With the advent of the data age, the continuous improvement and widespread application of medical information systems have led to an exponential growth of biomedical data, such as medical imaging, electronic medical records, biometric tags, and clinical records that have potential and essential research value. However, medical research based on statistical methods is limited by the class and size of the research community, so it cannot effectively perform data mining for large-scale medical information. At the same time, supervised machine learning techniques can effectively solve this problem. Heart attack is one of the most common diseases and one of the leading causes of death, so finding a system that can accurately and reliably predict early diagnosis is an essential and influential step in treating such diseases. Researchers have used various data mining and machine learning techniques to analyze medical data, helping professionals predict heart disease. This paper presents various features related to heart disease, and the model is based on ensemble learning. The proposed system involves preprocessing data, selecting attributes, and then using logistic regression algorithms as meta-classifiers to build the ensemble learning model. Furthermore, using machine learning algorithms (Support Vector Machines, Decision Tree, Random Forest, Extreme Gradient Boosting) for prediction on the Framingham Heart Study dataset and compared with the proposed methodology. The results show that the feasibility and effectiveness of the proposed prediction method based on group learning provide accuracy for medical recommendations and better accuracy than the single traditional machine learning algorithm

Keywords: heart attack prediction, machine learning, ensemble learning, stacking ensemble technique

UDC 1:371
DOI: 10.15587/1729-4061.2021.238528

DEVELOPMENT OF HEART ATTACK PREDICTION MODEL BASED ON ENSEMBLE LEARNING

Omar Shakir Hasan
Corresponding author
Assistant Teacher
Department of Computer Science*
E-mail: omarshakir06@gmail.com
Ibrahim Ahmed Saleh
Assistant Professor
Department of Software Engineering*
*College of Computer Science and Mathematics
University of Mosul
Al Majmoaa str., Mosul, Iraq, 41002

Received date 21.06.2021
Accepted date 06.08.2021
Published date 31.08.2021

How to Cite: *Hasan, O. S., Saleh, I. A. (2021). Development of heart attack prediction model based on ensemble learning. Eastern-European Journal of Enterprise Technologies, 4 (2 (112)), 26–34. doi: https://doi.org/10.15587/1729-4061.2021.238528*

1. Introduction

Doctors have many tools and methods to predict patients' health risks, but they still cannot cope with the complexity of the human body 100 % [1]. If a person has chest pain and other suspected heart disease symptoms, traditional detection methods sometimes do not necessarily detect whether the patient has a heart attack [2]. For better accuracy and results in medical diagnosis, the health field and artificial intelligence, especially in clinical diagnosis, have been linked [3]. The effect of different features and their weights on heart attack can be analyzed through machine learning, in that case, this will help in the prediction of such disease, reduce the risk, and prevention of heart attack [4, 5]. Heart attack is one of the most dangerous, difficult to predict, and common diseases these days, as millions of people are infected with this disease worldwide every year [6, 7]. If the disease is discovered early, it can be saved from death and serious complications [8]. Many healthcare organizations face the enormous challenge of providing quality services [9]. This study is used for the medical field to help doctors make accurate, fast, and error-free predictions and investigates the probability of a patient having a heart attack or not based on the patient's medical attributes such as age, blood pressure, gender, etc. The Framingham heart dataset was selected from the UCI repository and contained 16 features. Machine learning algorithms train these features for predicting heart attacks to improve the doctor's decisions. It is difficult or impractical to use traditional algorithms to perform the required tasks, the combination of several machine learning algorithms helps to improve the heart attack prediction.

Heart attack is a dangerous disease, most of the existing studies used traditional machine learning algorithms to predict the heart attack which in turn does not give the desired results. Therefore, it becomes necessary to propose the stacking technique to achieve better performance.

2. Literature review and problem statement

In recent years, several papers have been proposed for predicting heart attack; in this section, researchers' works for predicting heart attacks using machine learning algorithms are discussed and summarized.

The paper [10] used electronic clinical data, analyzed each feature in the dataset and its effect on the results, and plotted the dataset before wrangling and after the wrangling. The authors used the logistic regression algorithm for classification and random search technique to find the best parameters for building a prediction model. This study classifies the persons whether or not they have heart disease according to the clinical medical record. The "Sklearn" library is used to calculate the score. The accuracy showed 87 %, which is acceptable accuracy for predicting heart risk. Only one machine learning algorithm with a small dataset is used, no other machine learning algorithms are trained and tested

on the same dataset. This does not necessarily mean that this algorithm is the best classifier for that dataset, which is considered the disadvantage of this approach.

The paper [11] proposes a coarse group theory to select significant features and used the random forest algorithm for the classification process to predict heart disease. The Heart Stalog dataset from the UCI repository was used and contained 270 instances. The dataset has been preprocessed, namely noisy and irrelevant data were removed. The accuracy of this method reached 84 %. The disadvantage of this approach is that the parameters used were not mentioned, which in turn affects the performance of the algorithms directly and has more impact than the methods of extracting features, in which datasets often have limited features. Another disadvantage is that the random forest algorithm was not applied alone to compare its results with the random forest rough sets.

The paper [12] used the Cleveland dataset from the UCI repository, which consists of 303 states and 76 features, the pre-processing is made into a dataset such as processing the missing value, removing the noise, and extracting the most important features, then applying supervised machine learning algorithms on this dataset such as KNN, Decision trees, random forest, naive Bayes. The KNN algorithm achieved the highest accuracy of 90.78 %. The disadvantage of this approach is that the result is measured only by the accuracy and other measures, such as the ROC and confusion matrix, are not used.

The paper [13] suggested machine learning algorithms such as SVM, KNN, logistic regression, decision tree, random forest, naive Bays and applying them to the Cleveland dataset from the UCI repository to predict heart diseases. The KNN algorithm achieved the highest accuracy of 87 %. The disadvantage of this approach is that parameters selection is adopted using multiple values, while parameter optimization methods can be used such as grid search and random search to select the best parameters.

In [14], the author worked on the heart disease dataset from the UCI repository, then made data pre-processing, features selection and applied these features on the hybrid random forest with a linear model for heart disease prediction. The advantage of the proposed method is high accuracy equal to 92 %. The disadvantage of this method is that the proposed method is considered the best classifier, while the results showed that it had the lowest sensitivity compared to the rest of the algorithms.

In [15], the NN-FCA approach is proposed. FCA consists of two stages, the first stage is feature selection, and the second stage is feature correlation. Then, the Neural Network algorithm for classification on the KNHANES-VI dataset was used. The advantage of this method is high performance of the Neural Network algorithm for predicting heart disease. The disadvantage of the proposed methodology is that the dataset features are not large enough to conduct operations of correlation among the dataset features, and traditional machine learning algorithms can perform well on this dataset without this complication.

In [16], the authors suggested a fuzzy expert system for predicting heart disease, including three main steps, such as fuzzification, rule base, and defuzzification. For defuzzification, the centroid technique was applied. The system contains 13 input parameters and one output parameter, the dataset was taken from the UCI repository. The advantage of this approach is that the heart attack prediction system is simple in use, and the patients can use it by themselves directly. The accuracy is 93.33 %. The disadvantage of this approach is the complexity added by fuzzy logic. The results showed no significant difference compared to previous studies on the same dataset and earlier systems.

In [17], the authors in this method used a local dataset, selected the most significant features using a correlation matrix and applied three algorithms, first, the neural network, second, support vector machine, and third, KNN on the proposed dataset for heart attack prediction. The neural network algorithm showed the best performance compared to other algorithms used in this paper, and the neural network obtained an accuracy of 93 %. The advantage of the proposed approach is the stability of the three algorithms used, despite their implementation on the varying sizes of the dataset. The disadvantage of this work is that the local dataset used is not a certified global dataset.

In [18], the authors suggested a hybrid genetic neural network algorithm. ECG signal dataset taken from MIT-BIH arrhythmia was used. The dataset has been pre-processed, including data cleaning to remove noisy data and pattern identification to identify the pattern of ECG data. For the prediction process, the neural network is used with a genetic algorithm to optimize neural weights. The advantage of the proposed methodology is speeding up the neural network prediction of heart attacks by applying a genetic algorithm on the neural network. The disadvantage of this approach is the complexity added to the neural network.

The paper [19] proposed traditional machine learning algorithms such as naive Bayes classifier, logistic regression, random forest, support vector machine, decision tree classifier, and KNN to predict heart diseases. The classifier algorithms are trained and tested on the dataset available in the UCI repository. The accuracy results of the suggested algorithms are compared and showed that the random forest algorithm obtained the best accuracy, therefore is the best classifier for predicting heart disease. The advantage is that traditional machine learning algorithms are used without any complexity with an accuracy achieved of 91.17 %, which is an acceptable accuracy. The disadvantage is that there are no feature extraction techniques used in the dataset, which in turn helps give better results.

Previous studies rely on a small data set or local set, which makes the results unreliable. In addition, there is no comparison of the results with other algorithms to determine the efficiency of the proposed model. The prediction of diseases needs to use different measures to know the best model and that the results are not biased towards the majority category. Parameters are specified using multiple values, and this does not have to give the best values. Some methods increase the complexity of the system despite the lack of better performance.

All these give reason to develop a highly efficient and stable model for heart attack prediction.

3. The aim and objectives of the study

This work aims to determine a robust and accurate method for heart attack prediction using ensemble learning and compare it to single traditional classification algorithms.

To achieve the aim, the following objectives were set:

– to design a model using a stacking ensemble technique by combining the decision tree algorithm, logistic regression, SVM algorithm, XGboost algorithm, and use the logistic regression algorithm as a meta-classifier to predict heart attack with better accuracy;

– to investigate that the designed model has high prediction accuracy, compare this model with the models of single traditional algorithms, at the same time, to verify that the proposed method has a high prediction accuracy, comparing it with previous research on the same data set.

4. Materials and methods

The Framingham heart study dataset obtained from the UCI repository is used. This data is incomplete and cannot be trained directly. Hence, it needs pre-processing to solve this problem, such as missing value imputation using the mean method and removing noise. The dataset contains 16 attributes and 4.239 instances. The attributes are described in Table 1.

Table 1

Framingham dataset attributes

Attribute name	Type	Category
Gender	Int	The value of 1 refers to male, and 0 refers to female
Age	Int	Refers to the patient's age
Education	Int	The value of 1: indicates the high school. The value of 2: indicates the diploma. The value of 3: indicates the high college. The value of 4: indicates the higher degree
Current Smoker	Int	Currently smokes or not
Cigs Per Day	Int	Average number of cigarettes smoked per day
BP Meds	Int	Under blood pressure medication or not
Prevalent Stroke	Int	Had a stroke previously or not
Prevalent Hyp	Int	Has hypertension or not
Diabetes	Int	Has diabetes or not
Tot Chol	Int	Total cholesterol in the patient's body
SysBP	Numerical	Systolic blood pressure
DiaBP	Numerical	Diastolic blood pressure
BMI	Numerical	Body mass index
HeartRate	Int	Heart rate
Glucose	Int	Glucose level in the patient's body
Ten-YearCHD	Int	Will have a heart attack over the next ten years or not

In this paper, machine learning algorithms are used to build prediction models and select the models based on accuracy. The models' output chosen is used as new features for training the final ensemble model and compared with single traditional algorithms based on accuracy. The accuracy results of different machine learning classifications and proposed methods have been observed using the Python programming language. Research was performed on the 7th generation Intel Corei7 having an 8750H processor up to 4.1 GHz CPU and 16 GB ram. Below is a review of the algorithms used.

K-Nearest Neighbor (K-NN).

K-NN (K-Nearest Neighbor) is one of the most basic algorithms in machine learning [20]. K-NN is used for both classification and regression [21]. The idea of the KNN algorithm is that N-dimensional input vector corresponds to a point of the feature space, and the output value is the category label or a predicted value corresponding to the feature vector [22]. The KNN algorithm does not have an explicit learning process, it is very special [23]. It uses training data to divide the feature vector space and uses the result of the division as the final algorithm model [24].

Logistic Regression.

It is a generalized linear regression analysis model. It adds the sigmoid function to the original linear regression, thereby mapping the original positive and negative infinity interval to the range of 0 to 1, corresponding to the probability that the model is judged as a positive example, so it is often used in data mining, automatic diagnosis of diseases, economic trend prediction and other fields. This algorithm is also a common two-classification model in essence, and the category corresponding to the object is obtained by inputting the attribute feature sequence of the unknown category object [25, 26].

Random Forest.

It is an ensemble learning algorithm in supervised learning. Its essence is an ensemble classifier containing many randomly generated decision trees and combining several weak (base) classifiers to get a strong classifier with significantly superior classification performance [27]. The random forest algorithm requires an enormous difference between the decision trees with no correlation. If there is no strong dependency between the weak classifiers, the trees can be generated in parallel [28]. Random forest uses autonomous sampling to extract multiple samples from the original data. The extracted samples are first trained with a weak classifier-decision tree, and then these decision trees are combined to get the final classification or prediction result through voting [29].

XGBoost.

XGBoost is the abbreviation of "Extreme Gradient Boosting". It is an ensemble learning improvement method based on decision trees, which combines weak base classifiers into stronger classifiers. The algorithm consists of multiple decision trees, and the conclusions of all trees are added together as the final [30]. XGBoost uses Newton's method to solve the extreme value of the loss function, carries out a second-order Taylor expansion of the loss function, and adds a regular term outside the objective function to find the overall optimal solution. It is used to weigh the decline of the objective function and the complexity of the model to avoid overfitting [31].

Decision Tree.

A decision tree is a tree structure, in which each internal node represents a test on an attribute, each branch represents a test output, and each leaf node represents a category [32]. A classification tree is a kind of supervised learning. Generally, a decision tree contains a root node, several internal nodes, and several leaf nodes [33]. The leaf nodes correspond to the decision results, and each other node corresponds to an attribute test [34]. Each node contains the sample set divided into subnodes according to the results of the attribute test [35]. The full set of samples is contained in the root node [36]. The path from the root node to each leaf node corresponds to a decision test sequence [37, 38]. The purpose of decision tree learning is to produce a decision tree with strong generalization ability and a strong ability to deal with unseen strength [39]. Often, a decision tree is built based on a data set [40].

Support Vector Machine (SVM).

Support Vector Machine (SVM) is a linear classifier that implements binary classification of data through supervised learning [41]. The decision boundary of its classification is to

solve the maximum interval hyperplane for the learned data samples. The samples are divided into two categories by constructing this dividing hyperplane [42]. The sample points closest to the hyperplane are called support vectors, and the distance between these points and the segmentation plane is called the interval [43]. By maximizing the distance interval between the support vector and the segmentation plane, the algorithm performance is optimized, thereby enhancing the reliability of the classifier's prediction [44].

The evaluation of the proposed model was carried out using several criteria, namely, accuracy, recall, precision, *F*1 score, and ROC. The most important criterion is accuracy. Here are some assumptions for explaining the algorithm measurement tools [45]:

Assuming there are two classes *p*, *n* to be classified

TP : True Positives.

TN : True Negatives.

FP : False Positives.

FN : False Negatives

Accuracy: the accuracy is the part that the model predicts correctly.

The formula for accuracy is

$$\text{Acuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \quad (1)$$

When the data set is unbalanced, when the number of positive samples and negative samples is significantly different, the accuracy of the model alone cannot be used to evaluate the model performance. Precision and recall are better indicators for measuring unbalanced data sets [46, 47].

Precision: it refers to the effect of the degree of correctness of the prediction as a positive example in all classifications where the prediction is a positive example [48].

$$\text{Precision} = \frac{TP}{TP + FP}. \quad (2)$$

Recall Rate: recall rate refers to the classification sample of all positive predictions (correctly predicted to be true and incorrectly predicted but true). Recall rate refers to the degree of correctness of the prediction. It is also called sensitivity or true positive rate (TPR) [47].

$$\text{Recall} = \frac{TN}{TN \pm FP}. \quad (3)$$

F1 score: it is usually practical to combine the accuracy and recall rate into one index F1 value, especially when a simple method is needed to measure the performance of two classifiers. The F1 value is the harmonic average of precision and recall [48].

$$F1 = 2 \times \frac{\text{Precision} \times \text{recall}}{\text{precision} + \text{recall}}. \quad (4)$$

Receiver Operating Characteristic (ROC): it is a graph consisting of a false positive rate (FPR) as a horizontal axis and true positive rate (TPR) as a vertical axis, which shows the relationship between the true rate and the false positive rate of the classifiers. The ROC curve is a very important indicator to measure the classifier performance, and it represents the degree to which the model predicts accurately [49].

Confusion Matrix: the classification algorithm's performance is based on a confusion matrix. It can be said that the easiest way to regulate the performance of the classification model is comparing the number of positive cases that are correctly rated (true/false) and the number of negative cases that are correctly rated (true/false). In the confusion matrix, as shown in Fig. 1, the column represents expected labels while the rows represent the actual labels [50].

Stacking ensemble technique achieves better performance than any single traditional training model. It has been used for supervised learning tasks (including regression, classification, etc.) and unsupervised learning (density estimation), which can estimate the bagging error rate. A group of different classifiers is combined to produce a robust and high-level learner model. Usually, this technique performs better than a single learner model for making the final prediction; the following steps are used [51]:

– level zero data: all learners (classification algorithms) work on the dataset;

– level one data: it takes the prediction produced by the classification algorithms as new data;

– final prediction: it is another new learning process, it takes the level one data as new inputs and as output, and the final prediction is obtained.

In Fig. 2, the first layer is single models, cross-validation is used to produce a one-fold prediction result, and then all the prediction results are spliced into a completed result as the prediction feature of the model.

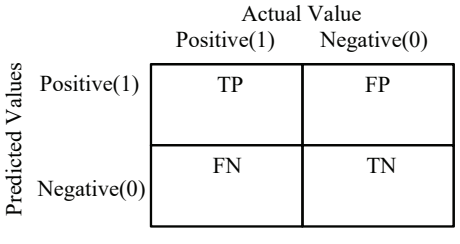

Fig. 1. Confusion matrix

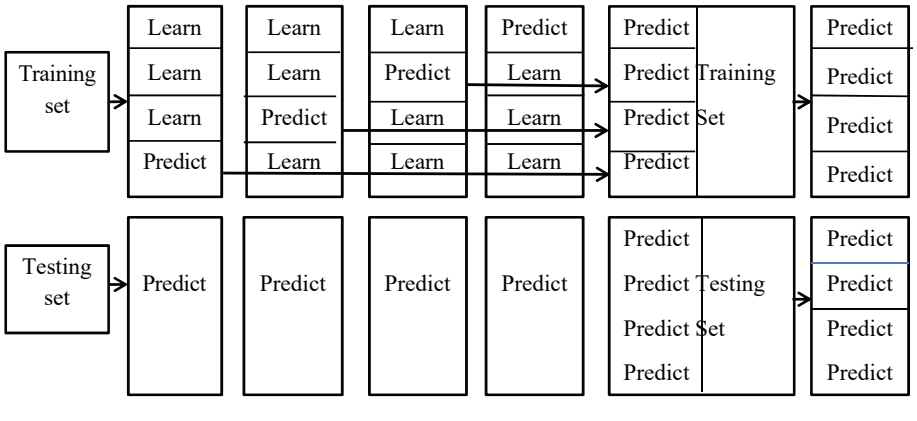

Fig. 2. Cross-validation

The second layer can use a classifier or traditional fusion methods such as averaging, voting, or weighting. In this study, the ensemble stacking technique improves the accuracy of various single classifier algorithms for heart attack prediction. The method used (KNN, Logistic regression, random forest, XGBoost, decision tree, and SVM) then combining the prediction models of SVM, XGBoost, decision tree, and random forest using the ensemble stacking technique. The output of prediction models generated new features and has been input for meta-classifier and appeared the final prediction; this method showed the highest prediction accuracy compared to other single classification algorithms. Fig. 3 shows the framework of the proposed method for heart attack prediction.

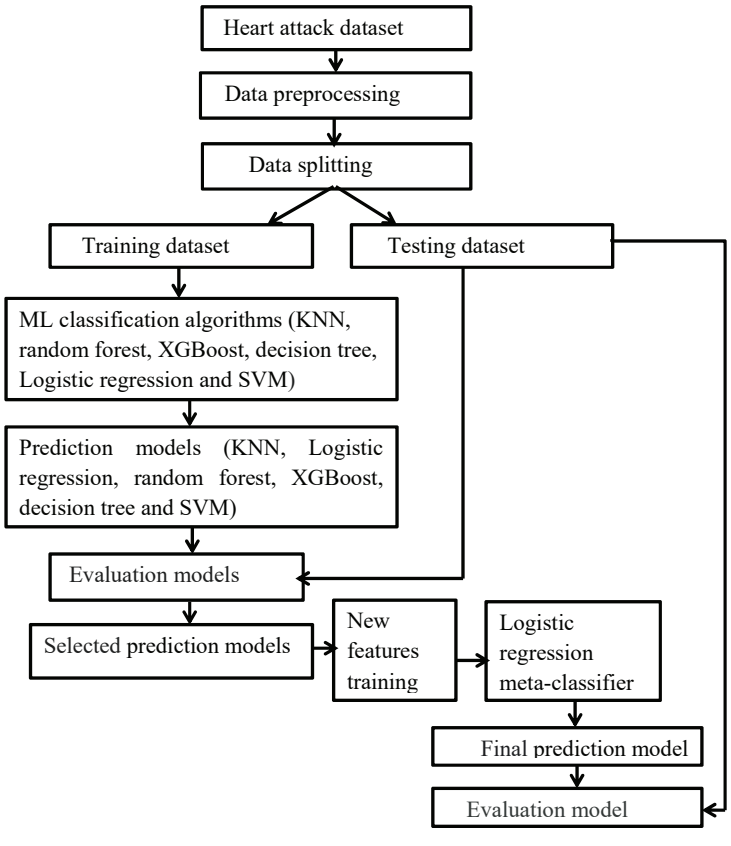

Fig. 3. Framework of the method

Stacking algorithm:
Input: training dataset $D=\{x_i,y_i\}_{i=1}^{m}(x_i\in\square,y_i\in\gamma)$.
Output: ensemble classifier H.
Step 1: learning a fist-level classifier (SVM, DT, RF, Boost).
for $i\leftarrow 1$ to T do
Learning base classifier h_i based on D.
End for.
Step 2: construct a new dataset from D.
for $i\leftarrow 1$ to m do
Construct a new dataset that contains $\{x_i',y\}$, where

$$x'=\{h_1(x_i),h_2(x_i),....,h_r(x_i)\}.$$

End for.
Step 3: learning a second-level classifier.
Learning a new classifier h' based on the newly constructed dataset
Return $H(x)=h'(h_1(x),h_2(x),.....h_r(x))$.

5. Results of studying heart attack prediction using ensemble learning

5. 1. Investigating the performance of the stacking model for heart attack prediction

Heart attack prediction model is based on ensemble learning using the stacking technique shown in Fig. 3. The performance of the model has been tested and measured using recall, F1 score, accuracy as shown in Table 2, confusion matrix as shown in Fig. 4, and ROC as shown in Fig. 5 and high performance was obtained.

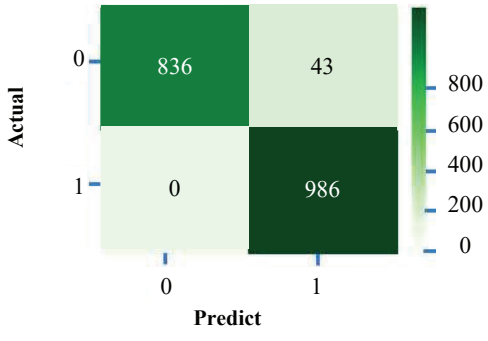

Fig. 4. Confusion matrix of the proposed methodology

Table 2

Stacking Ensemble Technique Performance Measures

Precision	Recall	F1 score	Accuracy	AUC
1.00	0.95	0.97	96.69	0.98
0.96	1.00	0.98		

Various measures are used to verify the method's performance since research in the medical field needs to present results that are not biased towards a specific scale. Our approach used different measures, as shown in Table 2, Fig. 4, and Fig. 5.

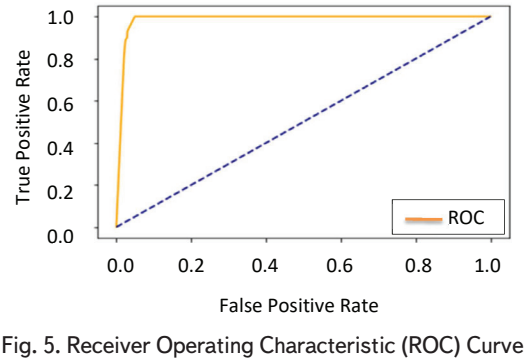

Fig. 5. Receiver Operating Characteristic (ROC) Curve

5. 2. Comparing the performance of the proposed method with other machine learning methods

The proposed method compared to single traditional algorithms achieved better accuracy as shown in Table 2. The results of the proposed method compared to the results of previous research methods as shown in Table 4 on the same data set showed better accuracy.

As shown in Table 4, many of the researchers relied on the Framingham heart study dataset and used different

machine learning methods to predict heart attacks. The proposed methodology showed better accuracy compared to previous research on the same dataset.

Table 3

Algorithm's accuracy

Algorithms	Accuracy
K-NN	69.6514745308311
Logistic regression	69.2225201072386
SVM	90.20554066130474
Decision Tree	92.70777479892761
XGBOOST	92.06434316353888
Random forest	93.69436997319035
Stacking Ensemble Technique	96.69436997319035

Table 4

Previous research on the Framingham heart study dataset

No.	Paper	Year	Dataset	Method	Accuracy
1	Paper [52]	2019	Framingham heart study dataset	Logistic regression	87 %
2	Paper [53]	2019	Framingham heart study dataset	Logistic regression	86.6 %
3	Paper [54]	2020	Framingham heart study dataset	Baseline, random forest, logistic regression, random forest, decision tree, and ensemble	Ensemble got high accuracy of 85.29 %
4	Paper [55]	2020	Framingham heart study dataset	Random forests models, decision trees, random forests, logistic regression and support vector machines	logistic regression got high accuracy of 84 %
5	Paper [56]	2020	Framingham heart study dataset	Random forests and logistic regression	logistic regression got high accuracy of 85.04 %
6	Paper [57]	2020	Framingham heart study dataset	Sparse autoencoder and ANN	90 %
7	Paper [58]	2020	Framingham heart study dataset	Logistic Regression, Random Forest, K-Nearest Neighbors (KNN), support vector machines (SVM), and Decision Tree	logistic regression got high accuracy of 88.86 %
8	Paper [59]	2021	Framingham heart study dataset	CBFS, PCA and MLP, ADboost, Naive Bayes, SMO	CBFS with MLP got high accuracy of 84.9 %

6. Discussion of experimental results of studying the heart attack prediction model based on ensemble learning

The results of the study show the superiority of the proposed method compared to the traditional methods as shown in Table 2 and compared to previous studies that used the same dataset as in Table 4 and different datasets, which also showed the preference of the proposed method.

The performance of the algorithms does not necessarily depend on the accuracy measure, so the performance of the proposed model was measured using the confusion matrix as shown in Fig. 4, ROC as shown in Fig. 5, recall, F1 score, and accuracy as shown in Table 2.

The method of aggregate results from more than one learning model and building a predictive model using the results shown in Fig. 3, has high performance because the final results will depend on the majority vote that the meta-classifier trained on.

The limits and disadvantages of the study are that most of the existing research to predict heart attacks depends on two separate data sets, one of which uses the electronic record of the patient and the other uses the ECG scheme, but the machine learning algorithms, including the proposed method, do not deal with the ECG directly. That is why we suggest using deep learning to extract important features from an ECG scheme and combine them with the electronic patient record dataset.

This study can be developed by applying it to a different dataset, comparing the results among them in addition to including ECG with data of the electronic patient record dataset, and finding a mathematical model that explains these results.

7. Conclusions

1. The presented work mainly uses machine learning algorithms based on electronic medical record data. A heart attack prediction learning model was developed by combining the algorithms of the decision tree, logistic regression, SVM, XGboost based on the ensemble learning technique, and using logistic regression as a meta-classifier. Thereby making up for the limitations of traditional single machine learning algorithms for heart attack prediction. The model performance was measured using recall, F1 score, accuracy, and ROC and high performance of 0.95 %, 0.97 %, 96.69 %, and 0.98 %, respectively, was achieved.

2. The accuracy of the stacking ensemble technique reached 96.69 with stability, as shown in Table 3, compared to the proposed methodology with models of single traditional machine learning algorithms, at the same time compared to methods of previous research that used the Framingham heart study dataset and various datasets. The experimental results of the proposed method showed higher prediction accuracy, better performance, proved the superiority and reliability.

That allows this model to help doctors to be used for practical tasks of heart attack prediction.

Acknowledgments

First of all, I would like to thank Associate Professor Ibrahim Ahmed Saleh for his meticulous care and help in my life and academics. Teacher Ibrahim has noble morals, kindness to others, rigorous scholarship, and profound knowledge. He not only taught me the skills of learning, but also the principles of life, which will benefit me for life. At the end of this topic, I would like to extend my sincerest gratitude to Teacher Ibrahim again. Thanks to the University of Mosul, college of computer science and mathematics for their care for my daily experiments and life.

References

1. Waqar, M., Dawood, H., Dawood, H., Majeed, N., Banjar, A., Alharbey, R. (2021). An Efficient SMOTE-Based Deep Learning Model for Heart Attack Prediction. Scientific Programming, 2021, 1–12. doi: https://doi.org/10.1155/2021/6621622
2. Muhammad, Y., Tahir, M., Hayat, M., Chong, K. T. (2020). Early and accurate detection and diagnosis of heart disease using intelligent computational model. Scientific Reports, 10 (1). doi: https://doi.org/10.1038/s41598-020-76635-9
3. Roth, G. A., Mensah, G. A., Johnson, C. O., Addolorato, G., Ammirati, E., Baddour, L. M. et. al. (2020). Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019: Update From the GBD 2019 Study. Journal of the American College of Cardiology, 76 (25), 2982–3021. doi: https://doi.org/10.1016/j.jacc.2020.11.010
4. Ramdurai, B. (2020). How AI (Artificial Intelligence) can improve Patient Experience in OPD (Out-Patient Dept.). doi: https://doi.org/10.13140/RG.2.2.23267.17440
5. Keya, M. S., Shamsojjaman, M., Hossain, F., Akter, F., Islam, F., Emon, M. U. (2021). Measuring the Heart Attack Possibility using Different Types of Machine Learning Algorithms. 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS). doi: https://doi.org/10.1109/icais50930.2021.9395846
6. Rincy, T. N., Gupta, R. (2020). Ensemble Learning Techniques and its Efficiency in Machine Learning: A Survey. 2nd International Conference on Data, Engineering and Applications (IDEA). doi: https://doi.org/10.1109/idea49133.2020.9170675
7. Virani, S. S., Alonso, A., Aparicio, H. J., Benjamin, E. J., Bittencourt, M. S. et. al. (2021). Heart Disease and Stroke Statistics – 2021 Update. Circulation, 143 (8). doi: https://doi.org/10.1161/cir.0000000000000950
8. Nurmamadovna, I. N. (2021). Coronary Heart Disease. The American Journal of Medical Sciences and Pharmaceutical Research, 03 (02), 31–36. doi: https://doi.org/10.37547/tajmspr/volume03issue02-04
9. Dash, S., Shakyawar, S. K., Sharma, M., Kaushik, S. (2019). Big data in healthcare: management, analysis and future prospects. Journal of Big Data, 6 (1). doi: https://doi.org/10.1186/s40537-019-0217-0
10. Saw, M., Saxena, T., Kaithwas, S., Yadav, R., Lal, N. (2020). Estimation of Prediction for Getting Heart Disease Using Logistic Regression Model of Machine Learning. 2020 International Conference on Computer Communication and Informatics (ICCCI). doi: https://doi.org/10.1109/iccci48352.2020.9104210
11. Yekkala, I., Dixit, S. (2018). Prediction of Heart Disease Using Random Forest and Rough Set Based Feature Selection. International Journal of Big Data and Analytics in Healthcare, 3 (1), 1–12. doi: https://doi.org/10.4018/ijbdah.2018010101
12. Shah, D., Patel, S., Bharti, S. K. (2020). Heart Disease Prediction using Machine Learning Techniques. SN Computer Science, 1 (6). doi: https://doi.org/10.1007/s42979-020-00365-y
13. Kamboj, M. (2019). Heart Disease Prediction with Machine Learning Approaches. International Journal of Science and Research, 9 (7), 1454–1458. Available at: https://www.ijsr.net/get_count.php?paper_id=SR20724113128
14. Bindhika, G. S. S., Meghana, M., Reddy, M. S., Rajalakshmi (2020). Heart Disease Prediction Using Machine Learning Techniques. International Research Journal of Engineering and Technology (IRJET), 07 (04), 5272–5276. Available at: https://www.researchgate.net/publication/344557562_Heart_Disease_Prediction_Using_Machine_Learning_Techniques
15. Kim, J. K., Kang, S. (2017). Neural Network-Based Coronary Heart Disease Risk Prediction Using Feature Correlation Analysis. Journal of Healthcare Engineering, 2017, 1–13. doi: https://doi.org/10.1155/2017/2780501
16. Kasbe, T., Pippal, R. S. (2017). Design of heart disease diagnosis system using fuzzy logic. 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS). doi: https://doi.org/10.1109/icecds.2017.8390044
17. Salhi, D. E., Tari, A., Kechadi, M.-T. (2021). Using Machine Learning for Heart Disease Prediction. Lecture Notes in Networks and Systems, 70–81. doi: https://doi.org/10.1007/978-3-030-69418-0_7
18. Kshirsagar, P. (2020). ECG Signal Analysis and Prediction of Heart Attack with the Help of Optimized Neural Network. Alochana Chakra Journal, IX (IV), 497–506. Available at: https://www.researchgate.net/publication/340599087
19. Malavika, G., Rajathi, N., Vanitha, V., Parameswari, P. (2020). Heart Disease Prediction Using Machine Learning Algorithms. Bioscience Biotechnology Research Communications, 13 (11), 24–27. doi: https://doi.org/10.21786/bbrc/13.11/6
20. Lee, W.-M. (2019). Supervised Learning-Classification Using K-Nearest Neighbors (KNN). Python® Machine Learning, 205–220. doi: https://doi.org/10.1002/9781119557500.ch9
21. Lin, A., Wu, Q., Heidari, A. A., Xu, Y., Chen, H., Geng, W. et. al. (2019). Predicting Intentions of Students for Master Programs Using a Chaos-Induced Sine Cosine-Based Fuzzy K-Nearest Neighbor Classifier. IEEE Access, 7, 67235–67248. doi: https://doi.org/10.1109/access.2019.2918026
22. Jiang, L., Cai, Z., Wang, D., Jiang, S. (2007). Survey of Improving K-Nearest-Neighbor for Classification. Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007). doi: https://doi.org/10.1109/fskd.2007.552
23. García, V., Mollineda, R. A., Sánchez, J. S. (2007). On the k-NN performance in a challenging scenario of imbalance and overlapping. Pattern Analysis and Applications, 11 (3-4), 269–280. doi: https://doi.org/10.1007/s10044-007-0087-5
24. Khateeb, N., Usman, M. (2017). Efficient Heart Disease Prediction System using K-Nearest Neighbor Classification Technique. Proceedings of the International Conference on Big Data and Internet of Thing - BDIOT2017. doi: https://doi.org/10.1145/3175684.3175703
25. Hasija, Y., Chakraborty, R. (2021). Logistic Regression. Hands-On Data Science for Biologists Using Python, 183–196. doi: https://doi.org/10.1201/9781003090113-9-9

26. Roback, P., Legler, J. (2021). Logistic Regression. Beyond Multiple Linear Regression, 151–192. doi: https://doi.org/10.1201/9780429066665-6
27. Imamovic, D., Babovic, E., Bijedic, N. (2020). Prediction of mortality in patients with cardiovascular disease using data mining methods. 2020 19th International Symposium INFOTEH-JAHORINA (INFOTEH). doi: https://doi.org/10.1109/infoteh48170.2020.9066297
28. Casarin, R., Facchinetti, A., Sorice, D., Tonellato, S. (2021). Decision trees and random forests*. The Essentials of Machine Learning in Finance and Accounting, 7–36. doi: https://doi.org/10.4324/9781003037903-2
29. Singh, Y. K., Sinha, N., Singh, S. K. (2017). Heart Disease Prediction System Using Random Forest. Advances in Computing and Data Sciences, 613–623. doi: https://doi.org/10.1007/978-981-10-5427-3_63
30. Santhi, P., Ajay, R., Harshini, D., Jamuna Sri, S. S. (2021). A Survey on Heart Attack Prediction Using Machine Learning. Turkish Journal of Computer and Mathematics Education (TURCOMAT), 12 (2). doi: https://doi.org/10.17762/turcomat.v12i2.1955
31. Frery, J. (2019). Ensemble Learning for Extremely Imbalced Data Flows. HAL. Available at: https://tel.archives-ouvertes.fr/tel-02899943/document
32. Pathak, S., Mishra, I., Swetapadma, A. (2018). An Assessment of Decision Tree based Classification and Regression Algorithms. 2018 3rd International Conference on Inventive Computation Technologies (ICICT). doi: https://doi.org/10.1109/icict43934.2018.9034296
33. Kocarik Gacar, B., Deveci Kocakoç, İ. (2020). Regression Analyses or Decision Trees? Celal Bayar Üniversitesi Sosyal Bilimler Dergisi, 18 (4), 251–260. doi: https://doi.org/10.18026/cbayarsos.796172
34. Larose, D. T., Larose, C. D. (2014). Decision Trees. Discovering Knowledge in Data, 165–186. doi: https://doi.org/10.1002/9781118874059.ch8
35. Hasija, Y., Chakraborty, R. (2021). Decision Trees and Random Forests. Hands-On Data Science for Biologists Using Python, 209–217. doi: https://doi.org/10.1201/9781003090113-11-11
36. Thomas, T., Vijayaraghavan, A. P., Emmanuel, S. (2020). Applications of Decision Trees. Machine Learning Approaches in Cyber Security Analytics, 157–184. doi: https://doi.org/10.1007/978-981-15-1706-8_9
37. Larose, C. D., Larose, D. T. (2019). Decision trees. Data Science Using Python and R, 81–96. doi: https://doi.org/10.1002/9781119526865.ch6
38. Suthaharan, S. (2016). Decision Tree Learning. Integrated Series in Information Systems, 237–269. doi: https://doi.org/10.1007/978-1-4899-7641-3_10
39. Mrva, J., Neupauer, S., Hudec, L., Sevcech, J., Kapec, P. (2019). Decision Support in Medical Data Using 3D Decision Tree Visualisation. 2019 E-Health and Bioengineering Conference (EHB). doi: https://doi.org/10.1109/ehb47216.2019.8969926
40. Alsaleem, M. Y. A., Hasoon, S. O. (2020). Comparison of DT& GBDT algorithms for predictive modeling of currency exchange rates. EUREKA: Physics and Engineering, 1, 56–61. doi: https://doi.org/10.21303/2461-4262.2020.001132
41. Perros, H. G. (2021). Support Vector Machines. An Introduction to IoT Analytics, 279–302. doi: https://doi.org/10.1201/9781003139041-11
42. Nalepa, J., Kawulok, M. (2018). Selecting training sets for support vector machines: a review. Artificial Intelligence Review, 52 (2), 857–900. doi: https://doi.org/10.1007/s10462-017-9611-1
43. Vamshi Kumar, S., Rajinikanth, T. V., Viswanadha Raju, S. (2021). Heart Attack Classification Using SVM with LDA and PCA Linear Transformation Techniques. Algorithms for Intelligent Systems, 99–112. doi: https://doi.org/10.1007/978-981-33-4046-6_10
44. Kaestner, C. A. A. (2013). Support Vector Machines and Kernel Functions for Text Processing. Revista de Informática Teórica e Aplicada, 20 (3), 130. doi: https://doi.org/10.22456/2175-2745.39702
45. Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. Journal of Machine Learning Technologies, 2 (1). 37–63. Available at: https://www.researchgate.net/publication/276412348_Evaluation_From_precision_recall_and_F-measure_to_ROC_informedness_markedness_correlation
46. Alsaleem, M., Hasoon, S. (2020). Predicting Bank Loan Risks Using Machine Learning Algorithms. AL-Rafidain Journal of Computer Sciences and Mathematics, 14 (1), 159–168. doi: https://doi.org/10.33899/csmj.2020.164686
47. Gupta, A., Tatbul, N., Marcus, R., Zhou, S., Lee, I., Gottschlich, J. (2020). Class-Weighted Evaluation Metrics for Imbalanced Data Classification. arXiv.org. Available at: https://arxiv.org/pdf/2010.05995.pdf
48. Cutler, J., Dickenson, M. (2020). Introduction to Machine Learning with Python. Computational Frameworks for Political and Social Research with Python, 129–142. doi: https://doi.org/10.1007/978-3-030-36826-5_10
49. Gneiting, T., Vogel, P. (2018). Receiver Operating Characteristic (ROC) Curves. arXiv.org. Available at: https://arxiv.org/pdf/1809.04808.pdf
50. Piegorsch, W. W. (2020). Confusion Matrix. Wiley StatsRef: Statistics Reference Online, 1–4. doi: https://doi.org/10.1002/9781118445112.stat08244
51. Vasudev, R. A., Anitha, B., Manikandan, G., Karthikeyan, B., Ravi, L., Subramaniyaswamy, V. (2020). Heart disease prediction using stacked ensemble technique. Journal of Intelligent & Fuzzy Systems, 39 (6), 8249–8257. doi: https://doi.org/10.3233/jifs-189145
52. Ravi, S., Sambath, D. M., Thangakumar, D. J., Kumar, D., Naveen, G., Bramiah, M. (2021). Prediction of Heart Disease Using Machine Learning Algorithms. Alinteri Journal of Agriculture Sciences, 36 (1), 260–264. doi: https://doi.org/10.47059/alinteri/v36i1/ajas21039
53. Zhang, Y., Diao, L., Ma, L. (2021). Logistic Regression Models in Predicting Heart Disease. Journal of Physics: Conference Series, 1769, 012024. doi: https://doi.org/10.1088/1742-6596/1769/1/012024

54. Yadav, K. K., Sharma, A., Badholia, A. (2021). Heart disease prediction using machine learning techniques. Information technology in industry, 9 (1), 207–214. doi: https://doi.org/10.17762/itii.v9i1.120
55. Glienke, J. S. (2020). Life and death: Quantifying the risk of heart disease with machine learning. Honors Program Theses, 415. Available at: https://scholarworks.uni.edu/hpt/415
56. Latifah, F. A., Slamet, I., Sugiyanto (2020). Comparison of heart disease classification with logistic regression algorithm and random forest algorithm. International Conference on Science and Applied Science (ICSAS2020). doi: https://doi.org/10.1063/5.0030579
57. Mienye, I. D., Sun, Y., Wang, Z. (2020). Improved sparse autoencoder based artificial neural network approach for prediction of heart disease. Informatics in Medicine Unlocked, 18, 100307. doi: https://doi.org/10.1016/j.imu.2020.100307
58. Chauhan, Y. J. (2020). Cardiovascular Disease Prediction using Classification Algorithms of Machine Learning. International Journal of Science and Research (IJSR), 9 (5), 194–200. Available at: https://www.researchgate.net/publication/341235098
59. Kuruvilla, A. M., Balaji, N. V. (2021). Heart disease prediction system using Correlation Based Feature Selection with Multilayer Perceptron approach. IOP Conference Series: Materials Science and Engineering, 1085 (1), 012028. doi: https://doi.org/10.1088/1757-899x/1085/1/012028
60. Zaker, N. A., Alsaleem, N., Kashmoola, M. A. (2018). Multi-agent Models Solution to Achieve EMC In Wireless Telecommunication Systems. 2018 1st Annual International Conference on Information and Sciences (AiCIS). doi: https://doi.org/10.1109/aicis.2018.00061
61. Kashmoola, M. A., Alsaleem, M. Y. anad, Alsaleem, N. Y. A., Moskalets, M. (2019). Model of dynamics of the grouping states of radio electronic means in the problems of ensuring electromagnetic compatibility. Eastern-European Journal of Enterprise Technologies, 6 (9 (102)), 12–20. doi: https://doi.org/10.15587/1729-4061.2019.188976
62. Ahmed, M. K., Aziz, S. F., Alsaleem, N. Y. A., Sielivanov, K., Moskalets, M. (2020). Method for determining the responses from a non-linear system using the Volterra series. Eastern-European Journal of Enterprise Technologies, 4 (9 (106)), 34–44. doi: https://doi.org/10.15587/1729-4061.2020.210754