHR at 158.31 bpm, marked by the cross. Outliers in this class are observed at 96 bpm, 105 bpm, and 111 bpm. This indicates that individuals in the heart attack group generally have higher MaxHR values, although there are a few with notably lower values. Overall, the data reveal that individuals without heart attacks tend to have lower MaxHR values compared to those who experienced heart attacks, who generally have higher values with some notable low outliers.

The first box plot from (Figure 3) pertains to "STDp" for "No Heart Attack" and "Yes Heart Attack" classes. For the "No Heart Attack" class, the median ST depression value is 1.40, with an IQR from 0.60 to 2.55. The values range from 0.00 to 4.40, with mean ST depression at 1.60 as indicated by the cross mark. Notable outliers are present at 5.60 and 6.20, indicating a few individuals with significantly higher ST depression values. This suggests relatively widespread ST depression values among those who did not experience a heart attack. The "Yes Heart Attack" class has a lower median ST depression value of 0.00, with an IQR from 0.00 to 1.10. The values range from 0.00 to 2.60, with the mean ST depression at 0.60, marked by the cross. Outliers in this class are observed at 3.00, 3.50, and 4.20. This indicates that while most individuals in the heart attack group have low ST depression values, there are a few with much higher values. Overall, the data reveal that individuals without heart attacks tend to have higher and more variable ST depression values compared to those who experienced heart attacks, who generally have lower values with some notable high outliers.

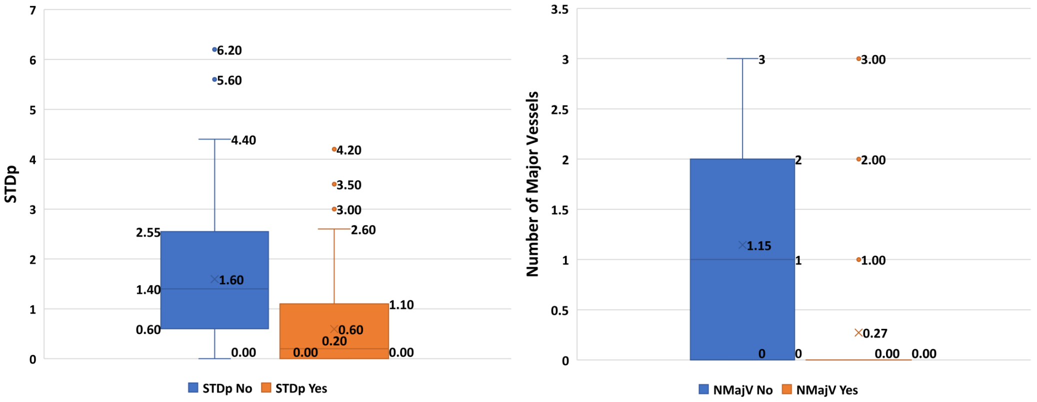

Figure 3. Box plot illustrating the ST depression induced by exercise relative to rest and number of major vessels within the dataset.

Finally, the second box plot from (Figure 3) illustrates the number of major vessels for the "No Heart Attack" and "Yes Heart Attack" classes. For the "No Heart Attack" class, the median number of major vessels is 0, with an IQR from 0 to 2. The minimum value is 0, and the maximum value is three major vessels. The mean number of major vessels, as indicated by the cross mark, is 1.15. There are no outliers present in this class, indicating that the data are relatively consistent within the observed range. In contrast, the "Yes Heart Attack" class shows a median of 0 major vessels and an IQR that also spans from 0 to 0, indicating no spread in the central values. The minimum and maximum values are both 0 major vessels. The mean number of major vessels is 0.27, as indicated by the cross mark. However, there are outliers at one, two, and three major vessels, suggesting some individuals in this class have more major vessels, but the overall distribution is concentrated around lower values. This figure highlights that individuals who had a heart attack generally had fewer major vessels compared to those who did not, with the latter group showing a wider and higher range of values.

3.1.2. Heart Attack Dataset Analysis

This subsection aims to provide a comprehensive analysis of the nominal features in relation to the target class. It will explore the distribution and potential correlation of these features, highlighting their value prevalence per state of the target class. By examining these aspects, we can gain deeper insights into how nominal features influence outcomes and inform more effective decision-making and predictive modeling.

Heart Attack Prevalence in Terms of Gender

Figure 4 illustrates the distribution of heart attack occurrences between genders. It shows that 24 females and 113 males did not experience a heart attack, while 72 females and 89 males did experience a heart attack. This comparison highlights the differences in heart attack occurrences between male and female patients. These data suggest that, in this sample, males have a higher overall count of heart attack cases compared to females. The disparity between genders might be influenced by various factors such as differences in lifestyle, genetic predispositions, hormonal influences, and healthcare access. Males might be at higher risk due to the higher prevalence of certain risk factors like smoking and hypertension [27].

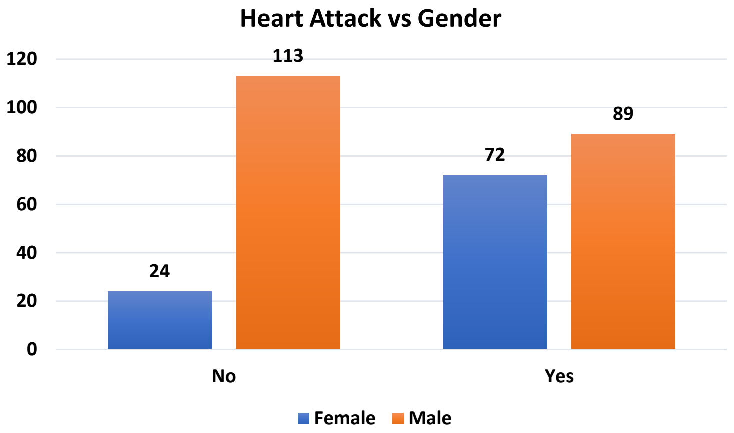

Figure 4. Heart attack prevalence among men (male) and women (female) (gender) within the dataset.

However, it is important to consider that females may experience heart attack symptoms differently, which can lead to under-diagnosis or misdiagnosis. Women are more likely to have atypical symptoms such as nausea or fatigue rather than the classic chest pain often seen in men. This can result in delays in seeking treatment, potentially affecting outcomes. Therefore, awareness and education about gender-specific symptoms are crucial in improving diagnosis and treatment for heart attacks in both men and women.

Heart Attack Prevalence in Terms of Chest Pain Type

Figure 5 provides insights into how different types of chest pain correlate with the occurrence of heart attacks. It reveals that among individuals who experienced heart attacks, those with typical and atypical angina were significantly more prevalent. Specifically, 66 cases involved atypical angina, and 40 cases involved typical angina, highlighting these as key indicators of heart attack risk. In contrast, fewer heart attacks were noted among those with non-anginal pain and asymptomatic individuals, indicating that these types of chest pain are less predictive of heart attacks.

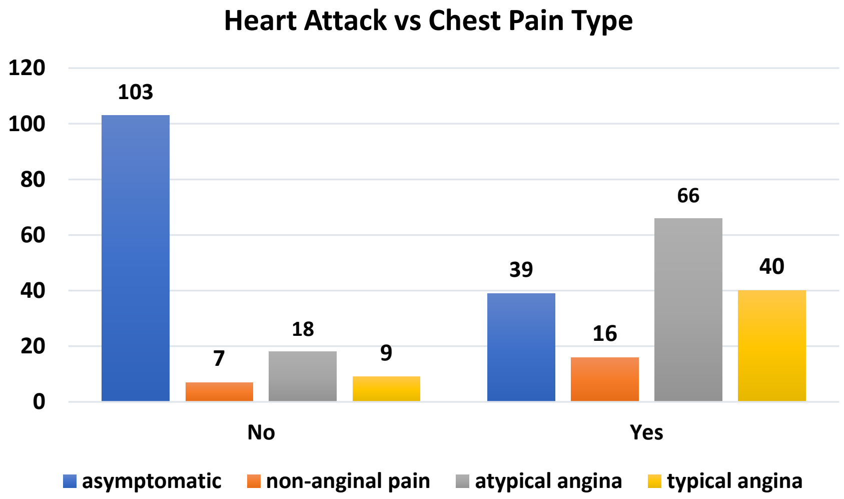

Figure 5. Prevalence of heart attack per chest pain type within the dataset.

Heart Attack Prevalence in Terms of Fasting Blood Sugar

Moreover, the bar chart in Figure 6 illustrates the relationship between the nominal variable FastBS (indicating whether fasting blood sugar is higher than 120 mg/dL) and the occurrence of heart attacks. It shows that among individuals with elevated FastBS (i.e., equal to Yes), the number of heart attack cases (23 subjects) is nearly equal to the number of non-heart attack cases (21 subjects). In contrast, among those with FastBS equal to No, there are more individuals with heart attacks (138 subjects) compared to those without heart attacks (116 subjects), highlighting a notable difference in heart attack incidence related to fasting blood sugar levels.

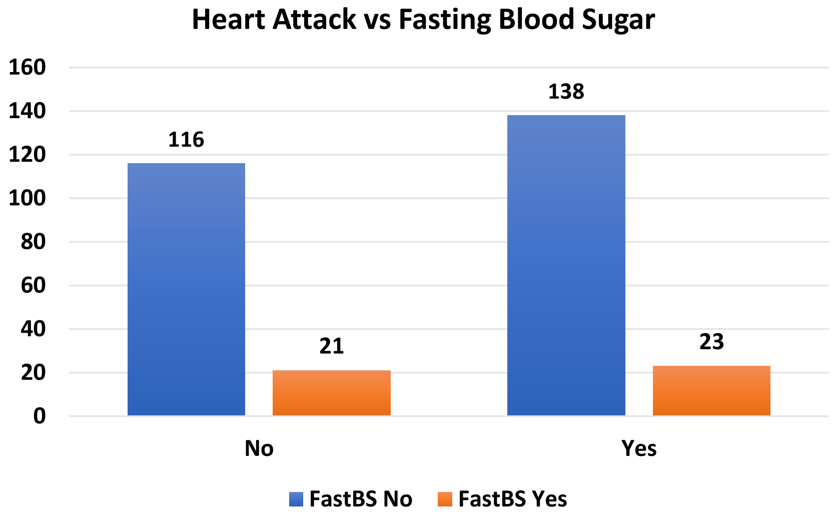

Figure 6. Heart attack prevalence per fasting blood sugar value within the dataset.

Heart Attack Prevalence in Terms of Exercise-Induced Angina

Focusing on Figure 7, as far as heart attack prevalence is concerned among those without exercise-induced angina, 138 out of 200 (69%) had a heart attack. Among those with exercise-induced angina, 23 out of 98 (23.47%) had a heart attack. It is observed that a significantly higher percentage of individuals without exercise-induced angina experienced heart attacks compared to those with exercise-induced angina. This could suggest that the presence of exercise-induced angina might be a protective factor or that those without it are at higher risk due to other underlying factors.

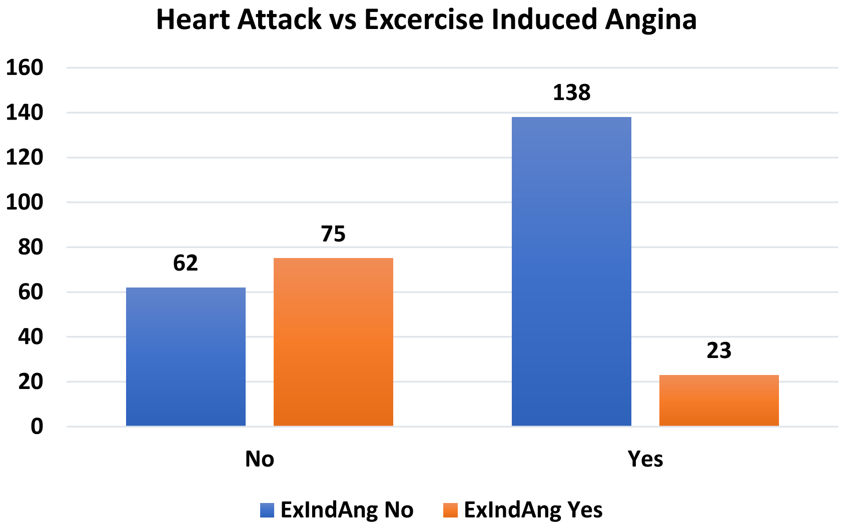

Figure 7. Prevalence of heart attack per exercise-induced angina type within the dataset.

Heart Attack Prevalence in Terms of Resting ECG

Figure 8 illustrates how heart attack prevalence varies among individuals with different resting ECG results. Individuals with normal resting ECG results show a high prevalence of heart attacks, with 62.16% experiencing a heart attack. This indicates that a significant portion of individuals with normal ECG results still experience heart attacks, highlighting the need for comprehensive risk assessment beyond ECG results. Individuals with hypertrophy on their resting ECG have a lower prevalence of heart attacks compared to those with normal ECG results, with 46.58% experiencing heart attacks. This suggests that while hypertrophy is associated with heart attacks, it may not be as strong an indicator as initially thought, or other factors may influence the risk. Individuals with ST-T wave abnormalities on their resting ECG show the lowest prevalence of heart attacks, with only 25% experiencing a heart attack. Despite the small sample size, this result indicates that ST-T wave abnormalities alone may not be a strong predictor of heart attacks in this dataset. The data indicate that individuals with normal ECG results and hypertrophy are at a higher risk for heart attacks compared to those with ST-T wave abnormalities.

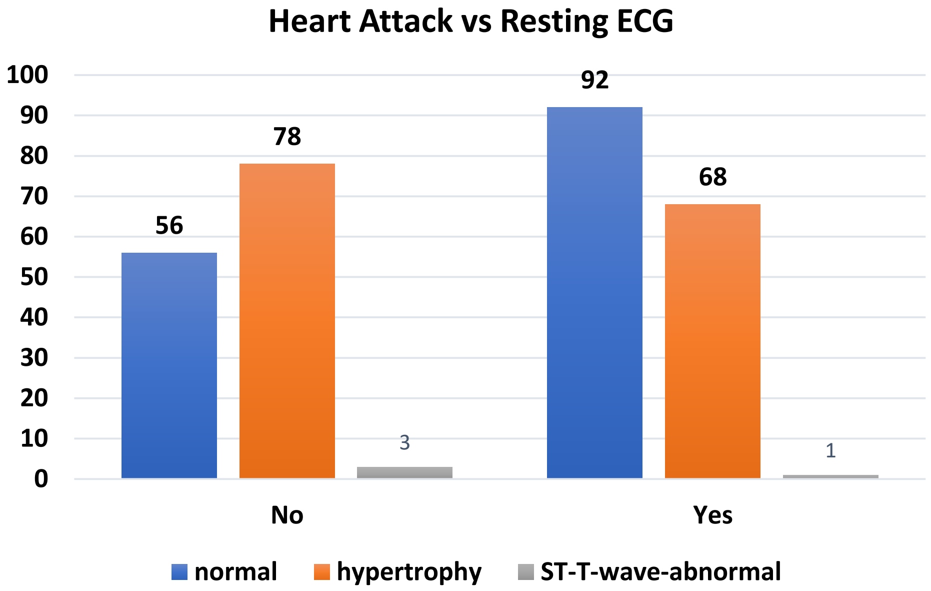

Figure 8. Prevalence of heart attack per resting ECG value within the dataset.

Heart Attack Prevalence in Terms of Sloping

Figure 9 illustrates a clear variation in heart attack prevalence based on ECG sloping types. Upsloping ECG results are strongly associated with a high risk of heart attacks, while downsloping and flat-sloping results show moderate and lower associations, respectively.

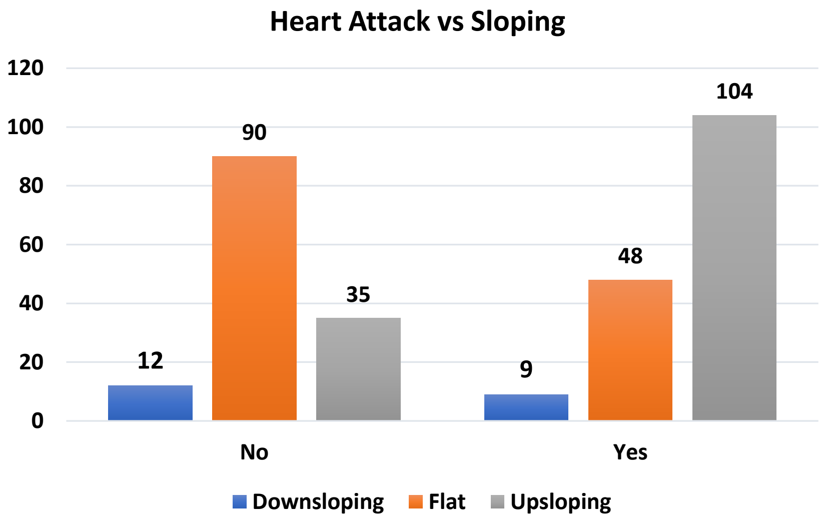

Figure 9. Prevalence of heart attack per sloping type within the dataset.

Individuals with downsloping ECG results have a heart attack prevalence of 42.86%. The majority (57.14%) of individuals with downsloping do not experience heart attacks, indicating a moderate association with heart attack risk.

Individuals with flat-sloping ECG results show a heart attack prevalence of 34.78%. A significant portion (65.22%) of individuals with flat sloping do not experience heart attacks, suggesting that flat sloping is less strongly associated with heart attack risk compared to other sloping types.

Individuals with upsloping ECG results have the highest prevalence of heart attacks, with 74.82% experiencing a heart attack. This indicates a strong association between upsloping ECG results and heart attack risk, and that individuals are at the highest risk for heart attacks, highlighting the importance of identifying and monitoring these individuals closely. Downsloping and flat-sloping ECG results are associated with lower heart attack risks, but they still warrant attention, especially in the presence of other risk factors.

Investigating the underlying mechanisms that contribute to the varying heart attack prevalences among different sloping types could lead to improved predictive models and treatment strategies.

Heart Attack Prevalence in Terms of Thalassemia

Observing Figure 10, individuals with normal thalassemia have the highest prevalence of heart attack, with 77.58% of individuals experiencing a heart attack. This suggests a strong association between normal thalassemia and the occurrence of heart attacks. The high prevalence of heart attacks among individuals with normal thalassemia highlights the need for careful cardiovascular risk assessment and management in this group. Regular monitoring and preventive measures may be crucial for individuals with normal thalassemia to reduce the risk of heart attack.

Individuals with fixed defect thalassemia have a heart attack prevalence of 33.33%, which is lower than that of normal thalassemia but higher than reversible defect thalassemia. This suggests that fixed defects have an intermediate level of association with heart attacks, and fall into an intermediate risk category, indicating that they require a balanced approach in terms of monitoring and preventive care.

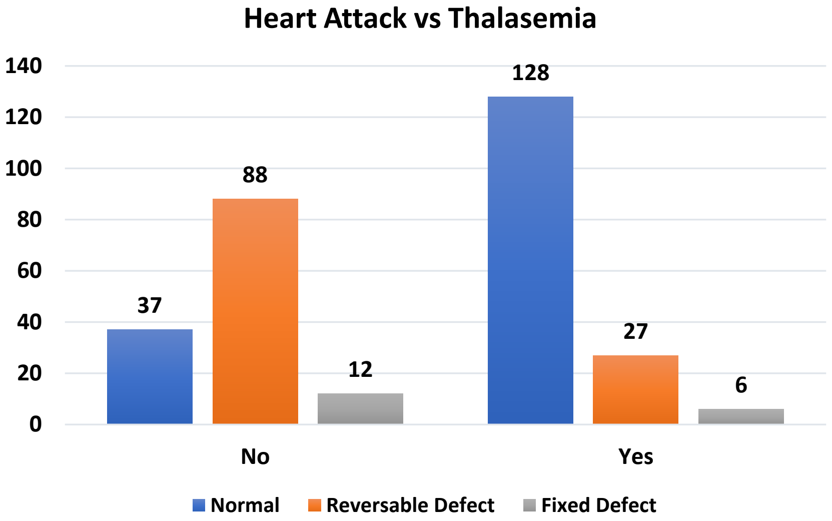

Figure 10. Prevalence of heart attack per thalassemia type within the dataset.

Individuals with reversible defect thalassemia have a significantly lower prevalence of heart attacks, with only 23.48% experiencing a heart attack. This indicates that they may be less associated with heart attacks compared to normal thalassemia. The lower prevalence of heart attacks in individuals with reversible defect thalassemia suggests that there might be protective factors associated with reversible defects. Understanding these factors could be beneficial in developing strategies to mitigate heart attack risks in other thalassemia conditions.

As a summary of the analysis of the above features, while the dataset provided some preliminary insights into differences among the features in heart attack incidence, (e.g., gender differences with a higher overall number of heart attack cases among males compared to females), the sample size was limited to generalize our conclusions. The complex interaction between multiple factors, such as age, cholesterol levels, and chest pain type, and how they influence heart attack risk, is also evident but requires further exploration. Due to this limitation, future studies with larger and more diverse samples are needed to fully understand potential biases, and feature correlations and confirm any of the observed (e.g., gender-based) differences. A more detailed analysis of how these factors interact with each other could add significant depth to our understanding of heart attack risks and will be a focus of future research.

3.2. Methodology

The comprehensive analysis of numerical features in Section 3.1 provided insights into data outliers and the necessity of normalization, particularly in ensuring that skewed distributions were properly scaled to prevent model bias. This approach stands in contrast to other studies [14,15,17–22] that either neglected or placed less emphasis on distributional analysis before preprocessing. Our findings underscore the importance of examining feature statistics and distributions in detail, highlighting that such careful consideration is essential for optimizing model performance, especially when working with small- to medium-sized datasets. The applied methodology involved the following tasks:

- The first step in data preprocessing was to check for feature columns with missing values and duplicate rows. It was observed that 6 out of the 303 records in the dataset had one empty-value feature. In the context of the current study, we selected to remove these records and not apply any data-imputation method (e.g., with the mean value of the numerical feature, or the mode of the categorical feature). Furthermore, there were no duplicate rows.
- Categorical variables were then encoded using one-hot encoding to transform them into a format suitable for machine learning algorithms. This process involved creating

binary columns for each category within a categorical feature, ensuring that the models could interpret and utilize this information effectively.

- Numerical features were normalized using min–max scaling to bring all features into the same range [0, 1], which helps accelerate the convergence of gradient-based optimization algorithms used in training the models.

Figure 11 illustrates a comprehensive workflow for developing and evaluating DL models for heart attack prediction. It begins with the collection and preprocessing of heart attack data, followed by splitting the dataset into training and testing sets. Various DL models are then built and trained using the training data. The performance of these models is evaluated on the test set using metrics such as accuracy, precision, recall, and AUC. To enhance interpretability, an XAI technique was utilized to interpret and understand the predictions made by the DL models. The primary method used in this study was SHAP. This technique is crucial for making the predictions of complex models (e.g., DL models) transparent and interpretable, especially in critical fields like healthcare. SHAP values are computed to understand the impact of different features on the model's predictions. A feedback loop allows for continuous improvement by feeding insights gained from the evaluation and SHAP analysis back into the initial data block for further data collection, enhancement, and model retraining. This iterative process ensures the development of robust and interpretable models for heart attack prediction.

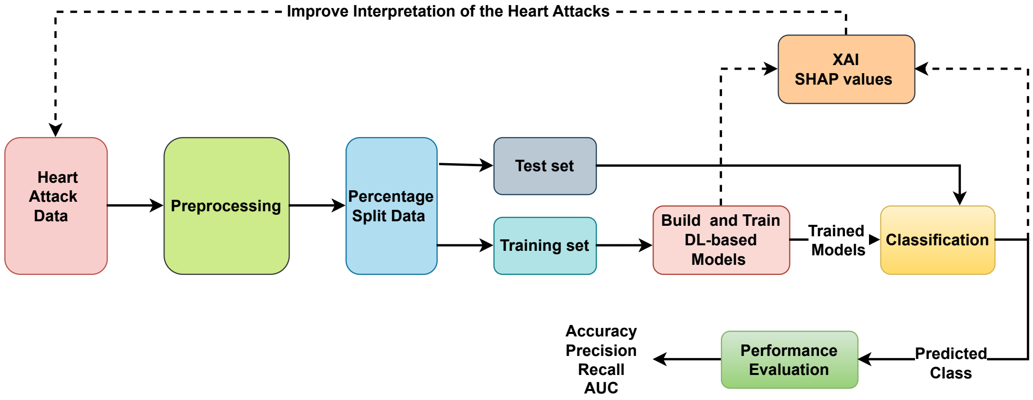

Figure 11. Workflow for heart attack prediction using DL, including preprocessing, model training, evaluation, SHAP interpretation, and iterative improvement.

SHAP values provide a unified measure of feature importance by connecting game theory with local explanations. The SHAP method assigns each feature an importance value for a particular prediction. The foundation of SHAP values is based on Shapley values from cooperative game theory, which fairly distribute the "payout" among features (or players) based on their contribution to the prediction. For a model M and a specific instance x, the SHAP value ϕ_i for feature i is calculated as follows:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} [f_x(S \cup \{i\}) - f_x(S)],$$

where N is the set of all features, S is a subset of N not containing feature i, $|S|$ is the number of elements in subset S, and $f_x(S)$ is the model's prediction for instance x with only the features in subset S. The term $\frac{|S|!(|N|-|S|-1)!}{|N|!}$ represents the weight of the contribution of the feature subset S in the coalition. This equation ensures that each feature's contribution is fairly weighted across all possible coalitions, making SHAP values consistent and interpretable.

In practice, calculating SHAP values directly from the equation can be computationally expensive. Therefore, approximation methods such as Kernel SHAP are often used. Kernel SHAP estimates SHAP values using a weighted linear regression in a transformed feature space, making it feasible to apply to complex models. In this study, SHAP values were computed for each model to quantify the impact of each feature on the predicted outcome.

3.3. Deep Learning Models

In this study, we employed various DL models to predict heart attack risk based on patient data. The models included MLP, RNN, LSTM, GRU, and a Hybrid model combining features of CNN and RNN.

The selection of DL models for this study was driven by their proven ability to efficiently process and analyze the intricate, high-dimensional data typical of medical datasets. CNNs were chosen for their exceptional capability in extracting spatial features, making them ideally suited for medical imaging data; however, in this study, they were adopted for use with tabular data. RNNs, along with LSTM networks and GRUs, were selected for their robust performance in handling sequential data, which are critical for capturing temporal patterns in patient histories.

The Hybrid model was strategically designed to combine the strengths of both CNNs and GRUs, aiming to enhance prediction accuracy by integrating spatial and temporal data-processing capabilities to address the limitations of other models where these features were inadequately captured. Combining CNN and GRU in a heart attack prediction model allows for the effective capture of both spatial and temporal features in the dataset. The CNN component excels at detecting complex spatial relationships between patient attributes, such as age, cholesterol levels, and blood pressure, which are crucial indicators of heart attack risk. Meanwhile, the GRU component can handle the temporal aspects, effectively analyzing how these features change over time, such as fluctuations in blood pressure or heart rate. By integrating these aspects, the hybrid CNN-GRU model can provide a more comprehensive and accurate prediction of heart attacks, identifying patients at risk by considering both their current health status and the progression of their health over time.

The following paragraphs provide a concise overview of each model's functionality and core attributes.

- **MLP** [28] is a feedforward neural network architecture. It consists of an input layer, multiple hidden layers, and an output layer. Each neuron in a layer is connected to every neuron in the subsequent layer through weighted connections. The output of each neuron is a weighted sum of its inputs, followed by a non-linear activation function. Mathematically, the activation $a_i^{(l+1)}$ of neuron i in layer $l+1$ is given by

  $$a_i^{(l+1)} = f\left(\sum_{j=1}^{n_l} w_{ij}^{(l)} a_j^{(l)} + b_i^{(l+1)}\right),$$

  where $a_j^{(l)}$ is the activation of neuron j in layer l, $w_{ij}^{(l)}$ is the weight-connecting neuron j in layer l to neuron i in layer $l+1$, $b_i^{(l+1)}$ is the bias of neuron i in layer $l+1$, and f is the activation function, typically the ReLU (Rectified Linear Unit), defined as $f(x) = \max(0, x)$. The learning process involves adjusting the weights and biases to minimize a loss function, often using backpropagation and gradient descent.
- **CNNs** [29] are specialized for processing data with a grid-like topology, such as images. The core operation in a CNN is the convolution, which applies a set of learnable filters to the input to extract local patterns. The convolution operation for an input X with a filter K is defined as

  $$(X * K)(i, j) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} X(i+m, j+n) \cdot K(m, n),$$

  where (i, j) are the spatial coordinates, and $M \times N$ is the filter size. The output is then passed through a non-linear activation function. Pooling layers, such as max pooling, reduce the spatial dimensions while retaining the most salient features. The final feature maps are flattened and fed into fully connected layers for classification.

- **RNNs** [30] are designed for sequential data by maintaining a hidden state that captures information about previous inputs. The hidden state h_t at time step t is computed as

  $$h_t = f(W_h h_{t-1} + W_x x_t + b),$$

  where x_t is the input at time step t, h_{t-1} is the hidden state from the previous time step, W_h and W_x are weight matrices, b is the bias vector, and f is the activation function, typically tanh. This recursive process allows the RNN to capture temporal dependencies in the sequence data.
- **LSTM** [31] networks are a type of RNN that mitigates the vanishing gradient problem by incorporating gating mechanisms. An LSTM cell comprises an input gate i_t, a forget gate f_t, an output gate o_t, and a cell state C_t, as follows:

  $$\begin{aligned} i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \\ f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \\ o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \\ \tilde{C}_t &= \tanh(W_C x_t + U_C h_{t-1} + b_C) \\ C_t &= f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \\ h_t &= o_t \odot \tanh(C_t) \end{aligned} \tag{1}$$

  where σ is the sigmoid function and $\odot$ denotes element-wise multiplication. These gates control the flow of information, enabling the network to retain or discard information as needed.
- **GRU** [32] is a simplified variant of LSTM that combines the forget and input gates into a single update gate z_t and introduces a reset gate r_t, as follows:

  $$\begin{aligned} z_t &= \sigma(W_z x_t + U_z h_{t-1} + b_z) \\ r_t &= \sigma(W_r x_t + U_r h_{t-1} + b_r) \\ \tilde{h_t} &= \tanh(W_h x_t + U_h (r_t \odot h_{t-1}) + b_h) \\ h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h_t} \end{aligned} \tag{2}$$

  The GRU architecture simplifies the gating mechanism while maintaining the ability to capture dependencies over long sequences.
- **Hybrid** model combines the strengths of both CNNs and GRUs to leverage spatial and temporal features effectively. The architecture involves first using CNN layers to process the input data and extract spatial features. These features are then fed into GRU layers to capture temporal dependencies.
  The convolution operation in the CNN layers is defined as

  $$(X * K)(i,j) = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} X(i+m, j+n) \cdot K(m,n)$$

  After passing through convolutional and pooling layers, the output feature maps F are flattened into a 1D vector F_{flat}, as follows:

  $$F_{flat} = \text{Flatten}(F)$$

This flattened vector F_{flat} is then reshaped to match the input format of the GRU layers. If F_{flat} has dimensions (N, C), where N is the number of time steps and C is the number of features per time step, it is fed into the GRU layer, as follows:

$$\begin{aligned} z_t &= \sigma(W_z F_{flat,t} + U_z h_{t-1} + b_z) \\ r_t &= \sigma(W_r F_{flat,t} + U_r h_{t-1} + b_r) \\ \tilde{h_t} &= \tanh(W_h F_{flat,t} + U_h(r_t \odot h_{t-1}) + b_h) \\ h_t &= (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h_t} \end{aligned} \tag{3}$$

Here, $F_{flat,t}$ represents the feature vector at time step t, and h_t is the hidden state at time step t.

Finally, the hidden states from the GRU are passed through fully connected layers for classification, as follows:

$$y = \text{softmax}(W_y h_T + b_y)$$

where W_y is the weight matrix, b_y is the bias vector of the fully connected layer, and h_T is the hidden state at the final time step T. The softmax function converts the logits into probabilities for each class, allowing for classification.

The Hybrid model benefits from the ability of CNNs to extract complex spatial features and the GRU's capability to capture temporal dependencies, resulting in improved predictive performance.

3.4. Evaluation Metrics

The effectiveness of the DL models is assessed through multiple essential metrics, including accuracy, precision, recall, F1-score, and AUC. These metrics provide a comprehensive understanding of the model's effectiveness in heart attack prediction [33].

- **Accuracy** is the proportion of correctly predicted instances among all instances. It is defined as
$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$
where TP is the number of true positives (heart attack cases), TN is the number of true negatives (non-heart attack cases), FP is the number of false positives (non-heart attack cases incorrectly classified as heart attack ones), and FN is the number of false negatives (heart attack cases incorrectly classified as non-heart attack ones).
- **Precision**, also known as positive predictive value, is the proportion of true positive predictions among all positive predictions. It is defined as
$$\text{Precision} = \frac{TP}{TP + FP},$$
indicating how many of the predicted heart attack cases were correct. High precision is crucial in avoiding unnecessary alarms or treatments.
- **Recall**, also known as sensitivity or true positive rate, is the proportion of true positive predictions among all actual positive instances. It is defined as
$$\text{Recall} = \frac{TP}{TP + FN},$$
and measures the model's ability to identify actual heart attack cases. It is important to achieve high recall to ensure that most heart attack cases are detected, minimizing the risk of missing a diagnosis.

- **F1-score** is the harmonic mean of precision and recall, providing a single metric that balances both. It is defined as

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$

The F1-score is especially useful when there is an uneven class distribution (e.g., many more non-heart attack cases than heart attack cases). It ensures that both false positives and false negatives are considered in the model's evaluation.

- **AUC** measures the ability of the model to distinguish between positive and negative classes. It is the area under the ROC curve that plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. Mathematically, it is often approximated using the trapezoidal rule, as follows:

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR})$$

The ROC curve itself is defined by

$$\text{TPR} = \frac{TP}{TP + FN} \quad \text{and} \quad \text{FPR} = \frac{FP}{FP + TN}$$

3.5. Experiment Environment and Hyperparameter Tuning

The experimental environment for this study was designed to ensure a robust and reproducible analysis of heart attack prediction using DL models. The experiments were conducted on a machine with an Intel Core i7 processor, 16 GB of RAM, and an NVIDIA GeForce GTX 1060 GPU to accelerate DL model training. The software setup included Python 3.8 as the primary programming language, utilizing several key libraries for data manipulation, model building, and visualization.

The data processing and manipulation were carried out using Pandas for efficient data handling, while Numpy was employed for numerical computations. Scikit-learn provided tools for data preprocessing, including standardization and splitting the dataset into training and testing sets. Additionally, Scikit-learn was instrumental in model evaluation, offering metrics such as accuracy, precision, recall, F1-score, and AUC to assess model performance comprehensively.

DL models were implemented using TensorFlow and Keras, which facilitated the construction, training, and evaluation of neural network architectures, including MLP, CNN, RNN, LSTM, GRU, and Hybrid models. For visualizations, Matplotlib and Seaborn were utilized to create plots. The experimental environment was set up to ensure that all experiments were repeatable, and the results were consistent, contributing to the overall validity and reliability of the study.

The dataset was split into an 80–20 train-test split. This split ensures that a substantial portion of the data are used for training the models, while the remaining data are reserved for testing, allowing for an independent evaluation of model performance. To optimize model performance, hyperparameter tuning is conducted using grid search and cross-validation. Grid search involves systematically evaluating combinations of hyperparameters by training models on each combination and selecting the best-performing set. Cross-validation involves splitting the training data into $k = 5$ subsets and training the model $k = 5$ times, each time using a different subset as the validation set and the remaining data for training. This process helps to ensure that the model generalizes well to unseen data. Each model was then trained on the training portion (80%) of the dataset using the tuned hyperparameters. Table 3 summarizes the optimal parameters with which the selected DL models have tuned during 5-fold cross-validation, aiming to maximize the previously presented metrics.

Table 3. Best parameters tuned for DL models.

Model	Parameter	Best Value
MLP	Number of Hidden Layers Units Per Layer Activation Function Learning Rate Batch Size	3 [64, 128, 64] ReLU 0.001 32
CNN	Number of Convolutional Layers Filters Per Layer Kernel Size Pooling Size Activation Function Learning Rate Batch Size	3 [32, 64, 128] (3, 3) (2, 2) ReLU 0.001 32
RNN	Number of Recurrent Layers Units Per Layer Activation Function Learning Rate Batch Size	2 [128, 64] tanh 0.001 32
LSTM	Number of LSTM Layers Units Per Layer Activation Function Learning Rate Batch Size	2 [128, 64] tanh 0.001 32
GRU	Number of GRU Layers Units Per Layer Activation Function Learning Rate Batch Size	2 [128, 64] tanh 0.001 32
Hybrid	Number of Convolutional Layers Filters Per Layer Kernel Size Pooling Size Number of GRU Layers Units per GRU Layer Activation Function Learning Rate Batch Size	2 [64, 128] (3, 3) (2, 2) 2 [128, 64] ReLU for CNN, tanh for GRU 0.001 32

4. Results and Discussion

The performance of each DL model was assessed using five crucial metrics: accuracy, precision, recall, F1-score, and AUC. These metrics offer a thorough insight into the models' proficiency in predicting heart attacks. The outcomes of these assessments are encapsulated in Table 4. The in-depth comparison highlights both the strengths of each model and the potential areas for enhancement.

Table 4. Performance metrics for DL models in heart attack prediction.

Model	Accuracy	Precision	Recall	F1-Score	AUC
MLP	0.85	0.83	0.84	0.84	0.89
CNN	0.87	0.85	0.86	0.85	0.91
RNN	0.84	0.82	0.83	0.82	0.88
LSTM	0.88	0.86	0.87	0.87	0.92
GRU	0.89	0.87	0.88	0.87	0.93
Hybrid	0.91	0.89	0.90	0.89	0.95

The MLP model achieved an accuracy of 0.85, a precision of 0.83, a recall of 0.84, an F1-score of 0.84, and an AUC of 0.89. Although MLP demonstrated balanced performance across all metrics, its slightly lower AUC suggests that while it can correctly identify most cases, it is not as effective in distinguishing between positive and negative cases. This limitation could be attributed to the model's architecture, which cannot capture complex patterns in the data as effectively as other models.

Moreover, the CNN model showed notable improvement over MLP, achieving an accuracy of 0.87, precision of 0.85, recall of 0.86, an F1-score of 0.85, and an AUC of 0.91. The CNN's ability to extract and leverage spatial hierarchies within the data likely contributes to its enhanced performance. This model's higher AUC indicates a better overall ability to distinguish between classes, making it particularly useful in scenarios where spatial feature extraction is critical.

In comparison, the RNN displayed adequate performance with an accuracy of 0.84, precision of 0.82, recall of 0.83, an F1-score of 0.82, and an AUC of 0.88. The RNN's architecture, designed to handle sequential data, performs well but falls short of the CNN and more advanced recurrent models. This suggests that while RNNs can capture temporal dependencies, their simpler structure might not be sufficient for the complex nature of heart attack prediction data.

The LSTM network, designed to address the limitations of standard RNNs, achieved an accuracy of 0.88, precision of 0.86, recall of 0.87, an F1-score of 0.87, and an AUC of 0.92. The LSTM's ability to manage long-term dependencies through its gating mechanisms is reflected in its superior recall and F1-score. This indicates that the LSTM is particularly effective at identifying true positive cases, making it a reliable choice for medical diagnostics where missing a positive case could have severe consequences.

In particular, the GRU model demonstrated the best performance among the individual models, with an accuracy of 0.89, precision of 0.87, recall of 0.88, an F1-score of 0.87, and an AUC of 0.93. The GRU's efficient gating mechanism, which simplifies the flow of information compared to LSTM, allows it to retain important features while discarding irrelevant ones. This balance between simplicity and functionality results in superior overall performance, making GRU an excellent choice for tasks requiring nuanced temporal data processing.

The Hybrid model surpassed all individual models, achieving the highest metrics across the rest DL models: an accuracy of 0.91, precision of 0.89, recall of 0.90, an F1-score of 0.89, and an AUC of 0.95. This model leverages the feature extraction capabilities of CNNs and the temporal processing strengths of GRUs, effectively capturing both spatial and temporal dependencies. The Hybrid model's superior performance underscores the potential of combining different DL architectures to address the multifaceted nature of medical data. By integrating the spatial sensitivity of CNNs with the temporal efficiency of GRUs, the Hybrid model provides a robust framework for predicting heart attacks with higher accuracy and reliability.

An important goal of this work was to provide a comprehensive comparison of five prominent DL models in the context of heart attack prediction. This was achieved, as evidenced by the experimental results where the Hybrid model demonstrated superior performance across all metrics, particularly in achieving the highest accuracy (91%) and AUC (0.95). This finding validated our approach of integrating CNN and GRU to leverage both spatial and temporal data, surpassing the performance of traditional and some DL models discussed in Section 2.

In addition to this objective, we aimed to ensure robust model performance across a range of metrics. The Hybrid model's consistently high scores across accuracy, precision, recall, and F1-score reflected the successful realization of this goal, demonstrating the model's reliability in predicting heart attack risks. The superior performance of the Hybrid model not only highlights the effectiveness of combining CNN and GRU architectures but also suggests the potential for these methods in other medical prediction tasks. This contribution is particularly significant when considering the limitations of existing models, such as the ones discussed in Section 2, where either temporal or spatial data handling was suboptimal.

Another objective was to enhance the interpretability of the DL models. Therefore, to further interpret these models and understand the underlying decision-making process, SHAP values were computed. SHAP answers the question of how much each feature contributes to the heart attack prediction. In Table 5, the average SHAP values (among all subjects) are demonstrated, revealing that, across all the DL models, the features of Max Heart Rate, Chest Pain Type, and thalassemia consistently showed high importance and, therefore, were the most influential across all models. These features had the most significant impact on the predictions made by the DL models in the context of heart attack prediction.

Table 5. Average SHAP values for features importance per DL model.

Feature	MLP	CNN	RNN	LSTM	GRU	Hybrid
Age	0.12	0.14	0.13	0.11	0.12	0.13
Gender	0.07	0.08	0.07	0.09	0.08	0.07
ChestP	0.18	0.21	0.20	0.19	0.20	0.22
RestBP	0.11	0.12	0.11	0.13	0.12	0.11
Chol	0.10	0.11	0.09	0.10	0.11	0.10
FastBS	0.04	0.05	0.04	0.06	0.05	0.04
RestECG	0.06	0.07	0.08	0.07	0.06	0.07
MaxHR	0.20	0.19	0.21	0.22	0.20	0.21
ExIndAng	0.08	0.07	0.09	0.10	0.08	0.09
STDp	0.13	0.12	0.14	0.13	0.12	0.14
Slope	0.09	0.10	0.09	0.11	0.10	0.10
NMajV	0.11	0.12	0.10	0.11	0.12	0.11
Thal	0.15	0.16	0.14	0.15	0.16	0.15

By integrating the SHAP values with the evaluation metrics, we gained a comprehensive understanding of not only how well the models performed but also why they made certain predictions, thereby enhancing the transparency and interpretability of the DL models in the context of heart attack prediction. By making the models' predictions transparent, healthcare professionals can better understand and trust the predictive system, leading to more informed decision-making. This study demonstrates the value of combining advanced DL models with XAI techniques to develop reliable and interpretable predictive systems for heart attack diagnosis, ultimately improving patient outcomes through early and accurate prediction. This goal was successfully met, as SHAP analysis provided clear insights into the feature importance, with Max Heart Rate, Chest Pain Type, and thalassemia being consistently influential across models. This not only aligns with our objective to make the models more clinically applicable but also differentiates our approach from other studies where interpretability was less emphasized.

5. Conclusions and Future Works

This study demonstrates the significant potential of DL models in predicting heart attacks by comparing the performance of five prominent models, namely MLP, CNN, RNN, LSTM, GRU, and a Hybrid one that integrates CNN and GRU architectures. Our findings indicate that advanced DL techniques can substantially improve the accuracy and reliability of heart attack predictions, which is critical for early diagnosis and timely intervention. Moreover, achieving the highest performance across all evaluated metrics—accuracy (91%), precision (89%), recall (90%), F1-score (89%), and AUC (0.95)—the effectiveness of the Hybrid model was confirmed, thereby contributing a robust solution to the field of medical diagnostics. The high AUC results indicated the Hybrid model's superiority in distinguishing between heart attack and non-heart attack subjects. This model's gating mechanism, which efficiently captures and retains important features while discarding irrelevant ones, likely contributed to its robust performance. In addition, the integration of the SHAP technique provided a deeper understanding of model predictions, enhancing their transparency and clinical applicability. The primary goal of enhancing model performance and interpretability was successfully achieved, as demonstrated by the Hybrid model's superior accuracy and the insights provided by SHAP analysis. These outcomes not only met but exceeded the initial objectives, offering a practical tool for clinicians that balances predictive power with transparency. The implications of this research are significant, particularly for the development of more reliable and interpretable diagnostic tools in healthcare.

While the findings of this study are promising, it is essential to acknowledge the limitations of the dataset used. The limited size of the dataset (instances and features) may not fully represent the broader population due to its limited size and specific demographic scope. The relatively small sample size raises concerns about the potential for overfitting, where the model may perform well on the training data but not generalize effectively to new data. Additionally, the distribution of the features within the dataset, such as the age range and gender representation, may introduce biases that affect the model's predictions. Moreover, the dataset lacks detailed information on factors such as socioeconomic status, medical history, the co-existence of other comorbidities, and lifestyle variables, which could influence heart attack risk but are not captured in this analysis. The absence of such factors that are known to significantly impact heart attack risk further constrains this study. These limitations highlight the need for future research using larger, more diverse datasets to enhance the robustness, accuracy, applicability, and generalizability of the predictive models developed in this study in varying clinical settings.

Collaboration with clinical experts is essential to enhance the practical relevance of our findings in real-world settings. These collaborations could be structured in a phased manner, beginning with retrospective validation using historical patient data and progressing to prospective studies where model predictions influence clinical decision-making. Challenges in these collaborations include integrating the model into clinical workflows, ensuring compliance with data privacy regulations, and addressing the initial scepticism that may arise from clinicians. A continuous feedback loop, where clinical outcomes are used to refine the model, will be critical in making the models more robust and reliable for real-world applications. Such interdisciplinary partnerships will be key to transitioning from research to practical implementation, ultimately improving patient outcomes.

The practical implications of this research are profound. By integrating DL models into clinical decision-making processes, healthcare providers can benefit from more accurate and timely predictions of heart attack risk. This integration can lead to personalized and proactive patient care, ultimately improving patient outcomes and reducing the strain on healthcare systems. Moreover, our comparative analysis provides valuable insights for researchers and practitioners, guiding future developments in the field of medical diagnostics.

In a future extension of the current study, a more detailed analysis of how the dataset features as risk factors interact with each other could add significant depth to our understanding of heart attack risk. Moreover, apart from the traditional evaluation metrics currently considered, a challenging alternative would be to assess the computational complexity of each model by measuring training time and resource utilization (e.g., memory footprint and GPU usage). This analysis would help to understand the practical feasibility of deploying these models in real-world settings, where computational resources may be limited and larger datasets are common. While the selected dataset is appropriate for model development, it is acknowledged that DL models generally benefit from larger datasets, which allow for better generalization and robustness. In future work, scaling the models to larger datasets will be essential to validate the findings and ensure that the models can handle the increased complexity and variability present in more extensive datasets. Moving forward, future research will focus on exploring the integration of other advanced ML/DL techniques with different XAI methods to further enhance the robustness and applicability of predictive models in clinical settings.

This study serves as a foundational exploration, with the expectation that the techniques and insights developed here will be scalable to larger data environments to include more diverse patient populations and incorporate additional relevant features to further enhance model performance.

Author Contributions: E.D. and M.T. conceived the idea, designed and performed the experiments, analyzed the results, drafted the initial manuscript, and revised the final manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Cardiovascular Diseases. Available online: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1 (accessed on 31 July 2024).
2. Mullainathan, S.; Obermeyer, Z. *Who Is Tested for Heart Attack and Who Should Be: Predicting Patient Risk and Physician Error*; National Bureau of Economic Research: Cambridge, MA, USA, 2019.
3. Gong, F.F.; Vaitenas, I.; Malaisrie, S.C.; Maganti, K. Mechanical complications of acute myocardial infarction: A review. *JAMA Cardiol.* **2021**, *6*, 341–349. [CrossRef] [PubMed]
4. Mechanic, O.J.; Gavin, M.; Grossman, S.A.; Ziegler, K. Acute Myocardial Infarction (Nursing). In *StatPearls [Internet]*; StatPearls Publishing: St. Petersburg, FL, USA, 2023.
5. Salyer, J.; Flattery, M.; Lyon, D.E. Heart failure symptom clusters and quality of life. *Heart Lung* **2019**, *48*, 366–372. [CrossRef] [PubMed]
6. Han, C.H.; Kim, H.; Lee, S.; Chung, J.H. Knowledge and poor understanding factors of stroke and heart attack symptoms. *Int. J. Environ. Res. Public Health* **2019**, *16*, 3665. [CrossRef] [PubMed]
7. Libby, P. *Braunwald's Heart Disease-E-Book: A Textbook of Cardiovascular Medicine*; Elsevier Health Sciences: Amsterdam, The Netherlands, 2021.
8. Truby, L.K.; Rogers, J.G. Advanced heart failure: Epidemiology, diagnosis, and therapeutic approaches. *Heart Fail.* **2020**, *8*, 523–536.
9. Kaul, D.; Raju, H.; Tripathy, B. Deep learning in healthcare. In *Deep Learning in Data Analytics: Recent Techniques, Practices and Applications*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 97–115.
10. Ahsan, M.M.; Siddique, Z. Machine learning-based heart disease diagnosis: A systematic literature review. *Artif. Intell. Med.* **2022**, *128*, 102289. [CrossRef]
11. Van der Velden, B.H.; Kuijf, H.J.; Gilhuijs, K.G.; Viergever, M.A. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Med. Image Anal.* **2022**, *79*, 102470. [CrossRef]
12. Salih, A.M.; Raisi-Estabragh, Z.; Galazzo, I.B.; Radeva, P.; Petersen, S.E.; Lekadir, K.; Menegaz, G. A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME. *Adv. Intell. Syst.* **2024**, 2400304. [CrossRef]
13. Ahmed, S.; Kaiser, M.S.; Hossain, M.S.; Andersson, K. A Comparative Analysis of LIME and SHAP Interpreters with Explainable ML-Based Diabetes Predictions. *IEEE Access* **2024**. [CrossRef]
14. Ashraf, M.; Rizvi, M.; Sharma, H. Improved heart disease prediction using deep neural network. *Asian J. Comput. Sci. Technol.* **2019**, *8*, 49–54. [CrossRef]
15. Harkulkar, N.; Nadkarni, S.; Patel, B.; Jadhav, A. Heart Disease Prediction using CNN Deep Learning Model. *Int. J. Res. Appl. Sci. Eng. Technol.* **2020**, *8*, 875–881. [CrossRef]
16. Barhoom, A.M.A.; Almasri, A.; Abu-Nasser, B.S.; Abu-Naser, S.S. Prediction of Heart Disease Using a Collection of Machine and Deep Learning Algorithms. *Int. J. Eng. Inf. Syst. (IJEAIS)* **2022**, *6*, 1–13.
17. Panda, R.N.; Zaheera, F. Prediction of Heart Disease using Deep Convolutional Neural Networks. *Turk. J. Comput. Math. Educ. (TURCOMAT)* **2019**, *10*, 1141–1148.
18. Bharti, R.; Khamparia, A.; Shabaz, M.; Dhiman, G.; Pande, S.; Singh, P. Prediction of heart disease using a combination of machine learning and deep learning. *Comput. Intell. Neurosci.* **2021**, *2021*, 8387680. [CrossRef] [PubMed]
19. Sharma, S.; Parmar, M. Heart diseases prediction using deep learning neural network model. *Int. J. Innov. Technol. Explor. Eng. (IJITEE)* **2020**, *9*, 2244–2248. [CrossRef]
20. Sarmah, S.S. An efficient IoT-based patient monitoring and heart disease prediction system using deep learning modified neural network. *IEEE Access* **2020**, *8*, 135784–135797. [CrossRef]
21. Pasha, S.N.; Ramesh, D.; Mohmmad, S.; Harshavardhan, A. Cardiovascular disease prediction using deep learning techniques. In Proceedings of the IOP Conference Series: Materials Science and Engineering, Chennai, India, 16–17 September 2020; IOP Publishing: Bristol, UK, 2020; Volume 981, p. 022006.

22. Waqar, M.; Dawood, H.; Dawood, H.; Majeed, N.; Banjar, A.; Alharbey, R. An Efficient SMOTE-Based Deep Learning Model for Heart Attack Prediction. *Sci. Program.* **2021**, *2021*, 6621622. [CrossRef]
23. Latha, C.B.C.; Jeeva, S.C. Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Inform. Med. Unlocked* **2019**, *16*, 100203. [CrossRef]
24. Heart Disease Datasets. Available online: https://github.com/Abdulrakeeb/Heart-disease-dataset/tree/main (accessed on 16 August 2024).
25. Heart Disease Dataset from Five Databases. Available online: https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive (accessed on 16 August 2024).
26. Heart Disease Dataset from Four Databases. Available online: https://archive.ics.uci.edu/dataset/45/heart+disease (accessed on 16 August 2024).
27. Haider, A.; Bengs, S.; Luu, J.; Osto, E.; Siller-Matula, J.M.; Muka, T.; Gebhard, C. Sex and gender in cardiovascular medicine: Presentation and outcomes of acute coronary syndrome. *Eur. Heart J.* **2020**, *41*, 1328–1336. [CrossRef]
28. Kruse, R.; Mostaghim, S.; Borgelt, C.; Braune, C.; Steinbrecher, M. Multi-layer perceptrons. In *Computational Intelligence: A Methodological Introduction*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 53–124.
29. Véstias, M.P. Convolutional neural network. In *Encyclopedia of Information Science and Technology*, 5th ed.; IGI Global: Hershey, PA, USA, 2021; pp. 12–26.
30. Kanagachidambaresan, G.; Ruwali, A.; Banerjee, D.; Prakash, K.B. Recurrent neural network. In *Programming with TensorFlow: Solution for Edge Computing Applications*; Springer: Berlin/Heidelberg, Grmany, 2021; pp. 53–61.
31. Van Houdt, G.; Mosquera, C.; Nápoles, G. A review on the long short-term memory model. *Artif. Intell. Rev.* **2020**, *53*, 5929–5955. [CrossRef]
32. Dutta, A.; Kumar, S.; Basu, M. A gated recurrent unit approach to bitcoin price prediction. *J. Risk Financ. Manag.* **2020**, *13*, 23. [CrossRef]
33. Rainio, O.; Teuho, J.; Klén, R. Evaluation metrics and statistical tests for machine learning. *Sci. Rep.* **2024**, *14*, 6086. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.