Heart Attack Prediction Using Machine Learning

Raiyan Ahmed
Dept. of ECE
North South University
ID: 2013130042
raiyan.ahmed05@northsouth.edu

Mehedi Hasan
Dept. of ECE
North South University
ID: 2011425042
mehedi.hasan41@northsouth.edu

Asif Mahmud Sifat
Dept. of ECE
North South University
ID: 2014043042
asif.sifat@northsouth.edu

Md. Asibul Alam
Dept. of ECE
North South University
ID: 2013790642
asibul.alam@northsouth.edu

***Abstract*—Heart attack predictability is a crucial component of cardiology treatment since it identifies those who are at risk for a heart attack and allows for swift action and favorable outcomes for patients. Although substantial advances in machine learning approaches, several obstacles remain, such as algorithm constraints, interpretability issues, data dependency, and scalability concerns. These problems highlight the importance of having strong, interpretable, and generalizable predictive models that can effectively handle the intricacies of medical data. In this study, we have applied seven machine-learning models to the heart disease dataset. SVM outperformed on default parameters with 90% accuracy and an f1 score of 91% on class 0 of Output. We had to go through data preprocessing, feature mapping, feature encodings, feature scaling, and hyperparameter optimization approaches to obtain a highly accurate and comprehensible predictive model. The suggested model solves the current obstacles in heart attack prediction, providing an exciting opportunity to improve coronary healthcare outcomes. However, By giving understandable explanations for specific predictions, LIME exposes the key elements impacting the model's predictions, increasing the machine learning model's efficiency and confidence.**

***Index Terms*—Heart Attack Prediction, Heart Disease Prediction, Cardiology**

I. Introduction

Heart attack is one of the most common diseases in every country of the world. It is known as a heart attack when the heart can't pump adequate blood to the entire body due to the blockage of the coronary arteries. Smoking tobacco, consuming too much alcohol, leading unhealthy lifestyles, genetic reasons, and at a particular stage of life, having some other diseases like diabetes, high blood pressure, etc. are the main factors of this fatal illness [1]. Around 1 in 14 people can live globally with a heart disease. It's the leading cause of death worldwide, with an estimated death of 17.9 million, according to the WHO [2]. In Bangladesh, With estimated results, 21.1 percent of total deaths are caused by heart attacks. There are three types of heart attacks. STEMI, non–STSEME type, coronary artery disease, or unstable angina. The type and origin of heart damage determines treatment for heart disease. Quitting smoking, regular exercise, adequate sleep, and a low-fat and salt diet are essential to the therapy. In addition to a chest X-ray and blood tests, other diagnostic procedures for heart illness include an electrocardiogram (ECG or EKG) [4]. So, ECG helps record the heart's electrical impulses.

From now on, we will describe the literature review briefly; the work of Gnaneswari G [5] utilized machine learning algorithms to predict the risk of heart diseases and heart attacks. The collected dataset contains 919 instances with 13 features, but some were reduced for less impact, as seen in the correlation matrix. The author experimented with various classifiers, and random forest gained an accuracy of 89.13%, making it the most efficient model. Alternatively, the decision tree and GNB got overfitted on the dataset with an accuracy of 100%.

Another work of Anup Lal and his team [1] employed machine-learning approaches to predict different types of heart diseases. They used the UCI Heart Disease Dataset containing 303 instances with 13 features, but there were many missing values on some features, so they had to fix this issue in data preprocessing. The authors used four different types of algorithms. By analyzing those algorithms, they achieved their best accuracy of 97.08% from their two algorithms, and their other two algorithms gave the lowest accuracy of 80.52% and 70.13%, respectively.

Then, we saw Shanbhag and his team [6] provide an inclusive overview of various machine-learning techniques used for heart disease and heart attack risk prediction. The authors used the dataset known as the 'Heart Disease Dataset,' which contains 14 attributes with 303 patients' records and 4000 tuples, and they have used the 'Smote' technique to balance their imbalanced dataset. The researchers have used machine learning techniques. However, in their study, SVM gave the highest accuracy with 85.7%.

To forecast the risk of heart disease and heart attacks, Harshit and his colleagues [4] used machine learning algorithms. It uses three models and has 303 instances with 14 characteristics. Using the dataset, logistic regression, and KNN, the accuracy was 85%. The best accuracy among the three algorithms was 88.52%. The project's accuracy rate was 87.5%.

In this article, we have employed machine learning and explainable AI techniques to predict heart attacks. We used the Heart Attack Analysis & Prediction Dataset from Kaggle, which contains 303 instances with 14 attributes. After

preprocessing the data, the dataset is ready to train with seven different machine-learning models. With the default hyperparameter, the SVM gained the best accuracy of 90% and an f1 score of 91%. Hyperparameter tuning has been performed on SVM to find the best hyperparameter values. Finally, explainable AI with the LIME library is applied to analyze what features play the most vital role in prediction.

The following paragraph is a breakdown of the report's structure. Section II discusses the proposed system of our work. This section also includes sub-sections, such as datasets, data preprocessing with equations, and machine learning models. Section III focuses on the results and discussion. Finally, Section IV includes a conclusion with remarks on our future work.

II. Proposed System

A. Dataset

The open-source dataset used in this work is acquired from the UCI Machine Learning Repository [7]. It contains 303 instances and 14 attributes. The independent features are Age, Sex, Cp, Trestbps, Chol, Fbs, Restecg, Thalach, Exang, Oldpeak, Sope, Ca, Thal, and num. It contains basic information about the patient's age, gender, chest pain type, resting blood pressure, cholesterol level, fasting blood sugar, resting electrocardiographic results, and maximum heart rate achieved. The dataset mentioned about four types of chest pain. They are typical angina, atypical angina, non-anginal pain, and asymptomatic. Finally, the target variable is output, defining the problem as a binary classification. Along with

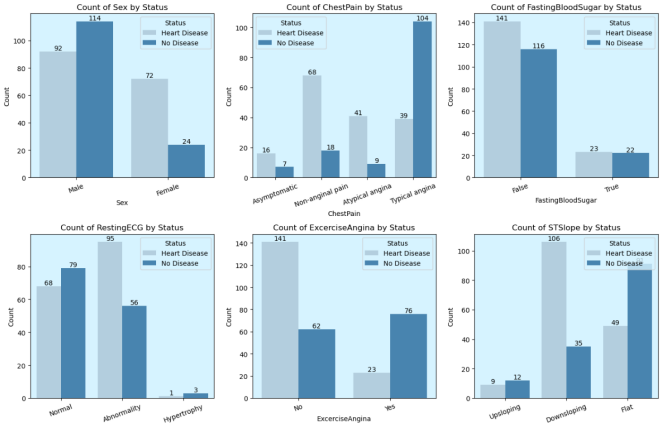

Fig. 1: Count of Sex, ChestPain, FastingBloodSugar, RestingECG, ExcerciseAngina, STSlope by status.

other features, cholesterol is an essential aspect of this dataset in As shown in Figure 5. In Figure 3, Maximum heart rate patterns are shown by a box plot and histogram. The box displays average rates (the middle 50%), with outliers indicated by whiskers. Frequency details are added via the histogram. A centered peak indicates that the heart rates are distributed normally.

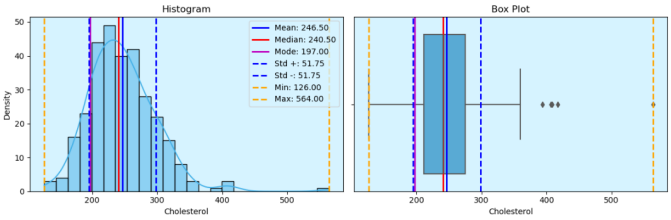

Fig. 2: Histogram and Boxplot of cholesterol feature.

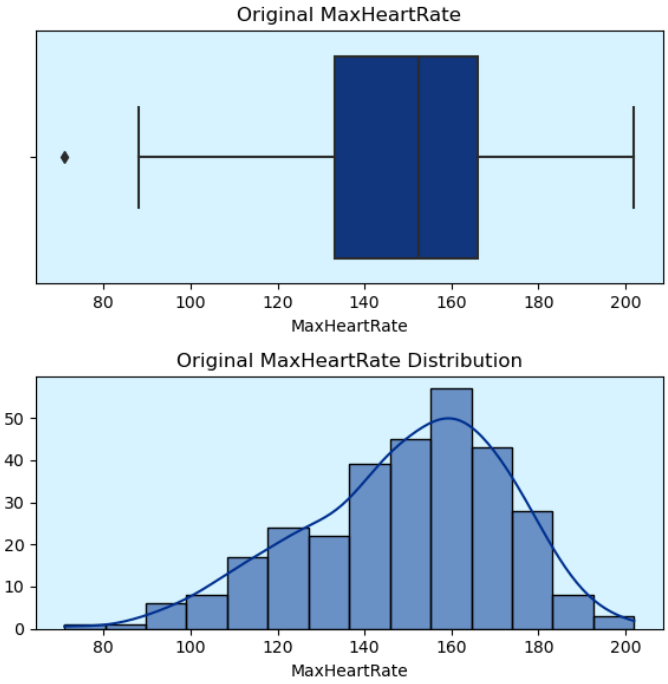

Fig. 3: The distribution of original maximum heart rate.

B. Dataset Preprocessing

Before utilizing the whole dataset for automatic heart attack prediction, some initial exploration and data preprocessing have been performed in this work. In the following order, we have described the process of data exploration and preprocessing.

1) **Mapping and Dropping Duplicates**:
   - At first, we applied mapping on all features and their values. Our dataset had no null values, but we had to drop one row as it contained duplicate values.
   - After dropping the rows with duplicate values, the size of the dataset became 302.

2) **Handling Categorical Variables**:
   - As some machine learning models cannot train on label data directly and require all the features in numerical values, we converted all the attributes with categorical values to numerical variables.
   - We applied one-hot encoding on features such as ChestPain, RestingECG, STSlope, and Thalium.

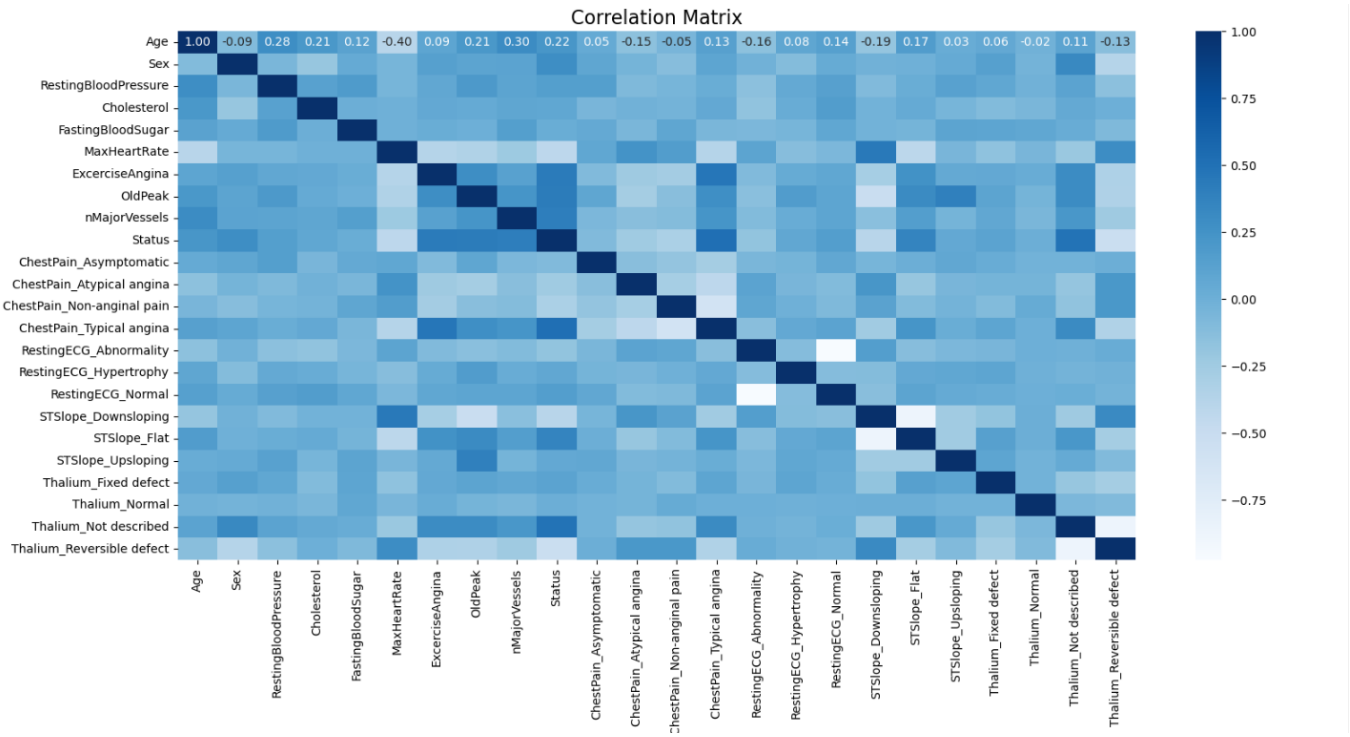

Fig. 4: The correlation table shows, Age positively correlates with resting blood pressure (0.28), cholesterol (0.21), and the number of major vessels (0.30), indicating older individuals tend to have higher values in these areas. Conversely, age negatively correlates with maximum heart rate (-0.40), so younger individuals have higher max heart rates.

- Label encoding was applied on features such as Sex, FastingBloodSugar, ExcerciseAngina, and Status.

3) **Outlier Detection**: By using the IQR method, we detect the outliers. We had to find lower and upper bounds to determine outliers using the IQR method.The equations are given below:

$$Q1 - 1.5 \times \text{IQR} \tag{1}$$

$$Q3 + 1.5 \times \text{IQR} \tag{2}$$

Here, eq (1) is about the lower bound, Q1 denotes the 25th percentile, and any data point that falls below the lower bound is considered an outlier. On the contrary, eq (2) is about the upper bound, Q3 denotes the 75th percentile, and any data point that exceeds the upper bound is also considered an outlier. To clean the dataset more precisely, we removed all the outliers, which will improve the robustness of our machine-learning models.

4) **Data Normalization**: For data normalization, we performed the standard feature scaling method on features.

$$X_s = \frac{X_i - \mu}{\sigma} \tag{3}$$

In eq (3) X_s is the values from scaling, X_i denotes an original value of a feature, σ is the standard deviation and μ is the mean value of the feature.

5) **Train-Test Split**:

- Finally, For training and testing, we divided the dataset into train 80:20 ratios.

C. Machine Learning Models

The (KNN) K-nearest neighbour classifier is a non-parametric and supervised learning model. In the KNN model, on training time, at first, it stores the training data, then during predicting time, the KNN mode calculates the distance between new data points and stored points. For this, we use Manhattan or Euclidean distance calculating method. This algorithm selects K neighbours based on the dataset and uses a majority vote to classify them. Here, K is the hyperparameter. In our work, we used K = 18 for binary classification.

The random forest learning model is called an ensemble learning model. This model builds multiple decision trees and combines their outputs to improve predictive performance and reduce overfitting. In our research, we applied the random forest model. In our model, we have applied random forest with estimators = 800, minimum samples leaf = 4, and minimum samples split = 5.

Support Vector Machines (SVM) are used in classification and regression. It is suitable for small and medium datasets. Support Vector is the training sample closest to the hyperplane. It works by choosing the best hyperplane. In our work, we have done with different types of SVM kernels in the training set. Later, we found the 'RBF' Kernel with the parameters c

= 1.0 and gamma = 0.01, which gave us the best result for this dataset. There are two types of SVM: Hard and Soft. In soft SVM, the number of misclassification errors means 'C', which is the hyperparameter.

The ZeroR classifier is the simplest classification method that relies on the target (output, y) and ignores all predictors (features, x). There is no predictability power in ZeroR. It made a frequency table from the output column of the training dataset. After that, identify the class with the highest frequency (the majority class). For binary classification, accuracy ¿= 50%. We use ZeroR to identify the imbalanced data; we use it for baseline performance.

logistic regression is based on the concept of probability. It's made a 'S' shaped curved function for predicting the output. We use cost function / j minimum in logistic regression to get the non-convex curve. In our model, we have applied logistic regression; for this, our penalty value was 12 and c = 0.01.

AdaBoost (Adaptive Boosting) is an ensemble machine learning algorithm combining multiple weak classifiers to form a robust classifier. It works on the original dataset. At first, the model will assign equal weight to all training samples and then train the weak classifier using the weighted training data. It will analyze the classifier's performance, compute its error rate and adjust the weights of improperly classified instances. In our work, we applied AdaBoost with the estimator value = 1.0, and we used 'gini' as the base estimation criterion.

Decision is applied in both classification and regression tasks. In this model, we have to find the root node. For this, we can use the Gini index or entropy. Gini or entropy coefficients are used to determine the information obtained, and nodes are selected based on them. These are expressed as,

$$\text{Gini}_i = 1 - \sum_{k=1}^{n} (p_{i,k})^2 \tag{4}$$

$$\text{Entropy} = \sum_{i=1}^{n} -p_i \log_2 p_i \tag{5}$$

Here in both equations, n represents the number of distinct class values. In our work, the minimum sample leaf was 4 and the maximum depth was 40, and the 'Gini' impurity works well in our dataset by using the random search CV hyperparameter tuning.

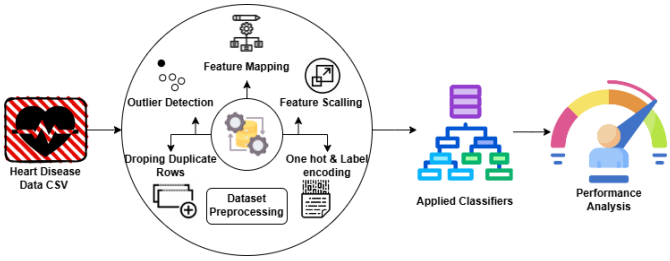

Fig. 5: Working sequences of the proposed Heart Attack prediction system

III. Result and Discussion

We have selected all 13 features and one label from the dataset and used them to predict the odds of heart disease. Hence, only SVM had a 90 percent accuracy, and other models achieved this accuracy without any hyperparameter fine-tuning. In this experiment, SVM accuracy improved through hyperparameter tuning by RandomizedSearchCV; thus, the accuracy is 85 percent, explaining the impact of hyperparameter tuning.

TABLE I: HYPERPARAMETER VALUES RANGES FOR VARIOUS ML MODELS

Model	Hyperparameter Value Range	Optimized Value
SVM	• C: [0.001, 0.01, 0.1, 1.0, 10.0, 100.0] • gamma: [0.0001, 0.001, 0.01, 0.1, 1.0, 10.0] • kernel: ['linear', 'poly', 'rbf', 'sigmoid']	(gamma: 0.01, C: 1.0, kernel: rbf)
KNN	• n_neighbors: [1, 2, ..., 20] • weights: ['uniform', 'distance'] • metric: ['euclidean', 'manhattan', 'minkowski']	k: 18 (weights: uniform, metric: manhattan)
Random Forest	• n_estimators: [50, 100, 200, 400, 600, 800, 1000] • max_features: ['auto', 'sqrt', 'log2'] • max_depth: [None, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100] • min_samples_split: [2, 5, 10] • min_samples_leaf: [1, 2, 4] • bootstrap: [True, False]	n_estimators: 800 min_samples_split: 5 min_samples_leaf: 4 max_features: log2 max_depth: 70 bootstrap: True
Decision Tree	• criterion: ['gini', 'entropy'] • splitter: ['best', 'random'] • max_depth: [None, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100] • min_samples_split: [2, 5, 10] • min_samples_leaf: [1, 2, 4] • max_features: ['auto', 'sqrt', 'log2', None]	splitter: best min_samples_split: 2 min_samples_leaf: 4 max_features: None max_depth: 40 criterion: gini
Logistic Regression	• C: [0.001, 0.01, 0.1, 1.0, 10.0, 100.0, 1000.0] • penalty: ['l1', 'l2']	penalty: l2 C: 0.01
AdaBoost	• n_estimators: [1, 50] • base_estimator__criterion: ['gini', 'entropy'] • base_estimator__splitter: ['best', 'random']	n_estimators: 1 _splitter: best _criterion: gini

The findings suggest that with default parameters, SVM is accurate, with the potential for even greater accuracy than models trained on the entire dataset. So, SVM is a suitable classifier for this case because it effectively solves complex problems with partitions of high dimensionality. More optimization of SVM and other models could be a potential way of getting even better results, which enforces the view on hyperparameter optimization in machine learning model creations.

TABLE II: PERFORMANCE METRICS OF VARIOUS ML MODELS WITH DEFAULT HYPERPARAMETERS

Model	Accuracy	Precision	Recall	F1-score
SVM	90%	0.90	0.90	0.90
KNN	89%	0.89	0.88	0.89
Random Forest	82%	0.82	0.82	0.82
Decision Tree	66%	0.66	0.66	0.66
Logistic Regression	84%	0.84	0.83	0.83
AdaBoost	82%	0.82	0.81	0.82
ZeroR	54%	0.29	0.54	0.38

Performance metrics of various ML models with default hyperparameters have been illustrated in Table II

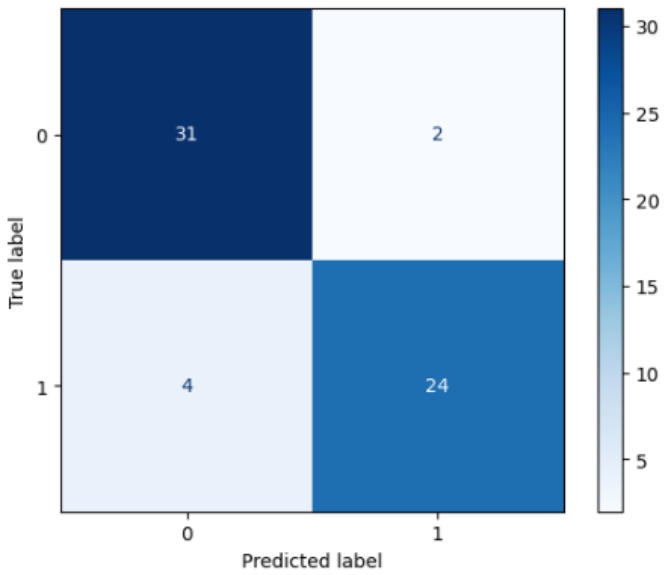

Fig. 6: Confusion matrix of SVM with default hyperparameter

In Figure 6 denotes the performance of an SVM model(best model) with default hyperparameters for this work. Out of 33 actual negative cases, 31 predictions were correct, and two were misclassified predictions. On the contrary, out of 28 actual positive cases, 24 predictions were accurate, and four were misclassified predictions.

TABLE III: PERFORMANCE METRICS OF VARIOUS ML MODELS WITH OPTIMIZED HYPERPARAMETERS

Model	Accuracy	Precision	Recall	F1-score
SVM	85%	0.86	0.85	0.85
KNN	85%	0.86	0.84	0.85
Random Forest	85%	0.85	0.85	0.85
Decision Tree	69%	0.69	0.69	0.69
Logistic Regression	85%	0.85	0.85	0.85
AdaBoost	69%	0.70	0.69	0.69

Performance metrics of various ML models with optimized hyperparameters have been illustrated in Table III.

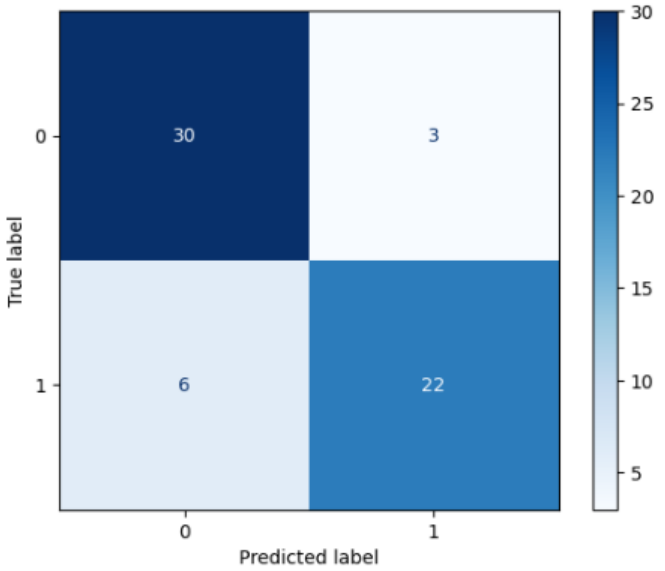

Fig. 7: Confusion matrix of SVM with optimized hyperparameter

In Figure 7 denotes the performance of an SVM model(best model) with optimised hyperparameters for this work. Out of 33 actual negative cases, 30 predictions were correct, and three were misclassified predictions. On the contrary, out of 28 actual positive cases, 22 predictions were accurate, and six were misclassified predictions.

TABLE IV: COMPARISON OF THE PROPOSED SYSTEM WITH EXISTING WORKS

Ref.	Model	Accuracy	Other metrics
[6]	SVM	85.7%	-
[4]	KNN & Logistic regression	88.5%	-
[8]	SVM	91.7%	F1 score 0.93
[9]	Decision Tree	85%	-
This work	SVM	90%	F1 score 0.90

Table IV illustrates a comparison of the proposed automatic heart attack prediction system with other existing works on the Heart Diseases dataset from UCI Machine Learning Repository.

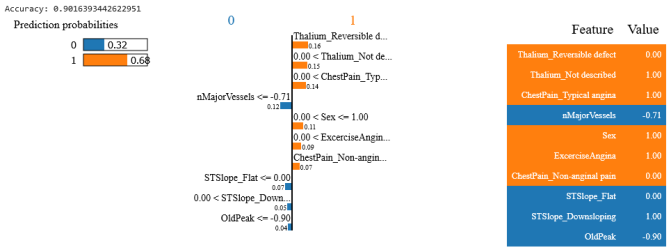

Fig. 8: Machine learning model prediction interpretation by LIME explainable AI library

LIME was applied in this work to improve model interpretability as we can see In Figure 8 It is one of the most used explainable AI libraries. It shows the accuracy of the model and its prediction probabilities for each class. This library also highlights the contribution of each feature to the model's prediction. Thus, it helped us to understand and determine which feature plays a vital role in the model's prediction.

IV. Conclusion

The primary cause of death is heart disease. Heart disease is estimated to become the top cause of mortality worldwide by 2020, killing 17.9 million people. Predicting or identifying this disease early on can improve survival rates while reducing the risk and effects of numerous other diseases. Our work has employed multiple machine-learning algorithms to forecast cardiac attacks. This work used an open-source, utterly balanced dataset from the UCI Machine Learning Repository [7]. This paper presents numerous performance indicators of machine learning approaches, including recall, precision, F1 score, and accuracy. SVM produced the best and highest accuracy in this article, with an F1 score of 0.90 and 90% accuracy. There is another extension for this work: we will apply more machine learning models and combine them, such as SVM and random forest, using stacking and bagging algorithms, which might improve accuracy and robustness against unexpected data fluctuations. Moreover, we may apply some deep learning algorithms to see the results.

References

[1] A. L. Yadav, K. Soni and S. Khare, "Heart Diseases Prediction using Machine Learning," 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 2023, pp. 1-7, doi: 10.1109/ICCCNT56998.2023.10306469

[2] V. Selvakumar, A. Achanta and N. Sreeram, "Machine Learning based Chronic Disease (Heart Attack) Prediction," 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA), Uttarakhand, India, 2023, pp. 1-6, doi: 10.1109/ICIDCA56705.2023.10099566.

[3] H. Agrawal, J. Chandiwala, S. Agrawal and Y. Goyal, "Heart Failure Prediction using Machine Learning with Exploratory Data Analysis," 2021 International Conference on Intelligent Technologies (CONIT), Hubli, India, 2021, pp. 1-6, doi: 10.1109/CONIT51480.2021.9498561.

[4] H. Jindal, S. Agrawal, R. Khera, R. Jain, and P. Nagrath, "Heart disease prediction using machine learning algorithms," IOP Conference Series: Materials Science and Engineering, vol. 1022, no. 1, pp. 012072, Jan. 2021. doi: 10.1088/1757-899X/1022/1/012072.

[5] G. G, "Analysis of The Diagnostic Parameters of Heart Diseases and Prediction of Heart Attacks," 2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT), Bangalore, India, 2022, pp. 1-6, doi: 10.1109/GCAT55367.2022.9972211.

[6] A. A. Shanbhag, C. Shetty, A. Ananth, A. S. Shetty, K. Kavanashree Nayak and B. R. Rakshitha, "Heart Attack Probability Analysis Using Machine Learning," 2021 IEEE International Conference on Distributed Computing, VLSI, Electrical Circuits and Robotics (DISCOVER), Nitte, India, 2021, pp. 301-306, doi: 10.1109/DISCOVER52564.2021.9663631.

[7] A. Janosi, W. Steinbrunn, M. Pfisterer, and R. Detrano, "Heart Disease," UCI Machine Learning Repository, 1988. Available: https://doi.org/10.24432/C52P4X.

[8] N. Bora, S. Gutta, and A. Hadaegh, "Using Machine Learning to Predict Heart Disease," WSEAS Transactions on Biology and Biomedicine, vol. 19, pp. 1-9, 2022. doi: 10.37394/23208.2022.19.1.

[9] V. Sabarinathan and V. Sugumaran, "Diagnosis of heart disease using decision tree," International Journal of Research in Computer Applications & Information Technology, vol. 2, no. 6, pp. 74-79, 2014.