ISSN: 0067-2904

Machine Learning for Real-Time Cardiovascular Disease Prediction Based on Cloud

Shahad Ali Ridha *, Mohammed Issam Younis
Department of Computer, College of Engineering, University of Baghdad, Baghdad, Iraq

Received: 15/1/2025 Accepted: 13/ 5/2025 Published: xx

Abstract

Diagnosing cardiovascular disease is an essential medical process to guarantee accurate classification, which aids cardiologists in treating patients appropriately. By employing machine learning for the cardiovascular illness classification, it is possible to decrease the occurrence of misdiagnosis and save patients' lives. This study has developed an effective Amazon Web Services (AWS) machine learning architecture model to predict cardiovascular diseases. It integrates a number of AWS services, including SageMaker, Lambda, API Gateway, and S3, which provide significant automation and real-time prediction. The cardiovascular dataset from Kaggle is classified using the ensemble algorithms, including CatBoost, XGBoost, and LightGBM. To improve classification accuracy, k-modes clustering with Huang initialization is used. Additionally, SageMaker Automatic Model Tuning is utilized to optimize a model's hyperparameters based on Bayesian optimization. The experimental results indicate that the CatBoost classifier outperformed with a slightly higher accuracy of 87.9% compared to the other applied algorithms. Furthermore, the model takes about 199 ms to respond to the prediction request, which makes it quick and appropriate for applications that require low latency.

Keywords: Machine Learning, XGBoost, K-mode Clustering, Cardiovascular Disease, AWS Services.

التعلم الآلي للتنبؤ بأمراض القلب والأوعية الدموية في الوقت الفعلي استنادًا إلى السحابة

شهد علي رضا*، محمد عصام يونس
قسم الحاسبات، كلية الهندسة، جامعة بغداد، بغداد، العراق

الخلاصة

يُعد تشخيص أمراض القلب والأوعية الدموية عملية طبية أساسية لضمان دقة التصنيف، مما يُساعد أطباء القلب على علاج المرضى بشكل مناسب. ومن خلال استخدام التعلم الآلي لتصنيف أمراض القلب والأوعية الدموية، يُمكن تقليل حدوث التشخيص الخاطئ وإنقاذ حياة المرضى. وقد طورت هذه الدراسة نموذجًا فعالًا لهندسة تعلم الآلة من Amazon Web Services (AWS) للتنبؤ بأمراض القلب والأوعية الدموية. ويدمج هذا النموذج عددًا من خدمات AWS، بما في ذلك SageMaker و Lambda و API Gateway و S3، مما يوفر أتمتةً عاليةً وتنبؤًا آنيًا. وتُصنف مجموعة بيانات أمراض القلب والأوعية الدموية من Kaggle باستخدام خوارزميات المجموعة، بما في ذلك CatBoost و XGBoost و LightGBM. ولتحسين دقة التصنيف، يُستخدم

*Email: gs22.saridha@coeng.uobaghdad.edu.iq

التجميع بأنماط k مع تهيئة Huang. بالإضافة إلى ذلك، يُستخدم ضبط النموذج التلقائي من SageMaker لتحسين معاملات النموذج الفائقة استنادًا إلى التحسين البايزي. وتشير النتائج التجريبية أن مصنف CatBoost تفوق بدقة أعلى قليلاً بلغت 87.9% مقارنةً بالخوارزميات الأخرى المُطبقة. علاوةً على ذلك، يستغرق النموذج حوالي ١٩٩ مللي ثانية للاستجابة لطلب التنبؤ، مما يجعله سريعًا ومناسبًا للتطبيقات التي تتطلب زمن وصول منخفضًا.

1. Introduction

The World Health Organization reports that heart disease is the leading cause of death globally [1]. A study on the global burden of disease found that about 43% of deaths are caused by heart disease [2] [3]. Cardiovascular disease (CVD) has become a significant concern globally because of its increasing spread and the increasing number of deaths it causes.

CVDs encompass a range of disorders that affect the blood vessels and heart, including vascular diseases like coronary artery disease, as well as heart conditions like heart failure, cardiomyopathy, rheumatic heart disease, heart attack, and arrhythmias [4][5]. Several risk factors cause heart disease to occur, some of which are controllable factors, such as smoking, alcohol, diabetes, excess LDL cholesterol, and insufficient physical activity [6]. The loss of clinical equipment, doctors, and other services makes heart diagnosis and treatment particularly difficult, especially in developing nations [7]. To lower the chance of serious cardiac issues and improve heart protection, a patient's heart attack risk must be correctly and accurately identified [8]. A clinical evaluation report, a review of the patient's medical history, and a medical expert's analysis of the patient's symptoms are the foundations for diagnosing invasive cardiac disease. Many of these methods provide inaccurate diagnoses and delay results because of human error. Additionally, it takes longer to determine, and it is more expensive and computationally complex [9].

The WHO estimates that by 2030, there will be 23.6 million deaths from CVDs overall, with heart disease and stroke accounting for the majority of these deaths [10]. Applying data mining and machine learning techniques to predict the likelihood of developing heart disease is essential to saving lives and reducing the financial burden on society. While deep learning has demonstrated remarkable success in domains such as image recognition, speech processing, and autonomous systems [11], its application is not always optimal, particularly in structured tabular data tasks like heart disease prediction. Machine learning can be used to identify, diagnose, and predict several diseases. Many methods, including DT, RF, and LR, can be employed in machine learning classification to overcome the difficulties associated with invasive-based heart disease detection [12]. The percentage of heart disease mortality has dropped due to these machine learning-based expert medical decision systems. [13].
Moreover, building machine learning models is simplified with the help of the cloud computing system Amazon Web Services (AWS), which enables a reliable environment and multiple services. Building, training, automating model tuning, and deploying machine learning models in the cloud is possible for data scientists and developers with Amazon SageMaker Studio. Which is an essential tool for automating machine learning workflows [14].

This study presents an evaluation of the effectiveness of the three ensemble algorithms (CatBoost, XGBoost, and LightGBM) in developing a cardiovascular prediction model. This study utilized the cardiovascular disease dataset that is publicly available on Kaggle. Additionally, all processing was performed on Amazon SageMaker using the Python programming language. In summary, this study provides the following significant contributions:

- Determine the most effective ensemble classifier for the classification of cardio diseases.
- Preprocessed the dataset using k-mode clustering to enhance the models' convergence.
- Utilize Amazon SageMaker to streamline the process of building, training, tuning, and deploying an ML model.
- Achieve integration with some Amazon services (Amazon S3, Amazon API Gateway, and AWS Lambda) to provide real-time prediction and scalable storage.
- Use a large, more diverse dataset to improve the generalizability of the results.
- Proof of the importance and impact of data preparation, hyperparameter optimization, and feature extraction on model performance.

2. Literature Review

Recently, the healthcare industry has witnessed notable progress in machine learning, especially in cardiovascular diseases. Since these diseases remain a leading cause of death in developing countries [15] [16] [17], researchers in this field have an unprecedented opportunity to develop and evaluate new preventative models, cardiovascular disease prognosis, and early signs of disease.

A. Dwivedi [18] developed a machine learning algorithm to predict the presence of cardiovascular disease. This study utilized KNN, Naive Bayes, classification trees, and logistic regression, SVM, and ANN algorithms. It uses the StatLog cardiovascular dataset (270 samples and 13 unique features) from the UCI Machine Learning Laboratory. The study findings confirmed that logistic regression achieved a maximum classification accuracy of 85%, with corresponding sensitivity and specificity values of 89% and 81%.

D. Shah et al. [19] proposed a model-based machine learning for cardiovascular disease prediction. This study classifies heart disease using supervised machine learning techniques, including Naïve Bayes, decision trees, KNN, and random forests. Data for this purpose was obtained from the Cleveland database of cardiac patients in the UCI repository. The dataset consists of 303 patients and 76 objects. Only 14 of the 76 parameters are tested to evaluate the performance of the algorithms. The findings show that KNN has the highest accuracy score of 90.78%.

R. Perumal and A. Kaladevi [20] created a model for predicting heart disease. Using the Cleveland dataset, which consisted of 303 data instances. This study used PCA for standardizing and reducing features. They used seven main components to train the machine learning classifiers. In contrast to KNN (69%), they found that LR and SVM offered nearly identical accuracy rates (87% and 85%, respectively).

A. Shimaa [21] carried out a study with the primary goal of determining the best techniques to increase the accuracy of cardiovascular disease prediction by evaluating a number of algorithms and training approaches. The study evaluated the performance of eight classification techniques (linear discriminant analysis, logistic regression, SVM, KNN, decision tree, naïve Bayes, random forest, and neural network). Additionally, it used two datasets and four types of cross-validation, including holdout, k-fold cross-validation, stratified k-fold cross-validation, and repeated random. The findings show that holdout cross-validation with neural networks achieves 71.82% better accuracy when used with the cardiovascular disease dataset from Kaggle (70,000 patients with 13 items). Moreover, random forest achieves better with a small dataset from the UCI repository (303 records with 14 attributes) at 89.01% accuracy.

In the study presented by F. Alotalibi [22], machine learning techniques were used for predicting cardiovascular disease. The study developed predictive models using the Cleveland Clinic Foundation dataset (303 patients with 14 features), which implements different ML methods such as Naïve Bayes, decision trees, SVM, logistic regression, and random forests.

The results showed that the decision tree algorithm outperformed the SVM method in cardiovascular disease prediction accuracy, with rates of 93.19% and 92.30%, respectively.
In a study presented by M. Rahma and A. Salman [23], a machine learning system is developed for the diagnosis of heart disease, which is based on the UCI database. This study compares three machine learning classifiers: SVM, Naive Bayes, and KNN. Additionally, five-fold cross-validation is utilized to prevent identical values during the model learning and testing phase. According to the experimental findings, the Naive Bayes algorithm achieved the highest accuracy of 97%.
In a study presented by N. Hasan and Y. Bao [24], a dataset of 70,000 patients with 13 variables from the Kaggle repository was used to develop an ML model to predict cardiovascular illnesses. This study compares different feature selection methods (filter, wrapper, and embedded) with random forest, SVM, KNN, Naive Bayes, and XGBoost classifier models. The results indicate that the wrapper feature selection technique with XGBoost achieves the highest accuracy at 73.74%.
A study by V. Shorewala [25] demonstrates how to increase the prediction accuracy of heart disease using ensemble techniques. The classifiers, such as KNN, binary logistic classification, and Naive Bayes, were compared against ensemble modeling techniques such as bagging, boosting, and stacking. The "Cardiovascular Disease Dataset" with 70,000 patient records is used. The models' performance is validated using data-analytic methods and K-folds cross-validation. The results indicate that the stacked model combining KNN, SVM, and random forest is the most effective, with an accuracy of 75.1%.
D. Waigi et al., [26] used the 70,000-record Kaggle cardiovascular illnesses dataset to develop a machine learning model for heart disease prediction. It utilized a decision tree model that achieved 72.77% accuracy.

A study by J. Maiga and G. Hungilo [27] compared ML algorithms for predicting cardiovascular illness applied to a dataset that consists of 70,000 patient records from Kaggle. The algorithms used are logistic regression, KNN, random forest, and Naïve Bayes. The findings demonstrate that Random Forest attains a high classification accuracy of 73%, with a sensitivity and specificity of 80% and 65%, respectively.

A study by SM Khazaal and H Maarouf [28] developed a machine learning model for predicting coronary artery disease (CAD) using a patient clinical factors dataset of 303 individuals with 56 variables. The SVM technique was used, which achieved a high prediction accuracy of 96.7%, with an AUC of 71.5%. This suggests that this technique can assist cardiologists in predicting CAD, which is one of the most prevalent cardiac conditions.
A study by V. Perrone et al. [29] compared two types of hyperparameter optimization techniques in terms of performance, including random search and Bayesian. This was implemented in Amazon SageMaker and used the automated model tuning service (AMT) for optimization. The results show that the Bayesian method is continuously superior to random search for a variety of hyperparameter assessments. Moreover, the study demonstrates that AMT service is an effective technology for hyperparameter optimization.
A study presented by [30] involved preprocessing and rearranging ECG data from three datasets—the Diagnostic ECG Database, the INCART 12 Lead Arrhythmia Database, and the PTB-XL ECG Database—to improve disease categorization and prediction performance during the learning phase. A CNN one-dimensional model based on the VGG-16 architecture was developed, and fold-cross validation was used to assess how well the proposed method works with datasets. According to the findings, the proposed model had a 93% accuracy rate and 0.069 losses. The total performance of all employees' datasets is reflected in these results. This

method leads to improved clinical results in heart care by improving the classification of ECG-related diseases.
To create scalable machine learning systems on AWS, A. Jana [31] provided a methodology to develop end-to-end ML tools. Many AWS services are used in this study, which help increase the scalability of machine learning functions, improve consistency, and reduce manual intervention.

3. Theoretical Background

3.1 Cardiac Disease Classification

Cardiac disease classification is a critical aspect that helps in the early detection and prevention of sudden cardiac death. Accuracy and efficiency in classifying heart diseases have increased with the development of computer-aided diagnostic methods. These methods decrease hospital stays and improve the effectiveness, affordability, and accessibility of healthcare. They also hold great promise for the bioinformatics and healthcare fields. Automated algorithms for classifying heart diseases can help physicians and other healthcare providers diagnose patients more accurately and provide better care. Once the data is prepared, the classification algorithm is used to train the model.

3.1.1 XGBoost Algorithm

The XGBoost classification algorithm is well known for its high performance and excellent classification and prediction results. The XGBoost is a gradient-boosted version of a decision tree [32]. This method sequentially generates decision trees. The decision tree then uses the weights assigned to each independent variable to provide predictions. The subsequent decision tree amplifies and incorporates the significance of the pertinent variables whenever the tree generates an incorrect forecast. The contributions of each of these classifiers are subsequently combined to create a more accurate model. Machine learning approaches have shown effectiveness in illness prediction and risk stratification using this integrated strategy [33] [34] [35].

3.1.2 Classification Evaluation Metrics

There are several performance indicators based on the confusion matrix used to evaluate the algorithm's performance. The binary classification prediction results are described by the confusion matrix presented in Figure 1 [36].

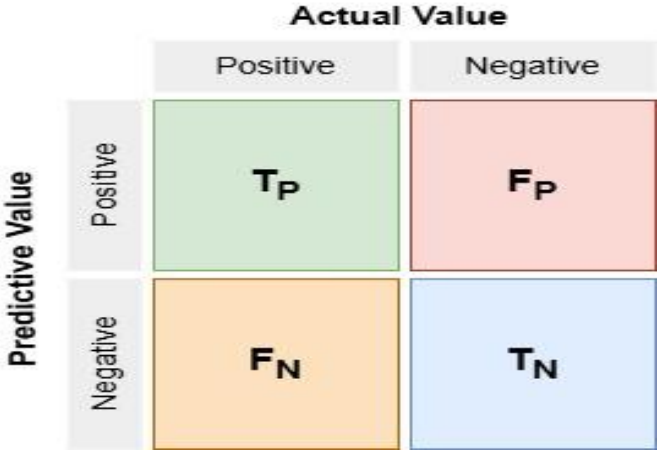


Figure 1: Confusion Matrix [36].

The variables TP and FP represent the proportion of true and false positives classified as subjects with heart disease, respectively. Similarly, TN and FN represent the number of subjects that were true negatives and false negatives classified as not suffering from heart disease, respectively. Several metrics that are derived from the confusion matrix and used to assess classification task performance are as follows [36]:

- Accuracy (A): is the most widely used indicator for evaluating a system's overall performance; it describes the degree to which a measured value matches the real value. It can be represented by the following formula:

$$\textbf{Accurecy (A)} = \frac{T_P + T_N}{T_P + T_N + F_P + F_N} \quad \text{Eq. (1)}$$

- Precision (P): is the proportion of participants correctly identified as positive out of all those identified as positive. It can be computed using the following formula:

$$\textbf{Precision } (\boldsymbol{P}) = \frac{T_P}{T_P + F_P} \quad \text{Eq. (2)}$$

- Recall (R): Also known as sensitivity, it shows the percentage of all positive samples that are real positives. The following formula can be used to compute it:

$$\textbf{Recall (R)} = \frac{T_P}{T_P + F_N} \quad \text{Eq. (3)}$$

- Specificity (S): is the percentage of true negatives among all samples in the negative category, often referred to as the true negative rate. The formula below can be used to compute it:

$$\textbf{Specificity (S)} = \frac{T_N}{T_N + F_P} \quad \text{Eq. (4)}$$

F1 Score (F): is the precision and recall harmonic mean that can be calculated using the formula:

$$\textbf{F1 score } (\boldsymbol{F}) = 2 * \frac{(P*R)}{(P+R)} \quad \text{Eq. (5)}$$

where the values of FP, FN, TP, and TN are described as follows:

- FP = the result when the model mistakenly predicts that the number of people who have never experienced a heart attack has one.
- FN is the number of people who have never experienced a heart attack, and the model forecasts people who have experienced one wrong.
- TP = the total number of people who have a heart attack.
- TN = Total number of heart disease and non-heart disease individuals.

Additionally, the ability to generalize the classification algorithm can be evaluated using the ROC curve, which represents the relationship diagram between the true positive rate (as the y-axis) and the false positive rate (as the x-axis) that depicts the relationship between recall and specificity. Furthermore, the area under the ROC curve (AUC) indicates the ability of the model to predict cardiac disease. A scalar statistic (AUC) is used to assess the overall performance, sensitivity, and specificity of the classification algorithm.

3.2 Dataset Overview

In this study, the publicly available cardiovascular disease dataset from the Kaggle repository is used [37]. This dataset is important for building, testing, and processing ML algorithms, in addition to increasing heart disease detection accuracy and improving patient care. The cardiovascular disease dataset contains 70,000 patient records with 13 different features that are described in Table 1. The last output feature, "cardio," indicates whether a patient is healthy (represented as 0) or has cardiovascular illness (represented as 1).

Table 1: Cardiovascular Disease Dataset Attributes

#	Feature	Parameters	Values Range/Type
1	Id	id	Integer (Min = 0 , Max = 69999)
2	Gender	gender	Integer (Male = 1, Female = 2)
3	Height	height	Integer (Min = 55 , Max = 250)
4	Weight	weight	Float (Min = 10 , Max = 200)
5	Age	age	Integer in days (Min = 10,798 , Max = 23,713)
6	Systolic blood pressure	ap_hi	Integer (Min = -150 , Max = 16,020)
7	Diastolic blood pressure	ap_lo	Integer (Min = -70 , Max = 11,000)
8	Cholesterol	cholesterol	Categorical values (Min = 1 , Max = 3)
9	Glucose	gluc	Categorical values (Min = 1 , Max = 3)
10	Smoking	smoke	Integer (Yes = 1 , No = 0)
11	Alcohol intake	alco	Integer (Yes = 1 , No = 0)
12	Physical activity	active	Integer (Yes = 1 , No = 0)
13	cardiovascular disease	cardio	Integer (Yes = 1 , No = 0)

3.3 Amazon SageMaker

SageMaker is a service provided by Amazon Web Services (AWS) that makes the machine learning model simpler to build, train, and deploy quickly. Moreover, it provides an automatic model tuning (AMT) service for a distributed, scalable, and fault-tolerant hyperparameter optimization [38]. AMT combines the most recent techniques for hyperparameter optimization, including Bayesian, random search, and grid search optimization. Moreover, it can handle large amounts of data by performing parallel tasks simultaneously. The optimal hyperparameters are automatically selected based on the objective (such as maximizing accuracy). Additionally, Amazon SageMaker offers a one-click model deployment feature for real-time forecasting. It acts as a point of contact for direct application implementation and real-time communication with the training model.

3.4 Amazon Simple Storage Service S3

S3 is a versatile cloud storage service that provides scalable storage for huge datasets. It offers enhanced performance and real-time data availability. Amazon S3 is used for backing up data, storing ML models, invoking Lambda functions, and retrieving data [39].

3.5 Amazon API Gateway

Amazon API Gateway is a completely managed service that allows developers to easily design, publish, maintain, monitor, and secure various types of APIs, such as REST, HTTP, and WebSocket APIs [40].

3.6 AWS Lambda

Amazon Web Services offers a server-less computing solution called AWS Lambda. It allows Python, Java, or Node.js programs to run without the need for infrastructure management [41]. It is an event-driven computing service that executes code in response to specified events, such as a request from an API gateway. Lambda leads to improved application performance and a significant decrease in infrastructure costs.

4. Methodology

This study aims to predict heart disease to assist patients and medical professionals with early detection. The proposed model uses an Amazon Web Services (AWS) machine learning architecture. The model depends on integrating various AWS services, such as Amazon SageMaker, AWS Lambda, API Gateway, and Amazon S3, all designed to provide significant automation and flexibility. The process begins with data preparation, which is followed by

stages of model training and automatic hyperparameter tuning. Then, the optimized classification model will be deployed as a SageMaker endpoint for prediction testing. Finally, performance evaluation metrics are calculated, as shown in the flowchart in Figure 2.

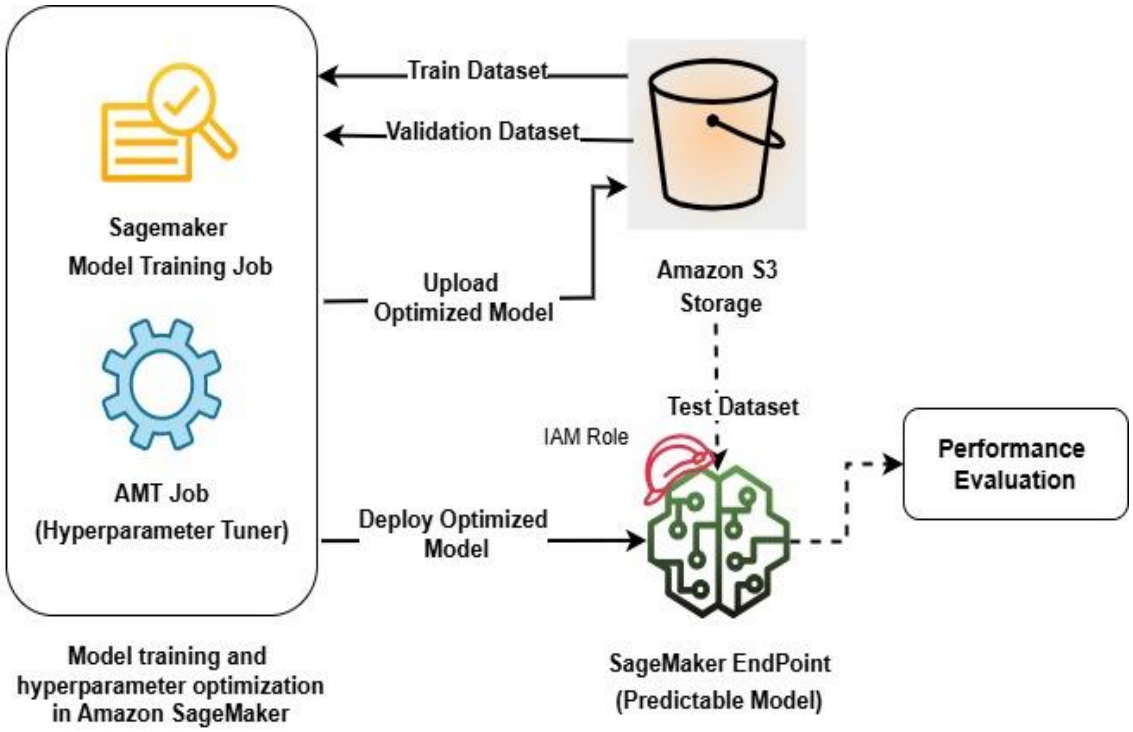


Figure 2: The flowchart of the proposed heart disease prediction model on AWS.

4.1 Data Preparation

In this work, data preparation was performed using conda-python3 in SageMaker's Jupyter notebooks. This includes preprocessing the original data, feature engineering, k-mode clustering, and data splitting into (80%) training dataset and (20%) testing dataset, as shown in Figure 3.

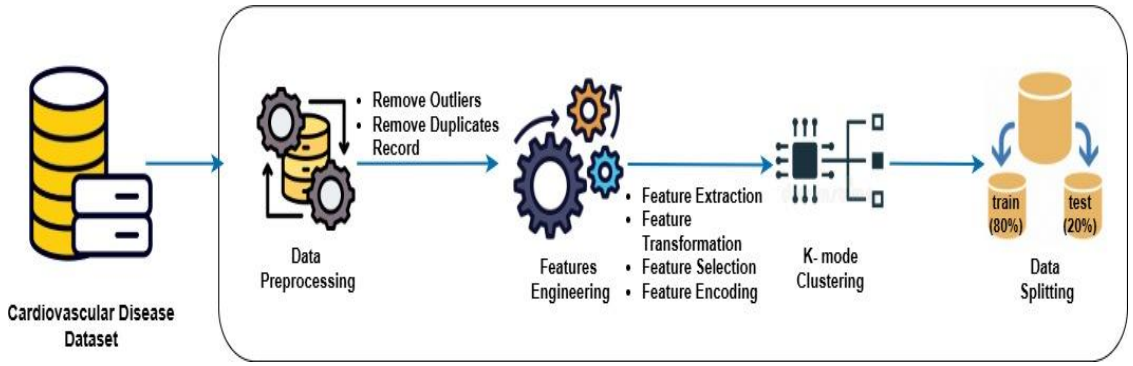


Figure 3: The Data Preparation flow diagram.

4.1.1 Preprocessing

The preprocessing is an essential step that ensures the original cardiovascular disease dataset is free of data outliers, such as noise or duplicate records. The presence of outliers that an error in data entry may have caused significantly impacts the classification model's accuracy.

The data must be examined and cleansed at this stage to ensure it is prepared for model building. As seen in Table 1, the dataset used contains 13 features. First, the parameter "id" is removed since it does not affect the model's development. In addition, the duplicate records are removed. Then, the patients' ages were recorded in days, but they were divided by 365 to convert them to years for better analysis and prediction. Finally, outliers are identified using a box plot, a widely utilized and precise method for detecting outliers. The minimum value, first quartile (Q1), median (Q2), third quartile (Q3), and highest value are all shown using a boxplot. Plotting these five points results in a graph resembling a box, and the point outside of this box is regarded as an outlier, as shown in Figure 4. In this study, outlier identification and removal were performed manually from the ap-hi, ap-lo, weight, and height attributes.

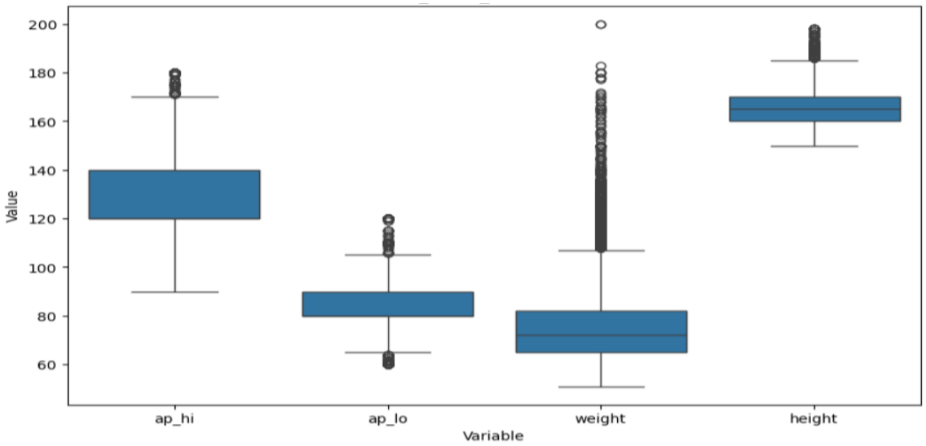


Figure 4: Boxplots of some attributes.

4.1.2 Features Engineering

The four feature engineering techniques—feature extraction, feature transformation, feature selection, and redaction, and categorical feature encoding—were applied in this study.

1. Feature Extraction:

In the extensive investigation by [42], US citizens initially without a history of clinical cardiovascular disease had a significant lifetime risk of the disease, which was further increased for those who were overweight or obese. Obese adults were demonstrated to have an earlier onset of cardiovascular disease when compared to those with a normal BMI. This implies that the features of weight and height can be converted into body mass index (BMI), which might enhance the overall performance of the proposed model for predicting heart disease.

$$\text{Body Mass Index (BMI)} = \text{Weight (kg/lb)}/\text{Height}^2(\text{m}^2/\text{in}^2) \quad \text{Eq. (6) [43]}$$

In medicine, a person's average blood pressure during one cardiac cycle is mean arterial pressure (MAP). The investigations by [44] and [45] found a relationship between severe CVD events and MAP, a measure of heart output and peripheral resistance. According to a study that includes individuals with type 2 diabetes, the risk of CVD increased by 13% for every 13 mmHg increase in MAP.

$$\text{Mean Arterial Pressure (MAP)} = (2\ \text{Diastolic BP} + \text{Systolic BP})/3 \quad \text{Eq. (7) [45]}$$

In this study, converting height and weight into the body mass index (BMI) was determined for each sample using Equation (6) [43]. In addition, using values of diastolic blood pressure and systolic blood pressure, mean arterial pressure (MAP) was calculated for each sample according to Equation (7) [45].

2. Feature Transformation:
The implementation of numerical values (continuous values) can be very difficult because the prediction algorithm should consider where to draw boundaries between specific classes [46]. To improve the efficiency of the classification algorithm, the binning method was utilized to convert continuous values into categorical values, such as age, BMI, and MAP. Furthermore, the data were separated into bins of four intervals, and each bin was assigned a category name as shown in Table 2.

Table 2: The categorical values for Age, MAP, and BMI.

Feature Values range	Category Name
(0 ≤ Age < 30) year	Young
(30 ≤ Age < 50) year	Middle-aged
(50 ≤ Age < 70) year	Older
(70 ≤ Age < 100) year	Senior
(0 ≤ MAP < 70) mmHg	Low
(70 ≤ MAP < 90) mmHg	Normal
(90 ≤ MAP < 110) mmHg	high
(110 ≤ MAP < 150) mmHg	Very high
(0 ≤ BMI < 19) kg/m^2	Underweight
(19 ≤ BMI < 25) kg/m^2	Normal
(25 ≤ BMI < 30) kg/m^2	Overweight
(30 ≤ BMI < 160) kg/m^2	Obesity

3. Feature Selection

Feature selection is essential for improving the efficacy of machine learning algorithms [47]. The important features are chosen to maximize algorithm efficiency and accuracy while lowering complexity and computation time. In this study, the feature importance technique is applied. It is an integrated method that includes tree-based classifiers. It can be utilized as a method of feature selection. This technique uses the F statistic to rank the features; a higher F value denotes the factor's more important or relevant for the final prediction. The characteristics with the highest F values are chosen as the most correlated with the class type.

4. Categorical Feature Encoding

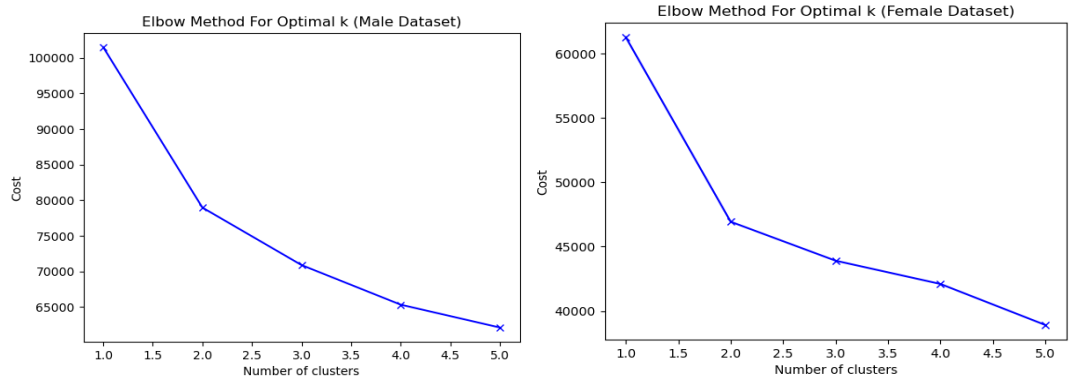
Label encoding transforms the string literals in the dataset into computer-understandable integer values [48]. In this work, the label encoding approach was used to convert text category variables (age, BMI, and MAP) into numerical variables used in the classification algorithm.

4.1.3 Clustering

Clustering is a technique that groups a collection of instances according to similarity measurements. The k-means approach is commonly used for clustering, but is ineffective for handling categorical data. The k-modes algorithm was created in order to get around this restriction. Similar to the K-means algorithm, the K-modes approach was introduced by Huang [49] in 1997. However, instead of using cluster means, it substitutes modes for them and uses dissimilarity measures for categorical data. As a result, the algorithm can process category data effectively. Moreover, gender-based dataset splitting can be beneficial for prediction because men and women differ significantly in terms of biology, which might affect how diseases manifest and progress. Additionally, the rate of cardiac disease varies between men and women. For instance, in terms of symptoms and risk factors, men tend to develop heart disease earlier than women [50].

In this study, the dataset was divided according to gender into male and female datasets, and then the K-Mode Clustering was performed. Additionally, the elbow curve was utilized with

Huang initialization to get the optimal number of clusters. The elbow curve generates a k-modes model with the specified number of clusters, fits it to the data, and estimates the cost (the distance between the cluster's attribute modes and allocated data points). The "elbow method" was then used to plot the costs on a graph to get the ideal number of clusters. By examining the costs plot, the elbow technique searches for a "knee" or inflection point, which is commonly understood to be the point at which adding more clusters does not materially improve the model's fit. Subsequently, the elbow curve approach was then employed to ascertain the ideal number of clusters. The knee point was found at 2.0 in each scenario, as shown in Figure 5, suggesting that 2 was the ideal number of clusters for both male and female datasets.



(a) (b)

Figure 5: The elbow curve approach for (a) Male dataset (b) Female dataset.

4.1.4 Data Splitting

The proposed study divided the data into training and testing using the "80-20 rule" [51]. This refers to the practice of dividing available data into 80% for training and 20% for testing. Which is used to train and test machine learning models. Additionally, this study used a stratified technique, which is a method of partitioning a dataset to obtain samples that accurately represent the class distribution in the population [52] [53], in which the dataset is divided into homogeneous subgroups, each with the same proportion of each category. This method has also been used in previous research in the healthcare sector [54] [55].

4.2 Classification Model Training, Tuning, and Deployment

The training dataset is pulled from Amazon S3, which served as the data source for the training model. The SageMaker model training job is started to perform the binary classification of the target variable 'cardio.' Additionally, the SageMaker AMT is used to optimize the model's performance based on the Bayesian optimization method. This assists in determining the optimal set of hyperparameters for increasing the model's expected accuracy.

After training and tuning, the optimized model is deployed as a SageMaker endpoint, which provides one-click model deployment capability for real-time predictions. It is a point of contact that enables real-time communication with the trained and predictable model. At this stage, it is possible to feed the model fresh data and obtain predictions in real time. Finally, the model's performance is tested using the test dataset to verify evaluation metrics.

4.3. Integration with AWS Services

To ensure that the model is used in real-time prediction, the model was connected to the AWS cloud services, such as Amazon REST API Gateway and Lambda. The Amazon API Gateway was configured, and it can be accessed using a networking protocol (HTTP post requests), through which the model input data (patient data) can be sent. Then, the AWS Lambda function was configured, which automatically executes the code upon receiving the anticipated request (patient data) from the REST API Gateway. Then, it communicates with the SageMaker

endpoint to retrieve the prediction result. Subsequently, it transmits the prediction result from the machine learning model to the API Gateway.

5. Results and Discussion

In this study, all ML processes were performed on the Amazon SageMaker platform Jupyter notebooks. The dataset was divided into two parts, where 80% of the models are used for training and 20% for testing. The cardiovascular disease dataset, which had 70,000 rows and 13 attributes, was reduced to roughly 65,485 rows and 11 features.

The proposed classifiers CatBoost, XGBoost, and LightGBM are compared based on the following evaluation metrics: accuracy, precision, recall, F1 score, and specificity. This comparison aims to determine which of the suggested classifiers performs the best in diagnosing heart disease. The ensemble classifiers' performance comparison results are displayed in Table 3. The results indicated that ensemble classifiers perform similarly in all evaluation metrics, with accuracy from 87.82% to 87.90%. It shows that they learn comparable decision boundaries and suggests that all three models are well suited for classifying heart disease. Although the differences are minimal, CatBoost gets the highest accuracy (87.90%), a slightly better overall forecast. It also has the highest recall (81.81%), which reflects better sensitivity by identifying positive cases. In addition, CatBoost's F1 score (87.01%) is the highest, which shows the best balance between accuracy and recall. Additionally, the XGBoost classifier achieved the following results: accuracy, precision, recall, F1 score, and specificity: 87.83%, 93.10%, 81.54%, 84.94%, and 94.04%, respectively. Finally, LightGBM does not outperform CatBoost in any evaluation metric, although it does perform quite similarly. Accuracy, precision, recall, F1 score, and specificity achieved were 87.82 %, 92.91 %, 81.79 %, 86.9 %, and 93.80%, respectively.

Table 3: Ensemble Classification models Results

Techniques	Accuracy	Precision	Recall	F1 Score	Specificity
CatBoost	87.90 %	92.90 %	81.81 %	87.01 %	93.82 %
XGBoost	87.83 %	93.10 %	81.54%	86.94%	94.04 %
LightGBM	87.82 %	92.91 %	81.79 %	86.9 %	93.80 %

Additionally, this study presents empirical findings that demonstrate the significance of data preparation and hyperparameter optimization. To demonstrate, the XGBoost classifier was applied to the cardiovascular diseases dataset, as shown in Table 4. The results outperformed all other experiments when all data preparation phases were performed, including data preprocessing, feature engineering, hyperparameter tuning, and clustering using k-mode, which achieved 87.83% accuracy, 93.10% precision, 81.54% recall, 86.94% F1 score, 94.04% specificity, and 0.962 mean AUC, respectively. Additionally, it can be noted that the lowest results have been recorded when using XGBoost with the original dataset without any data preparation techniques. Where accuracy (73.2%), precision (75.1%), recall (69.4%), F1 score (72.1%), specificity (77.0%), and mean AUC (0.795). This degradation indicates that noise or other imperfections in the original data limit the ability of the model to predict accurately. Furthermore, with hyperparameter optimization by the SageMaker AMT based on the Bayesian method, the model performed better overall; the accuracy enhanced from 87.4% to 87.83%, with improvement in all other metrics.

Table 4: The experimental results of the XGBoost model.

Techniques	Accuracy	Precision	Recall	F1 Score	Specificity	Mean AUC
Data Preparation	87.83 %	93.10 %	81.54%	86.94%	94.04 %	0.962
No Hyperparameter optimization	87.9 %	92.6 %	81 %	86.4 %	93.6 %	0.961
XGBoost with Original dataset	73.2%	75.1%	69.4%	72.1%	77.0%	0.795

Furthermore, the impact of the feature extraction (MAP and BMI) on the XGBoost model performance in terms of accuracy, precision, recall, F1 score, and specificity was evaluated, as indicated in Table 5. The results indicate that using these features leads to a significant improvement in the performance of XGBoost. It achieved the highest results in all performance metrics, with 87.83% accuracy compared to models that excluded one or both of these features. There were significant decreases in accuracy and other performance indicators when the MAP or BMI was not included. When using XGBoost without the MAP feature, accuracy significantly decreased to 78.4%. However, using XGBoost without BMI led to a drop in accuracy to 82.6%, indicating that both MAP and BMI are important for enhancing classification performance. The accuracy dropped to 80.3% when using XGBoost without BMI and MAP, which confirms that the integration of these two features enhances the XGBoost model's performance and increases its capacity for more accurate prediction.

Table 5: The experimental results of feature extraction techniques (MAP and BMI) on model performance.

Techniques	Accuracy	Precision	Recall	F1 Score	Specificity
XGBoost_with MAP,BMI	87.83 %	93.10 %	81.54 %	86.94 %	94.04 %
XGBoost_without MAP	78.4 %	82.5 %	71.8 %	76.8 %	85.0 %
XGBoost_without BMI	82.6 %	82.5 %	82.4 %	82.5 %	82.7 %
XGBoost_without MAP,BMI	80.3 %	85.3 %	72.8 %	78.6 %	87.6 %

Moreover, this study tested the overall system integrated with AWS services, including AWS Lambda, Amazon Gateway, and S3. A POST request sent sample data from the test dataset via the REST API to Lambda. Then, the model prediction response was received. The average response time of the model to the prediction request was around 199 ms, which ensured that it is fast, efficient, and suitable for applications requiring real-time data prediction, such as healthcare automation that requires low latency.

Table 6: Comparison with the literature review.

References	Best Method	Best Accuracy	Total Entries	Small or Large Dataset	Dataset Name
[18]	Logistic Regression	85%	270	small	UCI StatLog Heart Disease
[19]	KNN	90.78%	303	small	UCI Cardiovascular
[20]	SVM	87%	303	small	Kaggle Cleveland Clinic
[21]	Random Forest	89.01%	303	small	UCI Cardiovascular
[22]	Decision Trees	93.19%	303	small	Kaggle Cleveland Clinic
[23]	Naïve Bayes	97%	303	small	UCI Cardiovascular
[24]	XGBoost	73.74%	70,000	large	Kaggle Cardiovascular Disease
[25]	Stacking of KNN	75.1%	70,000	large	Kaggle Cardiovascular Disease
[26]	Decision tree	72.77%	70,000	large	Kaggle Cardiovascular Disease
[27]	Random Forest	73%	70,000	large	Kaggle Cardiovascular Disease
Proposed model	CatBoost	87.90 %	70,000	large	Kaggle Cardiovascular Disease

As shown in Table 6, the researcher compared the findings of this study with other articles. The researcher discovered that compared to the other study articles mentioned in the literature review section, the studies by [18-23] have demonstrated high accuracy rates for predicting heart disease using ML techniques. However, the principal limitation of this research is its restricted and small dataset, leading to a significant risk of overfitting. These models developed might not be suitable for large datasets. The presented study intends to overcome this constraint by utilizing a wider and varied dataset, which should improve the findings' generalizability. Therefore, the large Kaggle cardiovascular disease dataset was utilized. Furthermore, this dataset was also used by several previous investigations [24-27]. The researcher showed that the proposed model achieved better outcomes than the other studies that used the same large dataset. The maximum accuracy of 87.90 %, as determined by the CatBoost algorithm, showed improved accuracy.

6. Conclusion and Future Directions

The suggested architecture provided a strong basis for developing scalable machine learning systems to classify cardio prediction in real time. The main goal of this research is to classify cardiovascular disease using the best ensemble model that is implemented in Amazon Web Services (AWS). The cardiovascular diseases database was used, publicly available from the Kaggle platform. It comprises 70,000 patient records with 13 attributes. This study evaluated the CatBoost, XGBoost, and LightGBM models' performance for cardiovascular disease using evaluation criteria including accuracy, precision, recall, F1 score, and specificity. The findings showed that ensemble classifiers perform similarly across all evaluation measures, with accuracy ranging from 87.82% to 87.90%. Despite the slight differences, CatBoost achieves the highest accuracy (87.90%), indicating a marginally superior forecast overall.

Moreover, the significance of data preparation and hyperparameter optimization is demonstrated by experimental outcomes. To demonstrate, the XGBoost classifier is used as an

example. When using it with four data preparation techniques (data preprocessing, feature engineering, K-mode clustering, and hyperparameter tuning), it achieved the highest results at 87.83% accuracy, 93.10% precision, 86.94% F1 score, 94.04% specificity, and 0.962 AUC, respectively. Furthermore, how the feature extraction (MAP and BMI) affected the XGBoost model's performance was evaluated. The findings show that utilizing these features significantly improves XGBoost performance and expands its ability to make predictions with more accuracy.

Additionally, this study used Amazon's SageMaker platform to perform all ML processes. The ML model was deployed by SageMaker as an endpoint, and real-time prediction was obtained via an API gateway. This is different from many other environments that can be more complex. In addition, Amazon SageMaker offers automation and integration with several AWS services such as Amazon S3, Amazon API Gateway, and AWS Lambda. These services support machine learning stages, from scalable storage and data preparation to model deployment, in addition to the service of SageMaker AMT, which improves the performance of the classification model by determining the optimal hyperparameters based on the Bayesian method. Ultimately, using AWS Lambda can trigger machine learning, deploy models more efficiently, and reduce costs because it eliminates the need for server management and allows for faster tasks when needed. This work used different samples from the test dataset to assess the generalizability of the prediction model to other untested data. The results showed that the proposed model takes around 199 ms to respond to the prediction request. Because of this, it is fast, efficient, and suitable for smart healthcare applications that need low latency.

The future directions will focus on developing a smart healthcare web application that integrates an ML model for predicting and diagnosing cardiovascular disease. This study can be expanded to incorporate more patient data, such as other cardiovascular risk factors, which improve the model's accuracy, scalability, and comprehensiveness. Furthermore, it can integrate with other AWS services, offering an effective and reliable basis for developing scalable machine learning systems that successfully balance flexibility and powerful capabilities. Future studies can also focus on ECG analysis in real-time using portable devices, enabling early identification and prevention of heart disease. Furthermore, to gain a more profound knowledge of the k-modes clustering algorithm's performance, future studies could compare it with other well-known clustering algorithms, such as DBSCAN or k-means.

References

[1] World Health Organization (WHO), "The Top 10 Causes of Death," Americas, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[2] C. Estes et al., "Modeling NAFLD disease burden in China, France, Germany, Italy, Japan, Spain, United Kingdom, and United States for the period 2016–2030," *Journal of Hepatology,* vol. 69, no. 4, pp. 896-904, 2018.

[3] K. Drożdż et al., "Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: A machine learning approach," *Cardiovascular Diabetology,* vol. 21, no. 1, pp. 240, 2022.

[4] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access,* vol. 7, pp. 81542-81554, 2019.

[5] S. Shah, F. Shah, S. Hussain, and S. Batool, "Support vector machines-based heart disease diagnosis using feature subset, wrapping selection and extraction methods," *Computers & Electrical Engineering ,* vol. 84, pp. 106628, 2020.

[6] R. Das, I. Turkoglu, and Sengur, "Effective diagnosis of heart disease through neural networks ensembles," *Expert Systems with Applications,* vol. 36, no. 4, pp. 7675-7680, 2009.

[7] S. Ghwanmeh, A. Mohammad, and A. Al-Ibrahim, "Innovative artificial neural networks-based decision support system for heart diseases diagnosis," *Journal of Intelligent Learning Systems and Applications,* vol. 6, no. 3, pp. 176-183, 2013.

[8] Q. Al-Shayea, "Artificial neural networks in medical diagnosis," *International Journal of Computer Science Issues,* vol. 8, no. 2, pp. 150–154, 2011.

[9] K. Vanisree and J. Singaraju, "Decision support system for congenital heart disease diagnosis," *International Journal of Computer Applications,* vol. 19, no. 6, pp. 6-12, 2011.

[10] K. Saxena and R. Sharma, "Efficient Heart Disease Prediction System," *Procedia Computer Science,* vol. 85, pp. 962–969, 2016.

[11] T Kotsiopoulos, et al., "Machine learning and deep learning in smart manufacturing: The smart grid paradigm," *Computer Science Review,* vol. 40, pp. 100341, 2021.

[12] P. Shimpi, S. Shah, M. Shroff, and A. Godbole, "Machine Learning Approach for the classification of Cardiac Arrhythmia," *In Proceedings of the International Conference on Computing Methodologies and Communication (ICCMC)*, 2017.

[13] A. Methaila, P. Kansal, H. Arya, and P. Kumar, "Early heart disease prediction using data mining," *Computer Science & Information Technology Journal,* vol. 24, pp. 53–59, 2014.

[14] K. Venkateswar, "Using Amazon {SageMaker} to Operationalize Machine Learning," usenix.org, 2019. [Online]. Available: https://www.usenix.org/conference/opml19/presentation/venkateswar

[15] D. Waigi, D. Choudhary, D. Fulzele, and D. Mishra, "Predicting the risk of heart disease using advanced machine learning approach," *European Journal of Molecular & Clinical Medicine,* vol. 7, no. 7, pp. 1638-1645, 2020.

[16] L. Breiman, "Random forests," *Springer,* vol. 45, pp. 5-32, 2001.

[17] M. Gietzelt, K. Wolf, M. Marschollek, and R. Haux, "Performance comparison of accelerometer calibration algorithms based on 3D-ellipsoid fitting methods," *Computer Methods and Programs in Biomedicine,* vol. 111, no. 1, pp. 62-71, 2013.

[18] A. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease," *Neural Computing and Applications,* vol. 29, pp. 685-693, 2018.

[19] D. Shah, S. Patel, and S. Bharti, "Heart Disease Prediction using Machine Learning Techniques," *SN Computer Science,* vol. 1, no. 6, pp. 345, 2020.

[20] R. Perumal and A. Kaladevi, "Early Prediction of Coronary Heart Disease from Cleveland Dataset using Machine Learning Techniques," *International Journal of Advanced Science and Technology,* vol. 29, no. 6, pp. 4225–4234, 2020.

[21] A. Shimaa, "Proposed paradigm for intelligent heart disease prediction system using data mining techniques," *Journal of Southwest Jiaotong University,* vol. 56, no. 4, pp. 220-240, 2021.

[22] F. Alotaibi, "Implementation of Machine Learning Model to Predict Heart Failure Disease," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 6, 2019.

[23] M. Rahma and A. Salman, "Heart Disease Classification–Based on the Best Machine Learning Model," *Iraqi Journal of Science*, vol. 63, no. 9, pp. 3966-3976, 2022.

[24] N. Hasan and Y. Bao, "Comparing different feature selection algorithms for cardiovascular disease prediction," *Health and Technology,* vol. 11, no. 1, pp. 49-62, 2021.

[25] V. Shorewala, "Early detection of coronary heart disease using ensemble techniques," *Informatics in Medicine Unlocked,* vol. 26, pp. 100655, 2021.

[26] D. Waigi et al., "Predicting the risk of heart disease using advanced machine learning approach," *European Journal of Molecular & Clinical Medicine,* vol. 7, no. 7, pp. 1638-1645, 2020.

[27] J. Maiga and G. Hungilo, "Comparison of Machine Learning Models in Prediction of Cardiovascular Disease Using health record data," *In Proceedings of International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS). IEEE*, 2019.

[28] SM Khazaal and H Maarouf., " Predicting coronary artery disease utilizing support vector machines: optimizing predictive model," *Mesopotamian Journal of Artificial Intelligence in Healthcare,* vol. 2023, pp. 21-26, 2023.

[29] V. Perrone, Shen H, Zolic A, Shcherbatyi I, Ahmed A, Bansal T, Donini M, Winkelmolen F, Jenatton R, Faddoul JB, and Pogorzelska B, "Amazon Sagemaker automatic model tuning: Scalable gradient-free optimization," *In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* , 2021.

[30] G. Saad, "Design and implementation of heart diseases classification system based on ECG signal processing and deep learning,"2024

[31] A. Jana, "Framework for Automated Machine Learning Workflows: Building End-to-End MLOps Tools for Scalable Systems on AWS," *Journal of Artificial Intelligence, Machine Learning and Data Science (JAIMLD),* vol. 1, no. 3, pp. 575-579, 2023.

[32] T Chen, and C Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd International Conference on Knowledge Discovery and Data Mining*, 2016.

[33] Y. Chen, X. Wang, Y. Jung, V. Abedi, R. Zand, M. Bikak, and M. Adibuzzaman, "Classification of short single-lead electrocardiograms (ECGs) for atrial fibrillation detection using piecewise linear spline and XGBoost," *Physiological Measurement,* vol. 39, no. 10, pp. 104006, 2018.

[34] L. Torlay, M. Perrone, E. Thomas, and M. Baciu , "Machine learning–XGBoost analysis of language networks to classify patients with epilepsy," *Brain Informatics,* vol. 4, pp. 159-169, 2017.

[35] A.R. Rosendael et al., "Maximization of the usage of coronary CTA derived plaque information using a machine learning based algorithm to improve risk stratification," *Journal of Cardiovascular Computed Tomography,* vol. 12, no. 3, pp. 204-209, 2018.

[36] J. Yang and J. Guan, "A heart disease prediction model based on feature optimization and smote-Xgboost algorithm," *Information MDPI,* vol. 13, no. 10, p. 475, 2022.

[37] Kaggle, " Cardiovascular Disease Dataset," 2022.

[38] [Online]. Available: https://www.kaggle.com/datasets/sulianova/cardiovascular-diseasedataset

[39] E. Liberty et al., "Elastic machine learning algorithms in Amazon Sagemaker," *In Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, 2020.

[40] H. Singh , Practical Machine Learning with AWS: Process, Build, Deploy, and Productionize Your Models Using AWS, 2021.

[41] AWS, "Amazon API Gateway," 2017. [Online]. Available: https://aws.amazon.com/api-gateway/

[42] A.Works, "AWS Lambda," 2017. [Online]. Available: http://docs.aws.amazon.com/lambda/latest/dg/lambdaintroduction

[43] S.S. Khan et al., "Association of body mass index with lifetime risk of cardiovascular disease and compression of morbidity," *JAMA Cardiology,* vol. 3, no. 4, pp. 280-287, 2018.

[44] D. Mohajan, H. Mohajan, "Body mass index (BMI) is a popular anthropometric tool to measure obesity among adults," *Journal of Innovations in Medical Research,* vol. 2, no. 4, pp. 25-33, 2023.

[45] A.P. Kengne et al., "Blood Pressure Variables and Cardiovascular Risk," *Hypertension,* vol. 54, no. 2, pp. 399-404, 2009.

[46] D. Yu, Z. Zhao, and D. Simmons, "Interaction between Mean Arterial Pressure and HbA1c in Prediction of Cardiovascular Disease Hospitalisation: A Population-Based Case-Control Study," *Journal of Diabetes Research,* no. 1, 2016.

[47] R. Rivero and P. Garcia, "A Comparative Study of Discretization Techniques for Naive Bayes Classifiers," *IEEE Trans. Knowl. Data Eng.,* vol. 21, pp. 674-88, 2009.

[48] P.V. Balachandran et al., "Importance of feature selection in machine learning and adaptive design for materials," *In Materials Discovery and Design: by Means of Data Science and Optimal Learning,* vol. 280, pp. 59-79, 2018.

[49] X. Zeng, J. Huang, and C. Ding, "Soft-Ranking Label Encoding for Robust Facial Age Estimation," *IEEE Access,* vol. 8, pp. 134209-134218, 2020.

[50] Z. Huang , "A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining," *DMKD,* vol. 3, no. 8, pp. 34-39, 1997.

[51] A. Maas, Y. Appelman, "Gender differences in coronary heart disease," *Netherlands Heart Journal,* vol. 18, pp. 598-603, 2010.

[52] Y. Gao and Q. Liu, "An Over Sampling Method of Unbalanced Data Based on Ant Colony Clustering," *IEEE Access,* vol. 9, pp. 130990–130996, 2021.

[53] K. Sechidis, G. Tsoumakas and I. Vlahavas, "On the stratification of multi-label data," *In Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer: Berlin/Heidelberg, Germany, 2011.

[54] E. Liberty, K. Lang and K. Shmakov, "Stratified sampling meets machine learning," *In Proceedings of the International Conference on Machine Learning*, New York, NY, USA, 2016.

[55] H. Naz and S. Ahuja, "Deep learning approach for diabetes prediction using PIMA Indian dataset," *Journal of Diabetes & Metabolic Disorders,* vol. 19, pp. 391–403, 2020.

[56] S. Prusty, S. Patnaik and S. Dash, "SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer," *Front. Nanotechnol,* vol. 4, pp. 972421, 2022.