TOWARDS COMPARING MACHINE LEARNING MODELS TO FORESEE THE STAGES FOR HEART DISEASE

Khalid Amen[1], Mohamed Zohdy[1], and Mohammed Mahmoud[2]

[1]Department of Electrical and Computer Engineering,
Oakland University, Rochester, MI, USA
[2]Department of Computer Science and Engineering,
Oakland University, Rochester, MI, USA

ABSTRACT

With the increase in heart disease rates at advanced ages, we need to put a high quality algorithm in place to be able to predict the presence of heart disease at an early stage and thus, prevent it. Previous Machine Learning approaches were used to predict whether patients have heart disease. The purpose of this work is to compare two more algorithms (NB, KNN) to our previous work [1] to predict the five stages of heart disease starting from no disease, stage 1, stage 2, stage 3 and advanced condition, or severe heart disease. We found that the LR algorithm performs better compared to the other two algorithms. The experiment results show that LR performs the best with an accuracy of 82%, followed by NB with an accuracy of 79% when all three classifiers are compared and evaluated for performance based on accuracy, precision, recall and F measure.

KEYWORDS

Machine Learning (ML), Logistic Regression (LR), Naïve Bayes (NB), K-Nearest Neighbors (KNN).

1. INTRODUCTION

1.1. Machine Learning

Machine Learning (ML) is a branch of artificial intelligence (AI) that is increasingly utilized within the field of heart disease medicine. It is essentially how computers make sense of data and decide, or classify, a task with or without human supervision. The conceptual framework of ML is based on models that receive input data (e.g., images or text) and through a combination of mathematical optimization and statistical analysis predict outcomes (e.g., favorable, unfavorable, or neutral) [2]. We have used five ML algorithms in our previous work to predict multiple stage heart disease. The first one is SVM, it can recognize non-linear patterns for use in facial recognition, handwriting interpretation or detection of fraudulent credit card transactions. So-called boosting algorithms used for prediction and classification have been applied to the identification and processing of spam email.

The second algorithm is Random Forest (RF), it can facilitate decisions by averaging several nodes. The third algorithm is Gradient Tree Boosting (GTB), which is a ML technique for regression and classification problem that produces a prediction model in the form of an ensemble of weak prediction models. The fourth algorithm is Extra Random Forest (ERF), it is an

David C. Wyld et al. (Eds): ACSTY, AIBD, MLSC, CCCIOT, NATP - 2021
pp. 33-44, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.110304

ensemble learning technique which aggregates the results of multiple de-correlated decision trees collected in a "forest" to output its classification result. The fifth algorithm is Logistic Regression (LR), it is a classification algorithm, that is used where the response variable is categorical [3]. The idea of LR is to find a relationship between features and probability of particular outcome. We have previously described technical details of each of these algorithms, and found out that LR is the best algorithm in terms of accuracy, precision, recall and F measure to predict multiple stages of heart disease. We have also used several tools and methods such as Python libraries, graphs, and Pseudocodes to test the performance of each algorithm. In this paper, we are going to implement two more algorithms, Naïve Bayes (NB) and K-Nearest Neighbors (KNN) to predict multiple stage heart disease and compare with our winning algorithm, LR [1].

Machine Learning in healthcare is becoming more widely used and is helping patients and clinicians in many different ways. ML can play an essential role in predicting presence/absence of locomotor disorders, heart diseases and more. ML, when applied to health care, is capable of early detection of disease, which can aid to provide early medical intervention. Heart disease predication has been a very hot topic for ML, for example, the analysis of heart disease has become vital in health care sectors. The success of ML in the medical industry is its capability in analyzing the huge amount of data gathered by the health sector and its effectiveness in decision-making [4] [5].

As we have used in our previous work to conduct this prediction, a Jupyter notebook was constructed in Python using the publicly available Cleveland dataset for heart disease, which has over 300 unique instances with 76 total attributes. From these 76 attributes, only 14 of them are commonly used for research to this date. In addition, the libraries and coding packages used in this analysis are: SciPy, Python, NumPy, IPython, Matplotlib, Pandas, ScikitLearn and Scikit-Image.

1.2. Heart Disease

Heart disease is the major cause of morbidity and mortality globally and accounts for more deaths annually than any other cause. According to the World Health Organization (WHO), an estimated 17.9 million people died from heart disease in 2016, representing 31% of all global deaths. Over three-quarters of these deaths took place in low and middle-income countries [6] [14].

Heart disease is the number one killer of both men and women. Heart disease can happen at any age, but the risk increases as people get older. Children of parents with heart disease are more likely to develop heart disease themselves. African-Americans have more severe instances of high blood pressure than Caucasians and a higher risk of heart disease. Heart disease risk is also higher among Mexican-Americans, American Indians, native Hawaiians and some Asian-Americans. This is partly due to higher rates of obesity and diabetes. Genetic factors likely play some role in high blood pressure, heart disease and other related conditions [7] [8].

The silver lining is that heart attacks are highly preventable and simple lifestyle modifications (such as reducing alcohol and tobacco use, eating healthily and exercising) coupled with early treatment greatly improves its prognosis. It is, however, difficult to identify high-risk patients because of the multi-factorial nature of several contributory risk factors such as diabetes, high blood pressure, high cholesterol, et cetera. This is where ML and data mining come to the rescue [12].

2. RELATED WORK

In our previous work [1] and research, we have implemented five algorithms SVM, RF, GTB, LR and ERF to predict multiple stage heart disease using the Cleveland dataset. We concluded that the LR algorithm performed better in terms of accuracy, precision, recall and F measure as shown in table 1:

Table 1. Algorithms comparison

Algorithm	Accuracy	Precision	Recall	F Measure
SVM	80%	91%	78%	84%
LR	82%	91%	80%	85%
RF	77%	89%	76%	82%
GTB	74%	80%	76%	78%
ERF	79%	89%	78%	83%

Additionally, many researchers have completed a lot of work on data analysis and survivability analysis through ML and Data Mining (DM) approaches [7].

In [8], [10] the author applied Decision Tree (DT), LL, NB, SVM, KNN, PCA, ICA classifier respectively to analyze kidney disease data. Early detection and treatment of the diseases prevents it from getting to the worst stage, making it not only difficult to cure, but also impossible to provide treatment. Breast cancer affects many women, so researchers work on different classifiers such that DT, SMO, BF Tree and IBK help to analyze the breast cancer data and examine the performance of the related techniques in order to accurately predict breast cancer using DT and Weka software. RBF Network, Rep Tree and Simple Logistic DM techniques are used to predict and resolve the survivability of breast cancer patient. Simple Logistic is used for dimension reduction and proposed RBF Network and Rep Tree model used for fast diagnosis of the other diseases [19] [20] [21].

3. BACKGROUND OF CLEVELAND DATASET

In our previous work, we have used the Cleveland dataset for multiple stage heart disease prediction. We plan to use the same dataset to compare the results of the new algorithms we are adding with LR. Experiments with the dataset have concentered on 14 attributes that were used. The list is in Table 2.

Table 2. Cleveland dataset attributes

Name	Type	Description
Age	Continuous	Age in years
Sex	Discrete	0 = female 1 = male
Cp	Discrete	Chest pain: 1 = typical angina, 2 = atypical angina, 3 = non-angina pain, 4 = asymptom
Trestbps	Continuous	Resting blood pressure (in mm Hg)
Chol	Continuous	Serum cholesterol in mg/dl
Fbs	Discrete	Fasting blood sugar>120 mg/dl: 1 = true 0 = false

Restecg	Discrete	Resting Electrocardiograph
Thalach	Continuous	Exercise Max Heart Rate Achieved
Exang	Discrete	Exercise Induced Angina: 1=yes 0=no
Old peak ST	Continuous	Depression induced by exercise relative to rest
Slope	Discrete	The slope of the peak exercise segment: 1=up sloping 2=flat 3=down sloping
Ca	Continuous	Number of major vessels colored by fluoroscopy that range between 0 and 3
Tha	Discrete	3=normal 6=fixed defect 7=reversible defect
Class	Discrete	Diagnosis classes: 0=No Presence 1=Least likely to have heart disease 2=>1 3=>2 4=More likely have heart disease

The database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. In particular, the Cleveland database is the only one that has been used by DL researchers to date. The "num" field in the figure refers to the presence of heart disease in the patient. It is integer valued from zero (no presence) to four. Experiments with the Cleveland database have concentrated on attempting to distinguish presence (values 1,2,3,4) from absence (value 0) [16].

1. #3 (age)
2. #4 (sex)
3. #9 (cp)
4. #10 (trestbps)
5. #12 (chol)
6. #16 (fbs)
7. #19 (restecg)
8. #32 (thalach)
9. #38 (exang)
10. #40 (oldpeak)
11. #41 (slope)
12. #44 (ca)
13. #51 (thal)
14. #58 (num) (the predicted attribute)

4. BACKGROUND ON LOGISTICS REGRESSION ALGORITHM

Logistic regression is a type of regression analysis in statistics used for prediction of outcome of a categorical dependent variable from a set of predictor or independent variables. In LR the dependent variable is always binary. LR is mainly used for prediction and calculating the probability of success [17] [18]. An LR model specifies that an appropriate function of the fitted probability of the event is a linear function of the observed values of the available explanatory variables. The major advantage of this approach is that it can produce a simple probabilistic formula of classification. The weaknesses are that LR cannot properly deal with the problems of non-linear and interactive effects of explanatory variables. LR is a regression method for

predicting a dichotomous dependent variable. In producing the LR equation, the maximum likelihood ratio was used to determine the statistical significance of the variables. LR is useful for situations in which you want to be able to predict the presence or absence of a characteristic or outcome based on values of a set of predictor variables. It is similar to a LR model but is suited to models where the dependent variable is dichotomous [22]. The LR model for p independent variables can be written as: [1]

$$H(Y=1) = \frac{1}{1+e^{-(\beta 0+\beta 1x1+\beta 2x2+\cdots+\beta pxp)}} \quad (1)$$

where $P(Y=1)$ is the probability of the presence of CAD and $\beta 0,\ \beta 1,\ \beta 2, \ldots, \beta p$ are regression coefficients. There is a linear model hidden within the logistic regression model. The natural logarithm of the ratio of $P(Y=1)$ to $1-P(Y=1)$ gives a linear model in Xi: [2]

$$g(x) = \ln\left(\frac{P(Y=1)}{1-P(Y=1)}\right) \quad (2)$$

$$= \beta 0 + \beta 1x1 + \beta 2x2 + \cdots + \beta pxp$$

The $g(x)$, has many of the desirable properties of the LR model. The independent variables can be a combination of continuous and categorical variables.

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of the target or dependent variable is dichotomous, which means there will only be two possible classes [22] [23]. In simpler words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).

Mathematically, a LR model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, diabetes prediction, cancer detection etc. [22] [23].

4.1. Type of Logistics Regressions

Logistic Regression means binary LR having binary target variables, but there can be two more categories of target variables that can be predicted by it. Based on those number of categories, LR can be divided into following types [15] [24]:

- Binary or Binomial – In such a classification, a dependent variable will have only two possible types either 1 and 0. For example, these variables may represent success or failure, yes or no, win or loss etc.
- Multinomial – In such a classification, dependent variable can have three or more possible unordered types, or the types having no quantitative significance. As an example, these variables may represent “Type A” or “Type B” or “Type C”.
- Ordinal – In such a classification, dependent variables can have three or more possible ordered types, or the types having a quantitative significance. For example, these variables may represent “poor”, “good”, “very good” or “excellent” and each category can have scores such as 0, 1, 2 or 3.

5. METHODOLOGY

The proposed methodology using two classification techniques; NB and KNN. We use these two classifications to predict heart disease as the proposed methodology shown in Fig 2. These classifiers are used to improve the prediction. We applied the classifiers in Fig 5 to heart disease data that comes from the Cleveland dataset to predict in which of five stages a patient has heart problems. The performance of these classifiers are to evaluate on the bases of accuracy, precision recall and F measure, then we compare the results of these classifications with LR in terms of accuracy, precision recall and F measure.

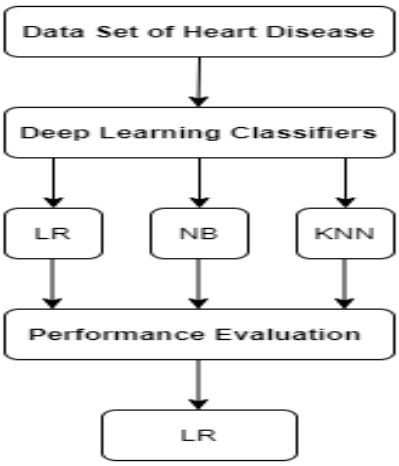

Figure 1: Proposed Methodology

The dataset of heart is taken from Machine LearningRepository UC Irvine, the classifier taking it as input for disease prediction. These classifiers are implemented in Python language. Python is a powerful interpreter language and a reliable platform for research [25]. The accuracy of prediction increased by comparing the results of these five classifiers using evaluation parameters. The experimental result describes which classifier is best between them.

5.1. Evaluation Parameters

- Accuracy is defined as the number of accurately classified instances divided by the total number of instances in the dataset as in (3).

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (3)$$

- Precision is the average probability of relevant retrieval as described in (4).

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4)$$

- The recall is defined as the average probability of complete retrieval as defined in (5).

$$\text{Recall} = \frac{TP}{TP + FN} \quad (5)$$

- F- Measure is calculated by using both precision and recall as shown in (6).

$$\text{F Measure } = \frac{2 * (Precision * Recall)}{Precision + Recall} \quad (6)$$

Some evaluation parameters in DM are accuracy, precision, recall and F measure. Where True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) [14].

Where all evaluation parameters accuracy, precision, recall and F measure are calculated from dataset when splitting the dataset into training data and test data. The Pseudocodes for the evaluation parameters are as follows:

```
Def evaluationParameters(X_train, y_train, X_test, y_test):
X_train ← fit_transform(X_train)
Classifier ← sklearn()
y_pred ← classifier.predict(X_test)
cm_test ← confusion_matrix(y_pred, y_test)
y_pred_train ← classifier.predict(X_train)
    cm_train ← confusion_matrix(y_pred_train,    y_train)
    training_accuracy=(cm_train[0][0]+cm_train[1][1])/  len(y_train)
    test_accuracy=(cm_test[0][0]+cm_test[1][1])/len(y_test)
training_percision = cm_train[0][0]/(cm_train[0][0] + cm_train[1][0])
test_percision=cm_test[0][0]/(cm_test[0][0]+cm_test[1][0])
training_recall = cm_train[0][0]/(cm_train[0][0] + cm_train[0][1])
test_recall = cm_test[0][0]/(cm_test[0][0] + cm_test[0][1])
training_f_measure ← (2 * training_percision * training_recall)/(training_percision +
training_recall))
test_f_measure ← (2 * test_percision * test_recall)/(test_percision + test_recall))

return (training_accuracy, test_accuracy, training_percision, test_percision, training_recall,
test_recall,  training_f_measure,  training_f_measure)
```

6. DATASET

To perform the research, the heart disease dataset is used. This heart disease dataset contains 14 attributes and 303 instances. This dataset is taken from UCL repository. It's an online repository that contains 412 diverse datasets. UCI provides data for ML to perform analysis in a different direction. The UCI database is known for its extensiveness in data, its completeness and accuracy [23].

6.1. Machine Learning Classifiers:

In this continuous research, two additional classification methods are implemented in python using the pandas and keras libraries. These models are used to improve prediction. These classifiers are compared with LR to find out which of the five stages best predicts the chance of heart disease in patients. In the next section, we briefly describe these classification techniques/classifiers.

1) Naïve Bayes (NB): are probabilistic classifiers based on Bayes theorem with naïve independence assumption between the predictors or features. NB classifier assumes, that the existence of a particular feature is not related to the existence of any other feature in a class [26].

For example, apples are considered a fruit, if it is red and round. Even if these are features related to each other or depend upon the presence of other features. NB classifier consider all these features to contribute independently to probability identifying that the fruit is an apple.

NB model is easy to construct and particularly valuable for large data sets and a Bayes theorem gives a way to calculate posterior probability P(c|x) from likelihood (predictor probability) P(x|c), class prior probability P(c) and predictor prior probability P(x) as shown in (7)

$$p(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad (7)$$

2)K-Nearest Neighbors (KNN): is used for regression and classification problems. KNN is commonly used for classification problems [26] [27]. KNN classifier store all the existing cases then classified new cases by the majority votes of its neighbors. The case is assigned to that class, which is most common to its k nearest neighbors, measured by distance function. These distance functions are Euclidean Eu, Manhattan Ma and Minkowski Mi calculated using (8), (9) and (10) respectively.

$$Eu = \sqrt{\sum_{k=1}^{n} (pk - qk)^2} \quad (8)$$

$$Ma = \sqrt{\sum_{k=1}^{n} |pk - qk|} \quad (9)$$

$$Mi = \left(\sqrt{\sum_{k=1}^{n} |pk - qk|} \right)^{1/r} \quad (10)$$

Whereas r is the parameter, n is the number of attributes or dimensions. pk and qk are respectively, the kth element of objects p and q [28].

6.1.1. Scaling Data

To accomplish the five stages output prediction for a patient to be diagnosed with one of five stages, it is important to scale the data so the machine learning algorithms do not overfit to the wrong features. Using the MinMaxScaler() method on Python, the values are scaled per features based on the minimum and maximum between 0 and 1. This keeps the information from being lost but allows the machine learning algorithms to correctly train with the data. The training data and test data are scaled between 0 and 1 and the output data is scaled between 0 and 1 as well. Then, the scaled output value is mapped as follows in table 3:

Table 3: Five Stages

Output Value	Stage
0	No disease presented
0 < and <= 0.25	Stage1
0.25 < and <= 0.5	Stage2
0.5 < and <= 0.75	Stage3
0.75 < and <= 1	Advance disease presented

7. EXPERIMENTAL RESULT

The experiment is conducted for the prediction of heart disease stages by applying two machine learning classifiers. From the experiment results, we have identified that LR performs better compared to the other four ML classifiers in the prediction of these diseases. In this experiment, we use multiple stages of heart disease prediction to forecast the stage at which a person is determined to have heart disease. In previous works [19] [20] [21], the study used two outcome predications, either a person has the disease or not; that is represented by (0 ,1) or (true, false). The Pseudocodes for the experiment are as follows:

```
data_frame ← read_CSV_file
X ← data_frame [column: 0 - 12]
y ← data_frame [column: 13]
target ← preprocessing.scale(y)
data ← preprocessing.scale(X)
        for k ← 0 to data - 1
        if data[k] = 0 then
        data[k] ← 'no disease'
        if data[k] > 0 && data[k] <= 0.25 then
        data[k] ← 'stage1'
        if data[k] > 0.25 && data[k] <= 0.5 then
        data[k] ← 'stage2'
        if data[k] > 0.5 && data[k] <= 0.75 then
        data[k] ← 'stage3'
        else
        data[k] ← 'disease presented'

X_train,X_test,y_train,y_test←train_test_split(X, y, test_size=0.2, random_state=0)
svm(X_train, y_train, X_test, y_test)
lr(X_train, y_train, X_test, y_test)
rf(X_train, y_train, X_test, y_test)
gtb(X_train, y_train, X_test, y_test)
erf(X_train, y_train, X_test, y_test)
```

The Figures 4, 5, 6, and 7 show the performance of various evaluation parameters in the prediction of heart disease. The experimental results show the comparison of LR, NB, and KNN classifiers and evaluate the performance on the bases of accuracy, precision, recall and F measure. In all classifiers, LR still performs the best with an accuracy of 82%, followed by NB with an accuracy of 79% and KNN with 70%. So, we can conclude that LR has better performance that ERF, GTB, SVM, RF, NB, andKNN, where LR is better than that ERF, GTB, SVM, and RF from previous evaluation, and it is better that NB and KNN in this evaluation.

8. CONCLUSIONS

The importance of extracting the valuable information from raw data has very good consequences in many fields of life such as the medical area, business area, and more. In this study, we proposed a multiple stage detection model of heart disease based on three algorithms, NB and KNN in this paper and LR from our previous work or evaluation to compare which one performs better. The proposed detection model was tested on well-known Cleveland dataset in order to provide a fair benchmark against existing studies. Based on the experimental results, our proposed model was able to outperform heart disease detection methods with respect to accuracy,

precision, recall and F measure. The result reflected the highest result obtained showed that Logic Regression has a better result comparing to the other two methods or algorithms. The performance was further enhanced using feature selection techniques. The feature selection techniques helped to improve the accuracy, precision, recall, and F measure of the ensemble algorithms. The experiment results show that LR performs the best with an accuracy of 82%, followed by NB with an accuracy of 79%, and KNN with an accuracy of 70% when all three classifiers are compared and evaluated for performance based on accuracy, precision, recall, and F measure.

ACKNOWLEDGEMENTS

This paper and the research behind it would not have been possible without the grace, the bounty, and the blessing of almighty Allah (God) first and foremost and the exceptional support of my professors, Mohamed Zohdy and Mohammed Mahmoud. Their enthusiasm, knowledge and exacting attention to detail have been an inspiration and kept my work on track from my first encounter with machine learning research to the final draft of this paper.

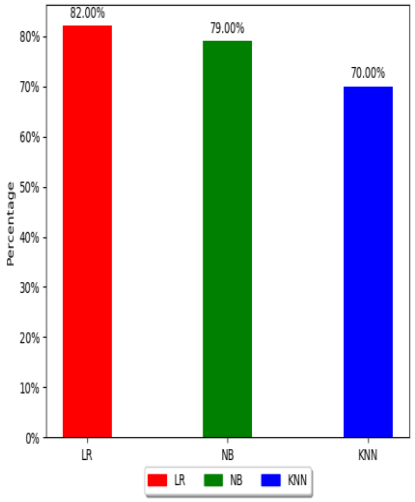

Figure 2: Heart Disease Accuracy

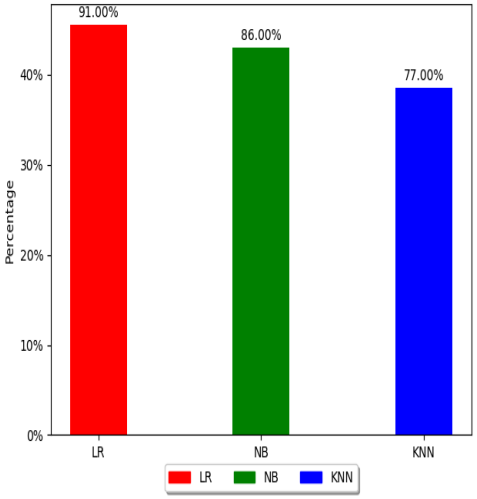

Figure 3: Heart Disease Precision

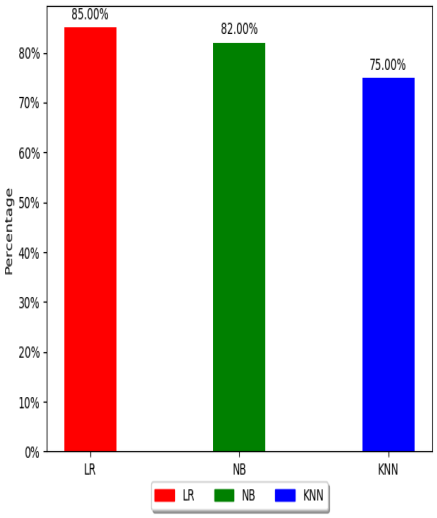

Figure 4: Heart Disease Recall

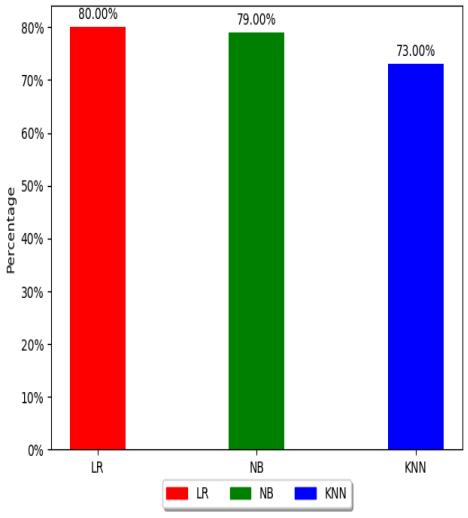

Figure 5: Heart Disease F Measure

Table 4: Five Stages

Algorithm	Accuracy	Precision	Recall	FM
LR	82%	91%	80%	85%
NB	79%	86%	79%	82%
KNN	70%	77%	73%	75%

REFERENCES

[1] K. Amen, M. Zohdy, M. Mahmoud, "Machine Learning For Multiple Stage Heart Disease Prediction". 7th International Conference on Computer Science, Engineering and Information Technology, pp. 205-223, September 26th, 2020.

[2] S. Riyaz, K. Sankhe, S. Ioannidis, K. Chowdhury, "Deep Learning Convolutional Neural Networks for Radio Identification". IEEE Communications Magazine. 56, 146–152 (2018).

[3] A Thompson, "Deep Learning on RF Data", March 29th, 2018

[4] N. Pasham, "Authenticating 'low-end wireless sensors' with deep learning + SDR", August 3rd, 2019.

[5] Z. L. Tang, S. M. Li, L. J. Yu, "Implementation of deep learning-based automatic modulation classifier on FPGA SDR platform". Electronics (Switzerland). 7 (2018), doi:10.3390/electronics7070122.

[6] Deep Learning in Healthcare, https://missinglink.ai/guides/deep-learning- healthcare/deep-learning-healthcare/

[7] Applied Deep Learning - Part 1: Artificial Neural Networks, https://towardsdatascience.com/applied-deep-learning-part-1-artificial-neural-networks-d7834f67a4f6

[8] Building A Deep Learning Model using Keras, https://towardsdatascience.com/building-a-deep-learning-model-using-keras-1548ca149d37

[9] B. Riyanto et al, "Software Architecture of Software-Defined Radio (SDR)", ITB Research Center on ICT, Institute Teknologi Banding, Indonesia.

[10] W. Liu et al., "Using deep learning and radio virtualization for efficient spectrum sharing among coexisting networks", Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST (Springer Verlag, 2019), vol. 261, pp. 165–174.

[11] Robert Sanders, "Distant galaxy sends out 15 high-energy radio bursts", https://news.berkeley.edu/2017/08/30/distant-galaxy-sends-out-15-high-energy-radio-bursts/, August 30th 2017.

[12] What Is Long-Term Care? https://www.nia.nih.gov/health/what-long-term-care

[13] Who Needs Care? https://longtermcare.acl.gov/the-basics/who-needs-care.html

[14] Bella Vista Health Center Blog, https://www.bellavistahealth.com/blog/2017/6/26/difference-between-short-term-care-and-long-term-care

[15] Emergency care, https://www.oregonlaws.org/glossary/definition/emergency_care

[16] Heart Disease in Cleveland, https://www.rpubs.com/aepoetry/log_reg_heart

[17] Mythili T., Dev Mukherji, Nikita Padalia, and Abhiram Naidu, "A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL)", International Journal of Computer Applications, 0975-8887, April, 2013.

[18] A. Khemphila, V. Boonjing, "Comparing performances of logistic regression, decision trees, and neural networks for classifying heart disease patients", IEEE CISIM, October 2010.

[19] S. A. Kaur Guneet, "Predict Chronic Kidney Disease using Data Mining Algorithms in Hadoop," international J. Adv. Comput. Eng. Netw. , vol. 5, no. 6, pp. 1–5, 2017.

[20] J. Joshi, R. Doshi, and J. Patel, "Diagnosis and Prognosis Breast Cancer Using Classification Rules," Int. J. Eng. Res. Gen. Sci., vol. 2, no. 6, pp. 315–323, 2014, [Online]. Available: www.ijergs.org.

[21] V. Chaurasia and S. Pal, "Data mining techniques: To predict and resolve breast cancer survivability," Int. J. Comput. Sci. Mob. Comput. IJCSMC, vol. 3, p. 15, 2017

[22] J. Hoffman, "Logistic regression is used for binary data", Chapter 33 - Logistic Regression, Academic Press, 2019.

[23] A. Yalcin, S. Reis, A.C. Aydinoglu, T. Yomralioglu, "A GIS-based comparative study of frequency ratio, analytical hierarchy process, bivariate statistics and logistics regression methods", Trabzon, NE Turkey, January 2011

[24] D. Speelman, "Logistic regression: A confirmatory technique for comparisons in corpus linguistics", Corpus Methods for Semantics: Quantitative studies in polysemy and synonymy, 2014.
[25] Json Brownlee, How to Develop an Extra Trees Ensemble with Python, https://machinelearningmastery.com/extra-trees-ensemble-with-python/, Apr. 2020
[26] M. Islam, J. Wu, M. Ahmadi, M. Sid-Ahmed, Maher, "Investigating the Performance of Naive-Bayes Classifiers and K- Nearest Neighbor Classifiers", 2008.
[27] A. L. Duca, C. Bacciu and A. Marchetti, "A K-nearest neighbor classifier for ship route prediction," OCEANS 2017 - Aberdeen, Aberdeen, 2017, pp. 1-6, doi: 10.1109/OCEANSE.2017.8084635.Evaluation ofk-Nearest Neighbor classifier performance for direct marketing
[28] Keller, James M., Michael R. Gray, and James A. Givens. "A fuzzy k-nearest neighbor algorithm." IEEE transactions on systems, man, and cybernetics 4 (1985): 580-585.

AUTHORS

Khalid Amen is a System Engineering and Computer Science PhD student in the Electrical and Computer Engineering department, Oakland University, Rochester, MI, USA.

Dr. Mohammed Zohdy is a professor in the Electrical and Computer Engineering department, Oakland University, Rochester, MI, USA.

Dr. Mohammed Mahmoud is a professor in the Computer Science and Engineering department, Oakland University, Rochester, MI, USA.