Dijlah Journal of Engineering Science (DJES) Vol. 3, No. 2, June, 2026, pp.346-358
ISSN: Printed: **3078-9656, Online: 3078-9664**, paper ID: 170

Machine Learning Based Heart Attack Risk Prediction Using Clinical and Lifestyle Features

Ammar Saleh Abd
Anbar Governorate Office, Anbar, Iraq

Article Info

Article history:

Received March., 25, 2026
Revised Apr.,11, 2026
Accepted May.,8, 2026

Keywords:

Heart Attack Prediction
Cardiovascular Risk
Healthcare Analytics
Machine Learning
Random Forest (RF)
Predictive Modeling

ABSTRACT

The research paper describes the machine learning-based prediction of heart attack to offer information on heart dynamics and the most important predictive variables. The dataset used was the Heart Attack Risk Prediction Dataset on Kaggle and it consisted of 8,763 records of patients with various patient attributes including age, cholesterol, blood pressure and lifestyle-related variables. Other machine learning models such as Support Vector Machine (SVM), Decision Tree and Random Forest (RF) were developed and tested. After preprocessing of the data, feature extraction and the model training, the SVM, Decision Tree, and the Random Forest models had the following accuracy: 0.86, 0.79, and 89 percent, respectively. Detailed classification reports and confusion matrices were used to support the performance evaluation to get a further understanding of the strengths and weaknesses of each model. The findings reveal that the Random Forest model was the most effective as compared to the other methods, which implies that it can be used to identify complicated trends in the data. Limitations of the models and future research directions are also addressed in the research with the emphasis on the opportunities of machine learning methods in preventing heart diseases in advance and risk assessment in individuals. On the whole, the given research can lead to the improvement of cardiovascular health prediction and can be a basis of the further development of predictive analytics in the healthcare system.

Corresponding Author:

Ammar Saleh Abd
Anbar Governorate Office

Alramdi, Anbar, Iraq
Email: ammarsalih307@gmail.com

1. INTRODUCTION

The Cardiovascular diseases are one of the most critical health issues of the modern world, which causes the high number of deaths in this area per year because heart attacks are one of the top causes of death [1]. Early detection of the risk of heart attack is important to enhance the likelihood of preventive and therapeutic intervention and minimize the long-term health-related complications. Nevertheless, conventional prediction and diagnosis mechanisms tend to be insufficient in managing the complexity of the situation arising due to the presence of numerous and intersecting risk factors. In this respect, the recent years have seen the growing popularization of using machine learning methods due to their sophisticated possibilities of processing massive and heterogenous medical data, including demographic data, clinical records, behavioral variables, and physiological measurements [2]. The special feature of machine learning algorithms is that they help to identify the hidden patterns and nonlinear relationships among variables and, therefore, are useful in supporting medical decision-making processes and enhancing the precision of forecasting the cardiovascular diseases [3]. The proposed study will design and test machine learning-based predictive models to predict heart attack risk with the use of patient data. After careful data

pre-processing, including missing data, coding of the various characteristics, and the standardization of the numerical features, a complete dataset consisting of numerous characteristics was utilized, including age, blood pressure, cholesterol, and lifestyle factors. Some typical machine learning algorithms were trained and tested and they included, Supporting Vector Machines (SVMs), Decision Trees, and Random Forests [4]. This paper covers the inconsistency of the functioning of models used in predicting heart attacks especially in the ability to discriminate high-risk and low-risk cases. To this end, the models were measured on a variety of statistical performance scores; accuracy, precision, recall, and F1, as well as classification reports and ambiguity matrices, to allow a thorough and trustworthy evaluation of the performance of each of the models. This research will offer a precise and efficient predictive model capable of aiding medical practitioners in determining patients who are highly likely to have heart attack so that preventive measures are achieved in time and help in reducing morbidity and mortality rates that come with heart disease [5]. Among the contributions of the study, there is the creation of a framework of predictive framework grounded on machine learning, systematic comparison of multiple algorithms on the basis of realistic and many-dimensional medical data, and a deep analysis of the model performance in the context of a clinical application. The scientific novelty of the study is seen in a systematic synthesis of a variety of machine learning algorithms into a single framework of predicting heart attacks, aimed at assessing their capabilities to identify cases at the risk and orient the outcomes towards the possible practice which can be incorporated into the smart healthcare systems.

2. BACKGROUNF AND MACHINE LEARNING MODELS

Machine Learning (ML) is a branch of artificial intelligence that allows the computer systems to learn using the information and enhance their operation without the required programming. ML techniques are popular in medical applications since they are capable of analyzing high-dimensional and complex clinical data and make predictive decisions [6]. This work dwells on the supervised learning algorithms because labeled patient data are provided.
Support Vector Machine (SVM) is a supervised learning algorithm that is normally applied in classification tasks. The primary task of SVM is to find an optimum hyper plane which classifies the data points in different classes having the highest distance [7]. The decision of the SVM classifier is provided as:

$$f(x) = \text{sign}(w \cdot x + b) \quad (1)$$

where w represents the weight vector, x denotes the input feature vector, and b is the bias term Fig. 1 illustrates the general architecture of the SVM model.

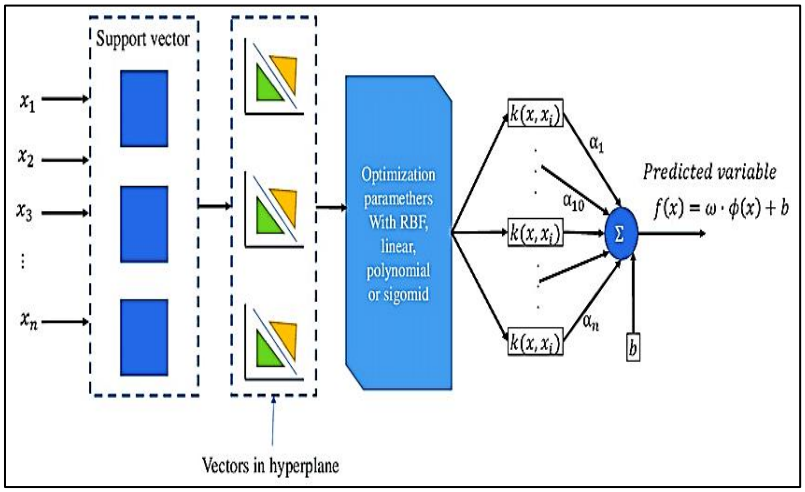

Fig1. SVM Architecture

Decision Tree (DT) is a supervised learning algorithm that follows hierarchical tree-like framework in making decisions based on feature values. The internal nodes are symbolized as decision rules and leaf nodes are represented by class labels [8]. The Fig. 2 indicates the structure of the Decision Tree classifier. The decision tree's formula is essentially a series of if-else statements based on feature values:

```
if (feature_value <= threshold_value)
go to left_child_node
else:
```

go to right_child_node

This is done until one gets to a leaf node and the outcome forecasted by the leaf node is returned [8].

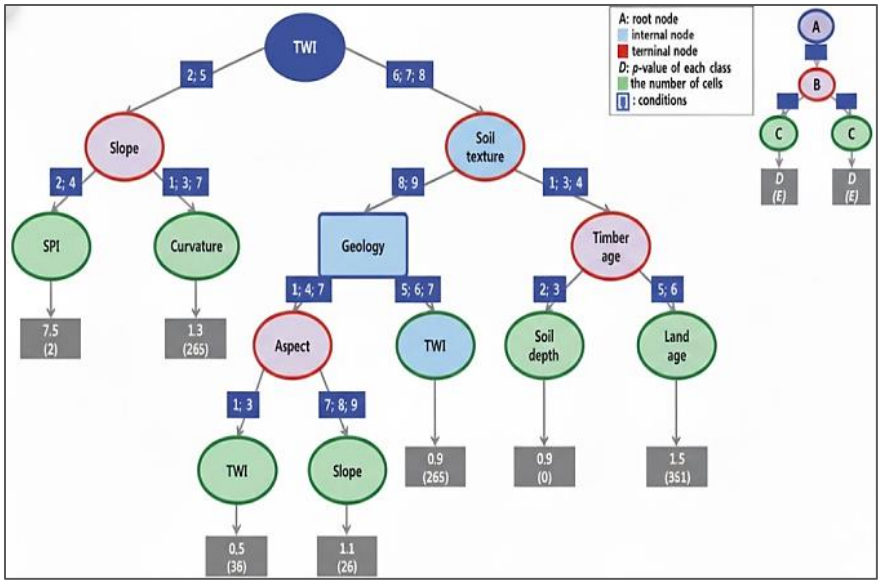

Fig2. Structure of the Decision Tree classifier

Random Forest (RF) is a learning method used as an ensemble that is known to build several decision trees during training and combine their predictions to enhance their classification accuracy and strength [7]. All trees are trained on a random sample of the training data and features. Random Forest is an ensemble based learning algorithm which fits several decision trees during training and averages or votes the prediction of the trees. At each decision tree, the training data and the training features are randomly selected and each decision tree is trained on a random sample of the training data and randomly on a portion of the training features. The last forecast of the Random Forest algorithm is often the average (in regression) or the majority vote (in classification) of the forecasts of the single trees. The prediction of the Random Forest may be written mathematically as:

Prediction = Average(Predictions of all decision trees)

Prediction = Majority Vote(Predictions of all decision trees)

The ensemble architecture of the Random Forest model is depicted in Fig. 3.

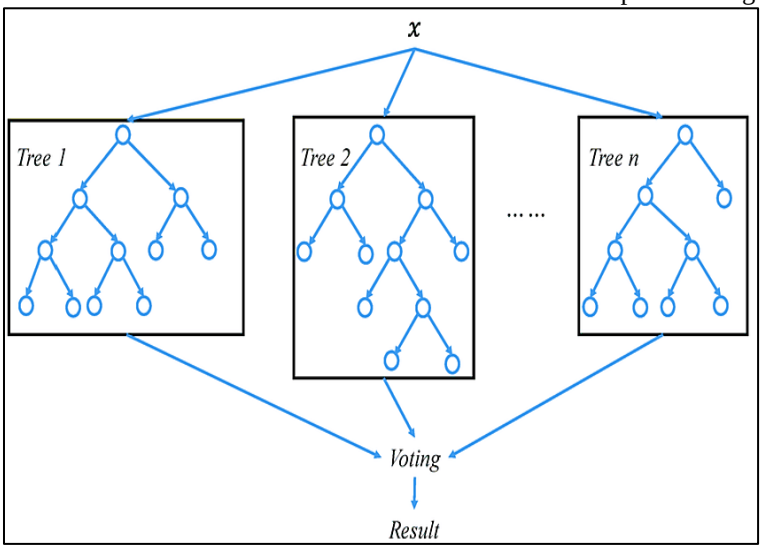

Fig 1. The ensemble architecture of the Random Forest model

2. LITERATURE REVIEW

A number of researches have examined how machine learning methods can be used to forecast cardiovascular events and heart attacks based on various types of data and modeling strategies. Weng et al. (2017) used over 378,000 electronic health records that were taken in the primary care facilities in the United Kingdom. Different machine learning regimes such as Random Forest, Gradient Boosting, and Neural Network, were compared to conventional cardiovascular risk prediction regimes. These findings indicated a favorable change in prediction performance and especially in AUC. Nonetheless, the research study was mainly on the risk of general cardiovascular disease and had poor interpretability of the model [1]. Mandair et al. (2020) used a massive harmonized electronic health record dataset of around 2.27 million patients in predicting the incidence of myocardial infarction. A number of models that included Logistic Regression, Random Forest, Gradient Boosting, and Deep Neural Networks were used. Although deep learning models have gotten high AUC scores, the research noted difficulty in the area of calibration and imbalance of classes that influenced the stability of predictions [2]. Salet et al. (2024) specially examined the issue of predicting acute myocardial infarction with data on 13, 218 cardiovascular patients. Random Forest model has been created and its comparison made with a classic SMART risk model. The suggested method was found to have high predictive performance with the AUC of 0.96, which proves to be promising in AMI detection. However, the validation of the model was done on one dataset alone, and this brings up the issue of generalization [3]. Commandeur et al. (2020) combined clinical data and imaging-based features, such as coronary artery calcium (CAC) and epicardial adipose tissue (EAT), and used an XGBoost model to estimate the risk of long-term myocardial infarction. The proposed model was better in performance than normal risk scoring methods. Although its performance was high, its use is restricted by the requirement to use sophisticated imaging information in the clinical routine practice [4]. Weiss et al. (2012) examined individualized myocardial infarction forecasting with the assistance of electronic health records of the healthcare systems in the United States. Personalized models with machine learning provided a better individual-level prediction than traditional models. Nonetheless, the methods used by the research were fairly old-fashioned machine learning methods, and could not make the most of new developments in the field [5]. Pezel et al. (2025) suggested machine learning fusion frameworks that integrate cardiac CT and MRI images and prediction of cardiovascular events in patients with obstructive coronary artery disease. The model had an AUC of around 0.86 which is good predictive power. However, the costliness and scarcity of the imaging data that are to be used in the proposed approach limit the scaling of the suggested method [6]. Recent work has demonstrated that combining NLP feature extraction with ensemble learning significantly improves phishing email detection accuracy [7]. Emakhu et al. (2022) trained ensemble machine learning models with patient data of the emergency department with a suspected acute coronary syndrome. The models demonstrated better decision support situations within emergency care settings. Nevertheless, it was carried out based on data of one center, which might be the limiting factor in the extrapolation of the results [8]. Sraitih et al. (2022) examined the ECG signal data in detecting the presence of myocardial infarction and tested multiple machine learning algorithms such as SVM, KNN, and Random Forest. The findings indicated high performance even in noisy situations. Although these are the merits, clinical or demographic characteristics were not taken into account in the study, which constrained the applicability of this study to provide a thorough assessment of risk [8].Boudali et al. (2024) used LightGBM and XGBoost, which are advanced boosting methods, on large multi-source clinical data to predict the risk of myocardial infarction. The best overall performance was obtained in LightGBM. Nevertheless, interpretation of the models has been considered as the main limitation to their clinical application[9

Table 1. Survey of Previous Studies

Author (Year)	Dataset	Methodology	Results	Strengths	Gaps / Limitations
Weng et al. (2017)	378,000+ EHR records (UK primary care)	RF, GBM, NN vs traditional models	Improved AUC over traditional risk models	Large-scale real-world data	Limited interpretability; general CVD risk
Mandair et al. (2020)	2.27M harmonized EHR records	LR, RF, GBM, DNN	High AUC for deep models	Very large dataset	Calibration and class imbalance issues
Salet et al. (2024)	13,218 cardiovascular patients	Random Forest vs SMART	AMI AUC = 0.96	Focused AMI prediction	Limited external validation
Commandeur et al. (2020)	Clinical + CAC + EAT imaging data	XGBoost	Outperformed standard risk scores	Multimodal data integration	Imaging data not always available
Weiss et al. (2012)	US EHR patient records	ML-based personalized models	Improved individual prediction	Early personalized approach	Outdated methods
Pezel et al. (2025)	CT and MRI imaging data	ML fusion models	AUC = 0.86	High predictive accuracy	High cost, limited scalability
Emakhu et al. (2022)	Emergency ACS patient data	Ensemble ML models	Improved emergency decision support	Clinical applicability	Single-center data
Sraitih et al. (2022)	ECG signals	SVM, KNN, RF	Robust MI detection	Noise robustness	No clinical data integration
Boudali et al. (2024)	Large multi-source clinical data	LightGBM, XGBoost	LightGBM best performance	Advanced boosting techniques	Limited explainability
Kurucz et al. (2024)	Emergency ACS visit counts	Random Forest Regression	Accurate trend prediction	Healthcare resource planning	Not individual-level prediction

Kurucz et al. (2024) directed their attention to forecasting the emergency department visit trends based on acute coronary syndrome in terms of the Random Forest regression. The given method offered correct predictions of short-term trends applicable in the planning of healthcare resources. However, the research failed to discuss the individual heart attack risk prediction at the level of table 1 presents the summary of the prior research [10]. Summary of Research Gap: Even though current literature shows that machine learning can be successfully used to predict heart attack, there are still several gaps, such as the inability to interpret models, the use of expensive or specialized sources of data, inadequate external validation, and the inability to predict risks at an individual level with the help of available clinical data. Such constraints inspire the creation of strong, interpretable, and generalizable machine learning models, which are the same as those suggested in this study.

3. RESEARCH METHODOLGY

The present research examines the issues of how offensive language can be identified on social media sites, especially Twitter. It employs the latest machine learning techniques such as the Logistic Regression and k-Nearest Neighbors to overcome the difficulties in identifying the offensive language. This is to be done with the aim of developing an adaptive system that will be able to maneuver through the intricacies of offensive expression. The research questions address not only the areas of understanding common forms of offensive language but also the efficacy of the algorithm. The research gives a detailed discussion of the available literature, research methodology, experiment findings, and interpretation [22].

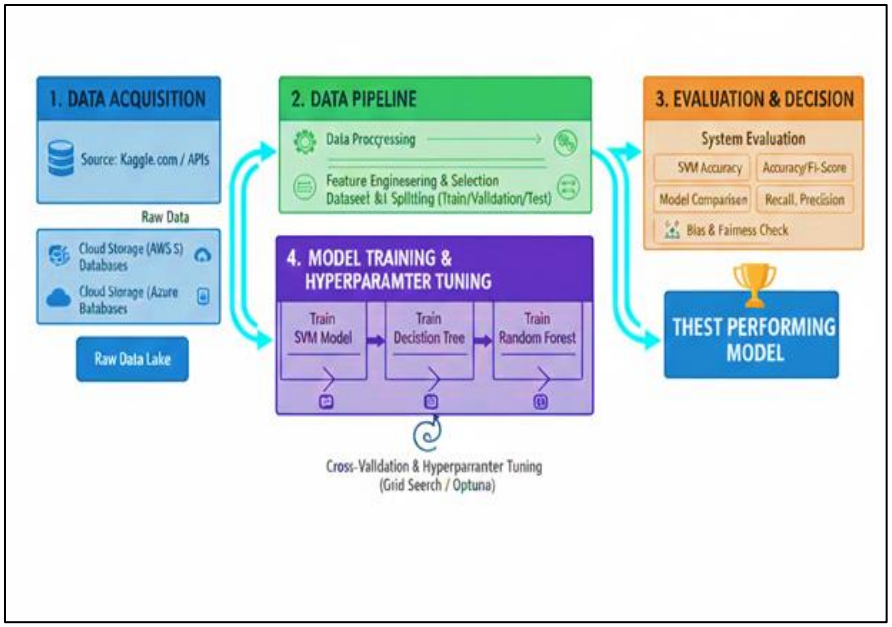

Fig4. Development Phases of Proposed model

3.1Phase One: Data Collection

In order to start the evaluation process, the initial step is the collection of data in Kaggle.com. The Heart Attack Risk Prediction Dataset that could be found in Kaggle is a very valuable source that will help to comprehend the intricate nature of heart health and its predictors. Since the problem of myocardial infarctions is a major health issue worldwide, the given dataset offers an extensive variety of characteristics concerning heart health and lifestyle habits. Attributes comprise patient specific like age, gender, cholesterol, blood pressure plus indicators like diabetes, smoking habits, and obesity. There are also lifestyle elements, medical factors, socioeconomic elements, and geographical elements and they end by creating a binary classification element that specifies whether heart attack risk is present or absent. This data can serve as a foundation in predictive analysis and studies on cardiovascular health and other studies that support preventive measures of heart disease and management.

3.2. Phase Two:

Data Preprocessing In the stage of data preprocessing and feature engineering, numerous steps were used to prepare the data to be used by training the models and its evaluation. These processes are essential in the quality of data, missing values, and feature transformation to learn models [23]. Handling Missing Values: There were missing values in numerical features that were filled with the median of each column. This method will provide resistance to outliers and maintain the distribution of the data [24]. In the case of categorical features, where there were missing values, they were either filled in with the mode (the most common value) or counted as an additional category, based on the nature and importance of the feature. 1.0) Categorical Variables: Given categorical variables, developing a codebook is the initial step to guarantee that users can access them. <|human|>Encoding Categorical Variables: |human|>1.0) Categorical Variables: Encoding Categorical Variables Given categorical variables, the first step is creating a codebook to ensure that the variables can be accessed by the user. One-hot encoding or label encoding methods were applied to categorical variables including gender, smoking status and dietary habits. This conversion is a conversion of categorical variables to numerical format, which is necessary to train a model [25].

Feature Scaling: Numerical characteristics were converted to normal range to make all features equally contribute to the formation of the model and to avoid dominance of those features with more significant values. Such standardization or normalization methods like Min-Max or Z-score were used. Feature Engineering: Existing features were modified to give new ones to represent more information or pattern in the data. As an example, the Body Mass Index (BMI) value was calculated based on height and weight measurements, which presented a more detailed measure of obesity [26]. Handling Outliers: Numerical features with outliers were detected and addressed through trimming, winsorization or transformation to reduce potential effect on the performance of the model. Feature Selection: The availability of relevant features was determined by the features that are important and contribute to the prediction task. Methods such as correlation analysis, ranking features in terms of their importance

as well as domain knowledge were used to select and maintain the most informative features. 3.3. Phase Three: Dataset Preparation.

a. Dataset Splitting:

During this step, the data is divided into the training and the testing data to support the training and testing of the model. The dataset will be segmented into two subsets namely the training set that will be used to train the machine learning models and the testing set that will be used to evaluate the performance of the models on unknown data.The common split ratio is 80-20, in which 80 percent of the data is set aside to the training group, and the other percentage to the testing group. Nonetheless, the split ratio may be changed depending on how big the dataset is and what are the exact demands of the analysis.b. Dataset Preparation: Once the data has been divided, further data manipulation steps are followed to make sure that the data is of the right format to be used to train the model. This may include: Feature scaling: A rescaling of numerical features in order to make models consistent. Encoding features: Converting the categorical variables into numbers with methods such as one-hot encoding or label encoding. Data normalization: Normalization of data distribution to enhance model convergence and model performance. Managing the issue of class imbalance: Managing the problem of class imbalance by using methods such as oversampling, under sampling, or by class weight in the model training. Data augmentation: Learners are enhanced to augment the dataset with synthetic data points to enhance the range and resilience of the training data. The division and proper preparation of the dataset is to guarantee that machine learning models are trained on representative data and tested with unbiased test data, which results in the reliable and truthful evaluation of its performance [27], [28].

3.3. Phase Four : Model Training

Support Vector Machine (SVM) is an effective supervised learning algorithm that is applied in classification and regression problems. In SVM, the algorithm aims at identifying the best hyper plane which best separates the classes in the feature space. Maximizing the distance between the hyper plane and the closest data points (support vectors) of each of the two classes defines the hyper plane. SVM has the capability to perform both the linear and nonlinear classification by applying varying kernel functions; e.g., linear, polynomial, or radial basis function (RBF) kernels. Model training: It is a procedure that is used to train the SVM classifier with respect to the training data in which the algorithm is learned to use the best parameters of the hyper plane in terms of the features and their associated class labels. Decision Tree:

- Decision Tree is a flexible supervised learning algorithm, which is capable of classifying, as well as regressing.
- In Decision Tree, the dataset is recursively broken into subsets according to the values of features and each subdivision is a decision node.
- Each decision node: The algorithm chooses the feature which splits the data best (the purpose is to maximize the purity or homogeneity of the resulting subsets).
- The process is repeated until some stopping criterion is reached, e.g. reaching a maximum tree depth or a minimum number of samples per leaf.
- Decision Trees can be interpreted and visually depicted and hence can be used in exploratory analysis and to gain an insight into the decision-making process of the model.

Random Forest:

- Decision Tree is a flexible supervised learning algorithm, which is capable of classifying, as well as regressing.
- In Decision Tree, the dataset is recursively broken into subsets according to the values of features and each subdivision is a decision node.
- Each decision node: The algorithm chooses the feature which splits the data best (the purpose is to maximize the purity or homogeneity of the resulting subsets).
- The process is repeated until some stopping criterion is reached, e.g. reaching a maximum tree depth or a minimum number of samples per leaf.
- Decision Trees can be interpreted and visually depicted and hence can be used in exploratory analysis and to gain an insight into the decision-making process of the model.

3.3 Phase Five: System Evaluation

In this phase, the performance of the proposed system is evaluated using standard classification metrics commonly applied in machine learning and medical prediction studies. These metrics provide a concise and reliable assessment of the model's effectiveness.

Accuracy measures the overall correctness of the model and is defined as:

$$\text{Accuracy} = \frac{Number\ of\ Correct\ Predictions}{Total\ Number\ of\ Predictions} \quad (1)$$

Precision evaluates the reliability of positive predictions and is defined as:

$$\text{Precision} = \frac{true\ positives}{True\ positives + false\ positive} \quad (2)$$

Recall (Sensitivity) measures the ability of the model to correctly identify actual positive cases and is defined as:

$$\text{Recall} = \text{True Positives} / (\text{True Positives} + \text{False Negatives}) \quad (3)$$

The F1-Score provides a balanced evaluation by combining precision and recall, particularly useful for imbalanced datasets. It is defined as:

$$\text{F1-Score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (4)$$

In addition, the Confusion Matrix is used to analyze classification performance by comparing predicted and actual class labels using True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), providing deeper insight into classification errors.

4. RESULT

This part provides and explains the performance of the proposed machine learning models to predict a heart attack. The assessment has been made based on the standard metrics of classification that comprised of accuracy, precision, recall, F1-score, and confusion matrices in order to have a complete account of every model. Comparative performance of models is described in Fig. 5 and the in-depth results of classification are presented in Table 3-5.

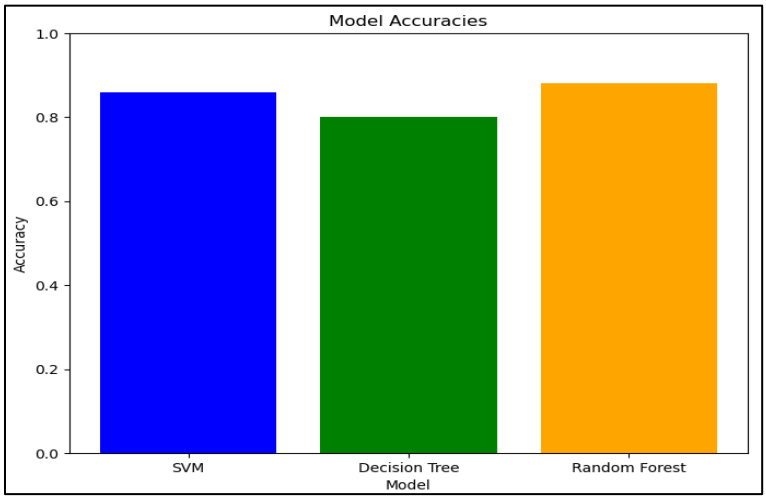

Fig5. Evaluation Performance Effects on RF, DT and SVM

Support Vector Machine (SVM) model was used and had an overall accuracy of 0.86 which showed balanced predictive ability across the classes. The model was found to be precise and recall 0.85 and 0.82, respectively, and recall 0.87 and 0.90, respectively, which yielded F1-scores of 0.83 and 0.88, respectively, as illustrated in Table 2.

Table 2. Classification Report for SVM

Class	Precision	Recall	F1-Score	Support
0	0.85	0.82	0.83	77
1	0.87	0.90	0.88	107
Accuracy	-	-	-	0.86 (184)
Macro Avg	0.86	0.86	0.86	184

Weighted Avg	0.86	0.86	0.86	184

Fig. 6 shows that the SVM model was able to properly recognize 63 true negatives and 96 true positives with a comparatively low set of misclassifications (14 false positives and 11 false negatives) which also indicates that it is fairly reliable in recognizing patients who are at risk of developing type 2 diabetes.

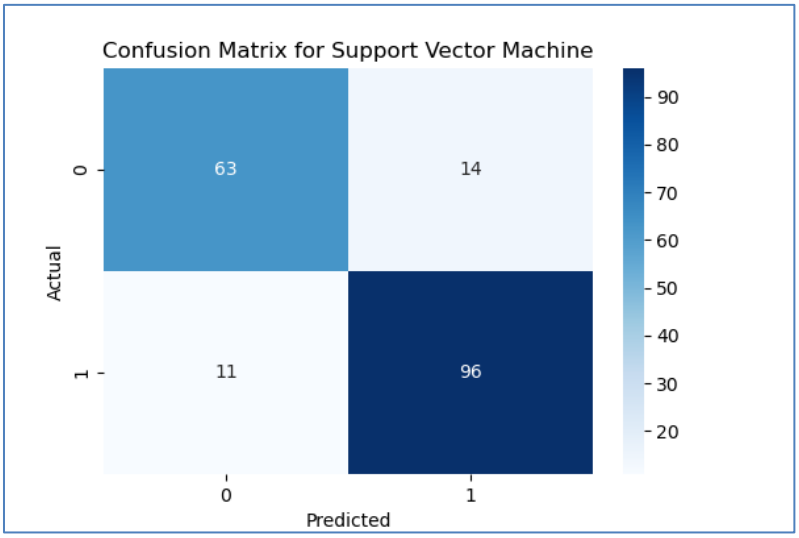

Figure 6. SVM Confusion Matrix

As it can be seen in Table 3, the Decision Tree (DT) model reached a smaller accuracy of 0.79. Despite the decent recall in the model (0.82 in class 0 and 0.79 in class 1), its precision and F1-scores were worse than those of SVM and RF, which means that it is less able to generalize.

Table 3. Classification Report for Decision Tree

Class	Precision	Recall	F1-Score	Support
0	0.72	0.82	0.77	77
1	0.86	0.78	0.81	107
Accuracy	-	-	-	0.79 (184)
Macro Avg	0.79	0.80	0.79	184
Weighted Avg	0.80	0.79	0.79	184

The confusion matrix in Fig.7 reveals a higher number of false negatives (22 cases), which may limit its suitability for critical medical decision-making despite its interpretability.

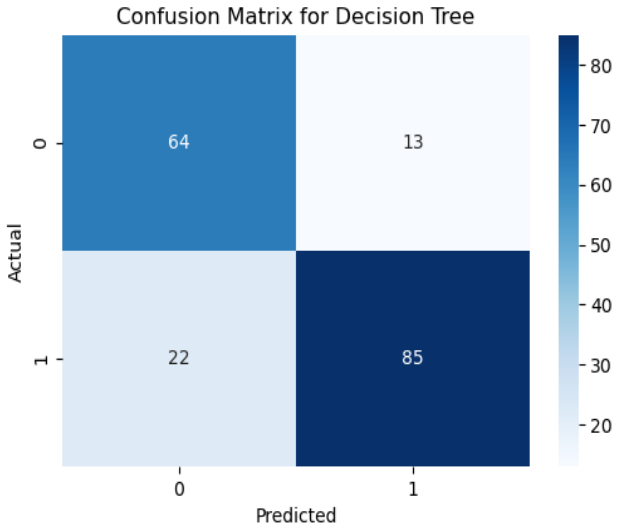

Fig7. DT Confusion Matrix

Among all evaluated models, the Random Forest (RF) classifier delivered the best performance, achieving an accuracy of 0.89, as detailed in Table 4..

Table 0. Classification Report for Random Forest

Class	Precision	Recall	F1-Score	Support
0	0.85	0.88	0.87	77
1	0.91	0.89	0.90	107
Accuracy	-	-	-	0.89 (184)
Macro Avg	0.88	0.89	0.88	184
Weighted Avg	0.89	0.89	0.89	184

Both classes were well-performing in the model, and class 1 was performing exceptionally good in terms of its precision (0.91), recall (0.89), and F1-score (0.90). The RF model (Fig.8) had the largest number of correct classification (68 true negatives and 93 true positives) with the lowest number of misclassification errors. Such high performance can be explained by the ensemble character of the Random Forest algorithm that increases robustness and decreases the overfitting because it combines several decision trees.

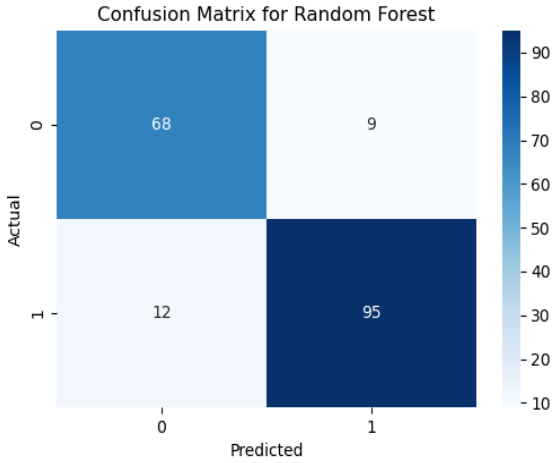

Fig8. RF Confusion Matrix

5. DISCUSSION

This paper has compared three heart attack/heart disease predictive models SVM, Decision Tree, and Random Forest. It was measured using common performance measures, like accuracy, precision, recall/sensitivity, F1 score and ambiguity matrices. According to the comparative results (Figure 4.1), it is clear that the overall performance of the Random Forest model was the highest as compared to the SVM and Decision Tree in most of the metrics. This indicates the power of prediction of patient characteristics and probability that is in complex and non-linear association, which is why ensemble learning techniques can be used to address it. This conclusion is supported by Table 4.3, which presents an Accuracy of 0.89 and high and balanced macro and weighted averages, which suggests the same performance in both categories and no preference to any one of them. Conversely, SVM also presented a balanced and consistent type of performance and was found to have a good capability of distinguishing the two categories and also a high recall of category 1 (presence of disease). This finds relevance in medical applications since the priority is minimization of false negatives to prevent the risk of overlooking at-risk patients. Nevertheless, the results of the SVM were still slightly less than that of the Random Forest which is expected since the SVM is sensitive to the choice of the kernel and the changes in the parameters. Decision Tree was simple to interpret, hence it was used to understand how decisions were made in a clinical setting, although it was the most volatile and inaccurate with regards to overfitting and changing data which was evidenced in the confusion matrix (Figure 4.3) that generated more errors than the Random Forest. The support values also indicate an apparent disproportion of the data (77 and 107 as the support to class 0 and 1 respectively), thus Precision/Recall/F1 and macro and weighted averages were required instead of merely the overall accuracy. In this respect, it can be stated that the most appropriate one to be used in practice as a clinical decision support model is Random Forest, which is the most appropriate in terms of balancing between accuracy and robustness, whereas SVM is a viable option when the sensitivity in disease detection is required, and Decision Tree can be used when the interpretability has the utmost importance, and the latter has to be improved or incorporated into a pooled model to enhance reliability.

The versatility of classifying algorithms in various issue domains has been established by the successful utilization of machine learning algorithms in protection areas, such as phishing attack recognition [20].This paper has compared three heart attack/heart disease predictive models SVM, Decision Tree, and Random Forest. It was measured using common performance measures, like accuracy, precision, recall/sensitivity, F1 score and ambiguity matrices. According to the comparative results (Figure 4.1), it is clear that the overall performance of the Random Forest model was the highest as compared to the SVM and Decision Tree in most of the metrics. This indicates the power of prediction of patient characteristics and probability that is in complex and non-linear association, which is why ensemble learning techniques can be used to address it [21]. This conclusion is supported by Table 4.3, which presents an Accuracy of 0.89 and high and balanced macro and weighted averages, which suggests the same performance in both categories and no preference to any one of them. Conversely, SVM also presented a balanced and consistent type of performance and was found to have a good capability of distinguishing the two categories and also a high recall of category 1 (presence of disease). This finds relevance in medical applications since the priority is minimization of false negatives to prevent the risk of overlooking at-risk patients. Nevertheless, the results of the SVM were still slightly less than that of the Random Forest which is expected since the SVM is sensitive to the choice of the kernel and the changes in the parameters. Decision Tree was simple to interpret, hence it was used to understand how decisions were made in a clinical setting, although it was the most volatile and inaccurate with regards to overfitting and changing data which was evidenced in the confusion matrix (Figure 4.3) that generated more errors than the Random Forest. The support values also indicate an apparent disproportion of the data (77 and 107 as the support to class 0 and 1 respectively), thus Precision/Recall/F1 and macro and weighted averages were required instead of merely the overall accuracy. In this respect, it can be stated that the most appropriate one to be used in practice as a clinical decision support model is Random Forest, which is the most appropriate in terms of balancing between accuracy and robustness, whereas SVM is a viable option when the sensitivity in disease detection is required, and Decision Tree can be used when the interpretability has the utmost importance, and the latter has to be improved or incorporated into a pooled model to enhance reliability.

6. CONCLUSTION

The paper has compared the three models SVM, Decision Tree and Random Forest of predicting the occurrence of heart attack/ heart disease. Common performance measures were used to measure it, such as accuracy, precision, recall/sensitivity, F1 score and ambiguity matrices. Going by the comparative findings (Figure 4.1), it is evident that the general performance of the Random Forest model was the best compared to the SVM and Decision Tree in the majority of metrics. This means that the prediction of patient characteristics and probability is in complex and non-linear association that is why ensemble learning techniques can be applied to solve this issue. Table 4.3 proves this conclusion as the Accuracy is shown as 0.89, high and balanced macro and weighted averages, which indicates the same results in both categories and does not favor any of the categories. On the other hand, SVM too exhibited a balanced and uniform form of performance and was established to possess a good performance of differentiating both categories as well as a high recall rate of category 1 (disease is present). This is applicable in medical practice because minimization of false negatives is the major focus at the expense of the risk of missing at-risk patients. However, the results of SVM were still a little bit lower than that of the Random Forest which is natural because the SVM is weak to the selection of the kernel as well as to the alterations of the parameters. Decision Tree was easy to interpret and thus it was used to learn the decision-making process in a clinical environment though it was the most unstable and inaccurate with reference to overfitting and manipulating data which was also illustrated in the confusion matrix (Figure 4.3) that produced more errors compared to the Random Forest. The support values also show that there is an apparent disproportion of the data (77 and 107 as the support to class 0 and 1 respectively), so Precision/Recall/F1 and macro and weighted averages were necessary as opposed to the accuracy in general. In this regard, one can say that the best one to be applied in practice as a clinical decision support model is the Random Forest one, which is the most appropriate when it comes to a balance between the accuracy and strength, when SVM may be applied when the sensitivity in the disease detection is needed, and when the Decision Tree can be used when the interpretability has the highest priority, and has to be improved or integrated into a pooled model to become more reliable.

REFERENCES

[1] M. Raihan et al., “Risk Prediction of Ischemic Heart Disease Using Artificial Neural Network,” 2nd International Conference on Electrical, Computer and Communication Engineering, ECCE 2019, Apr. 2019, doi: 10.1109/ECACE.2019.8679362.

[2] R. El-Bialy, M. A. Salamay, O. H. Karam, and M. E. Khalifa, “Feature Analysis of Coronary Artery Heart Disease Data Sets,” Procedia Comput Sci, vol. 65, pp. 459–468, 2015, doi: 10.1016/J.PROCS.2015.09.132.

[3] M. Djerioui, Y. Brik, M. Ladjal, and B. Attallah, “Neighborhood component analysis and support vector machines for heart disease prediction,” Ingenierie des Systemes d’Information, vol. 24, no. 6, pp. 591–595, 2019, doi: 10.18280/ISI.240605.

[4] “An Efficient SMOTE-Based Deep Learning Model for Heart Attack Prediction.” Accessed: Mar. 14, 2024. [Online]. Available: https://www.hindawi.com/journals/sp/2021/6621622/

[5] W. Zheng and M. Jin, “The Effects of Class Imbalance and Training Data Size on Classifier Learning: An Empirical Study,” SN Comput Sci, vol. 1, no. 2, Mar. 2020, doi: 10.1007/S42979-020-0074-0.

[6] Dai, S., Meng, F., Dai, H., Wang, Q., Chen, X., Bai, W., ... & Liu, J. (2026). Machine learning in peak demand forecasting: foundations, trends, and insights. Renewable and Sustainable Energy Reviews, 227, 116500.

[7] M. A. S. Al-Hitawi, A. L. Alzaidy, M. A. Alazaizi, and M. A. Tharthar, *“Toolkit for Generating and Augmenting Hungarian Handwritten Text Recognition Dataset,” Dijlah J. Eng. Sci*., vol. 3, no. 1, 2026.

[8] Yue, Y., Ji, S., Yin, W., Wang, Y., Huang, W., Pan, C., ... & Cheng, Y. (2026). Development of wear condition diagnosis model for main bearing with SVM algorithm. International Journal of Engine Research, 27(1), 139-155.

[9] Ferreira, R. C., Dunn-Lopez, K., Moorhead, S., Krupp, A., Zuchatti, B. V., Carvalho, L. A. C., ... & Duran, E. C. M. (2026). Decision Trees for Managing Impaired Physical Mobility in Multiple Trauma Patients. Journal of advanced nursing, 82(2), 1359-1370.

[10] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, “Can machine-learning improve cardiovascular risk prediction using routine clinical data?” PLOS ONE, vol. 12, no. 4, pp. 1–14, 2017, doi: 10.1371/journal.pone.0174944.

[11] G. Mandair et al., “Prediction of incident myocardial infarction using machine learning applied to harmonized electronic health record data,” BMC Medical Informatics and Decision Making, vol. 20, no. 1, pp. 1–12, 2020, doi: 10.1186/s12911-020-01268-x.

[12] N. Salet et al., "Using machine learning to predict acute myocardial infarction and ischemic heart disease in primary care cardiovascular patients," PLOS ONE, vol. 19, no. 4, pp. 1–16, 2024, doi: 10.1371/journal.pone.0307099.

[13] F. Commandeur et al., "Machine learning to predict the long-term risk of myocardial infarction and cardiac death based on clinical risk, coronary calcium, and epicardial adipose tissue: A prospective study," Cardiovascular Research, vol. 116, no. 14, pp. 2216–2225, 2020, doi: 10.1093/cvr/cvz321.

[14] J. C. Weiss et al., "Machine learning for personalized medicine: Predicting primary myocardial infarction from electronic health records," AI Magazine, vol. 33, no. 4, pp. 33–45, 2012, doi: 10.1609/aimag.v33i4.2438.

[15] T. Pezel et al., "A machine learning model using cardiac CT and MRI data predicts cardiovascular events in obstructive coronary artery disease," Radiology, vol. 305, no. 1, pp. 1–11, 2025, doi: 10.1148/radiol.233030.

[16] J. Emakhu et al., "Acute coronary syndrome prediction in emergency care: A machine learning approach," Computer Methods and Programs in Biomedicine, vol. 225, p. 107080, 2022, doi: 10.1016/j.cmpb.2022.107080.

[17] N. A. Mohammed, et al., "Recognizing Phishing in Emails by Using Natural Language Processing & Machine Learning Techniques," in 2025 3rd International Conference on Cyber Resilience (ICCR), 2025, doi:10.1109/ICCR67387.2025.11292212.

[18] M. Sraitih et al., "A robustness evaluation of machine learning algorithms for ECG myocardial infarction detection," Journal of Clinical Medicine, vol. 11, no. 17, p. 4935, 2022, doi: 10.3390/jcm11174935.

[19] I. Boudali et al., "A predictive approach for myocardial infarction risk assessment using machine learning and big clinical data," Healthcare Analytics, vol. 5, p. 100319, 2024, doi: 10.1016/j.health.2024.100319.

[20] M. A. S. Al-Hitawi and G. N. Máté, "Enhancing Transformer-Based Language Models for Hungarian Handwritten Text Recognition," *F1000Research*, vol. 15, p. 181, 2026, doi: 10.12688/f1000research.176408.1.

[21] V. C. Kurucz et al., "Prediction of emergency department presentations for acute coronary syndrome using a machine learning approach," Scientific Reports, vol. 14, no. 1, p. 23125, 2024, doi: 10.1038/s41598-024-73291-1.

BIOGRAPHIES OF AUTHORS

Ammar Saleh Abd received his academic qualifications and is currently affiliated with the Anbar Governorate Office, Anbar, Iraq. He works at the Anbar Governorate Office in Al-Ramadi, Anbar, Iraq.
He can be contacted at: ammarsalih307@gmail.com