Academic Journal of Nawroz University (AJNU), Vol.12, No.4, 2023

https://doi.org/10.25007/ajnu.v12n4a1974

A Machine Learning Model for the Prediction of Heart Attack Risk in High-Risk Patients Utilizing Real-World Data

Ridwan B. Marqas [1, 2], Abdulazeez Mousa [1], Fatih Özyurt [2] and Rojhat Salih [1]

1 Department of Computer Science, College of Science, Nawroz University, Duhok, Iraq
2 Department of Software Engineering, College of Engineering, Firat University, Elazig, Turkey

ABSTRACT: Heart disease is a significant global public health concern that impacts a vast number of individuals worldwide. The early identification of patients at risk of heart attack can significantly reduce mortality rates. In this research study, we employed machine learning methods to develop a model that predicts the likelihood of a heart attack. To create the model, we collected a real-world dataset of patient features, including demographic information, medical history, and lifestyle factors. We pre-processed the data to eliminate any missing values and standardized the features to ensure uniformity across the dataset. Additionally, we utilized feature engineering techniques to identify the most significant factors that contribute to the development of heart attacks. We evaluated several machine learning algorithms such as logistic regression, decision trees, and random forest to identify the most effective ones based on traditional metrics including accuracy, precision, recall, F1-score, Mathew correlation, ROC, and AUC. Our algorithm produced highly accurate predictions for heart attack risk. Our results demonstrate that machine learning algorithms can effectively predict heart attacks and identify high-risk patients. The model can be integrated into electronic health records to facilitate prompt identification and intervention by healthcare providers. However, our study has limitations that need to be addressed, based on the aforementioned results, it is apparent that the XGBoost Classifier demonstrates greater performance. This claim is supported by the significant test accuracy of 0.9191, a sensitivity of 0.943, a specificity of 0.89, a f1-score of 0.9243, and the minimum Log Loss value of 2079. The Extra Tree Classifier demonstrates the greatest mean area under the curve (AUC) of 0.950 when evaluated using the AUC metric. and more diverse dataset as well as the challenge of interpreting the model. Future research may incorporate additional data sources, advanced machine-learning techniques, and improved model interpretability. Our heart attack prediction model holds significant potential as a valuable tool for healthcare practitioners to identify high-risk patients and decrease heart attack rates.

Keywords: Heart Disease, Machine Learning, Prediction Model, High-Risk Patients, Mortality Rates.

1. Introduction

The World Health Organization (WHO) classifies and reports on various causes of death worldwide through its International Classification of Diseases (ICD) system [1]. Myocardial infarctions, also known as heart attacks, are a major cause of death and disability worldwide. According to the WHO, cardiovascular diseases (CVDs) are responsible for approximately 31% of global deaths, and over three-quarters of these deaths occur in low- and middle-income countries (LMICs) [2, 3]. In the United States, heart disease causes more than 600,000 deaths annually, accounting for approximately one in every four deaths [4]. Early identification and prevention of heart attacks can have a significant impact on patient outcomes and alleviate the burden on healthcare systems. Various demographic, lifestyle, and clinical factors have been found to be linked to an increased risk of heart attacks, including age, sex, smoking, hypertension, diabetes, and hyperlipidemia [5]. Machine learning has revolutionized disease detection by enabling the development of predictive models that analyze vast datasets to identify subtle patterns and anomalies, aiding in early diagnosis and intervention. In recent years, machine learning techniques have been increasingly utilized to predict the likelihood of heart attacks based on these factors [6].

Annually, the American Heart Association collaborates with the National Institutes of Health to disseminate the latest insights pertaining to heart disease, stroke, and cardiovascular risk factors. This document provides comprehensive information regarding crucial health behaviors, including smoking,

physical activity, nutrition, and weight management. Additionally, it addresses significant health indicators such as cholesterol levels, blood pressure, and glucose regulation, all of which exert an influence on cardiovascular well-being. The Statistical Update provides up-to-date data on many significant clinical illnesses linked to cardiovascular and circulatory diseases. Stroke, congenital heart disease, rhythm abnormalities, subclinical atherosclerosis, coronary heart disease, heart failure, valve disease, venous disease, and peripheral artery disease are among the various ailments that fall within this category. The report additionally includes details regarding interconnected outcomes, such as the quality of therapy, procedures, and economic costs [4]. In 2020, the age-adjusted death rate per 100,000 for non-Hispanic White males with coronary heart disease (CHD) was recorded at 128.5. Similarly, non-Hispanic Black males had a higher age-adjusted death rate of 153.6, while Hispanic males had a comparatively lower rate of 102.2. The prevalence was found to be 63.8 for non-Hispanic White females, 85.9 for non-Hispanic Black females, and 54.2 for Hispanic females [7].

Machine learning has had a profound impact on disease prediction, transforming the landscape of healthcare in recent years. By harnessing the power of advanced algorithms and large datasets, machine-learning models can analyze complex patterns and subtle correlations within medical information, allowing for early and accurate disease prediction [8]. This technology has significantly improved the ability to forecast a wide range of diseases, from cancer and cardiovascular conditions to infectious outbreaks. The impact is twofold: it enables healthcare providers to identify high-risk individuals for timely intervention, potentially saving lives, and it supports public health agencies in proactive surveillance and resource allocation, helping to mitigate the spread of diseases on a broader scale. As machine learning continues to evolve, its role in disease prediction promises to enhance healthcare outcomes and reduce the economic and human costs associated with preventable illnesses [9]. XGBoost, a powerful machine learning algorithm, has had a significant impact on heart disease detection. Its ability to handle complex, high-dimensional datasets and perform accurate classification has made it a valuable tool in this context [10].

Disease detection based on AI leverages sophisticated algorithms and machine learning models to analyze medical data, such as images, genomic information, and patient records, enabling the early identification of diseases with unprecedented accuracy [11]. The primary aims of this study are to investigate the association between diverse demographic, lifestyle, and clinical variables and the likelihood of experiencing myocardial infarctions. The objective of this study is to construct and assess a machine-learning model for the purpose of predicting heart attacks. This will be achieved by utilizing a dataset consisting of patient information and clinical measurements. In order to assess and contrast the efficacy of various machine learning algorithms and feature selection techniques in the domain of heart attack prediction, a comparative analysis is conducted. In light of the study's findings, it is imperative to offer valuable insights and evidence-based suggestions to healthcare providers and policymakers. The study's scope is constrained to the examination of a solitary dataset comprising patient information and clinical measurements. The model created in this investigation is specifically designed for research objectives and should not be employed as a replacement for clinical diagnosis or therapy. The structure of the paper will be as follows, the second part will contain some literature review while third part will include methodology, and the fourth one results. The last part will be the conclusion overall.

2. Literature Review

In recent years, there has been a notable increase in the prevalence of heart disease, leading to its prominence as a significant contributor to mortality on a global scale. Therefore, the investigation and mitigation of myocardial infarctions have become noteworthy focal points of inquiry within the healthcare domain [12]. The utilization of machine learning has become a highly promising methodology for effectively tackling the task of assessing and predicting occurrences of heart attacks through the examination of patient data [13].

This section provides an extensive assessment of the existing research pertaining to the application of machine learning techniques in the domain of heart attack prediction. The study primarily examines three key research areas, including clinical-based techniques, feature-based approaches, and hybrid approaches.

Methods that are based on clinical practice make use of clinical data and records of medical treatment in order to make predictions regarding the occurrence of heart attacks. In most cases, these methods make use of common statistical models, such as logistic regression, decision trees, and support vector machines (SVM), in order to identify relevant attributes and build prognostic models [13]. In-depth information about deep learning's use in cardiovascular medicine is given in this article. It goes over the useful applications of machine learning for heart attack prediction and other cardiovascular problems. The authors emphasize the potential of deep learning models in the analysis of intricate medical data to enhance risk assessment [14]. The primary objective of this research is to investigate the application of artificial intelligence techniques in the identification of individuals with atrial fibrillation (AF) based on electrocardiograms (ECGs) displaying normal sinus rhythm. The research showcases the capacity to identify atrial fibrillation (AF) during standard electrocardiogram (ECG) examinations through the development of an artificial intelligence (AI) system. The precise identification of atrial fibrillation (AF) is of utmost importance in mitigating cardiovascular consequences, such as myocardial infarctions [15].

Feature-based strategies in the domain of healthcare involve the strategic utilization of patient data by meticulously identifying and constructing meaningful features. These methodologies then utilize machine learning methods, such as neural networks, random forests, and gradient boosting, to formulate predictive models. The likelihood of getting a heart attack was forecasted in a study conducted by Ebrahim Zadeh et al. (2018) using a decision tree and random forest. The aforementioned forecast was derived from an amalgamation of demographic, behavioral, and clinical attributes. The researchers achieved a significant level of accuracy, with their forecasts yielding a respectable rate of 84% [16].

Cai et al. (2020) employed a convolutional neural network (CNN) for the purpose of forecasting cardiac ailments by examining electrocardiogram (ECG) signals, thereby aligning with the aforementioned methodology. The investigation carried out by the researchers yielded a notable degree of precision, exhibiting an accuracy rate of 97% [17]. Hybrid techniques endeavor to integrate clinical and feature-based approaches, capitalizing on the distinct advantages offered by each approach. Ahmadian and colleagues (2019) employed a hybrid research methodology in their investigation to generate prognostications pertaining to cardiovascular disease. The methodology utilized in this research entailed integrating clinical characteristics, including age, gender, and smoking behavior, in conjunction with electrocardiogram (ECG) measurements, notably QRS duration and ST segment deviation. The researchers utilized a random forest model and obtained a significant accuracy rate of 89% [18].

Overall, the existing evidence suggests that machine learning has shown effectiveness in predicting heart attacks. Moreover, there is evidence to suggest that the combination of clinical and feature-based techniques in hybrid methodology may result in higher levels of accuracy when compared to using either strategy independently. However, it is crucial to recognize that there are multiple barriers and limitations that require additional examination in subsequent research endeavors the utilization of machine learning techniques in the prediction of heart attacks shows potential, however, it is crucial to acknowledge and address numerous limitations associated with this approach. Significant hurdles in this sector encompass various aspects, including data quality and quantity, model interpretability, concerns over data privacy and security, potential biases, fairness issues in algorithms, clinical validation, and ethical considerations. Numerous studies have exhibited notable levels of precision, such as the 97% accuracy attained by Cai et al. (2020) through the utilization of a convolutional neural network, as well as the 84% accuracy achieved by Ebrahimzadeh et al. (2018) employing a decision tree and random forest. However, it is imperative to ascertain the clinical validation, ethical implementation, and absence of bias in these models to optimize their efficacy in prognosticating and averting heart attacks, while simultaneously safeguarding patient data and upholding fairness in the realm of healthcare. These encompass concerns related to the quality and quantity of data, the capacity to understand and explain models, and the ethical considerations that need to be considered. The next sections will provide a more in-depth examination of these challenges.

3. Methodology

Data mining in heart disease detection involves the application of advanced analytical techniques to

extract valuable patterns, insights, and trends from large datasets related to heart health [19]. This chapter presents a comprehensive account of the practical implementation of our heart attack prediction model. The initial focus of our discussion pertains to the data preprocessing procedures that were undertaken to adequately prepare the dataset for the purpose of training the model. Subsequently, we explicate the process by which the machine learning algorithm(s) were chosen for our model, as well as the manner in which we conducted training and evaluation to assess the efficacy of our model. Ultimately, we proceed with an examination of the efficacy of our implemented model and offer suggestions for enhancement. The results of our study indicate the potential efficacy of employing machine learning algorithms in the prediction of heart attack risk. Furthermore, our findings highlight the need for additional investigation and advancement in this particular domain. The machine learning algorithm The Random Forest Classifier is commonly used for classification. An ensemble technique predicts using several decision trees. Each choice tree This model is known for its data analysis accuracy. It handles categorical and numerical data well, making it versatile. [20] noted that this model resists overfitting. Limitations: This method may be harder to interpret and explain, reducing interpretability. When utilized with large datasets, this method may be computationally intensive. Popular machine learning algorithms for classification and regression include KNN. It uses data similarity to forecast non-parametrically.

The Cover and [21] approach is user-friendly due to its simplicity and intuitiveness. It also handles non-linear data nicely. Another benefit is that it requires no training. KNN has various drawbacks. First, KNN is sensitive to outliers, therefore extreme values in the dataset can dramatically affect prediction accuracy. Since it calculates the distances between the query instance and all other instances in the training set, KNN can be computationally expensive, especially during prediction. Finally, the value of K, which represents the number of nearest neighbors, can greatly affect the K results. Neural Networks, or Multi-layer Perceptron, are computer models with numerous layers of interconnected neurons. Application to complex and non-linear problems is a strength of this method. It can also become hierarchical, according to [22]. Limitations: The model performs poorly when data is scarce since it requires a lot of data. To get optimal results, the model's hyperparameter tuning sensitivity must be adjusted carefully. Overfitting occurs when the model becomes too specialized to the training data and performs poorly on new data. Extreme Gradient Boosting (XGBoost) is a machine-learning technique.

In [23] model has excellent projected accuracy, good missing data management, feature selection, and regularization. Limitations: Without proper calibration, the model may overfit and use a lot of processing power. The Extra Tree Classifier is an ensemble machine-learning technique. It can handle high-dimensional classification jobs. This approach reduces overfitting better than Random Forest. It handles high-dimensional data well and trains faster [24]. Weaknesses: The model is less interpretable and may perform worse than Random Forest on some datasets. SVC is a machine learning classification algorithm.

This method excels with high-dimensional data. Cortes and [25] noted that the kernel method improves efficiency with non-linear data. Limitations: Large datasets require careful kernel selection and processing costs. SGD is a popular machine learning and statistical modeling optimization approach. The system under examination excels at managing and interpreting large datasets. Good performance in linear issues. Limitations: Feature scaling may affect algorithm performance. The method may also need careful hyperparameter tweaking for optimal results. The AdaBoost classifier uses machine learning to combine weak classifiers into one strong one [26]. Technique [27] can integrate weak learners into a powerful model. Machine learning is prone to overfitting, but our method resists it. The technique can handle category and numerical data, making it versatile and applicable in many fields. The model is vulnerable to noise and outliers. The Decision Tree Classifier classifies using machine learning. It's popular in data mining and predictive analytics.

Algorithm constructs tree easy interpretation is a strength of the [28] method. This approach handles numerical and categorical data well. Additionally, this approach is robust to outliers. Limitations: Possible overfitting, and inability to capture complex data links. Gradient Boosting Machine (GBM) is a sophisticated machine learning technique used in many industries. An ensemble learning method that combines several High anticipated accuracies is a strength of the [29] method. This method also handles missing data, feature selection, and regularization well. Limitations: Without proper calibration, the model may overfit and use a lot of processing power. The algorithm chosen depends on the task, dataset, and

trade-offs between accuracy, interpretability, and processing resources. Experimenting with multiple algorithms and carefully adjusting hyperparameters is often recommended to find the best solution.

In this chapter, a comprehensive examination is presented on the implementation of our heart attack prediction model. The discussion encompasses the various stages involved, such as data preprocessing, model training and evaluation, and deployment without a graphical user interface (GUI).

3.1 Data Collection (Datasets)

The dataset utilized in this study consists of a combination of three well-established datasets, namely Hungary, Cleveland, and Statlog [30]. The dataset comprises a total of 1190 records, whereby each record encompasses 11 distinct attributes along with a target variable. The dataset comprises six variables that are of a nominal type and five variables that are of a numeric nature. In the next section, each characteristic will be explained thoroughly in a comprehensive manner.

a. Age: The age of the patients, measured in years (numerical value).
b. Gender: The variable indicating the gender of the patient, with a value of 1 representing male and 0 representing female. This variable is considered nominal.
c. Classification of Chest Pain Type: The type of chest pain reported by the patient is grouped into four distinct groups: 1) typical chest pain, 2) typical angina, 3) non-anginal pain, and 4) asymptomatic pain (nominal).
d. Resting Blood Pressure (BP): The numerical measurement of blood pressure in millimeters of mercury (mm/HG) at a state of rest.
e. Serum cholesterol levels are measured in milligrams per deciliter (mg/dL).
f. A fasting blood sugar level greater than 120 mg/dl is considered 1 for true and 0 for false (nominal) in terms of representation.
g. The resting electrocardiogram (ECG) yields three separate values to indicate the results obtained during a period of rest. 0: Normal 1: ST-T wave abnormality 2: Nominal left ventricular hypertrophy
h. The maximum heart rate reached, denoted by a numeric value, is referred to as the maximum heart rate.
i. Exercise-induced angina: Angina 0 showing NO 1 showing Yes (Nominal)
j. The variable "old peak" refers to the magnitude of ST-segment depression generated by exercise, as compared to the resting state. This variable is measured numerically.
k. The ST slope refers to the measurement of the ST segment's slope during the peak workout phase, specifically at time 0. Upsloping refers to an upward incline or slope. The second data point represents a flat slope, while the third data point represents a downsloping (nominal) slope.
l. The dependent variable target: The aim variable in this study pertains to the prediction of heart risk in patients. A value of 1 indicates that the patient is experiencing heart risk, while a value of 0 indicates that the patient is considered normal.

Hungarian heart disease data, this study used the Hungary Heart Disease Dataset, a subset of the heart disease dataset. The dataset includes medical data from Hungarian individuals, including age, gender, blood pressure, cholesterol, and heart disease, the target variable. The dataset description may not necessarily specify data collection methods. The dataset likely uses patient medical records and exams as its main sources. Academic research may have selection, confirmation, publication, and researcher biases. Biases can affect Source-related biases and may impair the dataset's representation of Hungarian demographics. The data-gathering method and any pre-processing or selection criteria might also introduce bias. The study has flaws. Absent or insufficient data is a major limitation. The approach may also not account for many demographic and geographical characteristics, limiting its application to different populations.

Cleveland Heart Disease Dataset, also known as Cleveland Clinic Heart Disease Dataset. The Cleveland

Heart Disease Dataset was obtained from the acclaimed Cleveland Clinic Foundation in the US. Diagnostic tests, clinical examinations, and medical record retrieval are common data collection methods. Age, sex, chest pain kind, cholesterol levels, and cardiac disease are included in this criterion. Potential biases: Due to the Cleveland Clinic's patient demographic, biases may arise, limiting the generalizability of findings. Due to data gathering throughout time, temporal biases may have occurred. Missing and incomplete data limit the dataset. The model may also overlook confounding variables or comorbidities that may affect heart disease risk. The Statlog (Heart) Dataset was used in this investigation. The Statlog project provides benchmark datasets for machine learning research, including the Statlog Heart Dataset. The dataset does not specify data collection methods. However, it typically includes age, gender, cholesterol levels, and heart disease diagnosis information. Biases may include: Since the dataset was developed for benchmarking, it may not fully capture clinical complexity and subtleties. Data selection and compilation for Statlog may introduce biases. Limitations: The dataset used in this work may lack the complexity of medical data, omitting important factors that affect heart disease prediction. Additionally, this study's sample size may be modest compared to bigger clinical datasets.

3.2 Preprocessing data

In the preprocessing step, the data is encoded and checked for the presence of duplicated or null values. If a small number of such values are found, they are eliminated from the dataset. However, if a large number of such values are present, they are filled with the average value of the corresponding feature column [31]. In our specific scenario, the decision was made to drop the values rather than fill them. Next, the categorical values are encoded and distinguished from the numerical values. Next, we proceed to quantify the prevalence of a specific symptom within our patient population. The dataset presented in the image below illustrates the outcomes pertaining to the symptoms observed.

```
dt['chest_pain_type'].value_counts()
```

```
asymptomatic        625
non-anginal pain    283
atypical angina     216
typical angina       66
Name: chest_pain_type, dtype: int64
```

```
dt['rest_ecg'].value_counts()
```

```
normal                          684
left ventricular hypertrophy    325
ST-T wave abnormality           181
Name: rest_ecg, dtype: int64
```

```
dt['st_slope'].value_counts()
```

```
flat           582
upsloping      526
downsloping     81
0                1
Name: st_slope, dtype: int64
```

```
#dropping row with st_slope =0
dt.drop(dt[dt.st_slope ==0].index, inplace=True)
#checking distribution
dt['st_slope'].value_counts()
```

```
flat           582
upsloping      526
downsloping     81
Name: st_slope, dtype: int64
```

Figure 1: Number of presenting symptoms in the patient.

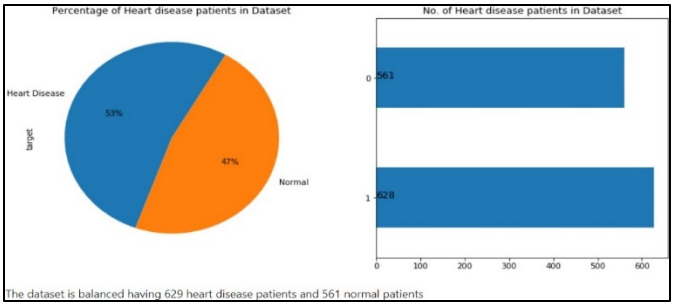

Figure 2: Number of presenting symptoms in the patient.

In this analysis, we investigate a fundamental aspect in the field of machine learning, namely, the equilibrium of our dataset. The existence of an imbalanced dataset may result in a reduction in the precision of our model. According to Figure 2, the dataset is balanced. Therefore, we will proceed with the analysis of the dataset. Next, we will examine the distribution of individuals based on their sex and age.

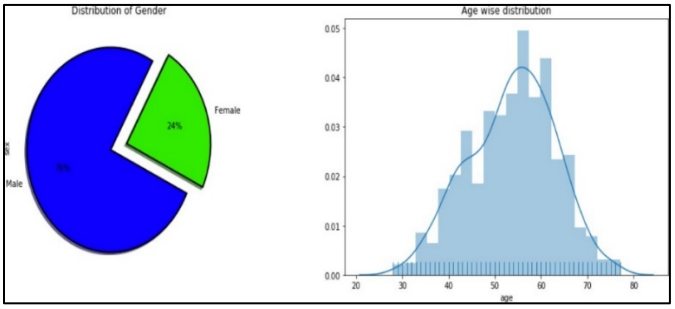

Figure 3: illustrates the distribution of gender and age groups.

Based on the depicted plot, it is evident that the proportion of males in the dataset significantly exceeds that of females. Additionally, the average age of the patients is approximately 55 years. Subsequently, distinct data frames were generated to represent individuals without any cardiac ailments and those diagnosed with heart disease.

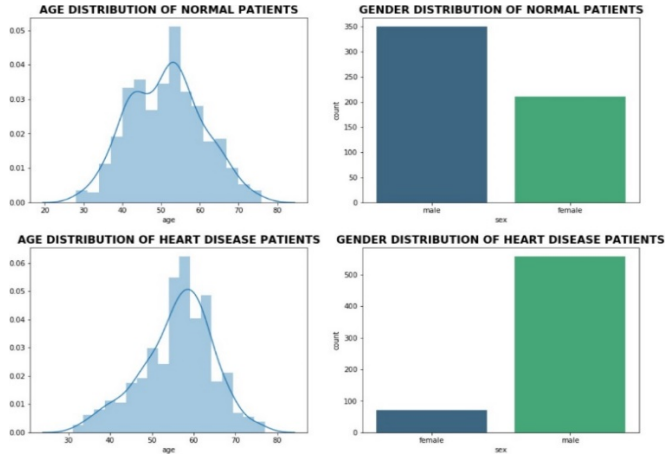

Figure 4: Comparing the outcomes of those with normal cardiovascular health vs those diagnosed with heart disease.

The data presented in Figure 4 indicates a higher prevalence of heart disease among male patients as compared to female patients. Additionally, the average age of individuals diagnosed with heart disease falls within the range of 58 to 60 years. Next, data frames are generated for further features. Classification of Chest Pain.

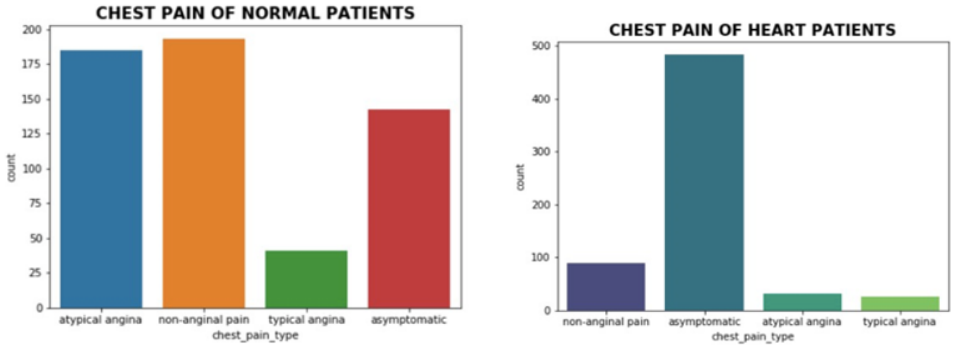

Figure 5: Chest pain types.

As indicated by the aforementioned figure, it is observed that 76% of individuals with heart disease have asymptomatic chest pain. Silent myocardial infarction (SMI), commonly referred to as asymptomatic heart attack, constitutes around 45-50% of cardiac morbidities and premature fatalities in India on an annual basis. The prevalence of severe mental illness (SMI) is twice as high among middle-aged males compared to females. The symptoms of Sudden Myocardial Infarction (SMI) are very modest compared to those of a full-fledged heart attack, leading to its characterization as a silent killer. In contrast to the symptoms commonly associated with a typical heart attack, such as severe chest pain, sharp pain radiating to the arms, neck, and jaw, quick onset of breathlessness, perspiration, and dizziness, the symptoms of SMI are of short duration, leading to confusion with ordinary discomfort and frequently resulting in disregard. Next, a data frame was generated for the electrocardiogram (ECG) in the following manner.

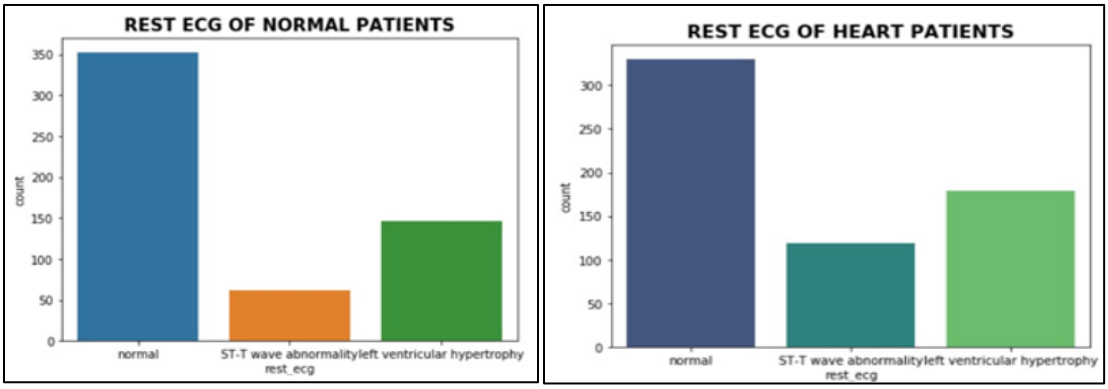

Figure 6: ECGs at rest of healthy and heart disease patients.

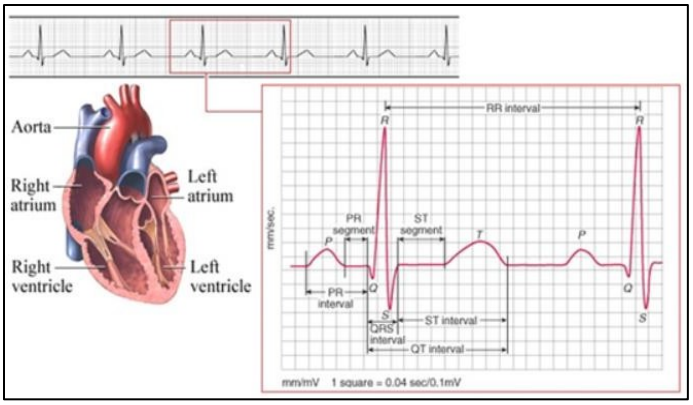

Figure 7: ECG waves at rest.

The electrocardiogram (ECG) is a diagnostic tool that captures and documents the electrical impulses generated by the heart. The test in question is widely employed for the purpose of identifying cardiac abnormalities and assessing the cardiovascular condition in various contexts. Electrocardiograms, commonly referred to as ECGs or EKGs, are diagnostic tools used in the field of cardiology. However, it is important to acknowledge that electrocardiography (ECG) does have certain limitations. The device is capable of assessing heart rate and rhythm; however, it does not inherently provide information regarding arterial blockages. Hence, within the confines of this dataset, it is observed that approximately 52% of individuals diagnosed with cardiac disease exhibit normal electrocardiogram (ECG) readings. Next, a data frame of ST_slope was generated.

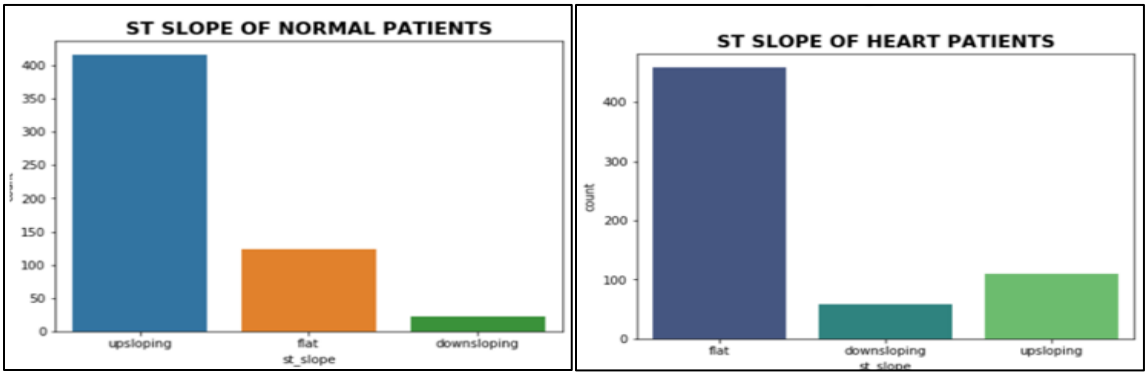

Figure 8: The findings pertaining to ST-segment slope.

The ST segment/heart rate slope (ST/HR slope) has been suggested as a more precise electrocardiogram (ECG) criterion for the diagnosis of severe coronary artery disease (CAD) in the majority of academic research publications. The data presented in the above plot indicates that an upsloping pattern is observed in a majority of normal patients, with approximately 74% exhibiting this characteristic. Conversely, a flat-sloping pattern is prevalent among cardiac patients, with approximately 72.97% displaying this feature. The current data frames consist of category values. In the subsequent steps, we will proceed to generate data frames specifically designed to accommodate numerical values.

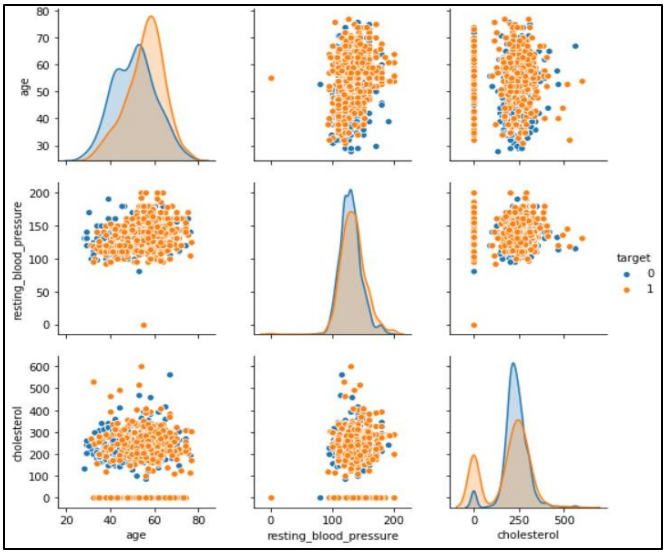

Figure 9: The numerical representation of data frames.

Based on the aforementioned plot, it is evident that there exists a positive correlation between age and the likelihood of developing heart disease. Based on the above plot, outliers are evident as certain patients exhibit a cholesterol level of 0. Additionally, one patient displays both a cholesterol level and resting blood pressure of 0, which might potentially be attributed to missing data entries. To address this issue, we will apply outlier filtering techniques. In order to accomplish this, the z-score equation has been employed.

3.3 Detecting outliers with the Z-score

An outlier refers to a data point that deviates significantly from the central tendency of a given dataset, either by being exceptionally large or very small. The observed discrepancy could perhaps be attributed to a data entering error, or it may indeed reflect authentic data. See Fig. 10.

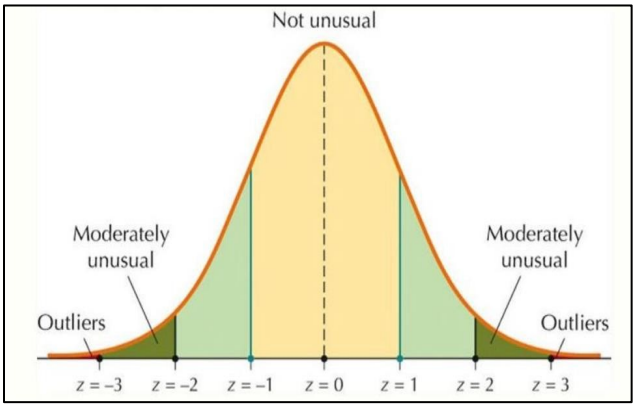

Figure 10: The z-score functionality.

$$Z = \frac{x - \mu}{\sigma} \qquad (1)$$

(Score = x; Mean = μ; SD = σ)

Additionally, a threshold was employed in order to obtain precise outcomes. Subsequently, any values that were identified as outliers will be excluded. This section outlines the methodology employed for data set preparation and preprocessing.

3.4 Machine Learning Models and Techniques

In order to identify the most effective machine learning algorithm, we conducted an extensive evaluation of several prominent algorithms, including the Random Forest Classifier, K Nearest Neighbors, Multi-layer Perceptron, XGBoost, Extra Tree Classifier, Support Vector Classifier, Stochastic Gradient Descent, AdaBoost Classifier, Decision Tree Classifier, and Gradient Boosting Machine. These algorithms were collectively employed in our machine-learning model.

3.5 Model Training

In this stage, we will construct various baseline models and employ a 10-fold cross-validation technique to identify the most effective baseline models for implementation in level 0 of the stacked ensemble approach. Which method yielded more precise and reliable outcomes.

3.6 Model Evaluation

In this stage, we will initially establish the assessment measures that will be employed to assess our model. The evaluation metrics that hold significant importance in this particular problem domain are sensitivity, specificity, precision, F1-measure, geometric mean, Mathew correlation coefficient, and the receiver operating characteristic (ROC) area under the curve (AUC). The Mathew Correlation Coefficient (MCC) is a statistical measure used to assess the quality of binary classification models [32]. The Matthews correlation coefficient (MCC) is a statistically robust measure that yields a high score only when the prediction demonstrates good performance across all four categories of the confusion matrix (true positives, false negatives, true negatives, and false positives), taking into account the relative proportions of positive and negative elements in the dataset.

$$MCC \frac{\mathrm{TP}.\mathrm{TN} - \mathrm{FP}.\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP}).(\mathrm{TP}+\mathrm{FN}).(\mathrm{TN}+\mathrm{FP}).(\mathrm{TN}+\mathrm{FN})}} \quad (2)$$

(worst value: -1; best value: +1)

3.7 Log Loss

The logarithmic loss metric evaluates the efficacy of a classification model by considering the prediction input as a probability value within the range of 0 to 1. The objective of our machine learning models is to minimize the aforementioned value. An ideal model would exhibit a log loss value of 0. The logarithmic loss metric exhibits an upward trend as the anticipated probability deviates from the true label. Predicting a chance of 0.012 when the actual observation label is 1 would be considered unfavorable and would provide a substantial log loss.

The following graph illustrates the spectrum of potential log loss values in relation to a verifiable observation (isDog = 1). As the expected probability tends towards 1, the log loss gradually diminishes. As the anticipated probability declines, the log loss exhibits a quick increase. The logarithmic loss metric imposes penalties on both sorts of errors, with a particular emphasis on predictions that exhibit high confidence but are ultimately incorrect.

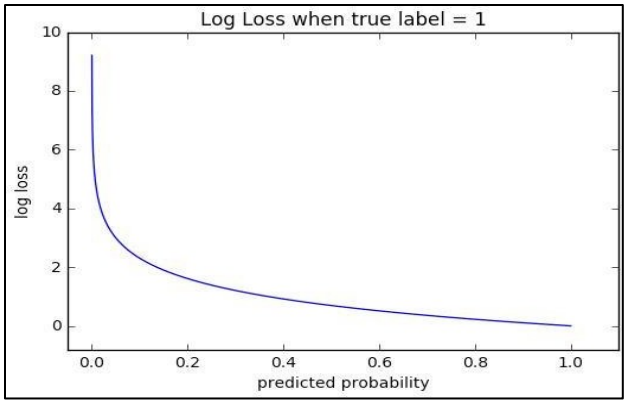

Figure 11: log loss graph.

3.8 F1 Score

The F1 Score can be defined as the arithmetic mean of Precision and Recall, with equal weights assigned to both metrics. Hence, this metric considers the inclusion of both erroneous positive and erroneous negative outcomes. From a cognitive perspective, comprehending F1 may be a greater challenge compared to accuracy [33]. However, F1 generally proves to be more advantageous than accuracy, particularly in scenarios where there exists an imbalanced distribution among classes. The optimal performance of accuracy is achieved when the costs associated with false positives and false negatives are comparable. In the context of this study, it is important to consider the potential financial implications associated with false positives and the distinction between false negatives is significant, hence it is preferable to consider both Precision and Recall. In the present scenario, the F1 score is calculated to be 0.701.

F1 Score = 2(Recall Precision) / (Recall + Precision)

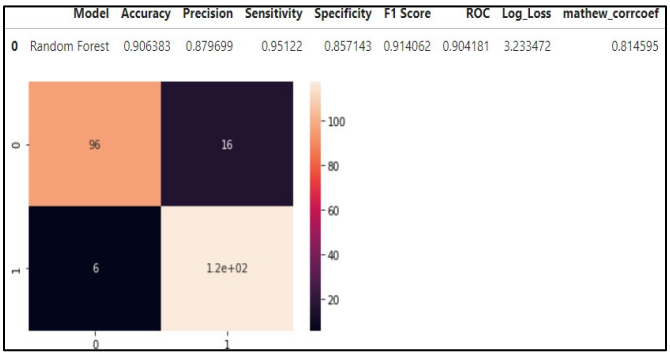

Figure 12: log loss graph.

Following the completion of our model evaluation, our objective is now to conduct a comparative analysis of the machine learning algorithms. In order to determine the most suitable machine learning algorithm for our model, it is necessary to do a comprehensive evaluation. Based on the aforementioned findings, it is evident that the XGBoost Classifier exhibits superior performance. This is substantiated by its notable test accuracy of 0.9191, sensitivity of 0.943, specificity of 0.89, f1-score of 0.9243, and the lowest Log Loss value of 2079.

	Model	Accuracy	Precision	Sensitivity	Specificity	F1 Score	ROC	Log_Loss	mathew_corrcoef
0	Random Forest	0.906383	0.879699	0.951220	0.857143	0.914062	0.904181	3.233472	0.814595
1	MLP	0.825532	0.801471	0.886179	0.758929	0.841699	0.822554	6.026006	0.652539
2	KNN	0.808511	0.786765	0.869919	0.741071	0.826255	0.805495	6.613907	0.618029
3	EXtra tree classifier	0.897872	0.883721	0.926829	0.866071	0.904762	0.896450	3.527415	0.795852
4	XGB	0.919149	0.906250	0.943089	0.892857	0.924303	0.917973	2.792538	0.838384
5	SVC	0.825532	0.801471	0.886179	0.758929	0.841699	0.822554	6.026006	0.652539
6	SGD	0.791489	0.862745	0.715447	0.875000	0.782222	0.795224	7.201750	0.595000
7	Adaboost	0.834043	0.813433	0.886179	0.776786	0.848249	0.831482	5.732052	0.668866
8	CART	0.834043	0.833333	0.853659	0.812500	0.843373	0.833079	5.732039	0.667176
9	GBM	0.851064	0.833333	0.894309	0.803571	0.862745	0.848940	5.144148	0.702485

Figure 13: The outcomes of the training program.

The Extra Tree Classifier achieves the highest average area under the curve (AUC) of 0.950 in the context of the AUC metric at AUC.

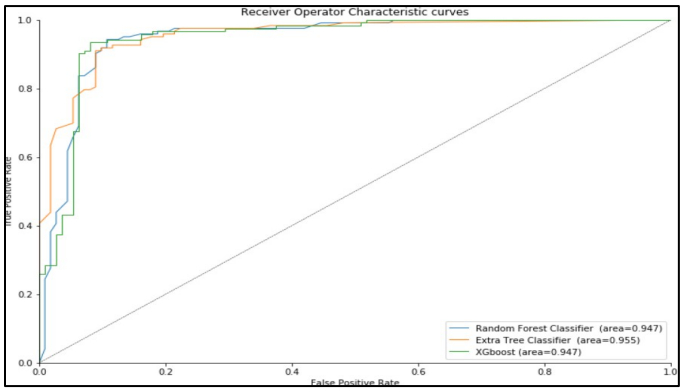

Figure 14: ROC data frames.

Next, we conducted the precession operation. It was determined that the random forest algorithm exhibits the highest precision.

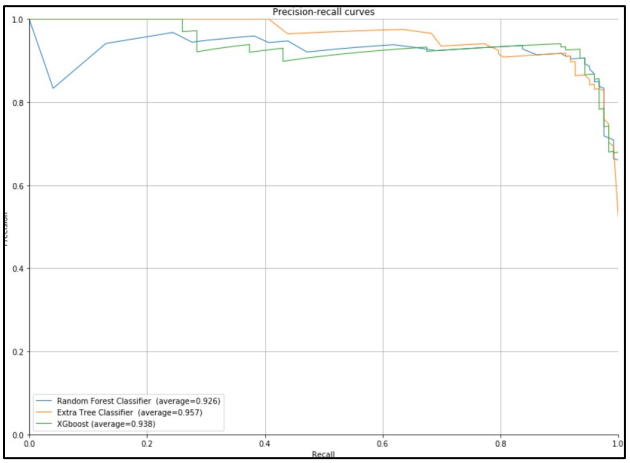

Figure 15: precision-recall curves.

Subsequently, a soft voting methodology was employed, which is a commonly utilized ensemble method in machine learning. This technique involves the aggregation of projected class probabilities from different models in order to generate a final prediction. Soft voting is a technique in which the projected class probabilities from many models are aggregated by taking their average. The class with the highest probability is then chosen as the ultimate prediction.

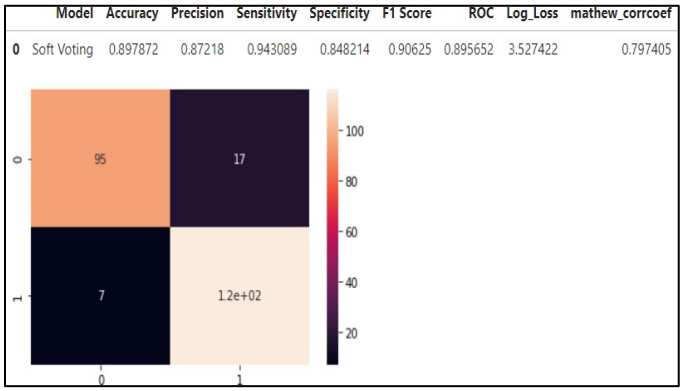

Figure 16: soft voting and its confusion matrix.

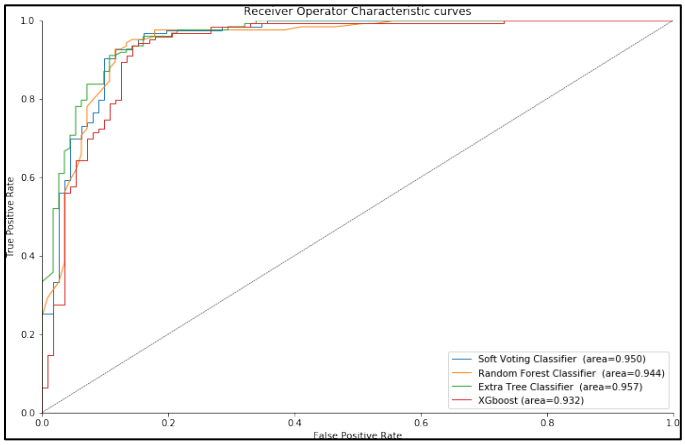

Figure 17: ROC Curve.

It is evident that the random forest algorithm demonstrates superior performance in terms of the receiver operating characteristic (ROC). This finding suggests that the Random Forest classifier is the most optimal method for our model. As demonstrated, the implementation of a stacked ensemble of multiple powerful machine learning algorithms yielded superior performance compared to any single machine learning model. Additionally, we have analyzed the second highest-performing method, namely the random forest algorithm. The five most significant contributing features are as follows: age, cholesterol, exercise-induced angina, maximum heart rate, and ST depression.

4. Conclusion

This study aimed to construct a heart attack prediction model by the utilization of machine learning algorithms. A comprehensive investigation was undertaken on a range of machine learning algorithms, with their performance evaluated using established metrics. The model underwent training and testing using a dataset derived from real-world observations. As a result, it exhibited a notable degree of accuracy in its ability to predict the probability of a heart attack occurrence. The utilization of machine learning in

the development of a heart attack prediction model presents an opportunity to improve patient care by effectively evaluating the likelihood of heart attacks, facilitating timely intervention, and optimizing the allocation of resources within healthcare systems. The integration of this technology into clinical practice is achieved through its deployment via an API, resulting in a seamless integration process. Additionally, continuous refinement is ensured through the identification of areas that require improvement. As an agent that stimulates additional investigation, it exhibits potential for the ongoing enhancement of cardiovascular disease prognostication and the advancement of health outcomes for the populace. Based on our comprehensive investigation, a number of significant risk variables associated with the occurrence of heart attacks have been found. These factors encompass age, gender, and a range of diverse medical problems. The model possesses the capability to consider these aspects and generate precise predictions, thereby assisting healthcare providers in identifying patients who are at a higher risk and enabling them to implement suitable preventative measures.

Although the development of our model did not involve a graphical user interface (GUI), we successfully deployed it through an application programming interface (API). This API enables users to engage with the model and generate predictions. Furthermore, various areas for enhancement were discovered throughout our analysis. These include the integration of supplementary data, the utilization of more sophisticated machine learning methodologies, and the enhancement of model interpretability. In general, the heart attack prediction model we have developed holds promise as a beneficial instrument for healthcare providers and patients alike. By effectively identifying patients who are at a higher risk of developing heart-related conditions and promptly implementing preventive treatments, it is possible to contribute to a reduction in the occurrence of cardiovascular diseases. Enhance patient outcomes through interventions and mitigate instances of attacks. It is our aim that the outcomes of our study will serve as a catalyst for additional scholarly investigations in this domain, ultimately culminating in enhanced health outcomes for the entire population.

5. References

[1] Rahman, A. U., Saeed, M., Saeed, M. H., et al. (2023). A framework for susceptibility analysis of brain tumours based on uncertain analytical cum algorithmic modeling. *Bioengineering*, 10(2), 147.

[2] C. O. S. Patricia, World health statistics 2021, vol. 3, no. 2. 2021.

[3] Asaad, R. R. (2022). Support vector machine classification learning algorithm for diabetes prediction. I*nternational Research Journal of Science, Technology, Education, and Management*, 2(2), 26-34.

[4] D. for H. D. and S. P. National Center for Chronic Disease Prevention and Health Promotion, "Heart Disease Facts," 2021. https://www.cdc.gov/heartdisease/facts.htm#:~:text=Coronary heart disease is the,killing 375%2C476 people in 2021.&text=About 1 in 20 adults,have CAD (about 5%25).&text=In 2021%2C about 2 in,less than 65 years old.

[5] M. N. Krishnan, "Coronary heart disease and risk factors in India - On the brink of an epidemic?," Indian Heart J., vol. 64, no. 4, pp. 364–367, 2012, doi: 10.1016/j.ihj.2012.07.001.

[6] Thirugnanam, T., Galety, M. G., Pradhan, M. R., Agrawal, R., Shobanadevi, A., Almufti, S. M., & Lakshmana Kumar, R. (2023). PIRAP: Medical Cancer Rehabilitation Healthcare Center Data Maintenance Based on IoT-Based Deep Federated Collaborative Learning. *International Journal of Cooperative Information Systems*, 2350005.

[7] L. W. & Wilkins, "Heart Disease and Stroke Statistics—2023 Update: A Report From the American Heart Association," AHA/ASA Journals, vol. 147, no. 8, doi: https://doi.org/10.1161/CIR.0000000000001123.

[8] Rukhsar, S., Awan, M. J., Naseem, U., et al. (2023). Artificial Intelligence Based Sentence Level Sentiment Analysis of COVID-19. *Computer Systems Science & Engineering*, 47(1).

[9] Aighuraibawi, A. H. B., Manickam, S., Abdullah, R., et al. (2023). Feature Selection for Detecting ICMPv6-Based DDoS Attacks Using Binary Flower Pollination Algorithm. *Computer Systems Science & Engineering*, 47(1).

[10] Ali, Z. A., Abduljabbar, Z. H., Taher, H. A., Sallow, A. B., & Almufti, S. M. (2023). Exploring the Power of eXtreme Gradient Boosting Algorithm in Machine Learning: a Review. *Academic Journal of Nawroz University*, 12(2), 320-334.

[11] Mohammed, M. A., Lakhan, A., Zebari, D. A., Abdulkareem, K. H., Nedoma, J., Martinek, R., ... & Tiwari, P. (2023). Adaptive secure malware efficient machine learning algorithm for healthcare data. CAAI Transactions on Intelligence Technology.

[12] Asaad, R. R. (2021). Review on Deep Learning and Neural Network Implementation for Emotions Recognition. *Qubahan Academic Journal*, 1(1), 1-4.

[13] C. Chen et al., "Deep Learning for Cardiac Image Segmentation: A Review," Front. Cardiovasc. Med., vol. 7, no. March, 2020, doi: 10.3389/fcvm.2020.00025.

[14] Krittanawong, C., Johnson, K. W., Rosenson, R. S., & Wang, Z. (2020). Deep learning for cardiovascular medicine: A practical primer. European Heart Journal, 41(22), 2058-2073. DOI: 10.1093/eurheartj/ehaa552. .

[15] Attia, Z. I., Kapa, S., Lopez-Jimenez, F., & Noseworthy, P. A. (2019). An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: A retrospective analysis of outcome prediction. The Lancet, 394(10201), 861-867. DOI: 10.1016/S0140-6736(19)31721-0.

[16] X. Su et al., "Prediction for cardiovascular diseases based on laboratory data: An analysis of random forest model," J. Clin. Lab. Anal.,

vol. 34, no. 9, pp. 1–10, 2020, doi: 10.1002/jcla.23421.
[17] X. Fan, Z. Hu, R. Wang, L. Yin, Y. Li, and Y. Cai, "A novel hybrid network of fusing rhythmic and morphological features for atrial fibrillation detection on mobile ECG signals," Neural Comput. Appl., vol. 32, no. 12, pp. 8101–8113, 2020, doi: 10.1007/s00521-019-04318-2.
[18] S. Ahmadian, S. M. J. Jalali, S. Raziani, and A. Chalechale, "An efficient cardiovascular disease detection model based on multilayer perceptron and moth-flame optimization," Expert Syst., vol. 39, no. 4, pp. 1–19, 2022, doi: 10.1111/exsy.12914.
[19] Rajab Asaad, R., & Masoud Abdulhakim, R. (2021). The Concept of Data Mining and Knowledge Extraction Techniques. *Qubahan Academic Journal*, 1 (2), 17–20.
[20] R. Alizadehsani et al., "A database for using machine learning and data mining techniques for coronary artery disease diagnosis," Sci. Data, vol. 6, no. 1, pp. 1–13, 2019, doi: 10.1038/s41597-019-0206-3.
[21] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.
[22] Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21-27.
[23] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning (Vol. 1). MIT press Cambridge.
[24] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 785-794).
[25] Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. Machine learning, 63(1), 3-42.
[26] Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.
[27] Bottou, L., Curtis, F. E., & Nocedal, J. (2018). Optimization methods for large-scale machine learning. SIAM Review, 60(2), 223-311.
[28] Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. In Proceedings of the Thirteenth International Conference on Machine Learning (pp. 148-156).
[29] Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and regression trees. CRC press.
[30] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of statistics, 29(5), 1189-1232.
[31] Ibrahim, D. A., Zebari, D. A., Mohammed, H. J., & Mohammed, M. A. (2022). Effective hybrid deep learning model for COVID-19 patterns identification using CT images. *Expert Systems*, 39(10), e13010.
[32] Kapoor, N. R., Kumar, A., Kumar, A., et al. (2022). Event-Specific Transmission Forecasting of SARS-CoV-2 in a Mixed-Mode Ventilated Office Room Using an ANN. *International Journal of Environmental Research and Public Health*, 19(24), 16862.
[33] Mohammed, H. J., Al-Fahdawi, S., Al-Waisy, A. S., et al. (2022) ReID-DeePNet: A Hybrid Deep Learning System for Person Re-Identification. *Mathematics,* 2022, 10, 3530.