FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 7 No. 2, April, 2023, pp 216 - 222
DOI: https://doi.org/10.33003/fjs-2023-0702-1392

BREAST CANCER DETECTION WITH MACHINE LEARNING APPROACH

Sunday Samuel Olofintuyi

Department of Computer Science, Achievers University, Owo.

*Corresponding authors' email: olofintuyi.sundaysamuel@mail.com

ABSTRACT

One of the most widespread diseases among women today is breast cancer. Early and accurate diagnosis is key in rehabilitation and treatment. The usage of mammograms has some uncertainties in the detection rate. To develop tools for physicians for effective and early detection and diagnosis, machine learning techniques can be adopted. The introduction of Machine Learning (ML) in developing the tool will increase the survival rate of patients with breast cancer. This research work proposed different six ML techniques; Logistic Regression, Linear Discriminant Analysis, Decision Tree (DT), KNN, Naïve Bayes (NB), and Support Vector Machine (SVM), and then recommended the model with the highest accuracy for breast cancer detection. The experiment was carried out in a python environment and all the aforementioned techniques were validated with Wisconsin Breast Cancer dataset and evaluated with accuracy, precision, and recall.

Keywords: Breast cancer, Machine Learning Algorithm, Detection

INTRODUCTION

Breast cancer (BC) is a tumor that operates in cells present in the breast. This malignant tumor can spread to another part of the body. It is one of the most occurring cancers in the world. Annually, over two million women are known to be affected. Out of women diagnosed with cancer in UAE, 43% have breast cancer (Lim *et al.* 2016), also breast cancer was diagnosed in 25% of females in the US at various stages of their life (WHO, 2020). About 268,600 new cases of invasive cancer and 62,930 new cases of non-invasive breast cancer were diagnosed in women in 2019 (Mohammed *et al.* 2020). To increase the chances of survivability from the second leading cause of death among women, early detection should be a welcome idea. Early detection and diagnosis remain the only chance of survival as there is no known prevention method for breast cancer. National Breast Cancer Formulation (NBCF) recommends that once a year, women above the age of 40 years should get a mammogram, this is because the symptoms are not presented well. Hence, there is a delay in diagnosis. Mammogram simply is defined as an X-ray of the breast, it is an approved medical technique for the detection of breast cancer in women. If the technique is well carry-out, there is no side-effect on the patients. In addition, women that get a mammogram once a year have a chance of high survivability than those that do not get one.

The killing disease remains a challenge for the physician. Recently, there is an improvement with the introduction of medical technologies. Also, new strategies are in place because of the availability of enormous patient data for the prediction and detection of the disease. In addition, accurate diagnosis has improved because of the availability of data from the patients and the physician. It has been identified by specialist doctors that there are some factors that can increase individual odds of developing BC, such factors include environmental factors, way of life, and hormonal imbalances. Other factors pointed out are gene mutation from the family record, postmenopausal hormonal imbalances, and obesity.

Machine learning (ML) is a branch of artificial intelligence that learns from a large set of dataset and then make a prediction based on the pattern learn. There are been widespread usage of predictive models virtually in all fields, this is because the predictive model has enhanced decision-making. ML has been used in Agriculture for prediction (Olofintuyi, 2022), computer networks for prediction (Olofintuyi, 2021), and cancer detection (Dana and Raed, 2016; Nahla *et al.* 2021 and Adebiyi *et al.* 2022). The other section in this research work discusses the Literature review in section, the methodology used in section three, section four discusses the results obtained after the experiment. Finally, section five gave a conclusion and recommendation about the research presented.

Recently, researchers around the globe have used several ML models for the detection and prediction of breast cancer. A neural network was adopted for predicting and classifying the invasive ductal carcinoma in the breast, and 88% accuracy was achieved by the researcher. The type of breast cancer was predicted by Silva et al., (2019) using NB, SVM, GRNN, and J48. GRNN and J48 achieved NB and SVM achieved an accuracy of 91% and 89% accuracy. Ojha and Goel, (2017) did a study on breast cancer by predicting its recurrence, the WPBC dataset was used during the experiment. Two major classes of algorithms were used namely classification algorithm and clustering algorithm. Experimental results show that the classification algorithm gave higher accuracy than the clustering algorithm. Pritom et al., (2016) also presented a paper on breast cancer recurrence, where it has been predicted using classification and selection techniques. WPBM dataset was used to validate NB, C4.5, and SVM algorithms. Experimental results indicated that SVM gave75%accuracy which is the highest result. In the same manner, Asri et al., (2016) presented a paper using machine learning to predict breast cancer and uses the WBC dataset for their work. Four different algorithms were used and after the experiment, SVM gave 97.13% which outperformed other techniques used. Hazra et al., (2016) presented an ensemble approach for detecting breast cancer. The author also used the WDBC dataset during the experiment. NB and SVM were ensembles for the experiment and the ensemble gave 97.3% accuracy. Rodrigues (2015) analyses the Wisconsin dataset with two machine learning models; NB and JB. The models gave 97.51% and 96.5% accuracy respectively. Saabith et al., (2014) did a comparative study of the three models; J48, MLP, and rough set. A breast cancer dataset was used during the experiment. J48 outperformed another model with 79.97% accuracy. SMO was presented by Chaurasia et al., (2017) for breast cancer detection, WBC dataset was used. The novel SMO presented outperforms other models by 96.19%. Anji et al., (2020) also use ML for breast cancer detection, the neural network presented gave a better accuracy as compared to other

methods. Finally, Siham *et al.* (2020) also presented a paper by analyzing various ML, in their work, SMO gave better accuracy.

MATERIALS AND METHODS

Dataset used

The WBC dataset was used in this research work because it is freely available at the UCI machine learning repository. All the models were trained and tested with the WBC dataset. The dataset has 699 instances with 10 attributes. From the instances of the dataset, 241 are malignant while 458 are benign. Table 1 depicts some information about the WBC dataset. There are two values for the class label which are 1 and 0. 1 means its malignant and 0 means benign. To manage the imbalanced data and missing values, data preprocessing was carried out on the dataset. Instances with missing values were removed so as to manage the missing value and for the imbalanced data, resample filter was used to rebalance the data. The following are the attribute of the WBC dataset; cell shape Uniformity, Cell Size Uniformity, Clump Thickness, Mitoses, Normal Nuclei, Bland Chromatin, Bare Nuclei, Single Epithelial Cell Size, Marginal Adhesion, and Class.

Table 1: Description of WBC dataset

Dataset	No of instances	No of attributes	Number of classes	Malignant	Benign
WBC	699	10	2	241	458

Data preprocessing

Basically, three steps were used in the preprocessing phase of the proposed model. Firstly, discretize filter was used to discretize the dataset, secondly, in order to base the class and maintain the class distribution towards uniform distribution, resample filters were used to resample the instances, and lastly remove the missing values from the dataset.

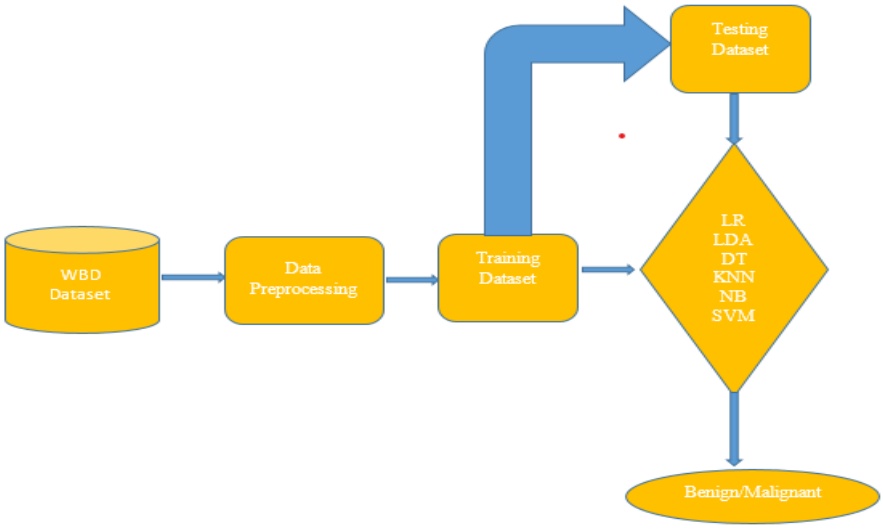

Figure 1: proposed workflow for breast cancer detection

Decision Tree (DT) Model

DT is a classifier that uses classification rules. It uses the pattern of a tree structure to operate, it consists of three main components which are the decision node, branch node, and leaf node. The decision node identifies the test attributes of the WBC dataset. The decision based on the test value is represented by the branch node while the class the instance belongs to is represented by the leaf node (Olofintuyi and Olajubu, 2021). Assuming a tuple Y, the decision tree is tested against the attribute value of the tuple. To know the prediction class of the tuple, a path is traced from the root to the leaf node. Also, the Information Gain (IG) of each attribute is measured to determine the root node. Also, the root node is selected based on the detail with the highest gain value. Information gain of the attribute for the next split is calculated and the nomination is done based on the attribute with the highest IG. Figure 2 below depicts a decision classifier.

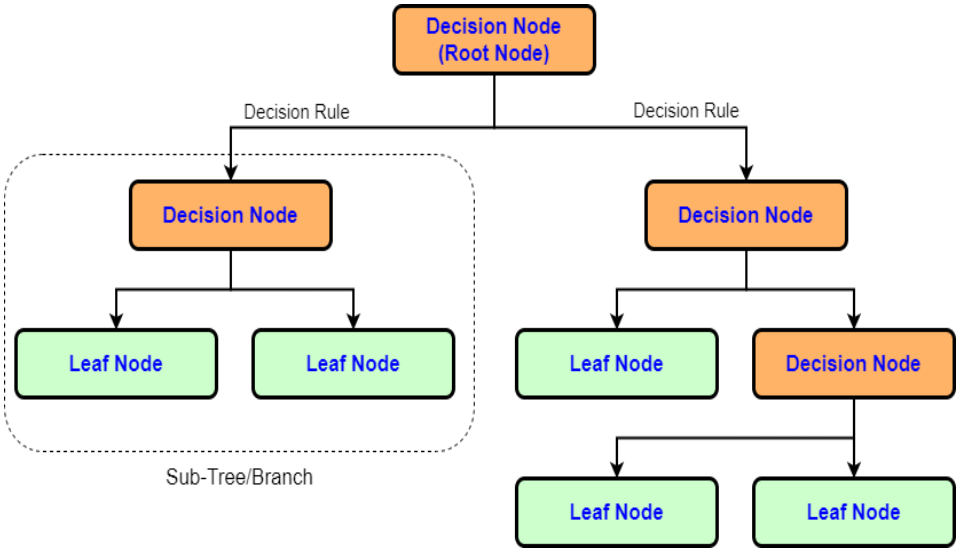

Figure 2: Decision Tree Classifier

Support Vector Machine (SVM) Model

Another ML model that uses a classification algorithm is SVM. It is suitable for two-group classification problems (Figure 3). It is a supervised ML that is able to categorize new text from a set of given labeled training of each category. Comparing SVM with other algorithms, SVM has a better performance and high speed as compared to other new algorithms like neural networks even with a limited number of samples, because of these advantages, it makes it suitable for text classification problems (Olofintuyi *et al.* 2019). It operates on statistical learning theory, and support vectors which are members of training data samples are used for data classification. SVM uses hyperplane for the classification of new data points. It is also suitable for the classification of linear and non-linear data (Tiwari and Ojha, 2019). The linear SVM uses a hyperplane to classify multi-dimensional data whereby the nearest training data point of each class is used by maximizing the margin between them. SVM classifies non-linear data by using the kernel function, the data used are mapped into higher dimensions for better classification. Radial Basis Function (RBF), polynomial kernel, and sigmoid kernel are some of the numerous kernel functions we have.

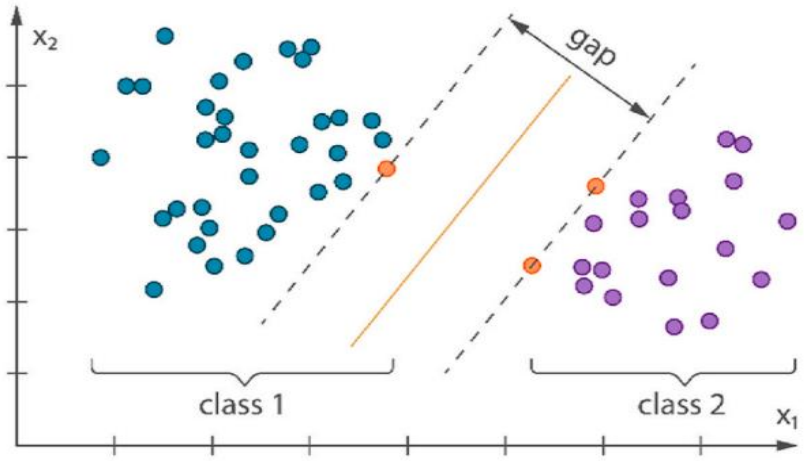

Figure 3: Support Vector Machine Classifier

K-Nearest Neighbor (KNN) Model

This algorithm uses Euclidean distance to measure the distance between neighbors and no parameter is needed when it comes to its usage. Figure 4 below depicts how KNN works. A new data instance is been classify into class A or class B based on the relative distance between the two classes. The red stars denote the malignant class while the green stars denote the benign class. The yellow box with a question mark indicates new instances which are classified into either class A or class B based on the maximum nearest neighbors. K denotes the number of nearest neighbors which is the core indicating factor in KNN. Generally, K is an odd number whenever we have two classes as depicted in the figure 4 below. The algorithm is known as the nearest neighbor when K=1.

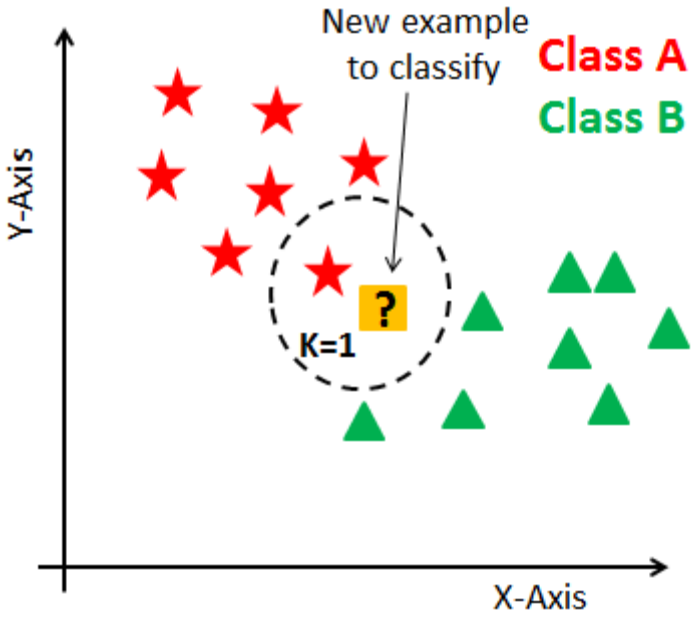

Figure 4: KNN classification principle

Naïve Bayes (NB) Model:
NB can be used to classify malignant class or benign based on previous records. NB is a good example of a supervised classification learning algorithm that adopts Bayes' theorem for its prediction (Olofintuyi et al., 2023). This algorithm is widely adopted because of its simplicity. For the training, a few samples are needed. It can also classify multi-label or binary datasets. A labeling decision is made to classify unlabeled traffic either as normal or anomalous based on the posterior probability. In order to detect the probability of the instance either as benign or malignant an independent set of features is used.

$$P(a|b) = \frac{P(b|a) * P(a)}{P(b)} \quad (1)$$

a represents the input variables from the dataset while b represents the output from the model. Where;

$$b = (b_1 b_2 b_3 \dots\dots\dots b_n)$$

Putting naïve assumption to Bayes' theorem, which is independent among the features, we then have;

$$P(a|b_1 b_2 \dots\dots b_n) = \frac{P(b_1|a)P(b_2|a) \dots\dots\dots P(b_n|a)P(a)}{P(b_1)P(b_2) \dots\dots\dots P(b_n)} \quad (2)$$

Which can be represented as

$$P(a|b_1 b_2 \dots\dots b_n) = \frac{P(a)\prod_{i=1}^{n} P(b_i|a)}{P(b_1)P(b_2) \dots\dots\dots P(b_n)} \quad (3)$$

We can remove the denominator, since it remains constant for a given input. We then have;

$$P(a|b_1 b_2 \dots\dots b_n) \alpha\ P(a) \prod_{i=1}^{n} P(b_i|a) \quad (4)$$

For the set of all the possible inputs, we picked the maximum probability when the set of inputs for all the possible values of the class variable y are picked. This can be represented mathematically as;

$$a = argmax_a\ P(a) \prod_{i=1}^{n} P(b_i|a) \quad (5)$$

Linear Discriminant Analysis Model (LDA)
LDA is a classification model used to classify the instance either as benign or malignant. LDA is suitable for dimensional reduction. i.e. from one dimension to the lower dimension of space. The component axis that maximizes the variant of space work is determined by the LDA classifier. Also, the axis which maximizes the separation among the different output classes is determined by the LDA. LDA does its classification based on the output class that has the highest probability. Again, it is a classification method with interpretation probability. Equation 6 below depicts how the maximum score function is calculated where φ represents the linear model coefficient, β represents the average vector and θ depicts the covariance matrix.

$$\varphi = \theta^{-1}(\beta_1 - \beta_2) \quad (6)$$

$$\theta = \frac{1}{n_1 + n_2}(n_1\theta_1 + n_2\theta_2) \quad (7)$$

The best discriminant between the two groups (benign and malignant) is calculated with the Mahalanobis equation as represented in Equation 8 below.

$$\Delta^2 = \varphi^T(\beta_1 - \beta_2) \quad (8)$$

$$\varphi^T(X - \frac{(\beta_1 + \beta_2)}{2} > log\frac{P(\theta_1)}{P(\theta_2)} \quad (9)$$

The difference between the two groups is represented by , and X represents the data vector while the class probability is represented by P. Finally, if the last Equation 9 is satisfied, the classification of a new feature is done.

Linear Regression (LR) Model
LR is very useful when it comes to finding the relationship between attributes. The relationship can either be dependent or independent which can be determined by the LR model. In the WBC dataset, the class attribute is the independent variable while other variables present in the dataset are the dependent variables. In this regard, LR determines the types of cancer based on the attribute that is highly related to the class variable.

Performance Metrics
This describes all the metrics used to evaluate all the six models used in this research work. The following terms False Negative (FN), True Negative (TN), False Positive (FP), and True Positive (TP) made our accuracy, precision, and recall.
Accuracy: is defined as the degree of correctness of classification among the wrong ones. Accuracy also talks

about the overall performance of a model generally. Equation 10 depicts the formula

$$\frac{TP + TN}{TP + TN + FP + FN} \quad (10)$$

***Recall*:** some researchers also call it sensitivity, it is defined as the ratio of true positive over all other observation.

$$\frac{TP}{TP + TN} \quad (11)$$

***Precision*:** This is defined as the ratio of true positive to all positive predictions as depicted in Equation 12 below.

$$\frac{TP}{TP + FP} \quad (12)$$

Experimental Setup

The programming aspect of the program was carried out in a python programming environment. Pandas, numpy, and Scikit-learn were also used in the research work. Jupyter Notebook which is an open-source web application was used to run our program. 10-fold cross-validation was used immediately after the data preprocessing phase. This was done so that we can minimize the bias that is connected with a random sampling of the training dataset. 10-fold cross-validation means that the dataset was partitioned into 10 places, where nine out of the dataset was used for training and the last used for testing the models.

RESULTS AND DISCUSSION

This section discusses the result obtained from each of the models. The six models were validated and the results were compared with each other. Table 2 below depicts the results after experimenting in a python environment. The results indicate that all the models perform well for breast cancer detection but Linear Regression (LR) gave the highest accuracy. LR gave 98.13% accuracy, 97.53% recall, and 98.94% precision. Support Vector Machine (SVM) gave the second detection accuracy of breast cancer with 97.90% accuracy, 98.24% recall, and 98.21 precision. From the table, it was observed that KNN gave the third accuracy level of breast cancer with an accuracy of 96.48%, 92.31% recall, and 95.17% precision. Finally, DT gave the least accuracy as compared with other models in this experiment. DT gave 92.48% accuracy, 97.53% recall, and 98.94 precision. Figures 5, 6, and 7 below depict the charts for accuracy, recall, and precision respectively.

Table 2: Depicts the performance metrics used to evaluate our models for breast cancer.

Classifiers	Precision	Recall	Accuracy
LR	0.9894	0.9753	0.9813
LDA	0.9827	0.9047	0.9578
KNN	0.9517	0.9231	0.9648
NB	0.9218	0. 9365	0.9454
SVM	0.9821	0.9824	0. 9790
DT	0.9894	0.9753	0.9248

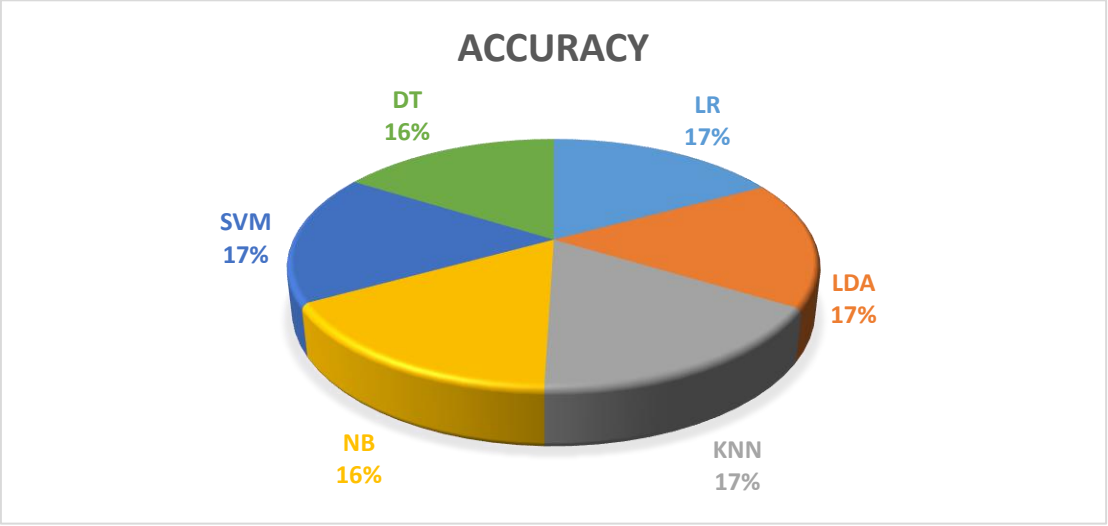

Figure 5: Depicts the accuracy level of all the six models

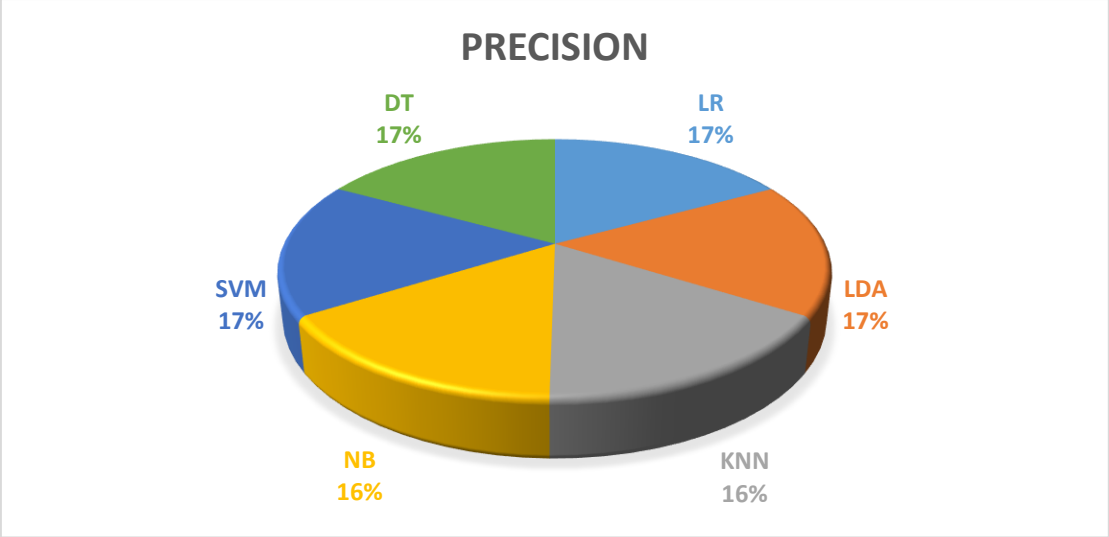

Figure 6: Depicts the precision level of all the six models

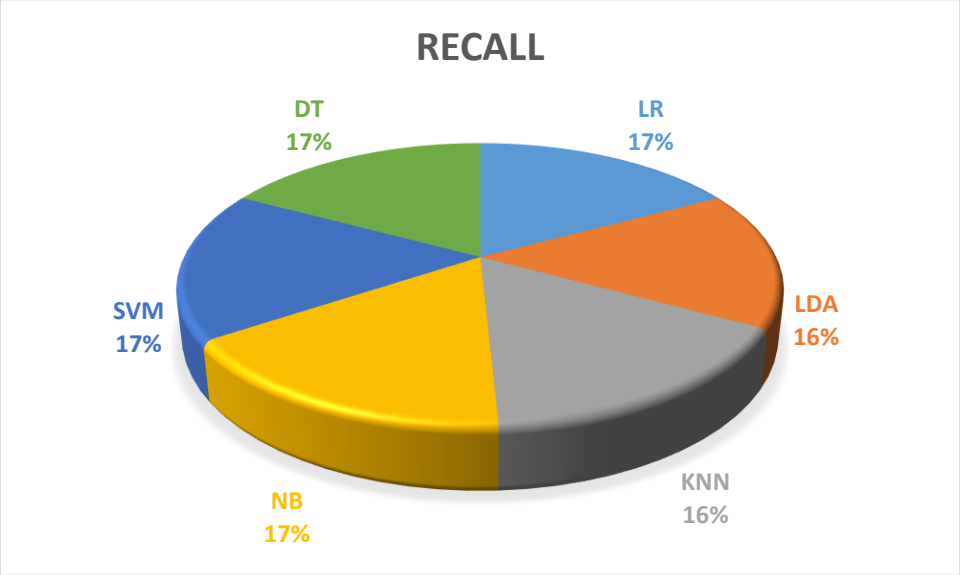

Figure 7: Depicts the recall level of all the six models

CONCLUSION

This research work was necessitated to help increase the detection rate of breast cancer among women in the world. Six different models were used after the WBC dataset has been preprocessed to evaluate the models. Experiment results indicate that LR gave the highest accuracy for breast detection. Our future research plan will be to use an ensemble model to improve the accuracy gotten so far. Also, this ensemble model will be used on different two or three datasets so as to ascertain the effectiveness of the model.

REFERENCES

Adebiyi, M.O.; Arowolo, M.O.; Mshelia, M.D.; Olugbara, O.O. (2022) A Linear Discriminant Analysis and Classification Model for Breast Cancer Diagnosis. Appl. Sci. 2022, 12, 11455. https://doi.org/10.3390/ app122211455

Asri, H., Mousannif, H., Al, M.H., Noel, T.(2016) Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Comput. Sci. 83, 1064–1069

Chaurasia, V., Pal, S. (2017). A novel approach for breast cancer detection using data mining techniques. Int. J. Innovative Res. Comput. Commun. Eng. 2 (An ISO 3297: 2007 Certified Organization)

Dana Bazazeh and Raed Shubair (2016). Comparative Study of Machine Learning Algorithms for Breast Cancer Detection and Diagnosis. 978-1-5090-5306-3/16/$31.00 c 2016 IEEE

Hazra, A., Mandal, S.K., Gupta, A. (2016) Study and analysis of breast cancer cell detection using Naïve Bayes, SVM and Ensemble Algorithms. Int. J. Comput. Appl. 145, 0975–8887

Mohammed, S.A., Darrab, S., Noaman, S.A., Saake, G. (2020). Analysis of Breast Cancer Detection Using Different Machine Learning Techniques. In: Tan, Y., Shi, Y., Tuba, M. (eds) Data Mining and Big Data. DMBD 2020. Communications in Computer and Information Science, vol 1234. Springer, Singapore. https://doi.org/10.1007/978-981-15-7205-0_10

Nahla F. Omran, Sara F. Abd-el Ghany, Hager Saleh, and Ayman Nabil (2021). Breast Cancer Identification from Patients' Tweet Streaming Using Machine Learning Solution on Spark. Hindawi Complexity Vol. 2021, Article ID 6653508

Olofintuyi S.S; Olajubu E.A; Olanike D. (2023). An ensemble deep learning approach for predicting cocoa yield. Heliyon. 2023 Apr 5;9(4):e15245. Doi: 10.1016/j.heliyon.2023.e15245.

Olofintuyi, S.S. (2021). Cyber Situation Awareness Perception Model for Computer Network. International journal of advanced computer science and application. 12(1), pp. 392-397.

Olofintuyi S.S and Olajubu E.A (2021). Supervised Machine Learning Algorithms for Cyber-Threats Detection in the Perception Phase of a Situation Awareness Model. International Journal of Information Processing and Communication (IJIPC) Vol. 11 No. 2 [December, 2021], pp. 61-74

Olofintuyi, S.S., Omotehinwa, T. O., Odukoya, O.H. and Olajubu, E. A. (2019). Performance comparison of threat classification models for cyber-situation awareness. Proceedings of the OAU Faculty of Technology Conference, 305-309.

O. S. Samuel (2022). "Early Cocoa Blackpod Pathogen Prediction with Machine Learning Ensemble Algorithm based on Climatic Parameters", *J. inf. organ. sci. (Online)*, vol. 46, no. 1, .

Ojha U., Goel, S. (2017). A study on prediction of breast cancer recurrence using data mining techniques. In: 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence, IEEE, pp. 527–530

Pritom, A.I., Munshi, M.A.R., Sabab, S.A., Shihab, S. (2016) Predicting breast cancer recurrence using effective classification and feature selection technique. In: 19th International Conference on Computer and Information Technology (ICCIT), pp. 310–314. IEEE

Rodrigues, B.L. (2015). Analysis of the Wisconsin Breast Cancer dataset and machine learning for breast cancer detection. In: Proceedings of XI Workshop de Visão Computational, pp. 15–19 (2015)

Saabith, A.L.S., Sundararajan, E., Bakar, A.A.(2014): Comparative study on different classification techniques for breast cancer dataset. Int. J. Comput. Sc. Mob. Comput. 3(10), 185–191

Siham A. Mohammed, Sadeq Darrab , Salah A. Noaman, and Gunter Saake (2020). Analysis of Breast Cancer Detection Using Different Machine Learning Techniques. DMBD CCIS 1234, pp. 108–117, https://doi.org/10.1007/978-981-15-7205-0_10

Silva, J., Lezama, O.B.P., Varela, N., Borrero, L.A. (2019). Integration of data mining classification techniques and ensemble learning for predicting the type of breast cancer recurrence. In: Miani, R., Camargos, L., Zarpelão, B., Rosas, E., Pasquini, R. (eds.) GPC 2019. LNCS, vol. 11484, pp. 18–30. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-19223-5_2

W. Lim, S. Hamid, and M. Grivna (2016). Breast cancer presentation delays among Arab and national women in the UAE, a qualitative study, SSM - Popul. Heal., Mar. 2016

WHO — Breast Cancer: Prevention and Control (2020) Retrieved 20 Jan 2023, from WHO — World Health Organization. http:// www.who.int/cancer/detection/breastcancer/en/index1.html