Breast Cancer Risk Prediction based on Six Machine Learning Algorithms

Md.Razu Ahmed*
Department of Software Engineering
Daffodil International University
Dhaka, Bangladesh
razu35-1072@diu.edu.bd

Md. Asraf Ali
Department of Software Engineering
Daffodil International University
Dhaka, Bangladesh
asraf.swe@diu.edu.bd

Joy Roy
Department of Software Engineering
Daffodil International University
Dhaka, Bangladesh
joy35-1706@diu.edu.bd

Shakil Ahmed
Department of Electrical and Mechanical Engineering
Massey University
Auckland, New Zealand
s.ahmed1@massey.ac.nz

N. Ahmed
Big Data Lab, SNCC
Massey University
Auckland, New Zealand
0000-0001-5663-0042

2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE) | 978-1-6654-1974-1/20/$31.00 ©2020 IEEE | DOI: 10.1109/CSDE50874.2020.9411572

***Abstract*—Breast Cancer is the second most important cause of death among women. As per the clinical expert, breast cancer is one of prominent cancers after lung cancer. However, early detection of this type of cancer in its initial stage helps to save lifes and increases lifespan. The survival chance of a patient can increase if there is a classifier t hat h elps w ith a quick prediction of breast cancer. Therefore, a smart framework is required that can effectively detect and predict with high accuracy early stage of breast cancer. In this article, six machine learning classification a lgorithms, n amely L ogistic Regression (LR), K-Nearest Neighbours (kNN), Decision Tree (DT), Support Vector Machine (SVM), Naive Bayes (NB), and Random Forest (RF) are implemented in order to evaluate the performance and the prediction power of the model. The main target of this work is to compare these algorithm performances using the Wisconsin Breast Cancer (original) dataset. The number of performance metrics such as accuracy, precision, recall, f-1 score, and specificity a re t aken i nto c onsideration O ur analysis of the results shows that the Support Vector Machine achieved the highest accuracy of 97.07% with the least error rate and Naive Bayes gives the lowest accuracy of 96%. All these experiments were carried out using SciKit.**

***Index Terms*—Breast Cancer, Machine Learning, Classification, Supervised learning, Computational Intelligence.**

I. Introduction

Chronic diseases are the leading causes of death and disability worldwide. Breast cancer is the second most dangerous disease for women around the world [1]. It is estimated that world-wide over 50,8000 women died of breast cancer in 2011 [2]. According to World Health Organization (WHO), more than one million women are diagnosed with breast cancer every year [3]. There are several factors that can influence the chances to get breast cancer such as genetics, family history, obesity, and lifestyle. In the early stages of cancer, women can often survive by carrying out a mastectomy. The treatment, diagnosis and surgery costs can be very high [4][5]. According to the American cancer society, there are different types of breast cancer and the most prominent types are ductal carcinoma in situ, invasive ductal carcinoma, and invasive lobular carcinoma [6]. The type of breast cancer is determined by the specific affected cells in the breast. Among the worse breast cancer is 'Carcinomas'. In addition, there are other less common types of breast cancer such as sarcomas, phyllodes, Paget disease, and angio-sarcomas which affects the connective tissues or cells of the muscles. For prediction, classification, and prognosis of the Breast Cancer, Machine Learning techniques can play an important role [7][8][9][10][11]. Using machine learning, practitioners are able to distinguish between benign and malignant tumours and are able to predict more accurately the classification of the breast cancer.

This is helpful for clinicians and can assist on the prescription of the most appropriate treatment. Machine learning can improve the detection and prediction of the disease [12]. There are many medical problems that can be solved by using machine learning techniques. Moreover, these technique can reduce the diagnosis cost significantly [13]. The main aspect of this study is to use machine learning to predict the breast cancer classification more efficiently. We used different machine learning techniques such as Logistics Regression (LR), Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbour (KNN) and Naïve Bayes (NB) The classification of breast cancer and the performance of these techniques were estimated using metrics such as accuracy, precision, recall, and f-1 score. Moreover, the performance was compared using the receiver operative characteristic (ROC) diagram to show the best possible tech-

Identify applicable funding agency here. If none, delete this.

nique for this purpose [14].
The remaining paper is organised as follows: section 2 presents the methods on breast cancer prediction. Section 3 describes the results and discussion including analysis of the results and performance evaluation. Finally, section 4 presents the conclusion of the study.

II. Method

A. Data Collection

In this study, we used the Wisconsin Breast Cancer data (Original) provided by the University of California, Irvine (also known as UCI Machine Learning Repository). In addition, this dataset is originally from the University of Wisconsin Hospitals Madison, Wisconsin, USA [15].This dataset contains 699 breast cancer patients' records whereas 458 (65.5%) samples are Benign and 241 (34.5%) samples are Malignant. This dataset contains 10 features and 2 target classes.

TABLE I
FEATURES DESCRIPTION ABOUT DATASET

No	Parameters	Information factor	Description
1	Id	Numerical	(1-10)
2	Clump Thickness	Numerical	(1-10)
3	Uniformity of Cell size	Numerical	(1-10)
4	Uniformity of Cell shape	Numerical	(1-10)
5	Marginal Adhesion	Numerical	(1-10)
6	Single Epithelial Cell size	Numerical	(1-10)
7	Bare Nuclei	Numerical	(1-10)
8	Bland Chromatin	Numerical	(1-10)
9	Normal Nucleoli	Numerical	(1-10)
10	Mitoses	Numerical	(1-10)
11	Class	Benign or Malignant	(2-Benign 4-Malignant)

B. Overview of the Classification Algorithms

1) Logistic Regression (LR): Logistic Regression (LR) was mostly used in biological research and applications in the early 20th century [16]. Logistic Regression (LR) is one of the most used machine learning algorithms where the features are categorical. Recently, LR is a popular method for binary classification problems. Logistic Regression computes the relationship between the feature variables by assessing probabilities (p) using an underlying logistic function. The logistic function is given by:

$$f(x) = \frac{C}{1 + e^{-}s(x - xi)}$$

where C= Curve's Maximum value, k= Steepness of the curve and xi= x value of Sigmoid's midpoint. A standard logistic function is called Sigmoid's function (s=1, xi=0,C=1)

2) Decision Tree (DT): Decision Tree (DT) is one of the well-known direct learning-based classification algorithms. DT can be utilized for tackling regression and classification problems. DT is also a classification based technique that breaks a dataset into smaller subsets, creating branches as the decision tree grows [18]. There are many types of Decision Trees like Classification and Regression Trees (CART), Iterative Dichotomise 3 (ID3), C4.5, etc. C4.5 is the upgrade version of ID3.

3) Random Forest (RF): Leo Breiman first introduced Random Forest (RF) which is one of the popular algorithm in machine learning [17]. RF works well to solve many clinical and biological problems. It solves both classification and regression problems in health care services. It creates a forest of decision trees. Each decision tree is trained independently with sampling from the original dataset using a bagging procedure.

4) Support Vector Machine (SVM): Support Vector Machine (SVM) is a supervised learning algorithm. SVM clustering strategy endeavours to pass a linearly separable hyperplane to divide the dataset into classes [19][20]. The main purpose of this algorithm is to determine the optimal hyperplane $f(i, j) = a.b + c$ which separates into two classes where input is $b \in R^p$ and the label is $u \in \{-1. + 1\}$. SVM learns through the constrained optimization problem of the following equation:

$$min\ a^s a + P \sum_{n=1}^{m} \theta_n$$

where $a^s a$ is the manhattan normalization, θ is the cost function, P is the penalty parameter.

5) K Nearest Neighbors (KNN): K Nearest Neighbours (KNN) is one of the most basic instance-based classification algorithms in machine learning. KNN works on the concept that samples are close enough to fit into the same class [21]. A KNN categorizes a sample to the class that is the closest among K neighbors. K is a constraint for fine tuning the classification algorithms [22].

6) Naïve Bayes (NB): Naive Bayes (NB) is one of the simplest, most effective and commonly used, machine learning techniques. It is a probabilistic classifier that classifies using the hypothesis of conditional independence with the pretrained datasets [23]. Henceforth, Naive Bayes classifiers are used for finding the traditional solution of classification problems, such as spam detection, and also well fit for medical problems. According to the Bayes theory

$$N(b|A) = \frac{N(A|b)N(b)}{N(A)}$$

Where b is the class variable and A is the parameters and A is given as $A = (a_1, a_2, a_3, ..., a_n)$ By substituting for A and expanding the chain rule

$$N(b|a_1, a_2, ..., a_n) = \frac{N((a_1|b)N(a_2|b).....N(a)n|b)}{N(a_1)N(a_2).....N(a_n)}$$

The denominator remain static after substitute them into the equation. Therefore probability is

$$N(b|a_1, a_2, ..., a_n) = N(b) \prod_{n=1}^{m} P(a_n|b)$$

If the class variable (b) has only binary (Yes or No) outcome. For the multivariate classification we need to find the class b with maximum probability

$$b = argmax_b N(b) \prod_{n=1}^{m} P(a_n|b)$$

C. Evaluation Matrices

In the work, we used some statistical analysis that measure the test performance of different metrics. The performance of the classification techniques was evaluated by different metrics such as accuracy, sensitivity, specificity, precision and f1 measure [24]. The performance assessment can also use a confusion matrix [27]. Some metrics used in machine learning are:

True Positive (TP): The result of prediction correctly identifies that a patient has breast cancer.

False Positive (FP): The result of prediction incorrectly identifies that a patient has breast cancer.

True Negative (TN): The result of prediction correctly rejects that a patient has breast cancer.

False Negative (FN): The result of prediction incorrectly rejects that a patient has breast cancer.

The accuracy provides the difference between healthy and patient ability ratio using the prediction model. To find the accuracy of the classifier, one can use the true positive, true negative, false positive and false negative.

$$accuracy = \frac{(TP + TN)}{TP + FP + TN + FN} \quad (1)$$

The sensitivity test gives the rate of correctly positive cases. It also is known as Recall and True Positive Rate (TPR).

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

Specificity is indicating the negative results. It gives a proportion of the absence of the disease in patients. It is also known as True Negative Rate (TNR).

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

Precision is also known as positive predictive value. It gives the ratio of a correctly predicted positive result by classifier algorithms

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

F1 measures the accuracy of the model by a combination of precision and recall. It gives the ratio both FP and FN of a model.

$$F1 = \frac{2 * (Recall * Precision)}{Recall + Precision} \quad (5)$$

III. Result and Discussion

A. Analysis of Result

In this dataset, there are some NaN values, it is denoted by '?'. Therefore, we used the dropna() function to remove the NaN values. After cleaning the datasets, we have 683 entries including 10 features. The heatmap is shown in figure (2) and it appears to have no correlated parameters.

The performance of six machine learning is shown in figure 3 SVM achieved the highest performance with a maximum classification accuracy of 97.07% while second highest classification accuracy was achieved by NB and RF (97%). Moreover, KNN, DT and LR shows the almost same performance by attaining 96% accuracy.

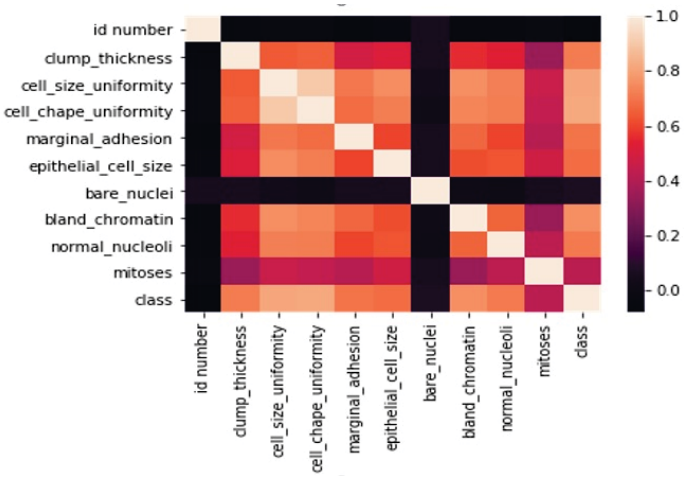

Fig. 1. The heatmap shown that the accuracy is affected for the 'ID number'.

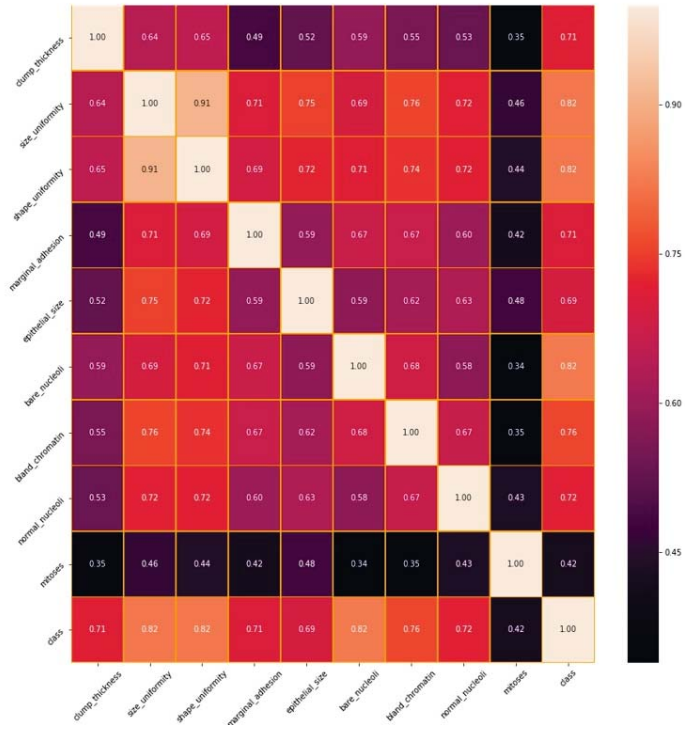

Fig. 2. Heat map for checking correlated columns for breast .

The performance measurements of six classification algorithms are presented is figure 4. The results clearly show that the DT and LR reached to the highest precision (97%). NB achieved the highest sensitivity, it's 100%. And NB also

achieved the worst specificity (92%). Considering f1 measure, all of classifiers show the same performance 97%. Figure 5 shows the prediction results for Naïve Bayes, Random Forest, Support Vector Machine, Decision Tree, KNN and Logistics Regression algorithms.

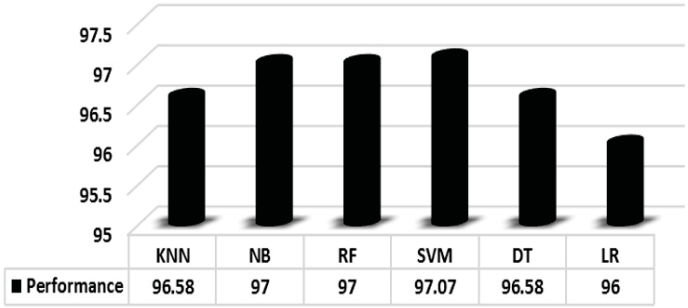

Fig. 3. The accuracy of six machine learning techniques.

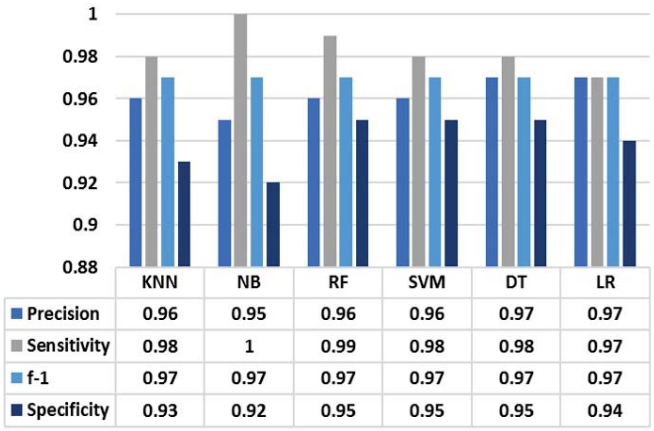

Fig. 4. Classification Performance Measurements (Breast cancer).

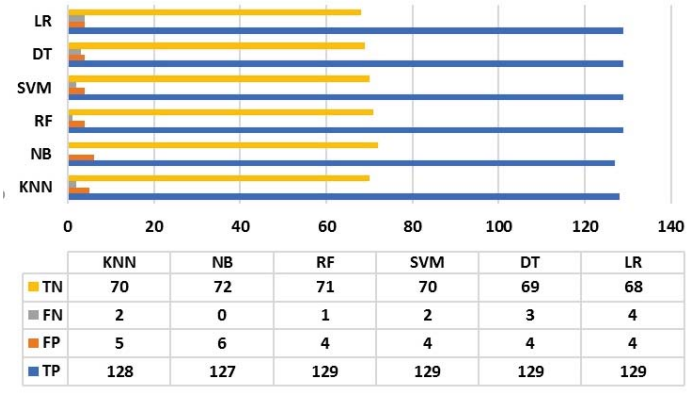

Fig. 5. Confusion matrix of classification techniques.

B. Performance Evaluation

All the machine learning classifiers show the accuracy level above 95% for breast cancer classification. This indicates that the performance of these classification techniques is excellent for breast cancer prediction. From the above discussion, it is very important to know about the receiver operating characteristics (ROC) curve, which is based on true positive rate (TPR) and false positive rate (FPR) of these classification results [25][26]. According to ROC curve and table 2, NB achieved highest AUC (area under curve) for ROC.

TABLE II
TRUE POSITIVE RATE AND FALSE POSITIVE RATE OF SIX SUPERVISED CLASSIFIERS

Algorithm	TPR (TP/TP+FN)	FPR (1-Specificity)
KNN	0.98	0.07
NB	1	0.08
rf	0.99	0.05
SVM	0.98	0.05
dt	0.97	0.05
LR	0.96	0.06

The ROC curve is presented in figure 6.

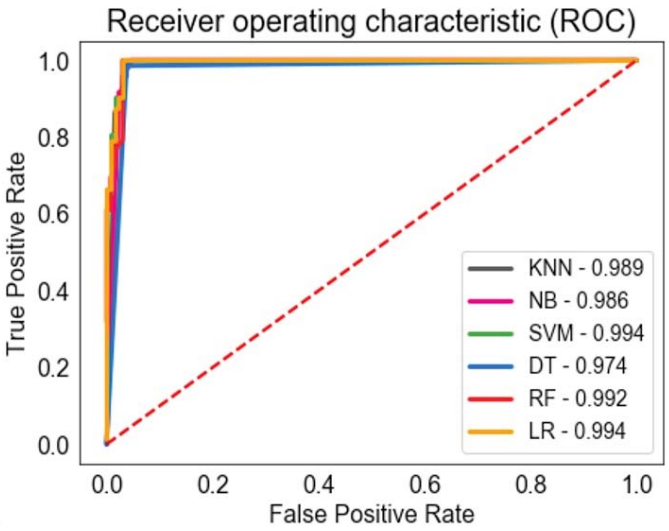

Fig. 6. ROC curve for six supervised classifiers.

IV. Conclusion

The workable aspect of this work is to make an effective diagnosis system for breast cancer patients. We investigated all classifier's performance on patient's data parameters and the SVM gives the highest classification accuracy of 97.07% to predict the breast cancer and NB gives the lowest accuracy of 96% among the classifiers. This application can be remarkably beneficial in low-income countries where fewer medical institutions as well as a shortage of specialized practitioners. Therefore, machine learning will play an important role in health care research and as well as medical centre's to early prediction of breast cancer.

V. Acknowledgement

We are grateful to Dr. Andre Barczak for his valuable comments, feedback, criticism and language correction.

REFERENCES

[1] M. R. Al-Hadidi, A. Alarabeyyat, and M. Alhanahnah, "Breast Cancer Detection Using K-Nearest Neighbor Machine Learning Algo-rithm," in 2016 9th International Conference on Developments in eSystems Engineering (DeSE), 2016, pp. 35–39.

[2] "WHO — Breast cancer: prevention and control," WHO, 2016.

[3] M. Amrane, S. Oukid, I. Gagaoua, and T. Ensari, "Breast cancer classification using machine learning," in 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), 2018, pp. 1–4.

[4] A. Almutairi, S. Tamrin, … R. W.-… in B. and, and undefined 2019, "Systematic Review on Knowledge and Awareness of Breast Cancer and Risk Factors Among Young Women," jour-nals.aiac.org.au.

[5] S. Neil-Sztramko, K. Winters-Stone, … K. B.-B. J. S., and unde-fined 2019, "Updated systematic review of exercise studies in breast cancer survi-vors: attention to the principles of exercise train-ing," bjsm.bmj.com.

[6] "Types of Breast Cancer — Different Breast Cancer Types." [Online]. Available: https://www.cancer.org/cancer/breast-cancer/understand-ing-a-breast-cancer-diagnosis/types-of-breast-cancer.html. [Accessed: 03-Jun-2019].

[7] D. Carvalho, P. R. Pinheiro, and M. C. D. Pinheiro, "A Hybrid Model to Support the Early Diagnosis of Breast Cancer," Procedia Comput. Sci., vol. 91, pp. 927–934, Jan. 2016.

[8] M. Kumari, "Breast Cancer Prediction system," Procedia Comput. Sci., vol. 132, pp. 371–376, Jan. 2018.

[9] H. Wang, B. Zheng, S. W. Yoon, and H. S. Ko, "A support vector machine-based ensemble algorithm for breast cancer diagnosis," Eur. J. Oper. Res., vol. 267, no. 2, pp. 687–699, Jun. 2018.

[10] N. Shukla, M. Hagenbuchner, K. T. Win, and J. Yang, "Breast cancer data analysis for survivabil-ity studies and prediction," Comput. Methods Programs Biomed., vol. 155, pp. 199–208, Mar. 2018.

[11] R. Jafari-Marandi, S. Davarzani, M. Soltanpour Gharibdousti, and B. K. Smith, "An optimum ANN-based breast cancer diagnosis: Bridging gaps between ANN learning and deci-sion-making goals," Appl. Soft Comput., vol. 72, pp. 108–120, Nov. 2018

[12] A. Abdelaziz, M. Elhoseny, A. S. Salama, and A. M. Riad, "A machine learning model for improv-ing healthcare services on cloud computing envi-ronment," Measurement, vol. 119, pp. 117–128, Apr. 2018.

[13] L. Abdel-Ilah and H. Šahinbegović, "Using ma-chine learning tool in classification of breast cancer," Springer, Singapore, 2017, pp. 3–8.

[14] "UCI Machine Learning Repository: Breast Can-cer Wisconsin (Original) Data Set." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(original). [Accessed: 04-Jun-2019].

[15] W. H. Wolberg and O. L. Mangasarian, "Multi-surface method of pattern separation for medical diagnosis applied to breast cytology.," Proc. Natl. Acad. Sci., vol. 87, no. 23, pp. 9193–9196, Dec. 1990.

[16] D. H. Jr, S. Lemeshow, and R. Sturdivant, Applied logistic regression. 2013.

[17] L. Breiman, "Random Forests," Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.

[18] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," IEEE Trans. Syst. Man. Cybern., vol. 21, no. 3, pp. 660–674, 1991.

[19] A. Y. Chervonenkis, "Early History of Support Vector Machines," in Empirical Inference, Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 13–20.

[20] V. Vapnik, I. Guyon, T. H.-M. Learn, and unde-fined 1995, "Support vector machines," stat-web.stanford.edu.

[21] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," Pattern Recognit., vol. 40, no. 7, pp. 2038–2048, Jul. 2007.

[22] G. Guo, H. Wang, D. Bell, Y. Bi, and K. Greer, "KNN Model-Based Approach in Classification," Springer, Berlin, Heidelberg, 2003, pp. 986–996.

[23] K. M. Leung, "Naive bayesian classifier," Poly-tech. Univ. Dep. Comput. Sci. Risk Eng., 2007.

[24] A. D.-N. C. and Applications and undefined 2016, "Performance eval-uation of different ma-chine learning techniques for prediction of heart disease," Springer.

[25] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes using Classification Algorithms," Pro-cedia Comput. Sci., vol. 132, pp. 1578–1585, Jan. 2018.

[26] M. Razu Ahmed, S. M. Hasan Mahmud, M. Altab Hossin, H. Jahan, and S. Rashed Haider Noori, "A Cloud Based Four-Tier Architecture for Early Detection of Heart Disease with Machine Learning Algorithms," 2018, IEEE 4th International Con-ference on Computer and Communi-cations.

[27] "Confusion Matrix." http://www2.cs.uregina.ca/ hamilton/courses/831 /notes/confusion$_m atrix/confusion_m atrix.html.[Accessed : 20 - Dec - 2018]$.