JURNAL EMACS
[Engineering, MAthematics and Computer Science] Vol.4 No.3 September 2022: 87-93
e-ISSN: 2686-2573
DOI: 10.21512/emacsjournal.v4i3.8683

Prediction of Heart Disease UCI Dataset Using Machine Learning Algorithms

Anderies[1*], Jalaludin Ar Raniry William Tchin[2], Prambudi Herbowo Putro[3]
Yudha Putra Darmawan[4], Alexander Agung Santoso Gunawan[5]

[1,2,3,4,5] Computer Science Department, School of Computer Science,
Bina Nusantara University,
Jakarta, Indonesia 11480
anderies@binus.edu; jalaludin.tchin@binus.ac.id; prambudi.putro@binus.ac.id;
yudha.darmawan@binus.ac.id; aagung@binus.edu

*Correspondence: anderies@binus.edu

Abstract *— Heart disease is inflammation or damage to the heart and blood vessels over time. the disease can affect anyone of any age, gender, or social status. After many studies trying to overcome and learn about heart disease, in the end, this disease can be detected using machine learning systems. It predicts the likelihood of developing heart disease. The results of this system give the probability of heart disease as a percentage. Data collection using secret data mining. The data assets handled in python programming use two main algorithms for machine learning, the decision tree algorithm, and the Bayes naive algorithm which shows the best of both for heart disease accuracy. The results we get from this study show that the SVM algorithm is the algorithm with the most excellent precision. and the highest accuracy with a score of 85% in predicting heart disease using machine learning algorithms.*

Keywords: Machine Learning; Hearth Disease; Classification; Feature Selection; Prediction

I. INTRODUCTION

One of the complicated and complex disease cases in the field of medical science is the prediction of heart disease. The heart is a very important organ for every human body[1]. Heart disease can accurately refer to conditions that exist, when the heart's abnormal function, can be caused by blood clots, heart arteries, or others. Health data is collected from several available sources, one of which is patient electronic devices stored in the format specified in the electronic health record (EHR). To detect or predict using this data and using AI or Machine Learning (ML) algorithms.

Heart disease is said to be the leading cause of death globally, according to the World Health Organization the deathrate from heart disease was around 17.7 million (31%) in 2015. Every year nearly 20 million people die, showing heart disease as the leading cause of death [2]. The group of diseases relatedto the heart and blood vessels is referred to as cardiovascular disease (CVD). CVD includes coronary heart disease (CHD) or known as coronary artery disease (CAD) which refers to disease of the heart arteries that supply oxygen and blood tothe heart and is associated with lifestyle conditions and age. Machine learning is a way of manipulating and extracting implicit, previously unknown or known, and potentially useful information about data [3]. Machine Learning is a very broad and diverse field and its scope from Supervised, Unsupervised, and Ensemble Learning is used to predict and determine the accuracy of a given data set.

Machine learning-based methods have been used in medical science. However, researchers continue to look for ways to optimize and improve the method. In machine learning, ensemble learning is one approach that has been proven to improve tasks in machine learning[3]. An ensemble classifier is using a set of individual classifiers together using a mechanism, such as a majority voting that incorporates a predictive component. Our purpose in this paper is to detect heart disease, using a machine learning system. This research has proven that ensemble classifiers often perform better than conventional classifiers.

In healthcare, machine learning can help doctors make more accurate predictions for patients, machine learning can increase the speed of processing and analyzing data. Using machine learning, predictive analytics algorithms can train large data sets and can perform deeper analysis on many variables with minimal changes when applied.

II. METHODS

2.1 Literature Review

Health is only one of the many fields where machine learning is employed extensively. Machine learning is the use of artificial intelligence (AI) that gives a system the capacity to automatically learn from experience and get better over time without having explicitly programmed[4]. Machine learning is frequently used to handle two main types of issues: regression and classification. Regression methods are specifically used to solve classification, binary, and multi-category problems involving numerical data. Unsupervised learning is one of the two divisions of machine learning. Unsupervised learning, which lacks a preconceived label and aims to identify natural structures in a dataset, differs from supervised learning in that it uses prior knowledge about the value of the output. Consequently, properly choosing machine learning algorithms is necessary.

2.1.1 Machine Learning Algorithm

There are some techniques of machine learning algorithms like classification, clustering, regression, association analysis, and outlier surveys. One is in predicting the heart by technical learning methods, classification. Multiple algorithms for machine studies used in predicting heart disease.

2.1.2 Support Vector Machine (SVM)

SVM (Support Vector Machine) is a sort of model used in regression and classification analysis to analyze data in order to find trends. Vector machine ratings are broken down into categories including linear, non-linear, RBF, sigmoid, and polynomial. With 13 properties, these approaches divide data points or vectors[5]. SVM, neural networks, Bayesian classification, decision trees, and logistic regression are tested for accuracy. SVM, that takes into account 102 instances, had the highest accuracy 90.5%, followed with neural networks 88.9%, Naive Bayes 82.2%, decision tree 77.9%, and logistic regression 73.9%.

2.1.3 Naïve Bayes Algorithm (NB Algorithm)

The Naive Bayes algorithm is a data mining technique that can be used to diagnose heart disease patients. A naive Bayes classifier is a method of applying the Bayes theorem with a strong independent model (Naive) that is easy to build, there is no time limit for difficult iterative parameters, so it is very helpful in the medical field to diagnose heart disease patients. 25 countries conducted a survey on heart disease, which included adults diagnosed with moderate to severe congenital heart disease, infective endocarditis/previous valvular intervention. Naive Bayes can classify 73.07% of input instances with precision. Thus, it can be shown that the average accuracy rate is 86%, precision is 73.07%, and recall is 73.7% using the UCI dataset repository based on 13 feature data [6]. Thus, the results show that the proposed method has a good performance quality, compared to other methods in the literature, after considering the lower factor that the attribute taken for analysis is not a direct indicator of heart disease. Bayes' theorem provides a way to calculate the posterior probability values, P(c—x), P(c), P(x), and P(x—c). Classifying Naive Bayes assumes that the effect of the value on the predictor (x) in a given class (c) is independent of the value of the other predictor.

2.1.4 Logistic Regression Algorithm and Artificial Neural Networks

Regression algorithms as well as artificial neural networks are models that contain many options regarding medical data classification tasks. Of the many options in the last category, the most well-known and well-known models are logistic regression (LR) and artificial neural networks (ANN). The heart disease dataset obtained from the UCI Machine Learning repository. It contains 13 attributes. These models come from different communities[7]. What is meant by the community, in this case, is the study of statistics and computer science mastered. In this case, there is also a summary that can be used to compare the discriminatory power of an artificial neural network (ANN). using a logistic regression model and achieve a fairly high percentage of 89% and stat. Testing 18% of this model is algorithmic, which means that both provide a functional form of f and a parameter vector a with the aim of expressing P(y-x) as P(y-x) = f(x, a) this parameter a is agreed based on the data set d, it is, therefore, possible to estimate the probability of reaching the maximum point.

2.1.5 Decision Tree Algorithm (DC Algorithm)

Decision tree classifiers are often used, and experiments are carried out to find the best classifier for the diagnosis of medical problems. conducted research on the application of decision trees and whether integrating voting with them could improve the accuracy of the decision trees themselves in the diagnosis of heart disease. The results showed that voting with decision trees showed a lower accuracy of 82%[6] by using 13 features of the UCI dataset repository.

2.1.6 K-Nearest Neighbor's Algorithm (KNN Algorithm)

KNN algorithm is one of the most popular machine learning algorithms, KNN is often used for the data classification process. Heart disease can be predicted also by analyzing the health parameters of each patient. KNN can be used with the parameter comparison method to increase its accuracy. This study uses a UCI machine learning dataset with 13 definite parameters, in the end getting an accuracy of 86% of the 13 parameters [8].

2.2 Data Collection and Features Selection

In this study, the heart disease data source used was UCI Cleveland data sourced from the UCI machine learning repository website, this dataset is a publicly available dataset source. The data set consists of 297 agencies with 14 attributes. The dataset used in diagnosing heart disease is the Heart Disease Dataset which is a dataset of 4 different combinations of datasets, but in some sources, only the UCI Cleveland dataset is used in this study. The basis of the data is that there are 76 attributes or features, but in published research, only 14 attributes have been processed [9]. The data is contained in theDataset repository, namely the UCI Cleveland dataset. In the dataset, there are 14 attributes and 1 attribute for prediction or known as the dependent variable name, the 'Diagnosis' attribute, and the rest will be entered as input or known as the independent variable. The attribute descriptions are shown in Table 1.

Table I. The Attribute Descriptions

No	Features / Attributes	Type Value	Description Features	Value
1	Sex	Discreate Variable	Male or female representative of your age	1: Male 0: Female
2	Age	Continuous Variable	Patient range shown by age	Multiple value in range 28 and 77 (age in years)
3	CP (Chest Pain) Type	Discrete Variable	Represent to chest pain type: typical angina, atypical angina, non-anginal pain, asymptomatic	0: typical angina 1: atypical angina 2: non-anginal pain 3: asymptomatic
4	Rest Blood Pressure (Trestbps)	Continuous Variable	represent the resting heart rate (in mm Hg on admission to the hospital)	Multiple continue value in mmHg
5	Serum Cholestoral (Chol)	Continuous Variable	represent the resting heart rate (in mm Hg on admission to the hospital)	Multiple Continuous value in mm/dl
6	Fasting Blood Sugar (FBS)	Discrete Variable	Represent the patient's fasting blood sugar level	0: false (FBS <120 mg/dl) 1: true (FBS > 120 mg/dl)
7	Max Heart Rate (Thalach)	Continuous Variable	represent the patient's maximum heart rate	Multiple values from 71 to 202 Low: under 50 beats/min Normal: 51 - 119 beats/min High: 120 - 180 beats/min [10][11]
8	Res Electrocardiographic (Restecg)	Discrete Variable	Represent the ECG's outcome. where each integer represents the level of pain.	0: normal 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
9	Exercise Induced (Exang)	Discrete Variable	Determine whether or not there is exercise-induced angina by representing.	1: yes 0: no
10	Oldpeak	Discrete Variable	Show how exercise-induced ST depression compares to rest	Multiple decimal number values between 0 and 6.2.
11	Slope	Discrete Variable	Describe the patient's state at the height of exercise. There are three sections in this paragraph.	0: upsloping 1: flat 2: down sloping
12	Major Vessels (Ca)	Discrete Variable	The quantity of main vessels that fluoroscopy can colour. This section displays the number of colored vessels.	number of major vessels (0-3) colored by fluoroscopy —-> (0, 1, 2, 3) value
13	Thal	Discrete Variable	Patients with chest pain or respiratory distress also need to have this parameter tested. Three different value types displaying Thallium test results are shown in this section.	0 = normal 1 = fixed defect 2 = reversible defect and the label
14	Target / condition	Discrete Variable	the data set's last column. The Class column or Label column are other names for this Target column. using the preceding 13 parameters for analysis, this column generates prediction results with two classes, i.e. class 0 and class 1. This indicates that the likelihood of avoiding developing heart disease is "0" if the class number. If the class displays the number "1," then the opposite is true, specifically with the potential for developing heart disease.	0: no disease 1: disease

In the study shows an analysis of some ai algorithm thatis, machine learning algorithm. Algorithm-algorithms used in this paper include support vector machine (SVM), naive bayes algorithm (NB algorithm), logistic regression algorithm and artificial neural networks, the decision tree algorithm (dc algorithm), and k-nearest neighbour's algorithm (KNN algorithm) that may help medical analysis in making a precise diagnosis of heart disease. This methodology is a process that includes the step-the converted data step that has been given into the new data used as user knowledge. The proposed methodology (Figure 1.) includes steps.

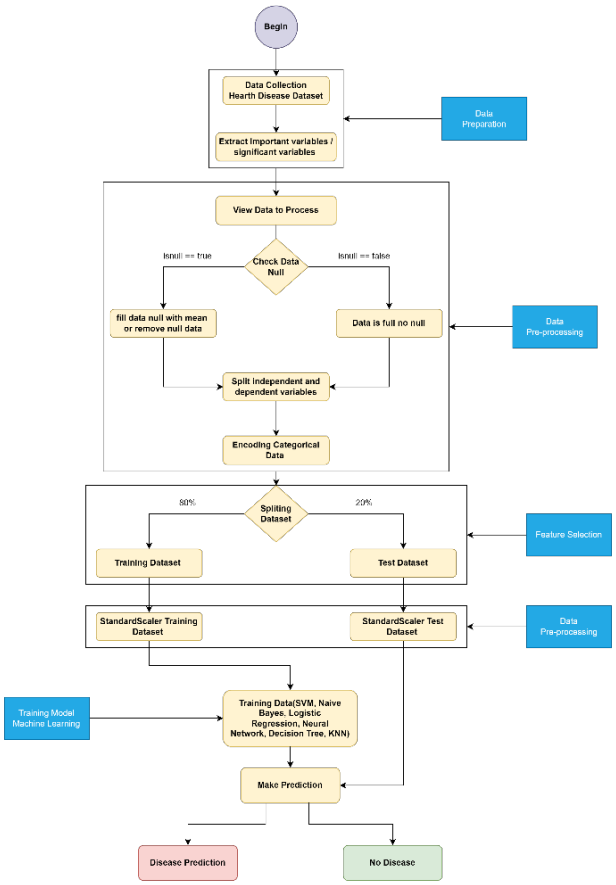

Figure 1. Proposed Models

After we have further data collection, an important data release that will then be used for the processof machine learning or known as preparation. Looking at the data chosen next, whether there isa missing value, if there is a missing value, then we can either eliminate the missing value by erasing or known by the cleaning data process, then we divide the independent variable and variable variables - Next turn the complex data into a number so the computer can read using encoding pre-processing data this process is as the pre-processing data. After encode data we go for dataset training and the division's dataset test have 80:20, 80% for training and 20% for dataset test, this process known as the feature selection process. We use the standard scaler to classify or diffuse the range value. After the pre-processing data through the selection features data then, further data training uses the model classifier algorithm.

Classifier used to classify data that had been processed earlier. The classifier algorithm includes the support vector machine (SVM), naive bayes algorithm (NB algorithm), logistic regression algorithm and artificial neural networks, decision tree algorithm (DC algorithm), and k-nearestneighbour's algorithm (KNN algorithm). And finally, from the top model, we should evaluate the model we made based on the accuracy of its prediction and its performance algorithm using the performance metrics or metrics indicators. This model can be said to be effective in predicting heart diseaseby using a different rating algorithm. In this model we use multiple compounds like chest pain type, rest blood pressure, Chol and others to predict.

III. RESULTS AND DISCUSSION

The results from the implementation of the machine learning classifier algorithm, Support Vector Machine (SVM), Naive Bayes Algorithm (NB Algorithm), Logistic Regression Algorithm, Decision Tree Algorithm (DC Algorithm), K-Nearest Neighbor's Algorithm (KNN Algorithm), and artificial neural networks are demonstrated in this section. The metrics Accuracy score, Precision (P), Recall/Sensitivity (R), Specificity (S), and F1 Score are being used to analyze the algorithm's performance. The precision parameter is the ratio of the positive correct predictions (TP) to the overall positive results which the model has predicted (Equation 1). Recall is calculated as the proportion of the correct positive prediction (TP) to the total positive data (Equation 2)[12]. Specificity Predicting the awful is accurate as compared to the whole bad data (Equation 3). The weighted comparison of the average precision and recall is defined by the F1 Score (Equation 4). Accuracy The correlation between the prediction value and the total quantity of data, or the model's SVM classifiers (Equation 5)[13]. Table 3 contains all of the results of the performance test of the machine learning algorithm.

$$\text{Precision} = \frac{(TP)}{(TP + FP)} \quad (1)$$

$$\text{Recall} = \frac{(TP)}{(TP + FN)} \quad (2)$$

$$\text{Specify} = \frac{(TN)}{(TN + FP)} \quad (3)$$

$$\text{F1 Score} = \frac{2 * (\text{Recall} * \text{Precission})}{(\text{Recall} + \text{Precission})} \quad (4)$$

$$\text{Accuracy} = \frac{(TP + TN)}{(TP + FP + FN + TN)} \quad (5)$$

- True Positive (TP): The model predicts the data is positive and it is indeed positive.
- True Negative (TN): The model predicts the data is negative and it is indeed negative.
- False Positive (FP): (Problem Type l) The model predicts the data is Positive, but the data is Negative.
- False Negative (FN): (Type 2 error, this typeof error is considered very dangerous) The model predicts the data is Negative, but the actual data is Positive.

To get the data above we use the confusion matrix. The Confusion Matrix is a performance measurement for machine learning classification problems. Confusion Matrix is a table with 4 different combinations of predicted values and actual values[14]. The terms of the confusion matrix are those already mentioned above such as TP, FP, TN, and FN shown in table2 and figure 2.

Table 2. Confusion Matrix.

		Predicted	
		has heart disease (Positive)	no heart disease (Negative)
Actual	has heart disease (Positive)	TP	FN
	no heart disease (Negative)	FP	TN

Figure 2 to figure 7 Confusion matrix obtained for Six classification methods with 13 attributes.

Confusion matrix for Support Vector Machine (SVM):

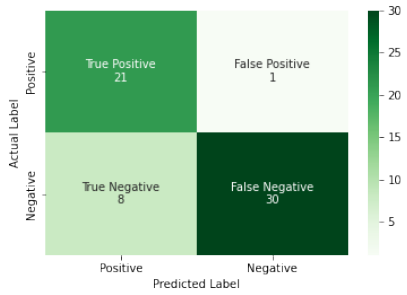

Figure 2. Confusion Matrix SVM

Based on an image above a comparison of predictions with the original (TP) + (FN) there are 51 data and (TN) + (FP) there are 9 incorrect data.

Confusion matrix for Naive Bayes Algorithm (NB):

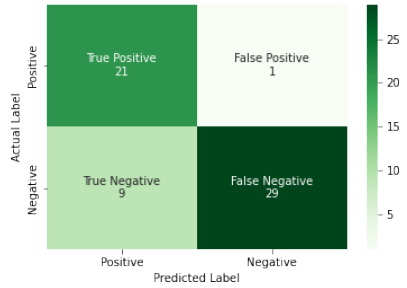

Figure 3. Confusion Matrix NB Algorithm

Based on an image above a comparison of predictions with the original (TP) + (FN) there are 50 data and (TN) + (FP) there are 10 incorrect data.

Confusion matrix for Logistic Regression Algorithm:

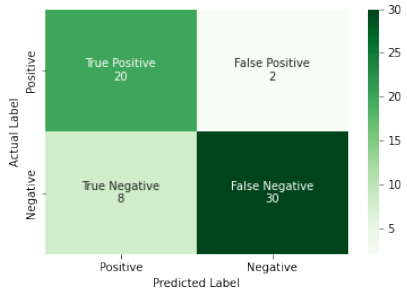

Figure 4. Confusion Matrix Logistic Regression

Based on an image above a comparison of predictions with the original (TP) + (FN) there are 50 data and (TN) + (FP) there are 10 incorrect data.

Confusion matrix for Neural Network Algorithm:

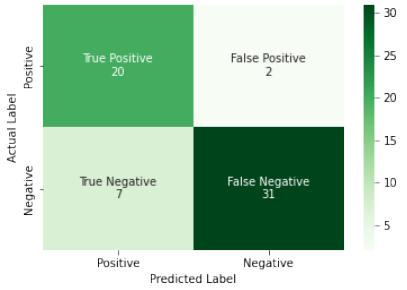

Figure 5. Confusion Matrix Neural Network Algorithm

Based on an image above a comparison of predictions with the original (TP) + (FN) there are 51 data and (TN) + (FP) there are 9 incorrect data.

Confusion matrix for Decision Tree Algorithm:

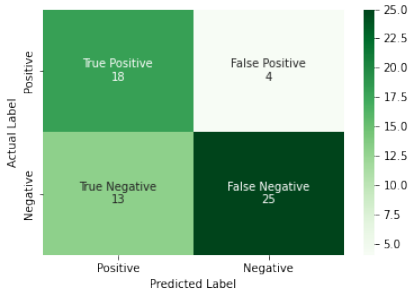

Figure 6. Confusion Matrix Decision Tree Algorithm

Based on an image above a comparison of predictions with the original (TP) + (FN) there are 43 data and (TN) + (FP) there are 17 incorrect data.

Confusion matrix for KNN Algorithm:

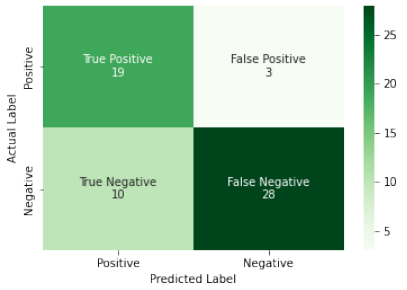

Figure 7. Confusion Matrix KNN Algorithm

Based on an image above a comparison of predictions with the original (TP) + (FN) there are 47 data and (TN) + (FP) there are 13 incorrect data.

Table 3. Analysis Performance Machine Learning Algorithm

Algorithm	Accuracy	Precision	Recall	Specifity	F1 Score
SVM	85%	0.97	0.79	0.79	0.87
Naïve Bayes	83.33%	0.96	0.76	0.76	0.85
Logistic Regression	83.33%	0.94	0.79	0.79	0.86
Neural Network	80%	0.91	0.76	0.76	0.83
Decision Tree	70%	0.86	0.63	0.63	0.72
KNN	78%	0.90	0.74	0.74	0.81

IV. CONCLUSION

In this study, we introduce the heart disease prediction system with different type of classifier techniques such as SVM, Naive Bayes, Logistic regression, Decision tree and KNN for the prediction of heart disease. The history of the medical person or patient that contains a collection of data that leads to heart disease, the data collection of the medical history includes CP (Chest Pain), Thalach (Max Hearth Rate), and others. In this study, we also compare several machine learning algorithms including SVM, Naïve Bayes, Logistic Regression, Neural Network, Decision Tree, and KNN to predict heart disease using the UCI Cleveland machine learning repository dataset[1][15][16]. The results we get from this research study show that the SVM algorithm is the algorithm with the greatest precision and highest accuracy with a score of 85% in predicting heart disease. In the future we will develop the algorithm to be more efficient and have a higher accuracy rate for predicting heart disease from previous research or studies.

REFERENCES

[1]C. S. Dangare, S. S. Apte, and M. E. Student, "Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques," International Journal of Computer Applications, vol. 47, no. 10, pp. 975–888, 2012.

[2]Chandra Reddy, N. S., Shue Nee, S., Zhi Min, L., & Xin Ying, C. (2019). Classification and Feature Selection Approaches by Machine Learning Techniques: Heart Disease Prediction. International Journal of Innovative Computing, 9(1), 39–46. *https://doi.org/10.11113/ijic.v9n1.210*

[3]Chandra Reddy, N. S., Shue Nee, S., Zhi Min, L., & Xin Ying, C. (2019). Classification and Feature Selection Approaches by Machine Learning Techniques: Heart Disease Prediction. International Journal of Innovative Computing, 9(1), 39–46. *https://doi.org/10.11113/ijic.v9n1.210*

[4]Nishadi, A. S. T. (n.d.). International Journal of Advanced Research and Publications Predicting Heart Diseases In Logistic Regression Of Machine Learning Algorithms By Python Jupyterlab. Retrieved March 22, 2022, from https://www.kaggle.com

[5]Mukherji, D., Padalia, N., & Naidu, A. (2013). A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL). International Journal of Computer Applications, 68(16), 975–8887.

[6]Dikananda, A. R., Ali, I., Sai, V., Reddy, K., Meghana, P., Reddy, S., & Rao, A. (2022). Prediction on Cardiovascular disease using Decision tree and Naïve Bayes classifiers Genre e-sport gaming tournament classification using machine learning technique based on decision tree, Naïve Bayes, and random forest algorithm Prediction on Cardiovascular disease using Decision tree and Naïve Bayes classifiers. Journal of Physics: Conference Series, 2161, 12015. *https://doi.org/10.1088/1742-6596/2161/1/012015*

[7]G, A., Ganesh, B., Ganesh, A., Srinivas, C., Dhanraj, & Mensinkal, K. (2022). Logistic regression technique for prediction of cardiovascular disease. Global Transitions Proceedings, 3(1), 127–130. *https://doi.org/10.1016/J.GLTP.2022.04.008*

[8]Kumar, Rs., & Fatima, D. (n.d.). Heart Disease Prediction Using Extended KNN(E-KNN). International Journal of Advanced Trends in Computer Science and Engineering, 9(5), 8799–8803. *https://doi.org/10.30534/ijatcse/2020/272952020*

[9]Rjeily, C. B., Badr, G., Hajjam, A., Andrès, E., Hajjarm, A., Hassani, E., & Andres, E. (2019). Medical data mining for heart diseases and the future of sequential mining in medical field. Springer, 149, 71–99. https://doi.org/10.1007/978-3-319-94030-4 4

[10]Ullah, F., Abdullah, A., Kaiwartya, O., systems, Y. C.-J. of medical,& 2017, undefined. (2017). Tra-Py-MAC: Traffic priority aware medium access control protocol for wireless body area network. Springer, 41(6), 93. https://doi.org/10.1007/s10916-017-0739-y

[11]Ullah, F., Abdullah, A. H., Kaiwartya, O., & Arshad, M. M. (2017). Traffic Priority-Aware Adaptive Slot Allocation for Medium Access Control Protocol in Wireless Body Area Network. Computers 2017, Vol. 6, Page 9, 6(1), 9. *https://doi.org/10.3390/COMPUTERS6010009*

[12]The relationship between recall and precision - buckland - 1994 ... (n.d.). Retrieved May 29, 2022, from *https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-4571%28199401%2945%3A1%3C12%3A%3AAID-ASI2%3E3.0.CO%3B2-L*

[13]Jiao, Y., & Du, P. (2016, December 23). Performance measures in evaluating machine learning based bioinformatics predictors for classifications - quantitative biology. SpringerLink. Retrieved May 29, 2022, from *https://link.springer.com/article/10.1007/s40484-016-0081-2*

[14]Xu, J., Zhang, Y., & Miao, D. (2019, July 11). *Three-way confusion matrix for classification: A measure driven view*. Information Sciences. Retrieved May 29, 2022, from *https://www.sciencedirect.com/science/article/pii/S00200255193060*

[15]Rajdhan, A., Agarwal, A., & Sai, M. (n.d.). Heart Disease Prediction using Machine Learning. IJERT Journal International Journal of Engineering Re-

search & Technology. Retrieved March 22, 2022, from *www.ijert.org*

[16]Subhadra, K., Innovative, B. V.-I. J. of, & 2019, undefined. (n.d.). Neural network based intelligent system for predicting heart disease. Researchgate.Net, 2278–3075. Retrieved March 15, 2022, from *https://www.researchgate.net/profile/Vikas-Boddu/publication/332035370_Neural_network_based_intelligent_system_for_predicting_heart_disease/links/601f7b36299bf1cc26ac05de/Neural-network-based-intelligent-system-for-predicting-heart-disease.*