Comparison of Some Machine Learning Algorithms for Predicting Heart Failure

Ramadan A.M. Elghalid
Department of Computer Science
College of Computer Technology
Libya, Benghazi
rama_b3@yahoo.com

Ahmed Alwirshiffani
Department of Computer Science
College of Computer Technology
Libya, Benghazi
ahmed.m.manae@gmail.com

Abdelhafid Ali I. Mohamed
Department of Computer Science
College of Computer Technology
Libya, Benghazi
hafithmathe@yahoo.com

Fatimah Husayn Amir Aldeeb
Department of Computer Science
College of Computer Technology
Libya, Benghazi
fatmaeldeib@yahoo.com

Aisha Andiasha
Department of Computer Science
College of Computer Technology
Libya, Benghazi
aisha2020ramadan@gmail.com

2022 International Conference on Engineering & MIS (ICEMIS) | 978-1-6654-5436-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICEMIS56295.2022.9914325

***Abstract*— In this modern era, people are working hard to meet their physical needs and non-effective their ability to spend time for themselves which leads to physical stress and mental disorder. Many reports state that heart failure is caused by many diseases that we ignore and chronic diseases as well as the global epidemic of the Coronavirus. Heart failure does not mean that it will stop at any moment but rather that the heart is not working as it should. Heart failure, also known as congestive heart failure, is a condition that develops when your heart does not pump enough blood for your body's needs. This paper aims to predict if someone is at high risk of being diagnosed as a heart patient using different machine learning methods. We have collected datasets to analyze data and mining using 7 algorithms of machine learning to predict whether the patient suffers from heart failure or not. This paper used a dataset retrieved from kaggle repository, which consists of 12 attributes (Features). This work is implemented using K-Nearest Neighbors (KNN), Naïve Bayes (NB), Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Decision Tree (TD) and Neural Network (NN) algorithms. Results showed that Logistic Regression, Support Vector Machine and Neural Network respectively gave the best result with an accuracy of up to 94.57%.**

Keywords—Heart Failure, Machine Learning, Prediction, Classification Algorithms

I. INTRODUCTION

Heart disease is the major cause of death globally. More people die annually from cardiovascular diseases (CVDs) than from any other cause. Each year 17.5 million people are dying due to cardiovascular disease according to World Health Organization reports[1]. Heart attacks are often tragic events and are the result of blocking blood flow to the heart or brain. People at risk of heart disease may show elevated blood pressure, glucose and lipid levels as well as stress. All of these parameters can be easily measured at home by basic health facilities. Coronary heart disease, Cardiomyopathy and cardiovascular disease are the categories of heart disease. The word "heart disease" includes a variety of conditions that affect the heart and blood vessels and how the fluid gets into the bloodstream and circulates there in the body. Cardiovascular disease (CVD) causes many diseases, disability and death. Diagnosis of the disease is important and complex work in medicine. Medical diagnosis is considered a crucial but difficult task to be done efficiently and effectively. The automation of this task is beneficial. Unfortunately, all physicians are not experts in any subject specialists and beyond the scarcity of resources, there are some places. Data mining can be used to find hidden patterns and knowledge that may contribute to successful decision-making[2]. This plays a key role for healthcare professionals in making accurate decisions and providing quality services to the public. The approach provided by the health care organization to professionals who do not have more knowledge and skills is also very important[3]. One of the main limitations of existing methods is the ability to draw accurate conclusions as needed. In our approach, we are using different data mining techniques and machine learning algorithms, KNeighbors, Naïve Bayes, Logistic Regression, Support Vector Machine, Random Forest, Decision Tree and Neural Network to predict heart failure based on some health parameters.

II. RELATED WORK

a short review of some recent papers that compare a number of machine learning algorithms and know their results.

Gnaneswar, B. and M.E. Jebarani[4], used five algorithms to predict heart failure, they are SVM, DT, K-Mean, NB, NN, and the authors suggested as a result of their work that it is preferable to develop algorithms that can work on linear datasets and non-linear datasets because there are algorithms

that gave good results with Linear data sets, while the results were weak with non-linear data sets.

Fahmy, A.S., et al [5], used four algorithms SVM, RF, LR, and NN in this research and the accuracy rates were 69%, 70%, 71% and 68%, respectively. It seems that the research agrees that the use of machine learning algorithms increases the chance of early detection of heart failure.

Wang, J. [6], checked various classification techniques like LR, KNN, NB, DT and RF were analyzed and compared, and results were obtained for accuracy as follows: 86%, 73%, 63%, 83% and 86%, respectively. The authors support the claim that finds other methods that can be ideal for performing heart failure prediction. For example, combining different models of machine learning.

Pushpavathi, T., S. Kumari, and N. Kubra [7], Some of the machine learning techniques were used in this work to process the data set which are KNN, NB and RF and it achieved the following accuracy 55%, 81% and 81.6% respectively.

Rairikar, A., et al [8], used three algorithms have been used such as Random Forest, Decision trees and Naive Bayes. the result was that Random Forest provides perfect results as compared to the Decision tree and Naive Bayes.

Amin, M.S., Y.K. Chiam, and K.D. Varathan [9], the authors said that the best performing classification modelling techniques that improve the accuracy of heart failure prediction were selected, These technologies are NB, SVM, LR, NN, KNN and DT. The top data mining techniques that produce high accuracy in prediction are identified in this research as Naïve Bayes and Support Vector Machine, Both of them get an accuracy of 78%, while the rest of the algorithms have a lower percentage.

Mansur Huang, N.S., Z. Ibrahim, and N. Mat Diah [10], applied some machine learning techniques in making early predictions of heart failure may have the potential to improve the healthcare management system. In this experiment, RF seems to achieve the best performance score compared to other techniques, where you get an accuracy of up to 88%. while the other techniques get SVM 85%, NB 86% and LR 88%.

The author in [1] examines the performance of the heart disease diabetes dataset by using several classification algorithms of machine learning and got the following results regarding accuracy SVM 83%, NB 83% and RF 81%.

It is anticipated that the use of our results and the previous findings will be useful to the machine learning community as it could be the basis of the prediction of heart failure on different clinical datasets.

III. MATERIALS AND METHODS

Several algorithms were utilized to predict heart failure among which K nearest Neighbors (KNN), Naive Bayes (NB), Decision Tree (TD), Neural Network (NN), Support Vector Machine (SVM), Logistic Regression (LR), and , Random Forest (RF). These algorithms are applied to a Heart Failure dataset taken from the kaggle repository including 918 samples (patients records). The dataset includes heart failure features. To enhance the performance of the algorithms, these features are analyzed, and the features' importance scores, Accuracy, Sensitivity, and Specificity are considered.

A. K-nearest neighbour classifier (KNN)

The k-nearest neighbours (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. It's easy to implement and understand but has a major drawback of becoming significantly slows as the size of that data in use grows[11]. KNN calculations use the data and characterize new data points dependent on resemblance measures (e.g., distance function). The KNN calculation accepts that comparative things are close to one another. In KNN, Classification occurs by considering the majority vote to its neighbours. The data point goes to the class that has the most intimate neighbours. As we increment the number of nearest neighbours, the estimation of k and accuracy may increment. The predictor variables anticipate the target variable or the dependent variable[12].

B. Naïve Bayes classifier

The next algorithm is known as Naïve Bayes. it's also a supervised learning classification model, which classifies the info by computing the probability of independent variables. After calculating the probability of every class, the high probability class does assign for the entire transaction[13]. Naïve Bayes may be a common approach used to predict classes for different types of datasets such as educational data mining[14] and medical data mining[15]. This model is also useful for classifying different quiet datasets like virus detection[16]. It works by using the values for independent variables and predicting a pre-defined class for every record.

C. Decision Tree classifier

The decision tree algorithm has a tree structure. It divides the dataset into smaller subsets. a choice node has two or more branches. A resolution may be a target node. the basis node is the top node of the choice tree. This algorithm uses entropy and knowledge gain. Entropy is employed to calculate the homogeneity of a sample. Building a choice tree is about finding the attributes that give the biggest information gain. Finally, select the attribute with the very best information gain as the decision node and the zero universe branch as the end node[4].

D. Neural Network classifier

The neural network is an iterative process. It uses nonlinear data. The most goal is to reduce the difference between the actual production and the cost of the forecast production. A random weight is assigned to each of the

entries. Then the corresponding performance is calculated and compared with the specified performance. The difference between them gives an error this algorithm minimizes the error with successive iterations by adjusting the input parameters. The advantages of neural network include adaptive learning, fault tolerance etc. Several Neural Network methodologies have been developed like the classification methodology called an artificial neural network, which may be a combination of a forward and backward propagation algorithm for predicting heart failure[17]. Many world problems can be solved using this methodology. The analysis is performed from a heart condition data set. Parallelism is implemented in each neuron throughout the hidden and output layers.

E. Support Vector Machine classifier (SVM)

Support Vector Machine (SVM) is one of the supervised learning methods, that can be widely used in statistical classification and regression analysis. SVM belongs to generalized linear classifiers, which are characterized by their ability to minimize empirical error and maximize geometric edge region at the same time. Therefore, another name for SVM is the maximum edge region classifier[6].

F. Logistic Regression classifier

Logistic regression is another kind of classification model, which learn and predict the parameters in the given dataset using regression analysis[18]. The learning and prediction processes are based on measuring the probability of binary classification. The logistic regression model requires class variables that should be binary classified. Likewise, in this dataset the ―target‖ column has two types of binary numbers, "0" for the patient who has no chances of heart failure, and "1" for the patients who have been predicted as heart failure patients. On the other side, the independent variables can be binary classified, nominal, or polynomial types.

G. Random forest classifier

The random forest is the next model chosen and implemented in this study. Since this model belongs to the classification family, it is also known as the teacher training algorithm. In the training phase, this model first generates a few random trees called a forest[19]. for example, if the dataset contains an "x" number of attributes, first select some features, randomly named "y". Using all possibilities; (ie "and"), create nodes using the simpler rift method. Also, the algorithm will work to create the entire forest by repeating the steps above. Then, in the forecasting process, the algorithm tries to shuffle the trees using the estimated result and the voting procedure[20]. the goal of combining random trees by voting in the forest is to eliminate the use of the highest predicted tree, which can improve the accuracy of forecasts for future data.

IV. DATASET DESCRIPTION

The dataset utilized in this research is collected from the Kaggle platform, the dataset is additionally known as Heart Failure Dataset[21]. Altogether, the info was a combination of five different datasets. It is an open dataset, having a variety of attributes, during this experiment we used all attributes that are most useful to predict heart condition in a patient. In addition the dataset file contains 918 records of patients. The entire description of each attribute and the number of values for each attribute is shown in table 1:

Table 1. Description of Features

N.	Features	Description
1	Age	The person's age in years
2	Sex	The person's sex (1 = male, 0 = female)
3	chest_pain_type	0: asymptomatic 1: atypical angina 2: non-angina pain 3: typical angina
4	RestingBP	The person's resting blood pressure (mm Hg on admission to the hospital)
5	Cholesterol	The person's cholesterol measurement in mg/dl
6	FastingBS	The person's fasting blood sugar (> 120 mg/dl, 1 = true; 0 = false)
7	RestingECG	resting electrocardiographic results 0: showing probable or definite left ventricular hypertrophy by Estes' criteria 1: normal 2: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
8	MaxHR	The person's maximum heart rate achieved
9	ExerciseAngina	Exercise induced angina (1 = yes; 0 = no)
10	Oldpeak	ST depression induced by exercise relative to rest ('ST' relates to positions on the ECG plot. See more here)
11	ST_Slope	the slope of the peak exercise ST segment 0 downsloping; 1 flat; 2 upsloping
12	HeartDisease	The Target (1 = no, 0= yes)

V. RESULTS AND DISCUSSIONS

In this paper, we have followed a series of steps to get to the best model. We first need to load the data and perform some necessary preprocessing on the data, such as the goal of converting the output column into a factor so that machine-

learning algorithms can work with it without any problems. Next, we trained the dataset and worked on it with the using mentioned machine learning algorithms. The results are as shown in a table 2:

Table 2. Results of used algorithms

Model Name	Accuracy %	Sensitivity %	Specificity %
K-nearest neighbour classifier (KNN)	91.30	89.09	94.59
Naïve Bayes	91.30	90.91	91.89
Decision Tree	82.61	78.18	89.19
Neural Network	94.57	96.36	91.89
Support Vector Machine (SVM)	94.57	96.36	91.89
Logistic Regression	94.57	96.36	91.89
Random Forest	92.39	94.55	89.19

Furthermore figure 1 shows the results of the classification algorithms used with respect to the accuracy, sensitivity, and specificity. The logistic regression algorithm, neural network and support vector machine (SVM) outperformed the others with an accuracy of 94.57%, and a sensitivity ratio of 96.36%. The Random Forest algorithm achieved an accuracy of 92.39%. Decision Tree, Naïve Bayes and K-nearest Neighbors all have an accuracy of less than 92%. While these algorithms achieved a sensitivity of less than 95%. Regarding specificity, the K-nearest Neighbors algorithm got the highest score of 94.59%, while the remaining algorithms got the lowest 92%:

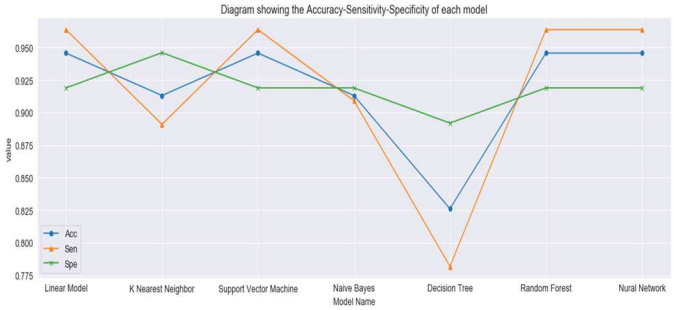

Fig 1. Show predicting results

In addition to that figures 2, 3 and 4 showed the attributes and their importance according to the classification algorithms used in this paper. These figures show the ranking of attributes according to the importance of attributes and coefficient scores for all applied evaluation algorithms. These figures also tend to represent the traits most responsible for heart failure.

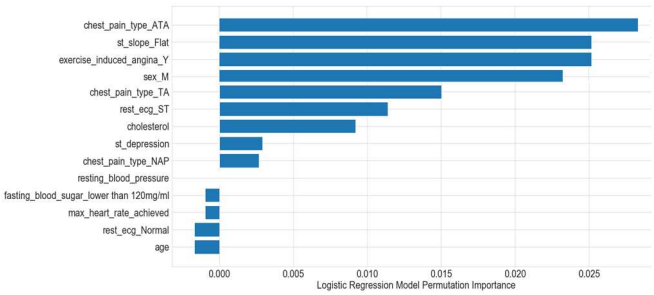

Figure 2. Visualizing important features for heart failure prediction produced by Logistic Regression

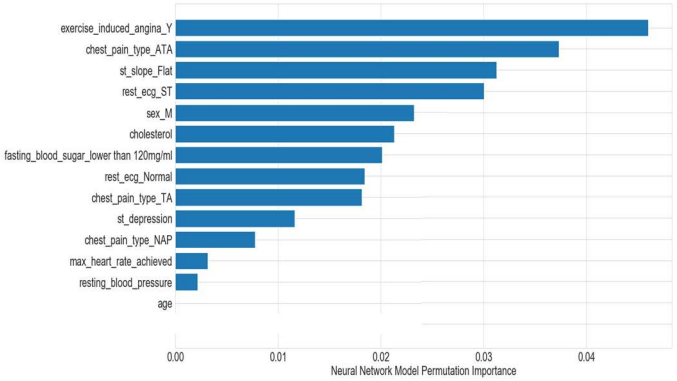

Figure 3. Visualizing important features for heart failure prediction produced by Neural Network.

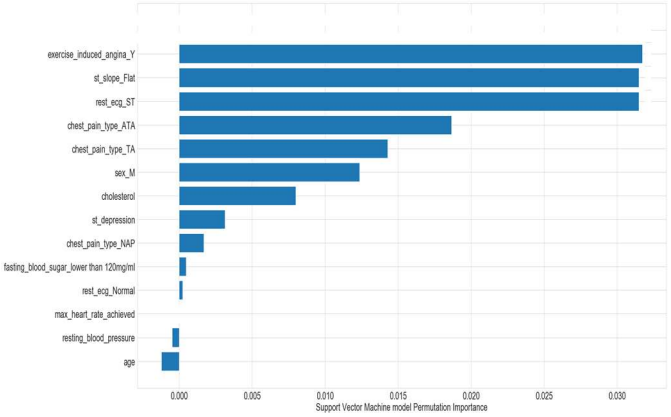

Figure 4. Visualizing important features for heart failure prediction produced by Support Vector Machine.

According to Figures 2, 3 and 4, exercise-induced angina is an important feature in the detection and prognosis of heart failure. In addition to the chest pain type. Table 3 showed the four main Features based on the importance of the features and the correlation value for the three best algorithms:

Table 3: Feature ranking for Heart Failure

Rank	Best Algorithms		
	Logistic Regression	**Neural Network**	**Support Vector Machine**
1ST	chest_pain_ type	exercise_induced_ angina	exercise_induced_ angina
2nd	St_slope	St_slope	chest_pain_type
3rd	exercise_induc ed_angina	Rest_ecg	St_slope
4th	sex	chest_pain_type	Rest_ecg

Note from the results obtained, compared with the results of previous studies that were referred to in the related work section, that the results of this research were better.

VI. THREATS TO VALIDITY

Construct validity, we have predicted heart failure by using some of the machine learning algorithms in a heart failure dataset, which contains the most important characteristics related to heart failure. There could have been a potential threat to validity if one algorithm was used without taking the rest into account, but in this paper, the results of a group of machine learning algorithms were compared and the best among them were determined. Internal Validity, the most important characteristics that affect the work of the heart have been identified, put into the data set and used to predict heart failure, and each algorithm identifies the most important characteristics in the data set and uses them to predict heart failure after the training process. External Validity, involves the extent to which the results of a study can be generalized and applied beyond the sample. In this research, a dataset from the Kaggle platform has been used, which contains as mentioned earlier, the most important features associated with heart failure. Therefore, we believe that the results may not differ significantly if the same algorithms are applied to another data set. Conclusion validity, one of the threats to conclusion validity was related to the dataset. To improve and ensure the validity of the results the dataset was preprocessed first

VII. CONCLUSION

Machine learning techniques help to reduce the effort and time for medical officers to conduct early predictions for healthcare management purposes. As the number of deaths increases due to heart failures, a machine learning technique system can help predict heart failure accurately and effectively. This paper showed that applying machine-learning techniques in making early predictions of heart failure may have the potential to improve the healthcare management system. After comparing the algorithms that were used in the process of predicting heart failure, Neural Network, Support Vector Machine and Logistic Regression seem to achieve the best performance score compared to other techniques. It can lead to a promising disease management strategy that may reduce the progression of the disease. As a scope of future work, a hybrid of machine learning techniques with optimization algorithms with more data will be examined. This will help in increasing the accuracy.

REFERENCES

1. Meshref, H., *Cardiovascular Disease Diagnosis: A Machine Learning Interpretation Approach.* International Journal of Advanced Computer Science and Applications, 2019. **10**(12).

2. Kumar, N.K., et al. *Analysis and prediction of cardio vascular disease using machine learning classifiers*. in *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*. 2020. IEEE.

3. Awan, S.M., M.U. Riaz, and A.G. Khan, *Prediction of heart disease using artificial neural network.* 2018.

4. Gnaneswar, B. and M.E. Jebarani. *A review on prediction and diagnosis of heart failure*. in *2017 international conference on innovations in information, embedded and communication systems (ICIIECS)*. 2017. IEEE.

5. Fahmy, A.S., et al., *Machine learning for predicting heart failure progression in hypertrophic cardiomyopathy.* Frontiers in cardiovascular medicine, 2021. **8**: p. 647857.

6. Wang, J. *Heart Failure Prediction with Machine Learning: A Comparative Study*. in *Journal of Physics: Conference Series*. 2021. IOP Publishing.

7. Pushpavathi, T., S. Kumari, and N. Kubra. *Heart Failure Prediction by Feature Ranking Analysis in Machine Learning*. in *2021 6th International Conference on Inventive Computation Technologies (ICICT)*. 2021. IEEE.

8. Rairikar, A., et al. *Heart disease prediction using data mining techniques*. in *2017 International conference on intelligent computing and control (I2C2)*. 2017. IEEE.

9. Amin, M.S., Y.K. Chiam, and K.D. Varathan, *Identification of significant features and data mining techniques in predicting heart disease.* Telematics and Informatics, 2019. **36**: p. 82-93.

10. Mansur Huang, N.S., Z. Ibrahim, and N. Mat Diah, *Machine learning techniques for early heart failure prediction.* Malaysian Journal of Computing (MJoC), 2021. **6**(2): p. 872-884.

11. Mahesh, B., *Machine learning algorithms-a review.* International Journal of Science and Research (IJSR).[Internet], 2020. **9**: p. 381-386.

12. Yousefi, S., *Comparison of the performance of machine learning algorithms in predicting heart disease.* Frontiers in Health Informatics, 2021. **10**(1): p. 99.

13. Bashir, S., et al. *Improving heart disease prediction using feature selection approaches*. in *2019 16th international bhurban conference on applied sciences and technology (IBCAST)*. 2019. IEEE.

14. Razaque, F., et al. *Using naïve bayes algorithm to students' bachelor academic performances analysis*. in *2017 4th IEEE International Conference on Engineering Technologies and Applied Sciences (ICETAS)*. 2017. IEEE.

15. Bou Rjeily, C., et al., *Medical data mining for heart diseases and the future of sequential mining in medical field*, in *Machine Learning Paradigms*. 2019, Springer. p. 71-99.

16. Qasim, O. and K. Al-Saedi, *Malware Detection using Data Mining Naïve Bayesian Classification* Comput. Commun. Eng, 2017. **6**(11): p. 211-213.

17. Rani, K.U., *Analysis of heart diseases dataset using neural network approach.* arXiv preprint arXiv:1110.2626, 2011.

18. Hosmer Jr, D.W., S. Lemeshow, and R.X. Sturdivant, *Applied logistic regression*. Vol. 398. 2013: John Wiley & Sons.

19. Donges, N., *The random forest algorithm.* Towards data science, 2018. **22**.

20. Bashar, S.S., et al. *A machine learning approach for heart rate estimation from PPG signal using random forest regression algorithm*. in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*. 2019. IEEE.

21. Alotaibi, F.S., *Implementation of machine learning model to predict heart failure disease.* International Journal of Advanced Computer Science and Applications, 2019. **10**(6).