Available online at www.asric.org
ASRIC Journal on Natural Sciences 1 (2021) 10-21

Machine Learning Techniques for improving study of Heart Disease Prediction System

Abdelmegeid Amin Ali[1], Ashraf Heikal[2], Eman M. Anwar[2,3], Shaimaa M Hussien[2,3]

[1]Faculty of Computers and Information, Department of computer science, Minia University, Minia, Egypt
[2]Al-Obour High Institute for Management and Informatics, Obour, Egypt, Department of information systems
[3]Faculty of Computers and Information, Department of Information System, Minia University, Minia, Egypt
Corresponding Authors Email: abdelmgeid@yahoo.com

Received 18 May 2021; revised 3 July 2021; accepted 23 August 2021

***Abstract* – The heart is an important organ in human beings. Because even a slight error might result in weariness or death, diagnosing and forecasting cardiac disorders requires increased precision, perfection, and accuracy. There are innumerable heart-related deaths, and the number is increasing significantly every day. To address the issue, researchers use a variety of data mining and machine learning approaches to evaluate massive amounts of complex medical data, assisting healthcare providers in the prediction of heart disease. Using various data mining approaches, the suggested research predicts the likelihood of heart disease and categorizes the risk level of patients. When compared to other machine learning algorithms, the trial results show that the bagging technique with decision tree algorithm has the highest accuracy of 88.56%.**

***Keywords – Heart* disease Classification, Machine Learning, Ensemble method**

I. INTRODUCTION

One of the most essential organs in humans is the heart. It is a muscular organ that pumps blood into the body and is an essential component of the cardiovascular system [1]. The cardiovascular system is made up of all blood vessels, such as arteries, veins, and capillaries, which form a complex network of blood vessels all over the body [2]. Any obstruction or abnormality in normal blood circulation or flow from the heart could result in a slew of pains caused by coronary heart disease. These are commonly known as cardiovascular diseases (CVDs) and are among the world's deadliest ailments. CVDs include diseases of the coronary heart, cerebral vascular diseases, and blood vessel diseases [3]. CVDs are the greatest cause of mortality and disability worldwide, according to the World Health Organization (WHO) Report international Atlas on upset interference and management [4]. Despite the fact that CVDs can be prevented through lifestyle changes and other related measures, they are nevertheless on the rise on a daily basis, according to several WHO publications.

However, multiple WHO investigations have shown a global increase in CVDs, which is extremely concerning. Cardiopathy affects both men and women in the same way. According to the World Health Organization, 17.9 million people died in 2016 as a result of heart disease, accounting for 31% of all fatalities worldwide. Stroke and coronary failure account for 85% of these deaths (WHO2016) [26]. Heart diseases occur when neither the heart nor the blood arteries work normally. Another issue with upset is induration, which is commonly defined as arterial hardness. In this instance, the arteries become thicker and more rigid. Arteriosclerosis is the narrowing of blood vessels, resulting in decreased blood flow through the buildups.

Heart attacks happen when blood clots, or blockages, form in the arteries, disrupting blood flow. These illnesses are also the result of a severe cellular physiological state disturbance produced by broad genetic and molecular abnormalities in cells. As a result, people who are afflicted with or at high risk of developing certain diseases would benefit from early detection and tailored medical treatment. Smoking is one of the most dangerous things that has an impact on heart health. High cholesterol, high blood pressure, physical inactivity, a poor diet, obesity, and poorly treated polygenic disease are only a few of the variables that contribute to the development of heart disease [27]. As a result, particular aspects related to mode must be addressed in order to examine the hazard of vascular disease. As a result, essential tests such as cholesterol, electrocardiograms, chest discomfort, blood pressure, highest heart rate, and excessive sugar levels should be performed by patients to swiftly expose and predict suitable guiding scenarios. Some estimates and circumstances make it even more difficult to compare medical practitioners' work to existing patient check outcomes [28]. Cardiopathy is usually diagnosed by collecting a medical history, using a stethoscope, ultrasound, and performing a diagnostic technique (ECG).

The doctor's guess, expertise, and experience are used to announce the conclusion and compare it to the prior information stored in the database to determine if a patient with a specific illness is normal or aberrant [29]. One of the disadvantages of the

previous methods is that clinicians often examine patients to diagnose their diseases, and this method is experimental in nature, therefore there's a chance that the diagnosis could be incorrect, and so the rising rate of heart diseases has been attributed to this. As a result, the healthcare industry must develop and improve how these diseases are addressed in order to reduce their social impact. Within the healthcare industry, there is a wealth of knowledge [5], most notably the gut illness data, which must be efficiently processed for effective call creation. According to data, statistics, clinical records, and hospital administration, medical data doubles every three years, making the health sector a multibillion-dollar realm [6]. Medical data analysis and information extraction rely heavily on machine learning and data processing techniques. The rising morbidity and mortality rates caused by cardiopathy around the world have prompted researchers to conduct a number of studies in an attempt to minimize the rates. In the deployment of clinical call support systems for cardiac disease prediction, data processing, and machine learning techniques are widely used. Information mining applications are utilized to improve health policy and reduce hospital errors, as well as early detection, disease prevention, and avoidable hospital fatalities [7].

The following is how the rest of the paper is organized: Section 2 outlines a literature review of current analysis proposals on this topic. Section 3 explains the proposed design and strategy. The findings of the experiment are provided in Section 4. Finally, in Section 5, the paper's conclusion is discussed.

II. LITERATURE REVIEW

There are various contributions to the literature on cardiopathy diagnostics, data mining, and machine learning techniques. in [8] compared machine learning algorithms using a variety of performance criteria. K-Nearest Neighbor (KNN), Random Forest (RF), and Artificial Neural Network (ANN) produce the best results. The algorithms were then combined, and the final results were applied to the gastrointestinal ailment knowledge set, resulting in better accuracy. The outcomes of merge algorithms (KNN, RF, ANN) were then evaluated in [9] and compared to algorithms with completely different performance metrics, which resulted in higher performance and effectiveness when compared to KNN, RF, Naïve Bayes (NB), Support Vector Machine (SVM), and ANN. The hybrid technique was proposed in [10] as a hybrid random forest with a linear model (HRFLM). It's a hybrid of Random Forest and linear models in which the fundamental traits of both techniques were integrated. In terms of heart disease prediction, the novel approach was quite accurate. 88% accuracy was attained.

Comparative Analysis

The work on machine learning for the prediction of heart problems is summarized in the table below. The work of various researchers is listed in table one, along with the accuracy of their predictions. It will aid medical practitioners in deciding on a more advanced machine learning strategy for effectively predicting heart disease dangers.

Table 1: Performance comparison of Machine Learning techniques in heart disease prediction

AUTHOR'S NAME	DESCRIPTION	ML TECHNIQUES USED	ACCURACY
G. Parthiban et al. [11]	Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients	NB and SVM	SVM has proven to be a categorized technique with excellent prognosticative performance, while NB achieved 74% accuracy.
Vikas Chaurasia et al. [12]	Data Mining Approach to Detect Heart Diseases	NB, J48, DT and Bagging,10-Fold cross validation method	In the diagnosis of heart diseases, the bagging algorithm gives a greater accuracy of 85.03 % with a total time of.05seconds to develop a model.
Boshra Brahmi et al. [13]	Prediction and Diagnosis of Heart Disease by Data Mining Techniques	KNN, SMO J48 and NB	It was discovered that J48 outperformed the other approaches.
K.Vembandasamy et al.[14]	Heart Diseases Detection Using Naive Bayes Algorithm	NB	The NB method provides 86.4198 % accuracy in the shortest amount of time.
Ahmed Fawzi Otoom et al. [15]	Effective Diagnosis and Monitoring of Heart Disease	Build an intelligent classifier using ML algorithms (NB, SVM and FT)	With cross validation testing, it achieves an accuracy of more than 85%, and the monitoring algorithm achieves a detection rate of 100%.
S. Seema et al [16]	Chronic Disease Prediction by mining the data.	SVM, DT and NB	In the instance of heart disease, SVM outperformed the other strategies in terms of accuracy.

K. Gomathi et al. [17]	Multi Disease Prediction using Data Mining Techniques.	J48 and NB	The accuracy of the NB algorithm for predicting cardiac disease is 79%. The accuracy of the J48 algorithm in predicting cardiac disease is 77%.
Ashwini Shetty A et al. [18]	Different Data Mining Approaches for Predicting Heart Disease	NN	The accuracy of NN is expected to reach 84 %.
Ayon Dey et al. [19]	Analysis of Supervised Machine Learning Algorithms for Heart Disease Prediction with Reduced Number of Attributes using PCA	SVM, NB and DT without PCA	The performance of DT is better.
		SVM, NB and DT with PCA	The SVM algorithm outperforms the other two.

As shown in the table above, there are a variety of attribute agents that can improve the effectiveness of machine learning algorithms for predicting heart disease.

III. PROPOSED TECHNIQUES

The goal of the suggested system technique is to employ machine learning techniques to improve the accuracy of heart disease prediction. The proposed system's design is depicted in Figure 1. It is divided into five stages: data collecting, data preprocessing, data splitting, training models, and model evaluation.
The steps of the suggested method are outlined below in detail.

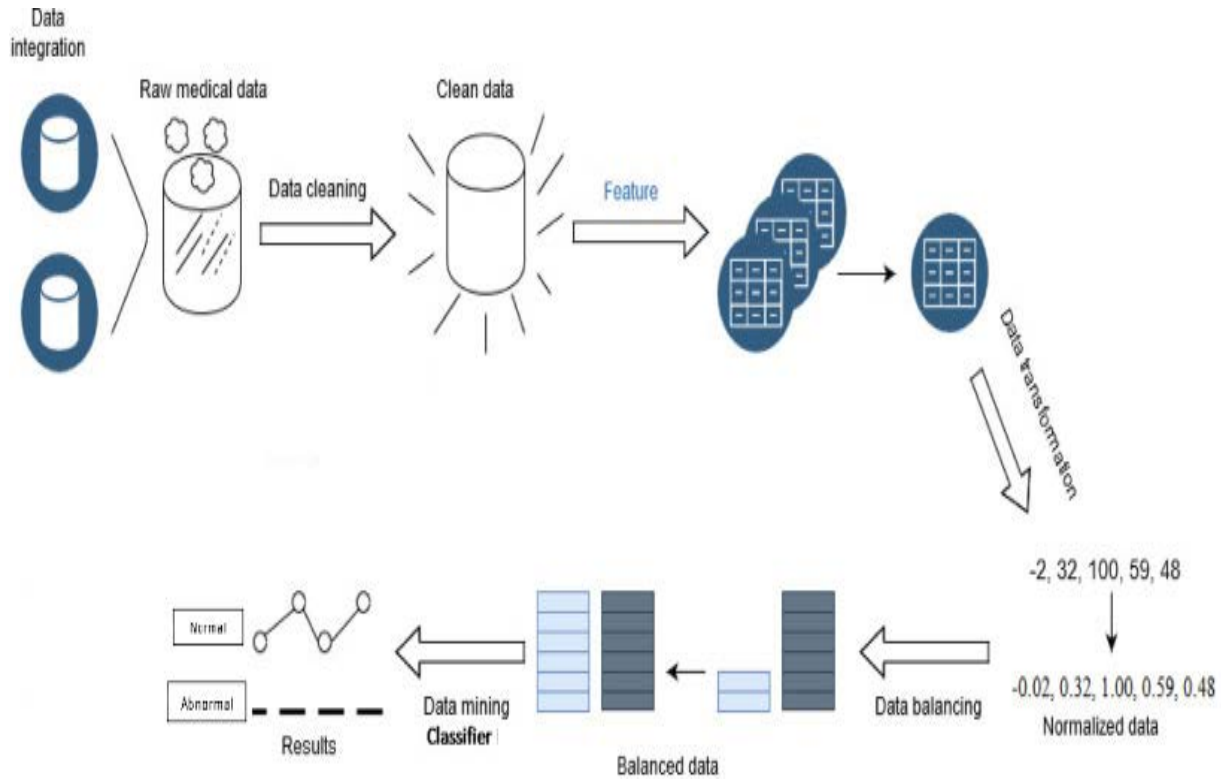

Figure 1: The structure of proposed system for heart diseases classification

A. Dataset Description

Although there are 76 attributes in this database, all published studies only use a subset of 14 of them. The Cleveland database, in particular, is the only one that has been used by machine learning researchers to yet. The "goal" field indicates

whether or not the patient has cardiac disease. As indicated in table 2, it has an integer value ranging from 0 (no presence) to 1 (presence). The dataset includes 165 people who are positive and 138 people who are negative.

B. Data Pre-processing

The data is divided into dependent and independent values, which correspond to features and targets, respectively. The features are then resized to be between [0, 1]. It's worth mentioning that the dataset is purged of missing values. Figure 2,3 show that Box plotting before and after standardization

Name	Type	Description
Age	Continuous	Age in years
Sex	Discrete	0 = female 1 = male
Cp	Discrete	Chest pain type: 1 = typical angina, 2 = atypical angina, 3 = non-anginal pain 4 =asymptom
Trestbps	Continuous	Resting blood pressure (in mm Hg)
Chol	Continuous	Serum cholesterol in mg/dl
Fbs	Discrete	Fasting blood sugar>120 mg/dl: 1-true 0=False
Exang Continuous Maximum heart rate achieved	Discrete	Exercise induced angina: 1 = Yes 0 = No
Thalach	Continuous	Maximum heart rate achieved
Old peak ST	Continuous	Depression induced by exercise relative to rest
Slope	Discrete	The slope of the peak exercise segment: 1 = up sloping 2 = flat 3 = down sloping
Ca	Continuous	Number of major vessels colored by fluoroscopy that ranged between 0 and 3.
Thal	Discrete	3 = normal 6 = fixed defect 7= reversible defect
Class	Discrete	Diagnosis classes: 0 = No Presence 1=Least likely to have heart disease 2= >1 3= >2 4=More likely have heart disease

Table 2: Heart Disease Classification Dataset Description

Box plotting before standardization

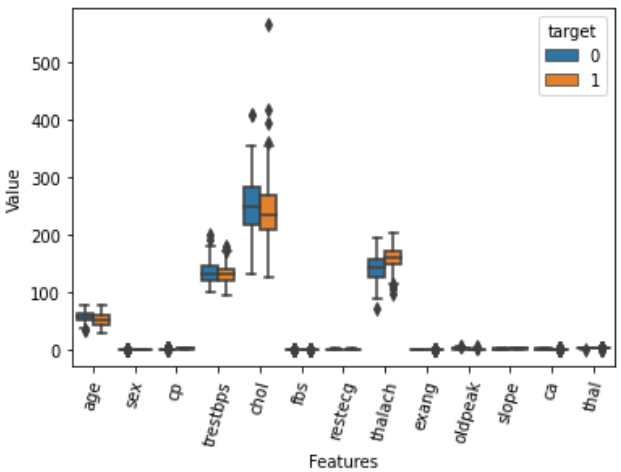

Figure 2: Box plot for features before standardization

Box plotting after standardization

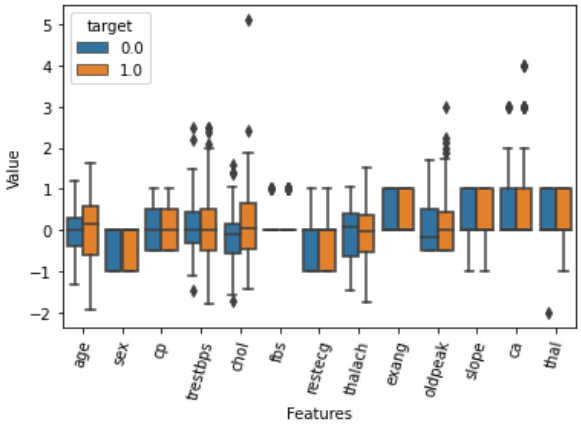

Figure 3: Box plot for features after standardization

1) Data Splitting

In this step, the heart disease dataset is separated into a training set of 75% and a testing set of 25%. The testing set is used to evaluate the models, while the training set is used to train them.

2) Training Model

The problem at hand is a categorization of heart disease based on a dataset attribute. For classification, we have a variety of Machine Learning algorithms.

(SVM) is a supervised machine learning technique that may be applied to classification and regression tasks. It is, however, mostly employed to solve categorization difficulties. Each data item is plotted as a point in n-dimensional space (where n is the number of features you have), with the value of each feature being the value of a certain coordinate in the SVM algorithm. The coordinates of each individual observation are simply referred to as SVM. The SVM classifier [20] is a frontier that separates the two classes (hyper-plane/line) the best.

(NB) Classifier is a classification algorithm based on the Bayes Theorem. This classifier is so named because it is based on a naive data assumption: the independence of feature distributions. Although this expectation is unrealistic in real-world applications, our classifier performs admirably in most classification challenges [21]. The Bayes theorem establishes a method for calculating the probability distribution P(c|x), the posterior distribution of class P(c), the previous probability of predictor P(x), and the possibility of predictor class P(x|c). The equation declared it the posterior probability of C: $pr(c|x) = (pr(x|c) * pr(c))/pr(x)$

(KNN) is a supervised machine learning method that is often used for classification and regression. It's a type of instance-based learning in which data from the training is stored as instances. In a nutshell, it consists of a training set of cases with the purpose of forecasting the label from the specified training data set that is closest in distance to the new instance. Depending on the location, the sample size. It could be a user-defined fixed point or a difference in this situation. The distance could be any measurement in total. A majority vote of the nearest neighbors to each point is used to classify fresh data [22].

(RF) is a learning algorithm that is supervised. It creates a "forest" out of an ensemble of decision trees, which are commonly trained using the "bagging" method. The bagging method's basic premise is that combining several learning models improves the overall output.

(DT): This is a supervised learning method that is commonly used to tackle classification difficulties. It works for both discrete and continuous output and input parameters. The algorithm learns simple judgment principles based on its data properties and then suggests target data values [23]. In other words, based on the most significant primary differences, the population or sample is divided into two or more homogeneous groups (or sub-populations). To decide whether to split a node into two or more sub-nodes, DT employs a number of algorithms. The existence of sub-nodes promotes the homogeneity of future sub-nodes.

Logistic regression (LR): it's a linear classifier. It is used to forecast the likelihood of occurrence by fitting experimental data to a sigmoid function. To put it another way, it forecasts values based on a set of independent factors. (Binary) discrete values were obtained [24].

Stochastic Gradient Descent (SGD) is a quick and easy method for fitting linear classifiers and regressors to convex loss functions like Support Vector Machines and Logistic Regression. Despite the fact that SGD has been present for a long time in the machine learning field, it has only recently gotten a lot of attention in the context of large-scale learning. SGD has been used to solve large-scale, sparse machine learning issues that are common in text categorization and natural language processing. Because the data is sparse, the classifiers in this module can easily scale to problems with more than 105 training samples and characteristics.

Ensemble techniques are strategies that can be used to improve a classifier's performance. It is a useful classification strategy that combines a weak and a strong classifier to improve the effectiveness of a weak learner [25]. In the suggested technique, the ensemble technique is employed to improve the accuracy of several algorithms for diagnosing cardiac disease. When compared to a single algorithm, the goal of mixing numerous algorithms is to improve performance. The ensemble approach is used to improve heart disease diagnosis, as seen in Figure 4.

There two types of Ensemble techniques: Boosting and Bagging.

- Boosting is the process of creating a model sequence with the goal of correcting model defects. The dataset is broken down into subsets in great depth. The classification algorithm is then trained on a sample to provide a series of average efficiency models, as illustrated in the Boost algorithm pseudo-code, where B is the number of base hypotheses and e is exp 1/e = 0.368. As a result, new samples are created based on the prior model's elements that were not accurately identified. The ensemble technique then improves its efficiency by mixing the weak models.

The pseudo-code of Boost algorithm.

Input: number of samples M, classifier C, number of iterations N
Output: result E

Training:

Normalize weight and the total weight w

Mi = sample from M

Ci = training classifier on Mi by C

$e_i = \frac{1}{w} \sum \text{weight}(X_i)$

$B_i = \frac{e_i}{1-e_i}$

Weight (X_i) = Weight $(X_i)B_i$, for all Xi where $C_i(X_i) = y_i$

End for

$E = \text{avg} \sum_{Ci(Xi)=y} \log(1/B_i)$

- Bagging: This is the process of training a model for each subset of a replacement training set with numerous subsets. The final performance forecast is based on the average of the forecast values of the sub-models. The voting mechanism for each categorization model is then carried out, as indicated in the Bagging algorithm pseudo-code. As a result, the bulk of the average values are used to decide the classification decision.

The pseudo-code of Bagging algorithm.

Input: training number of samples M, classifier C, number of iterations N

Output: result E

Training:

For i=1 to N

Mi = bootstrap sample from M

Ci = training classifier on Mi by C

End for

$E = \text{avg} \sum_{Ci(Xi)=y} Ci$

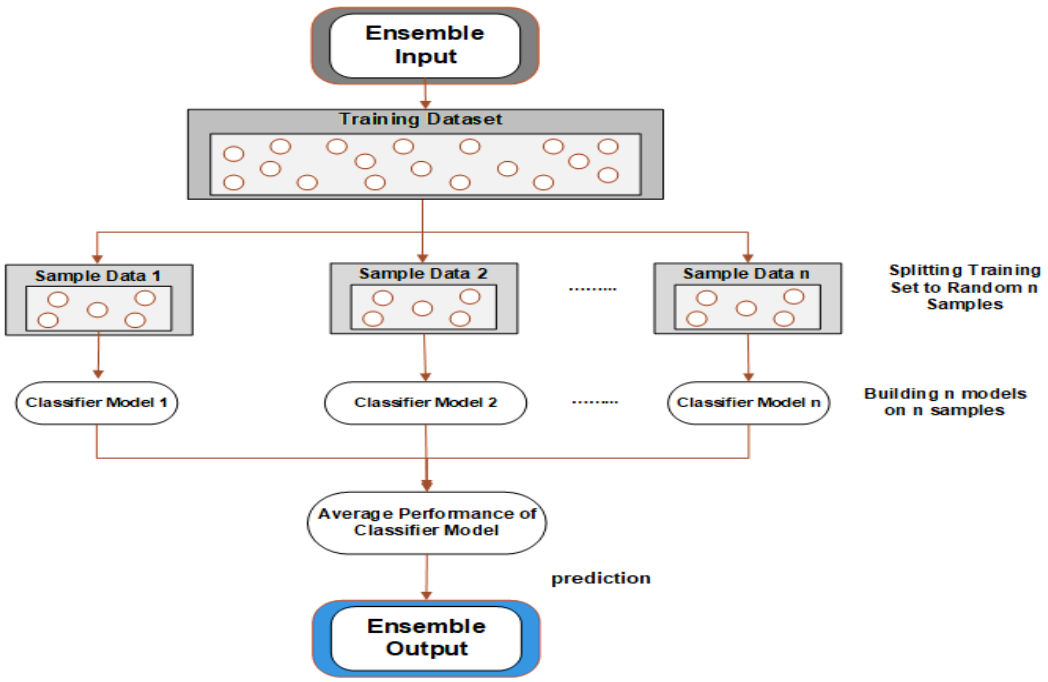

Figure 4: Ensemble Learning Architecture

3) Evaluating Models

The accuracy, ROC, and AUC criteria are used to evaluate the proposed model. One of the most critical classification performance measures is accuracy. As indicated in the equation below, it is defined as the ratio of correct classification to total sample:

$$Acc = \frac{(TP + TN)}{(TP + TN + FP + FN)}.$$

The receiver operating characteristic curve (ROC) is a graph that depicts a classification algorithm's efficiency over all categorization thresholds. This curve displays two parameters: true positive and false positive. The area under the curve (AUC) is a measure of a classifier's ability to distinguish between classes and is used to describe the ROC curve. The higher the AUC, the more effective the model is at distinguishing between positive and negative groups.

$$AU-ROC = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$$

[1] Experimental Results

The experimental results of classification algorithms are discussed in this section.

Experimental Setup

Python was used to implement the experimental results. They were also run on an Intel (R) Core i7 processor with 8 GB of RAM.

The features importance by different algorithms

The score of all extracted features to determine which ones are the most important. Figures 5,6,7 show this. The most essential feature for predicting heart disease is chest pain (CP), which has the greatest score by RF, light gradient boosting Machine (LGBM), and the worst feature score is fasting blood sugar (fbs), but chol has the best score and fbs has the worst score by XGBoost (XGB).

Correlation Matrix

As demonstrated in table 2 and figure 8, there was a correlation between all of the features in the dataset. For a better comprehension, we employ a correlation matrix and a feature importance graph.

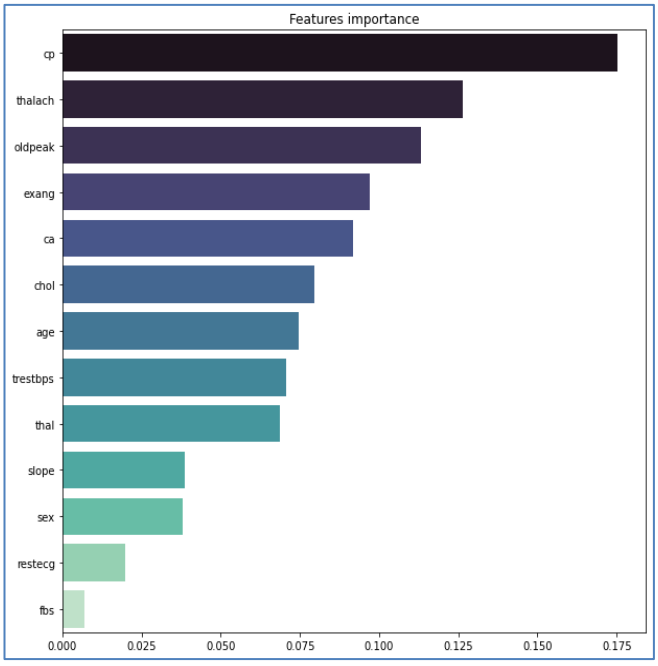

Figure 5: Features Importance for RF

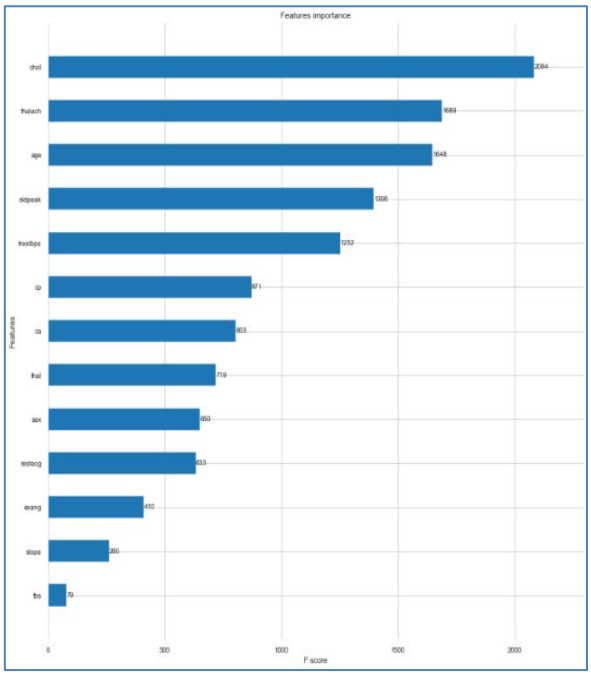

Figure 6: Features Importance for XGB

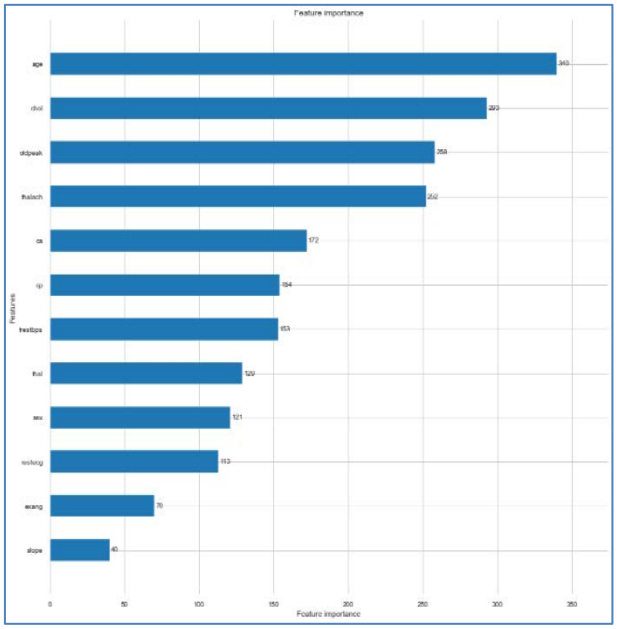

Figure 7: Features Importance for LGBM

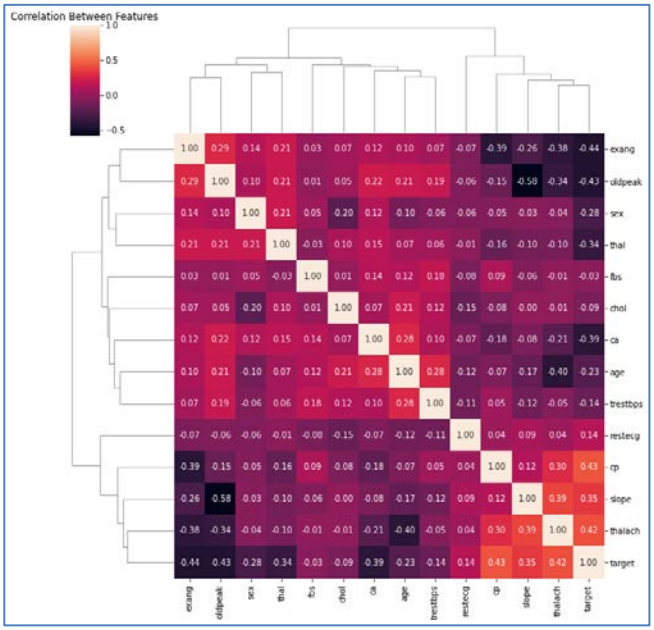

Figure 8: Correlation Matrix

Table 3: Correlation Values

Features	Score	Features	Score
cp	0.433798	chol	-0.085239
ca	-0.391724	trestbps	-0.144931
oldpeak	-0.430696	sex	-0.280937
thalach	0.421741	age	-0.225439
slope	0.345877	thal	-0.344029
restecg	0.137230	exang	-0.436757
fbs	-0.028046	chol	-0.085239

Results of Performance Classification

Table 4 and Figure 9 reveal that the LR and NB techniques have the best performance, with 85.25% accuracy, while KNN and SVM have the worst performance, with 68.85% accuracy and 72.13% accuracy, respectively. DT has a classification accuracy of 78.69%. Figure 11 shows the optimum N estimator value for RF, which achieved an accuracy of 83.6%, while linear SVM recorded the same accuracy of 83.6%. For the KNN, we ran tests with k values ranging from 1 to 20. Figure 10 shows that the optimal value of k = 7 yielded the best results, with an accuracy of 87.2%. The LR and NB performance algorithm outperform the eight classification algorithms, and RF, linear SVM are the second important classification algorithm.

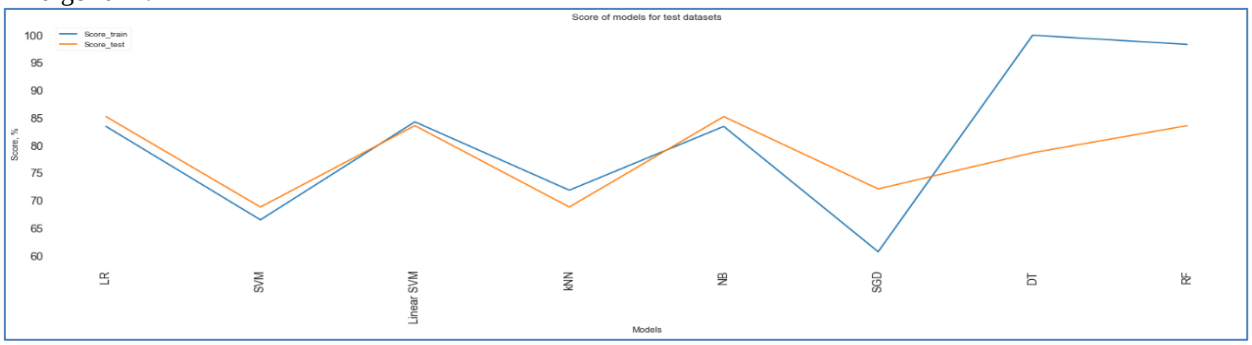

Figure 9: Accuracy of Machine Learning Algorithms

Table 4: Performance of Machine Learning Algorithms

Algorithms	SGD	DT	RF	Linear SVM	LR	NB	KNN	SVM
Accuracy	72.13	78.69	83.61	83.61	85.25	85.25	68.85	68.85

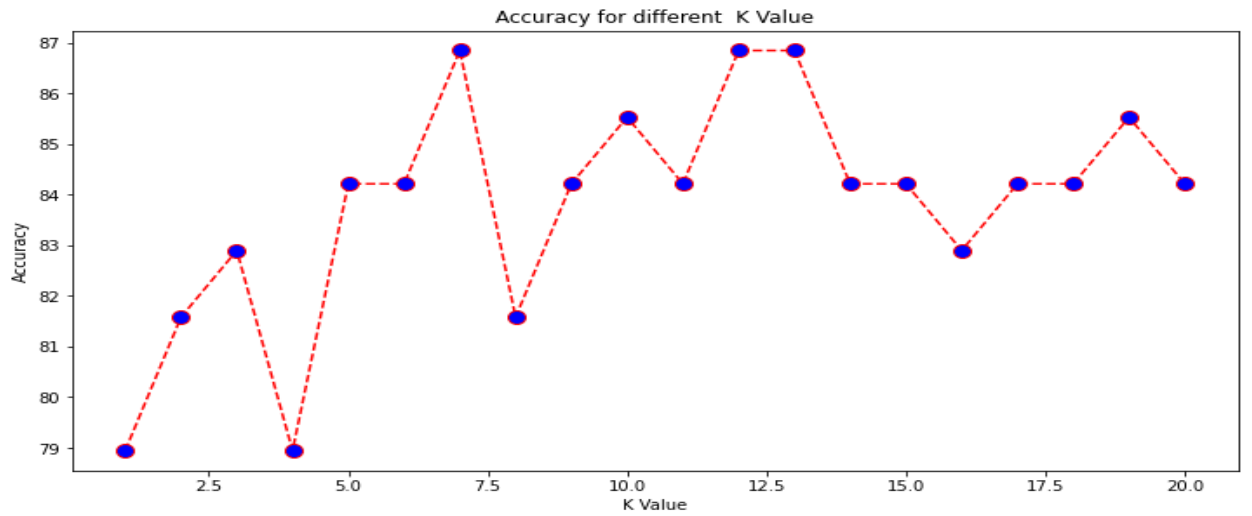

Figure 9: Accuracy of different K Values

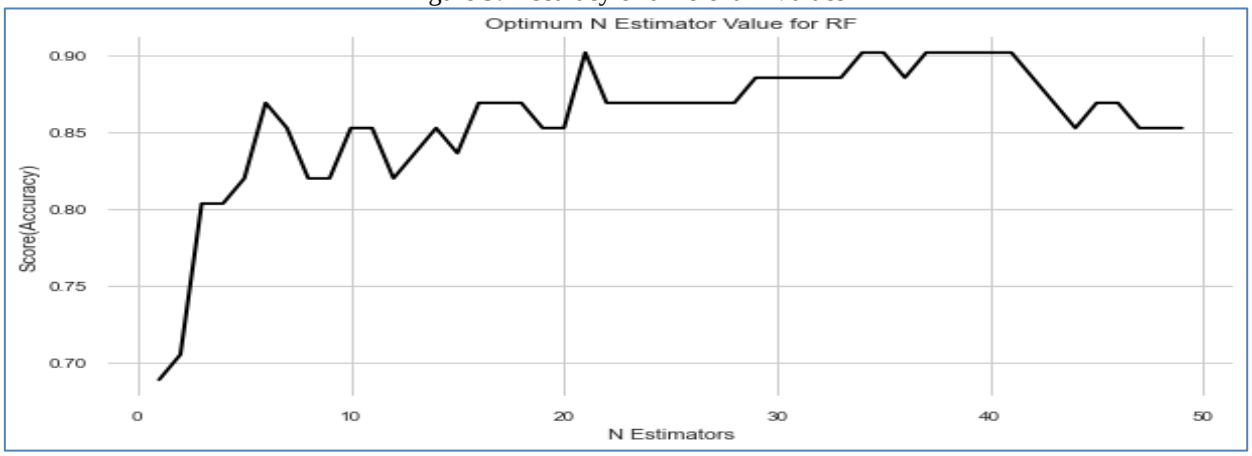

Figure 10: Optimum N Estimator Value for RF

Applying the ensemble Learning Algorithms

Table 5 and Figure 12 demonstrate that DT with bagging technique has the highest performance with 88.52% accuracy, while SGD with bagging has the poorest performance with 72.13% accuracy, and SGD with boosting has the best performance

with 73.77% accuracy. NB with boosting has a classification accuracy of 85.25%, whereas NB with bagging has a classification accuracy of 83.61%. SVM with bagging had 85.25% accuracy, but SVM with boosting had 85.25% accuracy. The accuracy of DT with boosting is 77%, but RF with bagging is 86.89%. Figure 9 shows that while using the hard voting method, the accuracy was 83.61%, but when using the soft voting algorithm, the accuracy was 85.25%. The DT performance with the bagging algorithm outperforms the four classification algorithms, and RF with bagging is the second important classification algorithm.

Table 5: The results of different algorithms with bagging and boosting techniques

Algorithms / **Accuracy**	**DT**	**RF**	**SVM**	**NB**	**SGD**
Bagging	**88.52**	86.89	85.25	83.61	72.13
Boosti ng	77.05	83.61	85.25	85.25	73.77

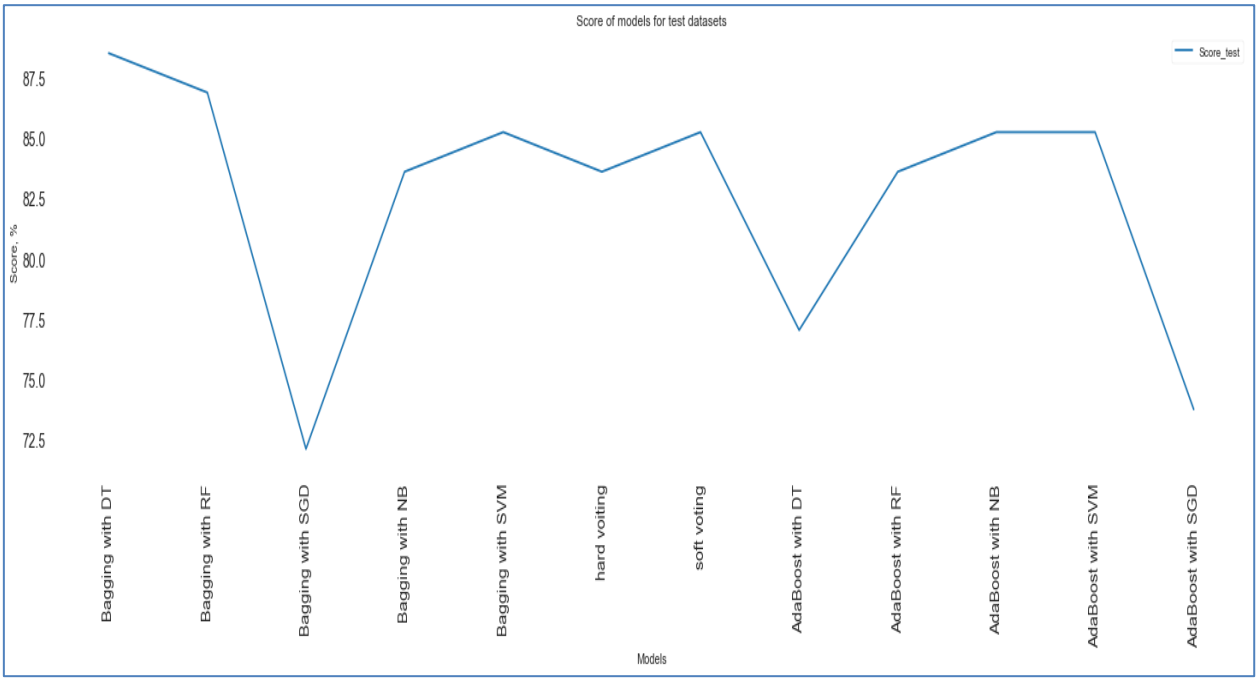

Figure 11: Accuracy of Ensemble Methods

Comparison between Results of machine learning Algorithms

The proposed model (the bagging ensemble learning approach with decision tree) is compared to several other state-of-the-art algorithms in Table 6 and Figure 13. The optimal performance of a state-of-the-art algorithm obtained an accuracy of 86.4% [14], as shown in Table 1. The proposed model, on the other hand, has an accuracy rate of 88.52%. As a result, it is obvious that the suggested model greatly outperforms other machine learning techniques.

Table 6: Comparison between Algorithms

Algorithms	**Accuracy**	**Algorithms**	**Accuracy**
SGD	72.13	AdaBoost with NB	85.25
DT	78.69	AdaBoost with DT	77.05
RF	83.61	**Bagging with DT**	**88.52**
Linear SVM	83.61	hard voting	83.61
LR	85.25	Bagging with RF	86.89
NB	85.25	soft voting	85.25
KNN	68.85	AdaBoost with RF	83.61
SVM	68.85	Bagging with SVM	85.25
Bagging with NB	83.61	Bagging with SGD	72.13
AdaBoost with SGD	73.77		

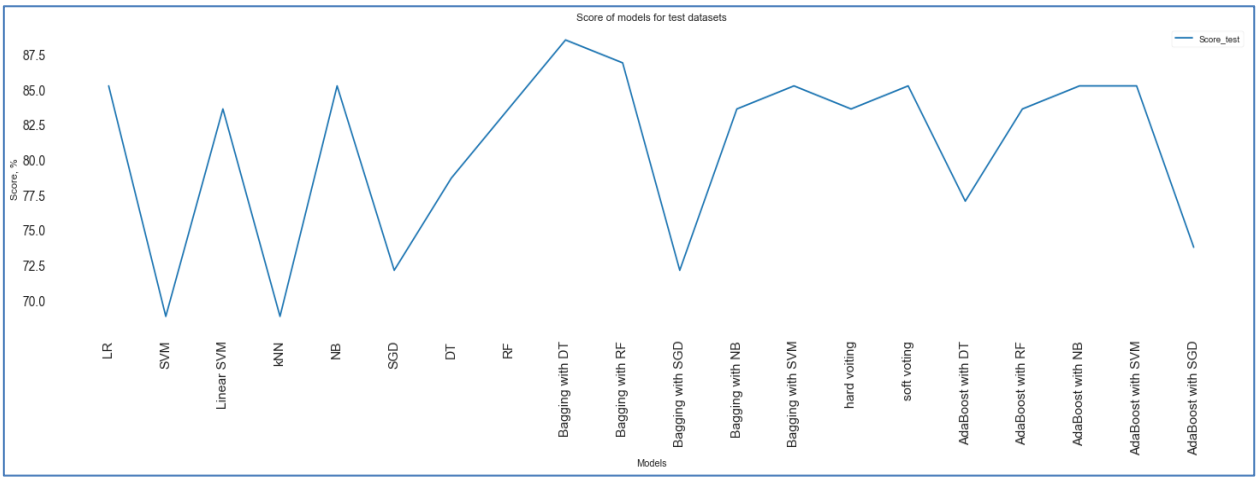

Figure 12: Comparison between Accuracy of Machine Learning Algorithm

IV. CONCLUSION

The proposed system for predicting heart disease is developed in this research. A comparison of five classifiers and ensemble approaches (boosting, bagging) (KNN, SVM, NB, DT, SGD and RF). For the bagging ensemble learning technique with DT method, the suggested ensemble approach achieved a classification accuracy of 88.52 percent, which is higher than any single classifier model. The major goal of the suggested ensemble model (boosting, bagging, and voting) is to increase the model's prediction accuracy, resilience, and reliability for heart disease, as well as to avoid any patient misdiagnosis. There is, however, still room for growth.

REFERENCES

[1] Nashif, S., Raiban, M., Islam, M., & Imam, M. H. (2018). Heart disease detection by using machine learning algorithms and a real-time cardiovascular health monitoring system. World Journal of Engineering and Technology, 6, 854-873.

[2] Hussein, M. U. (2017, October 29). Physics and the Cardiovascular System. Retrieved from ResearchGate: https://www.researchgate.net.

[3] Zhou, Ting, et al. "Using Published Health Utilities in Cost-Utility Analyses: Discrepancies and Issues in Cardiovascular Disease." Medical Decision Making (2021): 0272989X211004532.

[4] Nagendra, K. V., & Ussenaiah, M. (2018). A study on various data mining techniques used for heart diseases. International Journal of Recent Scientific Research, 24350- 24354.

[5] Solanki, A., & Barot, M. P. (2019). Study of heart disease diagnosis by comparing various classification algorithms. International Journal of Engineering and Advanced Technology, 8 (2S2), 40-42.

[6] Kashyap, A. (2018). Artificial intelligence and medical diagnosis. Scholars Journal of Applied Medical Sciences, 4982-4985. doi: 10.21276/sjams.2018.6.12.61.

[7] Patel, J., Upadhyay, T., & Patel, S. (2016). Heart disease prediction using machine learning and data mining techniques. IJCSC, 7 (1), 129-137.

[8] Youness Khourdifi and Mohamed Bahaji (2019). Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization, International Journal of Intelligent Engineering and Systems, Vol.12, No.1, 2019.

[9] Youness Khourdifi and Mohamed Bahaji (2019). The Hybrid Machine Learning Model Based on Random Forest Optimized by PSO and ACO for Predicting Heart, ICCWCS 2019, April 24-25, Kenitra, Morocco.

[10] Senthilkumar Mohan, Chandrasegar Thirumalai, and Gautam Srivasatava (2019). Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques, IEEE access volume7, 2019.

[11] G. Parthiban and S.K.Srivatsa (2012). Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients, International Journal of Applied Information Systems (IJAIS) – ISSN: 2249-0868.

[12] Vikas Chaurasia and Saurabh Pal (2012). Data Mining Approach to Detect Heart Diseases, International Journal of Advanced Computer Science and Information Technology (IJACSIT)Vol. 2, No. 4, Month Year, Page: 56-66, ISSN: 22961739.

[13] Boshra Brahmi, Mirsaeid Hosseini Shirvani (2015). Prediction and Diagnosis of Heart Disease by Data Mining Techniques, Journals of Multidisciplinary Engineering Science and Technology, vol.2, 2 February 2015, pp.164- 168.

[14] K.Vembandasamy, R.Sasipriyaand E.Deepa (2015). Heart Diseases Detection Using Naive Bayes Algorithm, International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 9, September 2015.

[15] Ahmed Fawzi Otoom, Emad E. Abdallah, Yousef Kilani, Ahmed Kefaye and Mohammad Ashour (2015). Effective Diagnosis and Monitoring of Heart Disease, International Journal of Software Engineering and Its Applications Vol. 9, No. 1 (2015), pp. 143-156.

[16] Dr.S.Seema Shedole, Kumari Deepika (2016). Predictive analytics to prevent and control chronic disease, https://www.researchgate.net/punlication/316530782, January 2016.
[17] K.Gomathi, Dr.D.Shanmuga Priyaa (2016). Multi Disease Prediction using Data Mining Techniques, International Journal of System and Software Engineering, December 2016, pp.12-14.
[18] Ashwini Shetty A, Chandra Naik (2016). Different Data Mining Approaches for Predicting Heart Disease, International Journal of Innovative in Science Engineering and Technology, Vol.5, May 2016, pp.277- 281.
[19] Ayon Dey, Jyoti Singh and Neeta Singh (2016). Analysis of Supervised Machine Learning Algorithms for Heart Disease Prediction with Reduced Number of Attributes using Principal Component Analysis), International Journal of Computer Applications (0975 – 8887) Volume 140 – No.2, April 2016.
[20] Eva Tuba et al. “Classification and Feature Selection Method for Medi- cal Datasets by Brain Storm Optimization Algorithm and Support Vector Machine”. In: Procedia Computer Science 162 (2019), pp. 307–315.
[21] Alexander Wood et al. “Private naive bayes classification of personal biomedical data: Application in cancer data analysis”. In: Computers in biology and medicine 105 (2019), pp. 144–150.
[22] Sandhya Harikumar. “Blended Models for Nearest Neighbour Algorithms for High Dimensional Smart Medical Data”. In: Smart Medical Data Sens- ing and IoT Systems Design in Healthcare. IGI Global, 2020, pp. 48–75.
[23] Diyang Xue, Adam Frisch, and Daqing He. “Differential Diagnosis of Heart Disease in Emergency Departments Using Decision Tree and Medi- cal Knowledge”. In: Heterogeneous Data Management, Polystores, and An- alytics for Healthcare. Springer, 2019, pp. 225–236.
[24] Changsheng Zhu, Christian Uwa Idemudia, and Wenfang Feng. “Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques”. In: Informatics in Medicine Unlocked 17 (2019), p. 100179.
[25] David, H. B. F. Impact of ensemble learning algorithms towards accurate heart disease prediction.
[26] Ali, A. A., Hassan, H. S., & Anwar, E. M. (2020, July). Improve the accuracy of heart disease predictions using machine learning and feature selection techniques. In International Conference on Machine Learning, Image Processing, Network Security and Data Sciences (pp. 214-228). Springer, Singapore.
[27] Gao, Xiao-Yan, et al. "Improving the Accuracy for Analyzing Heart Diseases Prediction Based on the Ensemble Method." Complexity 2021 (2021).
[28] Ali, A. A., Hassan, H. S., Anwar, E. M., & Khanna, A. (2021). Hybrid technique for heart diseases diagnosis based on convolution neural network and long short-term memory. In Applications of Big Data in Healthcare (pp. 261-280). Academic Press.
[29] Ali, A. A., Hassan, H. S., & Anwar, E. M. (2020, July). Heart diseases diagnosis based on a novel convolution neural network and gate recurrent unit technique. In 2020 12th International Conference on Electrical Engineering (ICEENG) (pp. 145-150). IEEE.