An Analysis On Breast Disease Prediction Using Machine Learning Approaches

F. M. Javed Mehedi Shamrat, Md. Abu Raihan, A.K.M. Sazzadur Rahman, Imran Mahmud, Rozina Akter

Abstract: The central aspect of this study is to evaluate the different Machine learning classifier's performance for the prediction of breast cancer disease. In this work, we have used six supervised classification techniques for the classification of breast cancer disease. For example, SVM, NB, KNN, RF, DT, and LR used for the early prediction of breast cancer. Therefore, we evaluated breast cancer dataset through sensitivity, specificity, f1 measure, and total accuracy. The prediction performance of breast cancer analysis shows that SVM obtained the uppermost performance with the utmost classification accuracy of 97.07%. Whereas, NB and RF have achieved the second highest accuracy by prediction. Our findings can help to reduce the existence of breast cancer disease through developing a machine learning-based predictive system for early prediction.

Keywords: Machine Learning, Classification, Breast Cancer, Prediction.

1. INTRODUCTION

Breast cancer is the notable cause of women's death and disability from a global perspective. A report shows that 508000 women died in 2011 by chronic disease, especially breast cancer [1]. In 2015, around 17.7 million people were global death caused by CHD [2]. The World Health Organization (WHO) estimated that above 23.6 million persons would be dead by 2030, because of such chronic disease [3]. Very few peoples can get their treatment, but most of the scenario affected by chronic disease treatment is very expensive and complicated [4]. Moreover, this reason to takes a long time, mistaken or delayed decisions are possible to cause for death. However, the cost of breast cancer diagnosis and replacement is very extreme, and it can be called an extreme level of financial expenses. A study reported that the cancer disease causes the commercial benefits with cost over $79 billion, and treating people with end-stage renal disease cost around $35 billion [5]. Breast cancer disease is chronic and takes a long time for curing. For these causes, most of the patients cannot afford the cost of the cure for cancer disease. Furthermore, chronic disease prediction is the most prominent matter for clinical practitioners and medical services
centers to take the accurate decision of such conditions. Therefore, a machine learning-based great platform can solve these kidney disease problems through early detection and diagnosis. The aspect of this main work is to improve the first treatment and diagnosis of kidney disease for peoples of low-income and developing countries. Hence, our study can be a significant approach for detecting kidney disease outbreak with machine learning algorithms. In the last ten years, the growth rate of medical data is going to a large amount from enormous arenas. From the art of Machine learning (ML) algorithms have portrayed that purpose to resolve various health and scientific problem [6][7]. An establishment of several studies shows that ML models already have obtained dramatically excessive accuracies in disease-based medical issues. However, supervised based models are one of the utmost operative methods for academic and health products in clinical fields. [8]. The aspect of this main work is to improve early treatment and diagnosis of chronic disease for peoples of low-income and developing countries. Hence, our study can be a significant approach for detecting persistent disease outbreak with machine learning algorithms. In this study, the main goal to achieve that the ML predictive model can detect the early breast cancer symptoms through prediction by the experimental model. The rest of the paper discusses the literature review in part 2, the Methodology (Experimental Setup, Data Collection, Data Preprocessing, Evaluation Criteria) consist of part 3, in part 4 discussed Result & Discussion, in part 5 short brief on Conclusion.

- *F. M. Javed Mehedi Shamrat is received Bachelor's degree program in Software Engineering at Daffodil International University, Bangladesh. E-mail: javedmehedicom@gmail.com*
- *Md. Abu Raihan is currently pursuing Bachelor's degree program in Computer Science and Engineering at Daffodil International University, Bangladesh. E-mail: mdraihansagor7@gmail.com*
- *A.K.M Sazzadur Rahman is completed Master's degree program in Computer Science and Engineering at Daffodil International University, Bangladesh. E-mail: sazzad433@diu.edu.bd*
- *Dr. Imran Mahmud is currently working as a senior Lecturer at the graduate school of business, universiti sains Malaysia. He worked as an assistant professor at Daffodil International University, Bangladesh. He has several publications and received many awards in the field of technology management. E-mail: imranmahmud@daffodilvarsity.edu.bd*
- *Rozina Akter completed Master's of Business Administration (MBA) from the University of Dhaka, Bangladesh. She is currently an Assistant Professor in the Department of Business Administration at the Daffodil International University, Dhaka, Bangladesh. She has published so many research articles, and she is an expert on the banking model and financial analytics. E-mail: rozina@daffodilvarsity.edu.bd*

2 LITERATURE REVIEW

Our central aspect is to develop a system using machine learning for the early forecast of cancer disease from the patient's data. Through related studies were done on applying and using several ML classifiers to determine early detection and prediction of breast cancer using ML techniques. However, the outcomes of the previous work on machine learning used in breast cancer prediction as follows: Jain et al. [12] presented a survey to attribute assortment and machine learning techniques for identification and forecast of chronic disease. This works focused on a comprehensive review of different feature selection methods and their advantage and limitation. The contribution of this study is to use adaptive with parallel classification techniques for chronic complaint prediction. Bartz-Kurycki et al.[12], introduced a new model to forecast "neonatal surgical site infections (SSI)" using diverse classification processes. The accuracy of the area under the curve for each model was similar. The contribution of this study is to examine the hybrid model and other models with fewer and more clinically relevant variables. Carvalho et al. [13] presented a new hybrid method to sustenance the early verdict of breast cancer. This study tries to find optimum accuracy to provide nutrition to decision in the circumstances, whereas Bayesian

Network does not provide a satisfactory outcome. The contribution of the study is to advance an automatic device to contribute a precise identification and prediction of breast cancer. Kumari [14] presented a new prediction system that can predict the occurrence of breast cancer at an early stage by analyzing the nominal set of attributes that has selected from medical datasets. The KNN classifier obtains the best performance (99.28%) than other classifiers. The contribution of this study is to use the proposed system to predict breast cancer at an early stage, which significantly reduces the cost of treatment and improves the quality of life. Tapak et al. [15] introduced a comparative study between Naïve Bayes, Random Forest, AdaBoost, Support Vector Machine, LSSVM, Adabag, Logistics Regression, and LDA to predict breast cancer survival and metastasis. LR and LDA achieved the highest accuracy (86%). The SVM and LDA have superior sensitivity in comparison to other classifiers. The contribution of this study is to use SVM to predict the existence of breast cancer. Asri et al. [16] presented a comparative study between SVM, DT (C4.5), NB, K-NN to predict the early stage of breast cancer. The intelligent techniques applied to the WEKA data mining tool. Experiment results show that the SVM has the best performance accuracy, it is 97.13% the contribution of this study to use SVM to predict the early stage of breast cancer. Chougrad et al. [16] developed a deep convolutional neural network-based computer-aided treatment system. The CNN model achieved the best performance, and it is 98.94%. And they tested the CNN model on an independent database, and they've got the accuracy 98.23% and 0.99 AUC. The contribution of this study to use the high performer classifiers within the proposed structure, and that can apply to forecast the patients are "benign or malignant." Wang et al. [17], presented a new model to use breast cancer diagnosis based on the patient's historical data from clinical data. The proposed WAUCE model reduces the variance by around 97.98% and increases accuracy by 33.34%. The contribution of this study can be further applied to a safer, more reliable illness diagnosis process. Madhuri & Bharat et. Al [18] presents a comparative study to diagnosis breast cancer patients through supervised machine learning techniques. They applied multiple machine learning algorithms, including LR, RF, DT, and Multi-layer perception. Multi-layer perception gives high performance compared to other algorithms. Layla & Hana [19], implement a feed-forward backpropagation network (FFBPN) to classify the "benign cancer or malign cancer." They showed that for an Artificial Neural network, the best design for classification is that three hidden layers and twenty-one neurons in every secret level. Propose plan gives the highest accuracy of 98%. Amrane et al. [20] presented a comparison between two machine learning classifiers NB and KNN, to provide an accurate diagnosis of breast cancer patients. The comparison result was KNN gives high accuracy at 97.51%, and NB has 96.19% accuracy. Al-Hadid et al. [21] proposed a model to detect breast cancer disease with a higher accurateness. Their model was divided into two-part, The first one was processing the images to extract the features, and the second one was to use two supervise machine learning techniques to get the accuracy. Xiao et al. [22], introduce a new strategy for gene expression analysis to five different classification algorithms with deep learning methods. The contribution of this study is shown to be accurate and useful prediction results for cancer prediction. To date, machine learning classification techniques have created a significant impact and obligation in the chronic disease research society for the initial discovery of the chronic disease. Moreover, ML algorithms are given more accurate results in constant disease prediction as compared to other data classification techniques [7][8]. Many of studies already show that the supervised based classification techniques have obtained excellent accuracies in the field of disease prediction [10][4][11] Motivated by this, the authors have used six prominent ML techniques for forecasting and proper treatment of chronic patients. The main goal of this study is to inspect the performance measurement of various leading supervised methods and gained more efficient outcomes by reducing extremely cost of diagnosis and dialysis of chronic diseases. For this study, six supervised learning techniques were used, including "KNN, Support Vector Machine, Decision Tree, Random Forest, Naïve Bayes, and Logistics Regression." Moreover, the performance of the classifiers of selected learning techniques is evaluated using the confusion matrix and different statistical methods. Henceforth, the outperform classification technique will donate for the decision support system and diagnosis of chronic disease.

3 METHODOLOGY

3.1. Environment Setup

The proposed experimental setup, including ML systems, has been presented in this segment. The intelligent breast cancer detection technique contains four steps to make this decision. Phase 1 focuses on extracting and combined data from diverse health systems with devices in phase 2 usages to store a massive amount of medical data. Therefore, phase 3 usages ML-based classifiers to training data of cancer disease datasets. Besides, phase 4 exemplifies the outcome of the breast cancer detection system for the clients. In this study, we have focused only on the machine learning phase (phase 3). For further study, in figure 1, we are currently developing in this full system architecture.

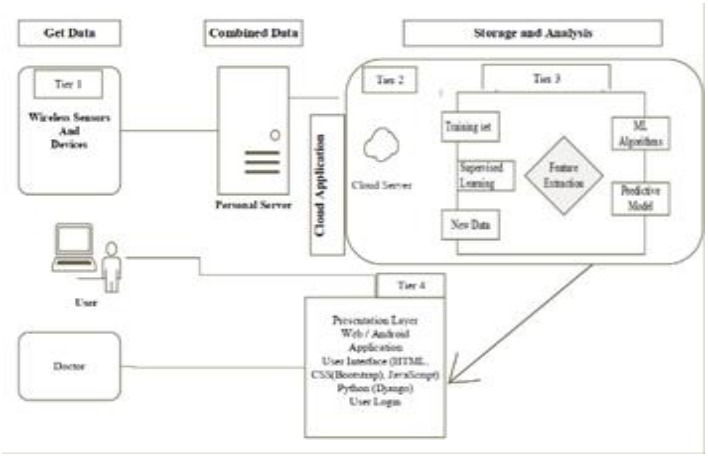

***Fig. 1**. The Experimental Setup.*

- **SVM**

SVM incorporates an administered learning system that looks and sorts it into one of two classes. An SVM yields a guide of the masterminded data with the edges between the two as far isolated as could sensibly be standard. SVMs are used in content requests, picture plan, handwriting affirmation, and specialized examinations.

For SVM, preparing includes the minimization of the mistake work:

$$\frac{1}{2}w^t w + c\sum_{i=1}^{N} \in i \quad (1)$$

Focus to the constraints:

$$y_i(w^t \emptyset(x_i) + b) \geq 1 - \zeta_i \, and \, \zeta_i \geq 0, i = 1, \ldots, N \quad (2)$$

Where C is the cutoff enduring, w is the vector of coefficients, b is a steady, and ζ_i addresses parameters for dealing with no recognizable data (inputs). The rundown I name the N planning cases. Note that y∈±1 addresses the class names, and xi addresses the free factors.

- **LR**

Logistic regression was used in the basic sciences in the mid-twentieth century. It was then used in various humanism applications. Strategic Regression is used when the dependent variable (target) is absolute.

The key twist relates the free factor, X, to the moving mean of the DV, P (Y). The formula to do so may be created,

$$p = \frac{e^{a+bX}}{1+e^{a+bx}} \quad (3)$$

- **KNN**

KNN makes expectations reliant on the aftereffect of the K neighbors closest to that point. As needs be, to make conjectures with KNN, we need to portray an estimation for evaluating the division between the inquiry point and cases from the model's test. One of the most standard choices to measure this detachment is known as Euclidean.

Euclidean Formula,

$$d(p,q) = d(q,p) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2} = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \quad (4)$$

3.2. Data Collection

In this study, we use the Wisconsin Breast Cancer data (Original) by the "University of Wisconsin Hospitals, Madison, Wisconsin, USA" [23]. Breast cancer dataset contains 699 breast cancer patients' records. Moreover, the datasets contain the Benign: 458 (65.5%) samples and Malignant: 241 (34.5%) samples. However, I chose the particular parameters for data analysis, which are summarized in table1.

TABLE 1

Parameters for Data Analysis

No	Factor	Information Factor	Description
1	Id	Numerical	Id
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 Nucleoli	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 for Benign, 4 for Malignant

3.3. Data Preprocessing

From the Wisconsin Breast Cancer data, it contains 699 breast cancer patient's data, including 11 parameters. We see that the column 'class' has a high correlation with all columns except ID Number, which has no significance and needs to be removed. Therefore, we removed the 'Id' parameter from the data set. If the 'ID number' column is not removed, the accuracy is affected when we conducted the analysis. In this dataset, there is no missing data shown in figure 2. But there are some NaN values, and it is denoted by '?'. Therefore, we used the dropna() function to remove the NaN values. After cleaning the datasets, we have 683 entries, including ten parameters. Hence, we didn't find any correlated column in this breast cancer dataset (figure 2).

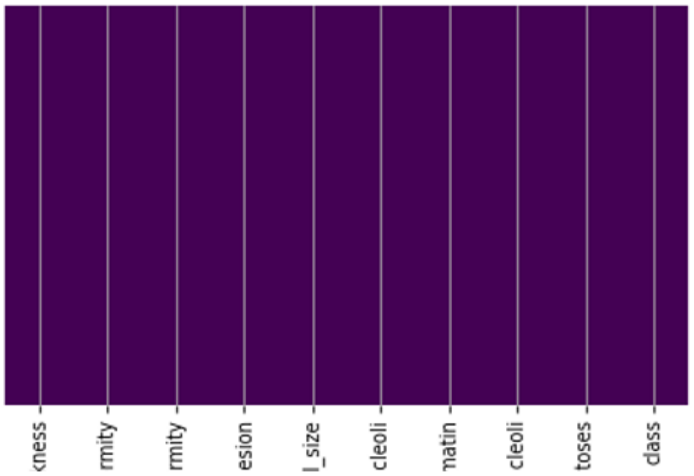

Fig. 2. No missing values on breast cancer datasets.

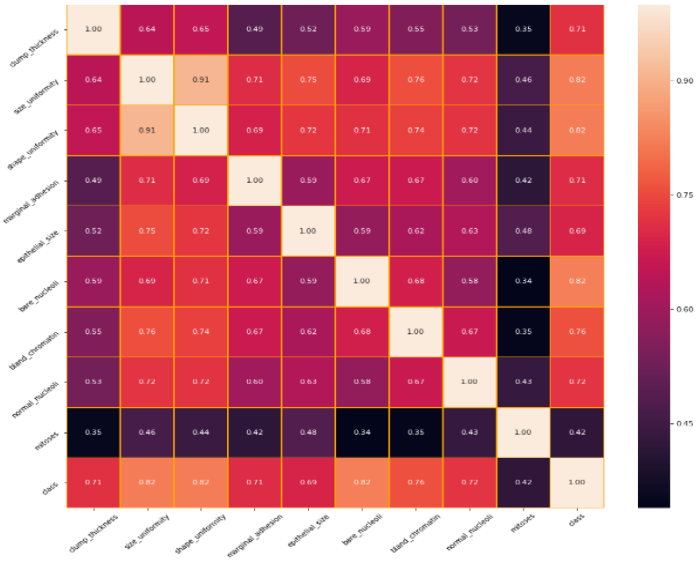

Fig. 3*. Heat map for checking correlated columns for breast cancers.*

3.4. Evaluation Criteria

In this work, we used six machine learning techniques for the early prediction of breast cancer disease. Therefore, the performance measurements of the classifiers are appraised by different statistical procedures. Such as confusion matrix (True

Positive, False Positive, True Negative, False Negative), Recall, Precision, f1- measure, etc. [24].

The computation method of the measurement considerations are as follows,

$Accuracy_i = (TP_i + TN_i) / (TP_i + FP_i + TN_i + FN_i)$ (5)

TPR_i or $Sensitivity_i$ or $Recall_i = TP_i / (TP_i + FN_i)$ (6)

$Specificity_i = TN_i / (TN_i + FP_i)$ (7)

$Precision_i = TP_i / (TP_i + FP_i)$ (8)

$f1_i = 2 * (Recall_i * Precision_i) / (Recall_i + Precision_i)$ (9)

False Positive Rate = $1 – Specificity_i$ (10)

The f1_measure is denoted by the weighted norm of the $recall_i$ and $precision_i$. To classify as a better classifier this the value will be 1 and for the lowest performance, it will be 0.

4 RESULT AND DISCUSSION

In this segment presents an analysis of cancer disease detection, we will discuss the general analysis process of the work. We have tested our models through several measurement experiments. Then we will present the performance of the models and compare them with other classifiers. I have accompanied several analyses to examine the six ML-based supervised techniques for diagnosis and forecast of cancer disease. The performance comparison and performance measure of six machine learning classifiers for chronic disease prediction. The performance of the selected ML classifiers shows in figure 4. The SVM classifier attained the uppermost performance with a supreme prediction accuracy of 97.07 percentage, whereas the next maximum classification accuracy is succeeded by NB and RF (i.e., 97%). Moreover, KNN, DT, and LR show the almost same performance by attaining 96% accuracy.

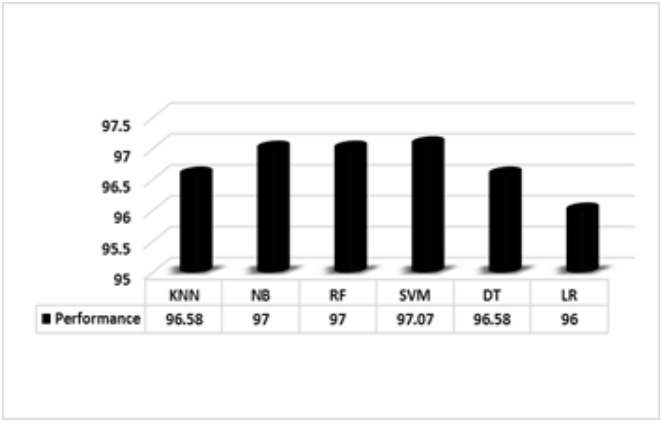

Fig. 4. *The Accuracy of Six Machine Learning (Breast Cancer).*

According to the performance measurements of six classification techniques are illustrated in figure 5. The results evidently show that the DT and LR reached to the highest precision (97%). NB achieved the highest sensitivity, and it's 100%. And NB also achieved the worst specificity (92%). Considering f1 measure, all of the classifiers show the same performance, and it's above 95%, respectively. Figure 6 demonstrated the confusion matrix of forecast results for "Naïve Bayes, Random Forest, Support Vector Machine, Decision Tree, KNN and Logistics Regression algorithms."

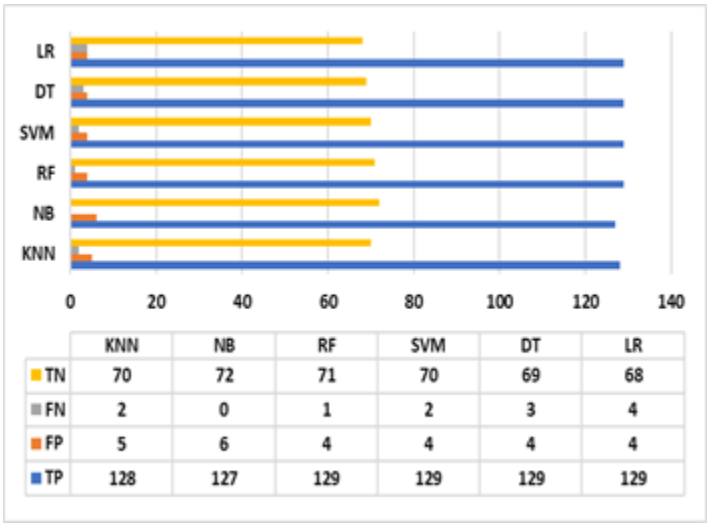

Fig. 5*. Classification Performance Measurements (Breast cancer).*

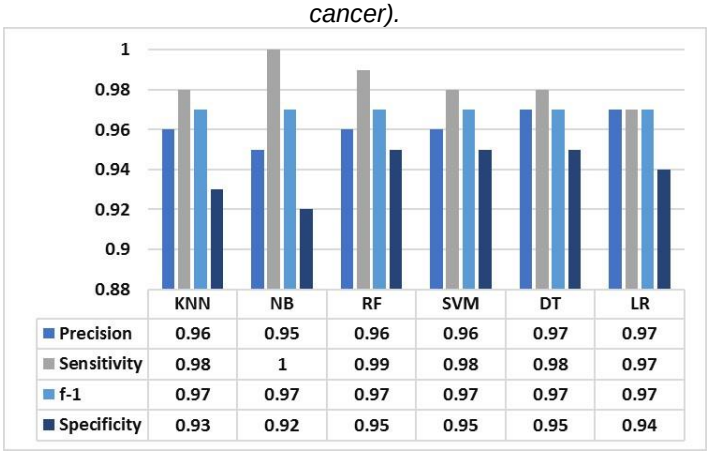

Fig. 6*. Confusion Matrix of Classification Techniques.*

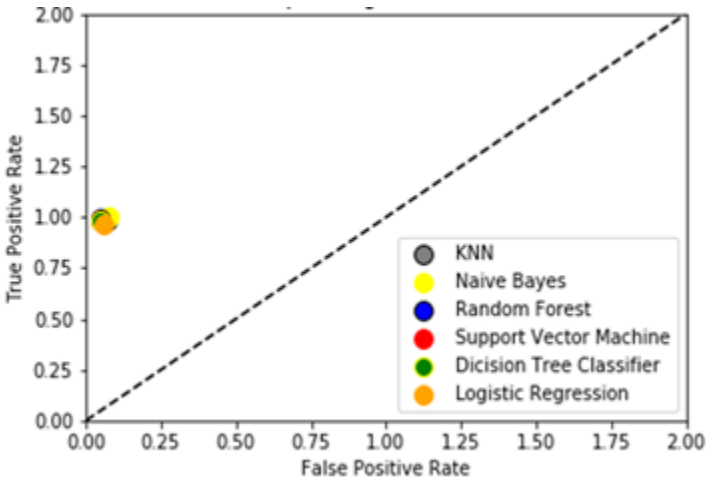

Fig. 7*. Receiver Operating Characteristics curve for Breast Cancer datasets. The prediction result shows the classifier's outcome above 95% for cancer disease detection.*

5 CONCLUSION

Moreover, the six machine learning algorithms are doing very good for cancer disease prediction. In our experiment, it is essential to recognize the "receiver operating characteristics (ROC) curve," which is grounded on the "true positive rate (TPR) and false-positive rate (FPR)" of these detection results. According to the ROC curve, RF and NB outperformed (Kidney Disease) all other techniques. Furthermore, KNN (Breast Cancer) and SVM (Liver Disease) achieved the highest AUC (area under the curve) for ROC. In this study, we have depicted several ML-based classification techniques. Therefore, we deliver an experimental process on ML-based system for the early prediction of breast cancer disease. Consequently, we competed the presentation of the six algorithms which are depleted in the forecast of cancer diseases and assessed by their results using a statistical technique, namely is the confusion matrix. The experimental performance shows that the Naïve Bayes and Random Forest has achieved outperform than the other classifiers within cancer datasets. This inspection has usage of six ML techniques for the prediction of cancer disease based on some attributes.

ACKNOWLEDGMENT

The authors are grateful and pleased to all the researchers in this research study.

REFERENCES

[1] "WHO | Breast Cancer: Preventation and contro", www.who.in/cancer/detection/breastcancer/en/index3.html [Accessed: 01-Nov-2019]

[2] "WHO | World Health Organization.", https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/ . [Accessed: 01-Nov-2019].

[3] K. Purushottam. Saxena, and R. Sharma, "Efficient Heart Disease Prediction System," Procedia Comput. Sci., vol. 85, pp. 962–969, Jan. 2016.

[4] P. Singh, S. Singh, and G. S. Pandi-Jain, "Effective heart disease prediction system using data mining techniques.," Int. J. Nanomedicine, vol. 13, pp. 121–124, 2018.

[5] "Chronic Kidney Disease Basics | Chronic Kidney Disease Initiative | CDC." [Online]. Available: https://www.cdc.gov/kidneydisease/basics.html. [Accessed: 12-Dec-2018].

[6] F.M. Javed Mehedi Shamrat, Md. Asaduzzaman, A.K.M. Sazzadur Rahman, Raja Tariqul Hasan Tusher, Zarrin Tasnim "A Comparative Analysis Of Parkinson Disease Prediction Using Machine Learning Approaches" International Journal of Scientific & Technology Research, Volume 8, Issue 11, November 2019, ISSN: 2277-8616, pp: 2576-2580.

[7] A.K.M Sazzadur Rahman, F. M. Javed Mehedi Shamrat, Zarrin Tasnim, Joy Roy, Syed Akhter Hossain "A Comparative Study On Liver Disease Prediction Using Supervised Machine Learning Algorithms" International Journal of Scientific & Technology Research, Volume 8, Issue 11, November 2019, ISSN: 2277-8616, pp: 419-422.

[8] A. K. Dwivedi, "Analysis of computational intelligence techniques for diabetes mellitus prediction," Neural Comput. Appl., pp. 1–9, Apr. 2017.

[9] Palaniappan S, Awang R. Intelligent heart disease prediction system using data mining techniques. Int J Comput Sci Net Secur. 2008;8:343–350.

[10] M. Heydari, M. Teimouri, Z. Heshmati, and S. M. Alavinia, "Comparison of various classification algorithms in the diagnosis of type 2 diabetes in Iran," Int. J. Diabetes Dev. Ctries., vol. 36, no. 2, pp. 167–173, Jun. 2016.

[11] M. Kukar, I. Kononenko, C. Grošelj, … K. K.-A. intelligence in, and undefined 1999, "Analysing and improving the diagnosis of ischaemic heart disease with machine learning," Elsevier.

[12] D. Jain and V. Singh, "Feature selection and classification systems for chronic disease prediction: A review," Egypt. Informatics J., Apr. 2018.

[13] 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.

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

[15] L. Tapak, N. Shirmohammadi-Khorram, P. Amini, B. Alafchi, O. Hamidi, and J. Poorolajal, "Prediction of survival and metastasis in breast cancer patients using machine learning classifiers," Clin. Epidemiol. Glob. Heal., Oct. 2018.

[16] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis," Procedia Comput. Sci., vol. 83, pp. 1064–1069, Jan. 2016.

[17] 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.

[18] M. Gupta and B. Gupta, "A Comparative Study of Breast Cancer Diagnosis Using Supervised Machine Learning Techniques," in 2018 Second International Conference on Computing Methodologies and Communication (ICCMC), 2018, pp. 997–1002.

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

[20] 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.

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

[22] X. Liu et al., "A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method.," Comput. Math. Methods Med., vol. 2017, p. 8272091, 2017.

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

[24] "ConfusionMatrix." http://www2.cs.uregina.ca/~hamilton/courses/831/notes/confusion_matrix/confusion_matrix.html. , [Accessed: 20-Dec-2018].

[25] Sayad AT, Halkarnikar PP. Diagnosis of heart disease using neural network approach. Int J Adv Sci Eng Technol. 2014;2:88–92.

[26] Gudadhe M, Wankhade K, Dongre S. Decision support system for heart disease based on support vector machine

and Artificial Neural Network. In: Computer and Communication Technology (ICCCT), 2010 International Conference on, 2010:741–745.

[27] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating error. Nature. 1986;323:533–536.