ISSN 2723-6471

Diagnosing Cardiovascular Diseases using Optimized Machine Learning Algorithms with GridSearchCV

Khalid Alemerien[1,*], Saleel Alsarayreh[2], Enshirah Altarawneh[3]

[1,2]*Department of Information Technology, Tafila Technical University, Tafila and 66110, Jordan*

[3]*Department of Computer Engineering, Hashemite University, Jordan*

(Received: August 14, 2024; Revised: September 13, 2024; Accepted: September 22, 2024; Available online: October 15, 2024)

Abstract

Accurate and timely diseases diagnosis is the most important responsibility in the healthcare industry for protecting the people lives. Many lives can be spared from death if their cases diagnosed accurately and early. One of the dangerous diseases is cardiovascular disease (CVD), is the leading cause of death worldwide, making it one of the hardest conditions to diagnose. Globally, about 17.9 million of people are died because of the cardiovascular disease. In order to assist physicians in this mission, automated solutions based on machine learning and deep learning techniques are introduced. Therefore, machine learning algorithms can diagnose diseases quickly and accurately, which adds a huge value to the medical industry. This gives physicians and patients plenty of time. To address this issue, we utilized several supervised machine learning (ML) techniques with GridSearchCV optimizer. Using the optimization techniques can enhance the performance and accuracy of proposed ML-based models. Therefore, we conducted a comparative analysis study to identify the most efficient classification model using two benchmark real datasets from the online Kaggle repository. Seven popular machine learning techniques were utilized: Decision Tree (DT), Support Vector Machine (SVM), Logistic regression (LR), K-Nearest Neighbor (KNN), Random Forest (RF), XGBoost and Naïve Bayes (NB). The findings revealed that both Random Forest and XGBoost classifiers yields highest results in both of the datasets used in our study in terms of accuracy 95.38% and 98.54%, respectively. The rest of ML algorithms showed less performance in predicting the CVD in terms of accuracy, where DT and RF achieved an accuracy of 98.53% and 98.52%, respectively, on the first dataset. Furthermore, employing the proposed ML-based model in the diagnosing CVD process shows the expected implications for patients and physicians. In addition, it shows the impact of constructing a real comprehensive dataset to enhance the performance of proposed solutions.

Keywords: Cardiovascular Disease, Heart Diseases, Machine Learning, Random Forest (RF), XGBoost, GridSearchCV Optimizer

1. Introduction

The human body consists of different organs, each with its unique functions. Among these organs is the heart, which is responsible for circulating blood throughout the body. Failure to perform this vital function can lead to life-threatening situations [1]. Presently, one of the leading causes of mortality is heart diseases [2]. Predicting heart diseases is considered one of the most challenging and intricate cases within the realm of medical science. The heart, being an essential organ, holds great significance for the overall wellbeing of the human body [3]. CVD is a global health concern, with approximately 17 million people dying from it every year worldwide.

Due to the heart disorders, the prevalence of cardiovascular disease is particularly high in Asian countries such as India and Pakistan [4]. Despite advancements in medical knowledge and technology, accurately predicting heart disease remains a significant hurdle. According to the World Health Organization (WHO), 37% of the deaths of CVD in poor nations due to late prediction of heart [5]. This highlights the need for high performance prediction methods and diagnostic tools to effectively address this critical health issue. Therefore, the primary goal of health care sector is to safeguard, improve, and promote population health, which must first be measured. Consequently, machine learning and deep learning techniques can assist in finding patterns, relationships, and models that support predictive and decision-making processes for diagnosis and treatment planning [6]. By leveraging large datasets and advanced analytical techniques, machine learning can uncover patterns and risk factors that may not be readily apparent to human

***Corresponding author: Khalid Alemerien (Khalid.alemerien@ttu.edu.jo)**
DOI: https://doi.org/10.47738/jads.v5i4.280

ISSN 2723-6471

clinicians [7]. This study aims to explore the effectiveness of machine learning algorithms in predicting heart diseases especially CVD. By analyzing a comprehensive dataset of patient records, we seek to develop a predictive model that can assist medical professionals in making accurate diagnoses and identifying individuals at high risk of developing cardiovascular complications. The findings of this research have the potential to significantly impact clinical practice and contribute to the prevention and early intervention of heart disease, ultimately improving patient outcomes and reducing mortality rates associated with CVD. Therefore, the following are the primary contributions of the suggested work: First, we try to draw attention to the dataset issues before standardizing and improving the datasets. After that, the selected ML classifiers were trained and tested using the datasets to see which ones produce the best performance in terms of accuracy scores. Second, the GridSearchCV optimizer was employed by the authors to determine the optimal values for the hyperparameter. Third, to attain the maximum accuracy through hyperparameter tuning, we employ the ML classifiers with the optimal hyperparameter indices. Finally, the proposed prediction models (XGBoost and RF) provide cutting-edge accuracy.

The remainder of the paper is organized as follows: In Section II, we present an appraisal of the literature on ML-based approaches. The machine learning algorithms of the research are outlined in Section III. We describe the proposed methodology followed in the study in Section IV. Section V demonstrates the experimental findings. Section VI concludes the paper and provides future work.

2. Literature Review

Machine learning (ML) and Deep learning (DL) approaches have advanced significantly in the healthcare sector in recent years. These methods have gained popularity and shown to be effective in a number of healthcare applications, most notably medical cardiology. Several researchers have developed techniques that use supervised ML-based approaches to predict the cardiac diseases. Some of these approaches are explained as follows:

In 2023, Roy et al., [8] used RF, Kstar, ZeroR, and voted perception algorithms to predict the heart diseases. The accuracy of these classifiers in this analysis study are as follows: RF had the highest accuracy (97.69%) and voted perception (94.39%). The study was conducted using UCI dataset. Bhatt et al., [9] proposed a model that used k-mode clustering to enhance the performance of classification with XGBoost, DT, RF, and Multi-layer perception (MLP). A dataset on Kaggle repository was used to train and test the classifiers. The multilayer perception classifier achieved the best evaluation in terms of accuracy (87.28%).

Arumugam et al., [10] used multiple classification algorithms (DT, NB, and SVM) to predict the likelihood of cardiac disease in diabetes patients. DT C4.5 algorithm showed the best accuracy (90%). The Cleveland dataset was used for conducting the study. The only accuracy metric was provided by authors. Furthermore, Aladeyelu and Adekunle [11] utilized and compared the performance of five algorithms: LR, DT, NB, KNN, and SVM. The study utilized a dataset obtained from the Kaggle repository. Upon analyzing the results, the SVM algorithm achieved the best accuracy score of 80% and outperformed all other models in all evaluation metrics. Additionally, it attained a precision score of 78%.

Kadhim and Radhi [12] proposed a model based on ML algorithms that has three stages. The first stage involves the collection and processing of patient data. The second stage focuses on training and testing the data using ML algorithms, including RF, SVM, KNN, and DT. Among these algorithms, the RF algorithm achieved the highest classification accuracy of 94.95%. In the third stage, the classification results were further optimized using a hyperparameter optimization technique called random search. The RF with the random search technique resulted in the best accuracy of 95.4%. These findings demonstrate the effectiveness of the used classifiers to accurately detect patient data, resulting in more effective decision-making.

Chandrasekhar and Peddakrishna [13] applied six classifiers (RF, KNN, LR, NB, Gboosting, and AdaBoost) on datasets that were collected from two popular data repositories: The Cleveland and IEEE Dataport. The objective of this study was to maximize model accuracy by employing GridsearchCV optimizer and five-fold cross-validation technique. LR surpassed the other methods on the Cleveland dataset, with an accuracy of 90.16%, while AdaBoost prevailed on the IEEE Dataport dataset, with 90% accuracy. Further accuracy improvement was achieved by using a soft voting ensemble classifier, which integrated all six algorithms, resulting in an accuracy of 93.44% for the Cleveland

ISSN 2723-6471

dataset and 95% for the IEEE Dataport dataset. On both datasets, these findings outperformed the performance of LR and AdaBoost techniques.

In 2022, Anderies et al., [14] utilized SVM, Naive Bayes, LR, DT and KNN classifiers to predict heart diseases using UCI dataset. A comparative study conducted to compare the outcomes of utilized classifiers. The SVM technique achieved the best results, with accuracy (85%) and precision (97%). In the work [15], nine proposed machine learning based models using multinomial Naïve Bayes (MNB), SVM, LR, XGBoost, Extra Tree (ET), RF, AdaBoost (AB), Linear discriminant analysis (LDA), and DT-CART. The highest accuracy was achieved with SVM (96.72%). Jiang [16] proposed models based on machine learning algorithms using LR, RF, XGBoost, and Neural Networks for predicting heart diseases. The RF based model achieved the highest accuracy (88.5%) and the other models achieved accuracy above 80%. The UCI was utilized in this study with 303 instances. Yilmaz and Yagin [17] proposed three distinct models that were developed using the RF, LR, and SVM to classify coronary heart disease. Hyperparameter optimization was performed using the 3-repeat 10-fold repeated cross-validation methods. The performance of the three models was examined using accuracy (RF (92.9%), SVM (89.7%), and LR (86.1%)) for classifying coronary heart disease.

In 2021, Alaawi and Alsuwat [18] utilized nine different machine learning algorithms: ANN, SVM, DT, LR, KNN, RF, Voting Classifier, GBoost, and NB. The RF model achieved the highest performance compared with other techniques employed in the study, achieving 94% accuracy with the cardiovascular disease dataset. The GBoost model achieved the highest performance at 73% in terms of accuracy. Katarya and Meena [19] conducted a comparative study to evaluate and analyze the performance of machine learning and deep learning algorithms for predicting heart diseases. In this study, the algorithms were utilized as follows: RF, KNN, SVM, NB, DT, MLP, DNN, and ANN. The performance of algorithms was measured using Root Mean Squared Error (RMSE), accuracy, precision, recall, and Mean Absolute Error (MAE). The RF algorithm attained the highest accuracy (95.6%) and lowest RMSE (0.0439). The study used the UCI dataset.

In 2020, Kumar et al., [20] examined machine learning algorithms such as, DT, SVM, RF, KNN, and LR. The UCI dataset with 304 instances and 10 features was used in this experiment. The results of this work showed that the high accuracy was 74.28% for Logistic Regression, 77.14% for SVM, and 85.71% for the Random Forest algorithm. KNN had the lowest accuracy (68.57%). Sampling techniques were utilized on the dataset. However, they didn't filter the dataset's data; instead, they applied machine learning methods directly. Singh and Kumar [21] examined the effectiveness of several machine learning algorithms for heart disease prediction, including KNN, DT, linear regression, and SVM. The UCI repository site is the source of all prediction datasets. Python software is utilized for algorithm implementation. Jupyter Notebook is used to process each and every algorithm. Based on the experimental results, the authors found that the KNN method had the highest accuracy (87%), followed by the SVM (83%), DT (79%), and linear regression (78%). These algorithms were all used to predict the risk of heart disease. Shah et al., [22] conducted an analysis study including KNN, RF, DT, and Naïve Bayes algorithms to predict heart diseases. The UCI dataset was selected for this study. The highest accuracy score was achieved with KNN (90.789%) followed by Naïve Bayes (88.157%).

In 2016, Saqlain et al., [23] ran an experiment to discover heart failure utilizing unstructured patient data. In this study, a number of ML algorithms were employed. The findings, in terms of accuracy, of this study are as follows: Neural network (84%), SVM (83%), RF (86%), DT (86%), and Logistic Regression (80%).

Machine learning-based techniques to detect heart disease, such as cardiovascular, would be excellent clinical tools, but these tools pose a high challenge to construct. Comparing with above mentioned research studies, our research focus on preprocessing the dataset and using GridSearchCV optimizer to enhance the performance of the proposed models. We performed the proposed model using two benchmark datasets to confirm its effectiveness in clinical fields. Furthermore, we achieved the best results with the proposed model using XGBoost in terms of accuracy (98.53%) and precision (98.58%) in a comparison with the state of art studies in predicting CVD.

 ISSN 2723-6471

3. Methods

Figure 1 shows the main steps of the proposed model, including dataset selection from Kaggle repository, dataset preprocessing, feature selection, splitting the dataset into training and testing datasets, dataset optimization with GridSearchCV, machine learning execution, and the model evaluation and results analysis.

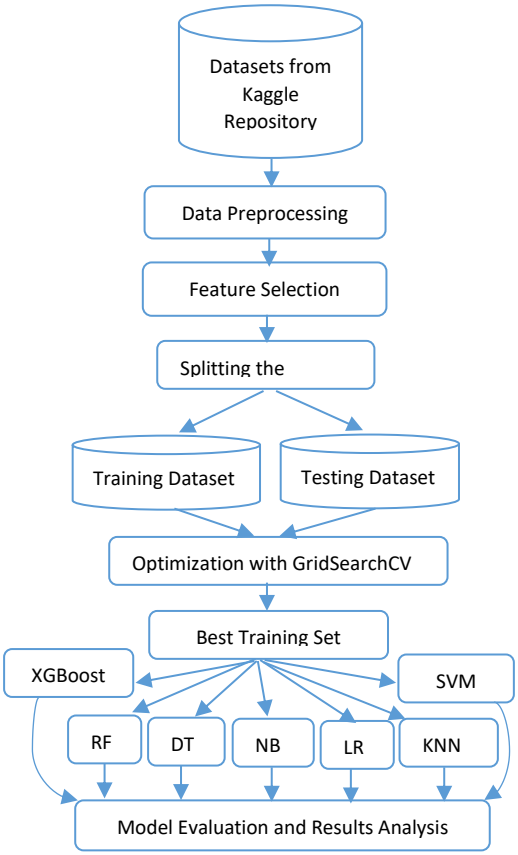

Figure 1. The proposed Model

3.1. Datasets Selection

We utilized two separate datasets from Kaggle repository for our study analysis [24], [25]. Both datasets underwent the same data preprocessing steps and feature selection procedures. Additionally, we employed the same machine learning techniques on each dataset individually, without merging them. While the two datasets share common features, one of them contains two additional features. Despite this, we treated and processed each dataset independently without merging them together. We were unable to merge the two datasets because they do not have the same features. There is a difference between them, particularly in a specific feature that is missing in one of the datasets. This missing feature has a significant impact, and it cannot be deleted under any circumstances. The datasets details are discussed as follows: The first dataset, as shown in in table 1, includes of four datasets: Cleveland, Hungary, Switzerland, and Long Beach V. This dataset has 76 features, including the predicted feature, but the majority of state-of-the-art experiments refer to using a subset of 14 features. The dataset has 1025 instances and 14 features, 13 of which are data features and one class feature [17].

Table 1. The Health Disease Dataset 1

No.	Feature	Description
1	age	Age(years)
2	sex	[discrete categorical factor], where female (0) and male (1)
3	CP	Pain of Chest [discrete categorical factor], where typical angina (0), atypical angina (1), non-anginal pain (2), and A symptomatic (3).

ISSN 2723-6471

4	trestbps	Blood pressure rate at rest (mm Hg)
5	chol	Cholesterol measurement (mg/dl)
6	fbs	A fasting blood sugar level of less than 120 mg/dl is measured as 0, and a level greater than 120 mg/dl is measured as 1.
7	Rest_ecg	Electrographic test at rest [discrete categorical factor], where the ST T wave abnormality (1) and normal (0)
8	Tha_lach	Highest heart rate obtained
9	Ex_ang	Exercise-induced angina [discrete categorical factor], where (No (0) and Yes (1))
10	Old_peak	Exercise-induced ST depression, in contrast with rest status V.
11	slope	ST segment's slope of exercise [discrete categorical factor] where (downslope (2), flat (1), and unslope (0))
12	ca	Number of substantial vessels with fluoroscopically color [0-3]
13	thal	Defect type [discrete categorical factor], where (reversible defect (2), fixed (1), and Normal (0))
14	target	Has cardiac disease, where (No (0) and Yes (1))

The second utilized dataset combines the five most widely used datasets for detecting cardiovascular diseases: Cleveland, Long Beach, Switzerland, Hungary, and the Statlog heart datasets. This dataset includes a total of 1190 cases and 11 attributes. The nature of the dataset is discussed in [26]. Table 2 provides the details of this dataset.

Table 2. The Heart Disease Dataset 2

No.	Feature	Description
1	age	Age (years)
2	sex	[discrete categorical factor], where female (0) and male (1)
3	Chest_pain_type	Chest pain [discrete categorical factor], where typical angina (1), atypical angina (2), non-anginal pain (3), and A symptomatic (4).
4	Resting_bps	Blood pressure rate at rest (mm Hg)
5	Cholesterol	Serum Cholesterol (mg/dl)
6	Fasting_blood_sugar	Fasting blood sugar rate (less than 120 mg/dl (0) and more than 120 mg/dl (1))
7	Resting_ecg	Electrographic test at rest [discrete categorical factor], where (normal (0) and has ST T wave abnormality (1))
8	Tha_lach	Maximum heart rate obtained
9	Exercise_ angina	Exercise-induced angina [discrete categorical factor] (0 = N, 1 = Y)
10	Old_peak	Exercise-induced ST depression, in contrast with rest status V.
11	ST_slope	ST segment's slope of exercise [discrete categorical factor] where (unslope (0), flat (1), and downslope (2))
12	target	Has cardiac disease, where (No (0) and Yes (1))

3.2. Preprocessing of Datasets

The two Kaggle datasets were utilized in order to predict cardiovascular disease. We utilized the Colab platform for data preprocessing. Colab, developed by Google, is a powerful and beneficial tool for executing advanced computations and analyses. Colab offers several notable features. Firstly, it is free to use and provides cloud computing resources without the need for installing software and tools on the local machine.

Secondly, Colab provides an integrated and robust development environment that allows easy coding and execution in the Python language, along with support for various useful libraries and extensions. By using Colab, we were able to efficiently and swiftly perform data preprocessing tasks. The optimized environment of Colab provides powerful computing resources and access to high-performance processing units (GPUs and CPUs), facilitating computationally intensive operations and analysis of large datasets.

 ISSN 2723-6471

In the data cleaning step, we first ensured that there were no missing values. To accomplish this, we utilized the isnull() function from the Python Pandas library, which allowed us to check for any null or missing data in the dataset. It is pertinent to remember that the dataset was purely numerical, containing only numeric data. The actions involved in implementation are: we imported the libraries that are needed to run the experiments: NumPy, Pandas, and Scikitlearn. Then, we scaled the features through the data normalization in the datasets by performing standardization using the StandardScalar() class and FitTransform() function of the scikit-learn library. FitTransform() function calculates the mean and standard deviation of such features in the datasets to scale the data for training and test processes. The FitTransform() function combines the fit() and transform() functions. The sklearn.preprocessing.StandardScaler() class includes these functions for performing the data transformation. Moreover, we applied the proposed ML-supervised models. In the last step, the proposed model with the best performance was deployed for production.

3.3. Model Development

We select the following machine learning algorithms that are the most effective algorithms to predict the heart diseases [8], [9], and [18]. These state-of-the-art studies employed these classification algorithms to predict the heart disease, which is the research issue that has been investigated in our study.

3.3.1. Decision Tree (DT)

DT is ranked among the most extensively utilized machine learning methods for classification purposes [27]. DT designs decision logic by evaluating and matching results to classify data elements into a tree structure. The DT classifier is straightforward to use, quick to comprehend, and easier to interpret. It does not require parameter setting or domain expertise [28]. DT creates a structure akin to a tree with potential solutions to an issue depending on specific limitations. DT obtains its name from the fact that it starts as a single, straightforward decision, or root, and grows into several branches until a decision or prediction is made, eventually forming a tree [29].

3.3.2. Random Forest (RF)

RF is an ensemble learning algorithm, which generates assortment decision trees using the training data. Each decision tree predicts a class as an output, and the final result is determined by taking the class that is predicted by the majority of the decision trees, in the case of classification problems [30]. The algorithm allows us to specify the number of trees to create. RF utilizes a technique called bootstrap aggregating or bagging, which helps reduce the variance in the results. By combining the predictions of multiple trees, RF improves the overall accuracy and generalization of the model [31]. RF is a powerful algorithm that leverages the collective knowledge of assortment decision trees to make robust predictions. RF is particularly effective in reducing overfitting and providing reliable results in classification tasks [30]. The Gini index or Gain information were used to judge which feature can be used to generate the decision tree. The equations (1) and (2) used to calculate the Gini index and Gain information, respectively.

$$\text{Gini} = 1 - \sum_{i=0}^{c-1} [\text{Pt}]^2 \qquad \text{c is the number of classes} \tag{1}$$

$$\text{Entropy} = \sum_i - \text{Pi} \, (\log_2 \text{Pi}) \qquad \text{Pi is probability of class i} \tag{2}$$

3.3.3. Support Vector Machine (SVM)

SVM is a potent mathematical computational algorithm for classification task. SVM is a supervised learning technique used in the domains of regression and classification. It is efficacious and has robust statistical basis. SVM excel not only in linear classification but also in non-linear classification through the use of kernel functions [10]. SVM are the latest advancements in supervised machine learning approaches [32]. The primary function of the SVM technique is to create a line of hyperplanes that separates a dataset into two groups. Support vectors are the data points which are closest to the hyperplane or points of the dataset that, if omitted, could shift the position of the hyperplane segmentation. As a result, support vectors might be considered as crucial elements of the dataset [33].

3.3.4. Logistic Regression (LR)

LR is a statistical method that is used commonly for classification problems. LR is used to predict a continuous dependent variable from a binary outcome [34]. LR and linear regression are very similar except for the way they are

 ISSN 2723-6471

applied [35]. LR is a popular method that is applied in across several fields, including medical research, social sciences, and marketing, for tasks such as assessing the probability that a customer making a purchase, or predicting the presence or absence of a disease based on certain risk factors. As a result, it is an effective tool for resolving issues involving binary classification. The binary classification can be carried out with the aid of equation (3). When a linear regression algorithm predicts values larger than zero, LR uses a logistic function to transform those values into discrete values such as 0 and 1.

$$P(x) = \frac{1}{1 + e^{-y}} \quad (3)$$

P is the probability of x occurrence, which the dependent variable that has a binary value (0 or 1). y is the independent variable that is used to predict dependent variable (x) and e is the exponential constant that has an approximately value of 2.71828.

3.3.5. Naïve Bayes (NB)

NB is a very simple Bayesian network consisting of directed acyclic graphs (DAGs) with a single parent that represents the undetected node and several children that correspond to detected nodes, with a strong assumption of independence among child nodes in the context of their parent. Moreover, NB is an independence model that relies on estimation [36]. Bayesian classifiers are generally less accurate than other sophisticated learning methods, such as ANNs. However, the application of the Bayes theorem with a strong independent model (Naive) is simple to construct and has no time limit for challenging iterative parameters is known as a NB classifier [28]. NB is highly beneficial for diagnosing people with heart problems in the medical field. NB is employed to calculate the posterior probability for each class. This probability is determined by conditional probability, and it serves as the basis for classifying datasets. The NB classifier is developed using equation (4).

$$P(B|A) * P(A) = P(A|B) * P(B) \quad (4)$$

P(B|A) represents the probability of B occurring when event A has already taken place, P(A|B) represents the probability of A occurring when event B takes place first. P(B) denotes the probability of event B happening on its own, P(A) denotes the probability of event E happening on its own

3.3.6. K-Nearest Neighbor (KNN)

KNN is a widely used algorithm for mining comprehensive information from medical databases [37]. The KNN algorithm is a method for quickly and effectively classifying new data into existing data. The KNN's idea is to locate data with the closest number of K (neighbors) to the training data and the closest distance between the evaluated data [38]. The output of the algorithm identifies the closest neighbor, which is considered the most similar case. Consequently, the new case is assigned to the class that contains the closest neighbors. The KNN algorithm consists of two main steps: determining the K training examples are most similar to the unidentified example. Selecting the most frequently occurring class labels among these K examples. The KNN algorithm is a classification technique that determines the class of a new instance based on the similarity to its K nearest neighbors. It is commonly used in medical data analysis to make predictions and identify patterns [36].

3.3.7. eXtreme Gradient Boost (XGBoost)

XGBoost is an ensemble ML algorithm. The gradient is mostly helpful in lowering the loss function, which is just the real difference between the original and projected values [30]. Using residuals to train the model is the core idea behind the XGBoost, the output of the most recent tree training is utilized as the input for the subsequent iteration, and the error is steadily minimized across several serial iterations. Ultimately, the ensemble learner is generated by linearly weighting each of the weak learners [39]. The XGBoost algorithm has the ability to boost the weak learner into a strong learner for each generated tree using its optimization step and allow the ML-based model to produce fewer false alarms. In addition, XGBoost labels data easier, resulting in a more accurate classification of data. Also, XGBoost has an important aspect called regularization to effectively avoid the data overfitting problem in both tree-based and linear-based models. Based on the algorithm's application, three boosting techniques are performed within the XGBoost-based classification, including stochastic, regularized, and gradient. Furthermore, using the XGBoost algorithm can reduce the execution time and lead to optimal use of memory [40].

3.4. The GridSearchCV Optimizer

In this research, the GridSearchCV optimizer is employed to identify the optimal values for hyperparameters across all algorithms. The exhaustive search capability of GridSearchCV makes it a powerful tool for hyperparameter tuning [15]. The objective is to maximize accuracy by using ML classifiers with the optimal hyperparameter values that have been identified through the GridSearchCV optimizer. This process enables the fine-tuning of various machine learning algorithms using GSCV. We proposed a GridSearchCV optimizer that is employed in six different algorithms, to get the best suitable hyperparameters values. We use the GridSearchCV optimizer, as a cross validation technique, from Scikit Learn library in Python. The GridSearchCV optimizer calculates the score of each hyperparameter combination on the grid. Then, these combinations are passed to the dictionary in order to evaluate the model in terms of accuracy and loss. Then, the hyperparameter combination with the best performance will be chosen for the model.

In general, there are four main internal steps that are been performed in GridSearchCV optimizer: estimator object includes the machine learning model such as DT, SVM, LR, KNN, RF, and NB; param_grid, which is a dictionary includes the hyperparameters for tuning such as C and gamma; scoring based on accuracy; and training the final model using the best hyperparameters.

3.5. Evaluation Measures

We evaluated the performance of proposed models using evaluation metrics such as precision as shown in equation 5, F1-score as shown in equation 6, recall as shown in equation 7, and accuracy as shown in equation 8. False Positive (FP) denotes that the algorithm predicts a positive result incorrectly, while False Negative (FN) denotes that a model predicts a negative result incorrectly. Conversely, True Positive (TP) denotes correct positive outcomes, and True Negative (TN) indicates correct negative outcomes.

$$\text{Precision} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{(\text{TP} + \text{FN})} \tag{6}$$

$$\text{Accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})} \tag{7}$$

$$\text{F1} - \text{Score} = \frac{(\text{Precision X Recall})}{(\text{Precision} + \text{Recall})} \tag{8}$$

4. Results and Discussion

4.1. Results of the Dataset 1

The most crucial evaluation metric for ML prediction is accuracy. Table 3 demonstrates that XGBoost, KNN, DT, and RF have the best results with an accuracy greater than 98%. The XGBoost algorithm outperformed the best accuracy (98.54%). On the other hand, the performance of NB algorithm was lower than other ML algorithms. Furthermore, figure 2 demonstrates that the algorithms for XGBoost, KNN, DT, and RF performed exceptionally well in terms of accuracy scores, precision scores, recall scores, and F-score. The ROC results for dataset 1 for the models based on DT, LR, KNN, RF, NB, XGBoost, and SVM are presented in figure 3.

ISSN 2723-6471

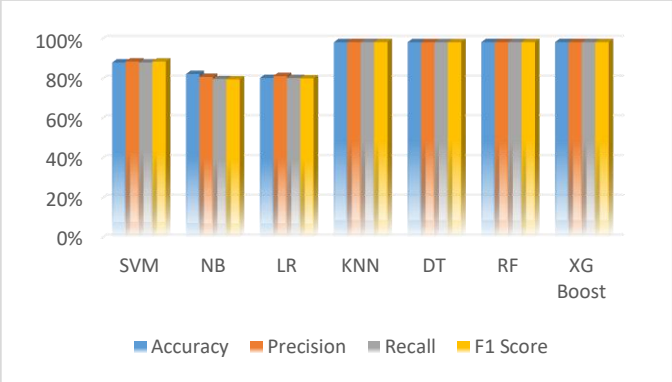

Figure 2. The accuracy, recall, precision, and F1-score results of the seven ML algorithms for dataset 1

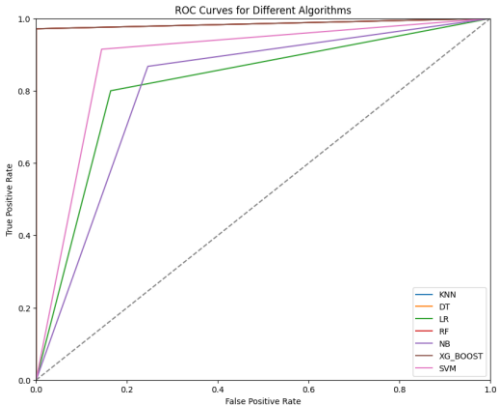

Figure 3. The ROC curve of DT, LR, KNN, RF, NB, XGBoost, and SVM based models on dataset 1.

Table 3. The Results of Seven Algorithms Using the Dataset 1

ML model	Accuracy	Precision	Recall	F1 Score
SVM	88.29%	88.78%	88.29%	88.78%
NB	82.49%	81.05%	80.00%	79.82%
LR	80.49%	81.44%	80.49%	80.33%
KNN	98.50%	98.57%	98.54%	98.54%
DT	98.52%	98.52%	98.52%	98.52%
RF	98.53%	98.57%	98.53%	98.57%
XGBoost	98.54%	98.58%	98.54%	98.58%

4.2. Results of the Dataset 2

Table 4 demonstrates that RF achieved better accuracy than other ML algorithms, providing the highest accuracy value (95.38%), and the precision value (95.40%). KNN was the only ML algorithm that performed poorly in this scenario when contrasted to other ML algorithms. The findings of the proposed model based on ML using the dataset 2 as shown in figure 4, where XGBoost and RF achieved the excellent performance using these metrics: accuracy, recall, precision, and F-score. Figure 5 demonstrates the ROC results of the DT, LR, KNN, RF, NB, XGBoost, and SVM based models on dataset 2. Figure 6 shows clearly the results of the DT, LR, KNN, RF, NB, XGBoost, and SVM based models in terms of accuracy, precision, recall, and F1-score values. As shown in the figure, the evaluation results are consistent for both datasets. This supports the robustness of proposed models.

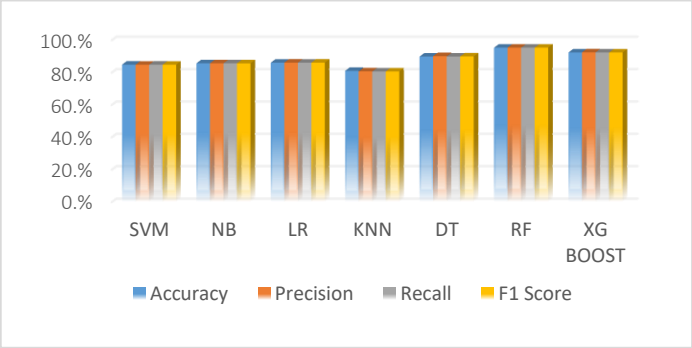

Figure 4. The accuracy, recall, precision, and F1-score results of the seven ML algorithms for dataset 2

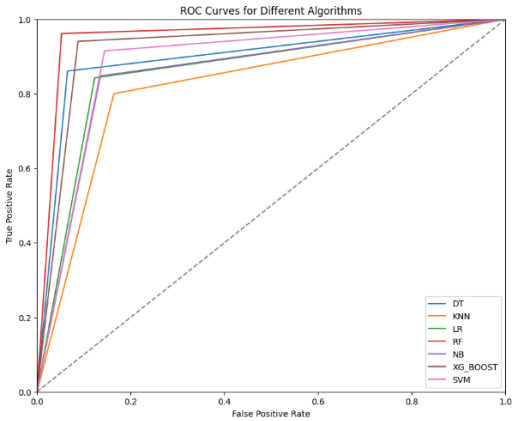

Figure 5. The ROC curve of DT, LR, KNN, RF, NB, XGBoost, and SVM based models on dataset 2.

ISSN 2723-6471

Table 4. The Results of Seven ML Algorithms Using the Dataset 2

ML model	Accuracy	Precision	Recall	F1 Score
SVM	84.87%	84.86%	84.87%	84.86%
NB	85.71%	85.71%	85.71%	85.70%
LR	86.13%	86.14%	86.13%	86.14%
KNN	81.07%	80.77%	80.67%	80.70%
DT	89.92%	90.16%	89.92%	89.94%
RF	95.38%	95.40%	95.38%	95.38%
XGBoost	92.44%	92.51%	92.44%	92.41%

4.3. A Comparative Analysis of the Performance

In table 5, we evaluate the performance of our proposed model to the models that employ machine learning approaches to predict cardiac diseases. Also, we focus on the models that used the cardiovascular disease datasets from the Kaggle repository. Our model achieved higher performance than other models as shown in the table 5. The RF and XGBoost based models with GridSearchCV optimizer achieved high accuracy scores.

The potential applications of ML models, including the proposed ML, in heart diseases diagnosis and treatment. ML-based techniques have the potential to significantly improve the early identification and diagnosis of cardiac problems. Here are several ways in which these technologies can aid in heart disease detection: (1) Risk prediction models: ML algorithms have the capability to analyze a variety of patient data, including demographic information, medical history, lifestyle aspects, and genetic predispositions, to determine the risk of developing heart diseases. The risk prediction models can assist in identifying people who may be more susceptible to heart issues, enabling early intervention and preventive measures [41]. (2) Electrocardiogram (ECG) analysis: ML models can analyze ECG data to detect patterns related with various cardiac conditions including, arrhythmias, ischemia, or heart attacks [42]. (3) Biomarker identification: ML algorithms can examine big datasets, including omics data (genomics, proteomics, and metabolomics), to recognize possible biomarkers associated with specific heart diseases. These biomarkers can be indicators of early detection and individualized treatment plans [41], [43]. (4) Remote patient monitoring: Continuous patient monitoring can be achieved through the utilization of wearable technology, ML, and sensors [43], which are used to collect and analyze patient data. Alerts can be triggered by anomalies or changes in vital signs, activity levels, or other pertinent parameters. This allows for prompt intervention and reduces the chance of cardiovascular events.

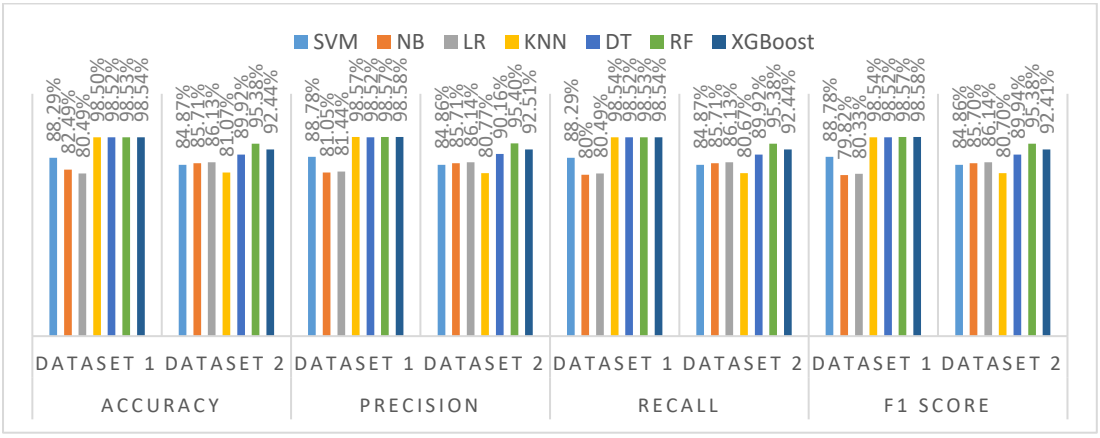

Figure 6. A comparison of evaluation metrics of DT, LR, KNN, RF, NB, XGBoost, and SVM based models on datasets.

(5) Treatment optimization: ML algorithms can assist in optimizing treatment plans by analyzing patient responses to various medications and interventions [44]. Based on individual patient characteristics and historical data, personalized

 ISSN 2723-6471

treatment recommendations can be generated. And finally (6) Applying Natural Language Processing (NLP) methods in Electronic Health Records (EHR): ML and NLP can be applied to analyze unstructured clinical notes and reports in EHRs to extract valuable information related to heart health [45]. This can boost the effectiveness of data analysis and improve the general comprehension of a patient's health status [42], [46]. To sum up, ML can significantly contribute in the early identification, diagnosis, and treatment of cardiovascular diseases by leveraging diverse data sets and offering insights that may not be immediately discernible through conventional techniques.

Table 5. The Results of the Comparative Study

Ref	Dataset	ML Algorithms	Accuracy
Bhatt et al., [9]	Kaggle heart disease dataset	MLP	87.28%
Aladeyel and Adekunl [11]		SVM	80.00%
Bora et al., [27]	UCI + Kaggle heart disease datasets	SVM (UCI)	92.00%
		RF (Kaggle)	94.13%
		RF (Combined)	93.31%
Arumugam et al., [10]	UCI Cleveland dataset	DT C4.5	90.00%
		RF	97.69%
Roy et al., [8]		VP	94.39%
		Kstar	94.05%
Kumar and et al., [20]		RF	85.71%
Singh and Kumar [21]		KNN	87.00%
Shah et al., [22]		KNN	90.789%
Yang and Guan, [39]	HDD dataset	RF	91.15%
		KNN	91.77%
		XGBoost	93.44%
The proposed model for dataset 1		DT	98.52%
		RF	98.53%
		KNN	98.5%
		XGBoost	98.54%
The proposed model for dataset 2		RF	95.38%
		XGBoost	92.44%

5. Conclusion

In this research, we employed seven distinctive ML algorithms, including DT, SVM, KNN, LR, NB, XGBoost, and RF, for the prediction of cardiovascular diseases. We employed hyperparameter tuning that plays a crucial role in improving the performance of the utilized ML models. We used two publicly available datasets from Kaggle. For dataset 1, XGBoost achieved the highest accuracy value of 98.54%. Furthermore, we achieved the highest accuracy of 95.38% and 92.44% using the RF algorithm and XGBoost, respectively, on dataset 2.

The lowest accuracy in dataset 1 was obtained by NB of 82.49% but in dataset 2 the lowest accuracy (81.07%) was achieved by KNN. In this study, we can conclude that the XGBoost and RF are the best nominated algorithms for constructing cardiovascular disease detection models based on machine learning algorithms. Each ML algorithm has strengths and weaknesses, depending on the specific methodology employed and the dataset it is applied to. A particular algorithm may yield high results in one dataset but may not perform as well in another dataset. In future work, we will conduct the proposed model on other diseases with varied datasets. Also, we will investigate the effectiveness of deep learning algorithms on large scale datasets for the disease diagnosis. We plan to construct a dataset by adding more features and instances to improve the performance of cardiovascular disease prediction systems. Finally, we continue improving the proposal model in order to make it reliable and easy to use in real applications.

6. Declarations

6.1. Author Contributions

Conceptualization: K.A., S.A., and E.A.; Methodology: S.A.; Software: K.A.; Validation: K.A. and E.A.; Formal Analysis: K.A. and E.A.; Investigation: K.A.; Resources: E.A.; Data Curation: E.A.; Writing Original Draft Preparation: K.A. and S.A.; Writing Review and Editing: E.A. and K.A.; Visualization: K.A.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Pouriyeh, S., Vahid, S., Sannino, G., De Pietro, G., Arabnia, H., and Gutierrez, J., "A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease," *In 2017 IEEE symposium on computers and communications (ISCC), July 3-6, 2017, Heraklion, Crete, Greece. IEEE*, vol. 2017, no. 7, pp. 204-207, 2017, doi: 10.1109/ISCC.2017.8024530.

[2] Zhang, D., Wang, F., Burgos, R., and Boroyevich, D. "Common Mode Circulating Current Control of Interleaved Three-Phase Two-Level Voltage-Source Converters with Discontinuous Space-Vector Modulation," *IEEE Energy Conversion Congress and Exposition, 20-24 September 2009, San Jose, California, USA. IEEE.* vol. 1, no. 6, pp. 3906-3912, 2009.

[3] Dangare, C. S., and Apte, S. S., "Improved study of heart disease prediction system using data mining classification techniques," *International Journal of Computer Applications*, vol. 47 no. 10, pp. 44-48, Jun. 2012.

[4] B. Fida, M. Nazir, N. Naveed and S. Akram, "Heart disease classification ensemble optimization using Genetic algorithm," *2011 IEEE 14th International Multi-topic Conference, Karachi, Pakistan*, vol. 2011, no. 12, pp. 19-24, 2011.

[5] AbuKaraki, A., Alrawashdeh, T., Abusaleh, S., Alksasbeh1, M., Alqudah, B., Alemerien, K., and Alshamaseen, H., " Pulmonary Edema and Pleural Effusion Detection Using EfficientNet-V1-B4 Architecture and AdamW Optimizer from Chest X-Rays Images," *Computers, Materials, and Continua*, vol. 80, no. 1, pp. 1055-1073, Jun. 2024, doi: 10.32604/cmc.2024.051420.

[6] Dehkordi, S. K., and Sajedi, H., "Prediction of disease based on prescription using data mining methods," *Health and Technology*, vol. 9, no. 1, pp. 37-44, Jan. 2019, doi: 10.1007/s12553-018-0246-2.

[7] Garg, A., and Mago, V., "Role of machine learning in medical research: A survey," *Computer science review*, vol. 40, no. 5, 100370, May 2021, doi: 10.1016/j.cosrev.2021.100370.

[8] Roy, D., Mahmood, M. A., and Roy, T. J., "An Analytical Model for Prediction of Heart Disease using Machine Learning Classifiers," *TechRxiv*, vol. 21, no. 6, pp. 1-6, Jun. 2021.

[9] Bhatt, C. M., Patel, P., Ghetia, T., and Mazzeo, P. L., "Effective heart disease prediction using machine learning techniques," *Algorithms*, vol. 16, no. 2, pp. 88, Feb. 2023, doi: 10.3390/a16020088.

[10] Arumugam, K., Naved, M., Shinde, P. P., Leiva-Chauca, O., Huaman-Osorio, A., and Gonzales-Yanac, T., "Multiple disease

 ISSN 2723-6471

prediction using Machine learning algorithms," *Materials Today: Proceedings*, vol. 80, no. 3, pp. 3682-3685, Mar. 2023, doi: 10.1016/j.matpr.2021.07.361.

[11] Aladeyelu, A. C., and Adekunle, G. T., "Predicting Heart Disease Using Machine Learning," *Journal of Multidisciplinary Engineering Science and Technology (JMEST)*, vol. 10, no. 4, pp. 15837-15841, Apr. 2023.

[12] Kadhim, M. A., and Radhi, A. M., "Heart disease classification using optimized Machine learning algorithms," *Iraqi Journal For Computer Science and Mathematics*, vol. 4 no. 2, pp. 31-42, Feb. 2023, doi:10.52866/ijcsm.2023.02.02.004.

[13] Chandrasekhar, N., and Peddakrishna, S., "Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization," *Processes*, vol. 11, no. 4, pp. 1210, Apr. 2023, doi: 10.3390/pr11041210.

[14] Anderies, A., Tchin, J. A. R. W., Putro, P. H., Darmawan, Y. P., and Gunawan, A. A. S., "Prediction of heart disease UCI dataset using machine learning algorithms," *Engineering, MAthematics and Computer Science (EMACS) Journal*, vol. 4, no. 3, pp. 87-93, Sep. 2022, doi: 10.21512/emacsjournal.v4i3.8683.

[15] Saboor, A., Usman, M., Ali, S., Samad, A., Abrar, M. F., and Ullah, N., "A method for improving prediction of human heart disease using machine learning algorithms," *Mobile Information Systems*, vol. 22, no. 3, pp. 1-9, Mar. 2022, doi: 10.1155/2022/1410169.

[16] Shu Jiang, "Heart disease prediction using machine learning algorithms," Master thesis, University of California, Los Angeles, California, USA, 2022.

[17] Yilmaz, R., and YAĞIN, F. H., "Early detection of coronary heart disease based on machine learning methods," *Medical Records*, vol. 4, no. 1, pp. 1-6, Jan. 2022, doi: 10.37990/medr.1011924.

[18] Alalawi, H. H., and Alsuwat, M. S., "Detection of cardiovascular disease using machine learning classification models, " *International Journal of Engineering Research and Technology*, vol. 10, no. 7, pp. 151-157., Jul. 2021.

[19] Katarya, R., and Meena, S. K., "Machine learning techniques for heart disease prediction: a comparative study and analysis," *Health and Technology*, vol. 11, pp. 87-97, Jan. 2021, doi: 10.1007/s12553-020-00505-7.

[20] Kumar, N. K., Sindhu, G. S., Prashanthi, D. K., and Sulthana, A. S., "Analysis and prediction of cardio vascular disease using machine learning classifiers," *In 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 6-7 March, 2020, Coimbatore, India*, vol. 2020, no. 3, pp. 15-21. IEEE. 2020, doi: 10.1109/ICACCS48705.2020.9074183.

[21] Singh, A., and Kumar, R., "Heart disease prediction using machine learning algorithms," *In 2020 international conference on electrical and electronics engineering (ICE3), 14-15 February, 2020, Gorakhpur, Uttar Pradesh, India*, vol. 2020, no. 2, pp. 452-457. IEEE. 2020, doi: 10.1109/ICE348803.2020.9122958.

[22] Shah, D., Patel, S., and Bharti, S. K., "Heart disease prediction using machine learning techniques," *SN Computer Science*, vol. 1, no. 10, pp. 1-6, Oct. 2020, doi: 10.1007/s42979-020-00365-y.

[23] Saqlain, M., Hussain, W., Saqib, N. A., and Khan, M. A. "Identification of heart failure by using unstructured data of cardiac patients," *In 45th International Conference on Parallel Processing Workshops (ICPPW), 16-19 August, 2016, Philadelphia, PA, USA*, vol. 2016, no. 8, pp. 426-431. IEEE. 2016, doi: 10.1109/ICPPW.2016.66.

[24] Lapp, D., "Heart Disease Dataset," [online]. Available: https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset, [Accessed January 13, 2024].

[25] Mahesh, B., "Machine learning algorithms-a review," *International Journal of Science and Research (IJSR)*, vol. 9, no.1, pp. 381-386, Jan. 2020, doi: 10.21275/ART20203995

[26] Siddhartha, M., "Heart Statlog Cleveland Hungary Final," [online]. Available: https://www.kaggle.com/datasets/sid321axn/heart-statlog-cleveland-hungary-final, [Accessed January 13, 2024].

[27] Bora, N., "Using Machine Learning to Predict Heart Disease," Ph.D. dissertation, Department of Computer Science and Information System, California State University San Marcos, San Marcos, California, USA. 2021.

[28] Sharma, H., and Rizvi, M. A., "Prediction of heart disease using machine learning algorithms: A survey," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 5, no. 8, pp. 99-104, Aug. 2017.

[29] Das, K., and Behera, R. N., "A survey on machine learning: concept, algorithms and applications," *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 5, no. 2, pp.1301-1309, Feb. 2017.

 ISSN 2723-6471

[30] Pal, M., and Parija, S., "Prediction of heart diseases using random forest," *In Journal of Physics: Conference Series*, vol. 1817, no. 1, pp. 012009. IOP Publishing, Mar. 2021, doi: 10.1088/1742-6596/1817/1/012009

[31] Katarya, R., and Srinivas, P., "Predicting heart disease at early stages using machine learning: A survey," *In International Conference on Electronics and Sustainable Communication Systems (ICESC), 2-4 July, 2020, Coimbatore, India*, vol. 2020, no. 7, pp. 302-305. IEEE. 2020, doi: 10.1109/ICESC48915.2020.9155586

[32] Osisanwo, F. Y., Akinsola, J. E. T., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., and Akinjobi, J., "Supervised machine learning algorithms: classification and comparison," *International Journal of Computer Trends and Technology (IJCTT)*, vol. 48, no. 3, pp. 128-138, Jun. 2017.

[33] Beyene, C., and Kamat, P., "Survey on prediction and analysis the occurrence of heart disease using data mining techniques," *International Journal of Pure and Applied Mathematics*, vol. 118. no. 8, pp. 165-174, Aug. 2018.

[34] Khanna, D., Sahu, R., Baths, V., and Deshpande, B., "Comparative study of classification techniques (SVM, logistic regression and neural networks) to predict the prevalence of heart disease," *International Journal of Machine Learning and Computing*, vol. 5, no. 5, pp. 414-419, Oct. 2015, doi: 10.7763/IJMLC.2015.V5.544.

[35] Thabtah, F., Abdelhamid, N., and Peebles, D. "A machine learning autism classification based on logistic regression analysis." *Health information science and systems*, vol.7, no.12, pp. 1-11, Jun 2019, doi: 10.1007/s13755-019-0073-5.

[36] Ayon, S. I., Islam, M. M., and Hossain, M. R., "Coronary artery heart disease prediction: a comparative study of computational intelligence techniques," *IETE Journal of Research*, vol. 68, no. 4, pp. 2488-2507, Jan. 2022, doi: 10.1080/03772063.2020.1713916.

[37] Cai, Z., Gu, J., Wen, C., Zhao, D., Huang, C., Huang, H., ... and Chen, H., "An intelligent Parkinson's disease diagnostic system based on a chaotic bacterial foraging optimization enhanced fuzzy KNN approach," *Computational and mathematical methods in medicine*, vol. 18, no. 6, pp. 1-24, Jun. 2018, doi: 10.1155/2018/2396952.

[38] Anggoro, D. A., and Kurnia, N. D., "Comparison of accuracy level of support vector machine (SVM) and K-nearest neighbors (KNN) algorithms in predicting heart disease," *International Journal of Emerging Trends in Engineering Research*, vol. 8, no. 5, pp. 1689-1694., May. 2020, doi: 10.30534/ijeter/2020/32852020.

[39] Yang, J., and Guan, J., "A heart disease prediction model based on feature optimization and smote-Xgboost algorithm," *Information*, vol. 13, no. 10, pp. 475, Oct. 2022, doi: 10.3390/info13100475.

[40] Budholiya, K., Shrivastava, S. K., and Sharma, V., "An optimized XGBoost based diagnostic system for effective prediction of heart disease." *Journal of King Saud University-Computer and Information Sciences*, vol. 34, no. 7, pp. 4514- 4523, Jul. 2022, doi: 10.1016/j.jksuci.2020.10.013

[41] DeGroat, W., Abdelhalim, H., Patel, K., Mendhe, D., Zeeshan, S., and Ahmed, Z., "Discovering biomarkers associated and predicting cardiovascular disease with high accuracy using a novel nexus of machine learning techniques for precision medicine," *Scientific reports*, vol. 14, no. 1, pp. 1-13, Jan. 2024, doi: 10.1038/s41598-023-50600-8.

[42] Kaur, D., Hughes, J. W., Rogers, A. J., Kang, G., Narayan, S. M., Ashley, E. A., and Perez, M. V., "Race, Sex, and Age Disparities in the Performance of ECG Deep Learning Models Predicting Heart Failure," *Circulation: Heart Failure*, vol. 17, no. 1, pp. 14-23, Jan. 2024, doi: 10.1161/CIRCHEARTFAILURE.123.010879.

[43] Bahbouh, N. M., Compte, S. S., Valdes, J. V., and Sen, A. A. A., "An empirical investigation into the altering health perspectives in the internet of health things," *International Journal of Information Technology*, vol. 15, no. 1, pp. 67-77, Jul. 2023, doi: 10.1007/s41870-022-01035-3.

[44] Dubey, A. K., Sinhal, A. K., and Sharma, R., "Heart disease classification through crow intelligence optimization-based deep learning approach," *International Journal of Information Technology*, vol. 16, no. 1, pp. 1-16., Jan. 2024, doi: 10.1007/s41870-023-01445-x.

[45] Thukral, A., Dhiman, S., Meher, R., and Bedi, P., "Knowledge graph enrichment from clinical narratives using NLP, NER, and biomedical ontologies for healthcare applications," *International Journal of Information Technology*, vol. 15, no. 1, pp. 53-65, Jan. 2023, doi: 10.1007/s41870-022-01145-y.

[46] Marelli, A. J., Li, C., Liu, A., Nguyen, H., Moroz, H., Brophy, J. M., ... and Li, Y., "Machine learning informed diagnosis for congenital heart disease in large claims data source," *JACC: Advances*, vol. 3, no. 2, pp. 1-12, Feb. 2024, doi: 10.1016/j.jacadv.2023.100801.