DAFFODIL INTERNATIONAL UNIVERSITY JOURNAL OF SCIENCE AND TECHNOLOGY, VOLUME 16, ISSUE 2, JULY 2021
ISSN 2408-8498 (Online)

A MODEL FOR THE CLASSIFICATION OF BREAST CANCER USING RANDOM FOREST ALGORITHM

Akinyemi Moruff Oyelakin
Department of Computer Science, Faculty of Natural and Applied Sciences, Al-Hikmah University, Ilorin, Nigeria.
Email:amoyelakin@alhikmah.edu.ng

Abstract: **Breast cancer is a common disease among women globally. Past studies have used Machine learning techniques to speed up the prediction of the disease using labeled datasets. This study proposed a supervised machine learning approach for the classification of breast cancer. The model was built using Random Forest Algorithm. The dataset chosen for this study is a Wisconsin breast cancer (Diagnostic) dataset. The breast cancer dataset was originally released by the University of Wisconsin Hospitals, Madison. Python programming language and some of its libraries were used for the experimental analyses. The dataset was split in the ratio 75:25 percent as training and testing sets respectively. The metrics used for the performance evaluation of the model built include: accuracy, precision, recall, f1-score, and Cohen's Kappa Statistics. In the experimental analyses, accuracy of 96% was recorded. 98% was obtained for the precision. For the recall, 96% was obtained. Moreso, 97% was obtained for F1-score while 91% was recorded for Cohen's Kappa Statistics. The model provides superior classification performance in terms of the chosen evaluation metrics.**

Keywords: *Breast Cancer Classification, Machine Learning, Feature Selection, Predictive Accuracy*

I. INTRODUCTION

The diagnosis of any disease early can make it to be more treatable with a little amount of human effort. In real life situations, diagnosis and detection of cancerous growth may be too late and as such the disease becomes chronic. Breast cancer is one of the diseases that are often not properly diagnosed at early stage. Several studies have investigated the detection or classification of breast cancer using Machine Learning algorithms [1, 2]. Breast cancer is a type of cancer that starts in the breast. The cancer starts when cells begin to grow out of control [3, 4]. Breast cancers come in different forms. The most common symptom of breast cancer is a new lump or mass, but other symptoms are also possible. It's important to have any breast change checked by a health care provider [3].

In medical domain, the diagnosis of breast cancer is carried out when an abnormal lump is observed either from self-examination or through X-ray images [3]. In breast cancer examination, if a suspicious lump is found, the medical doctor is expected to conduct a diagnosis with a view to determining whether it is cancerous or not. He will equally check if the cancer has spread to other parts of the body or not. This study proposes to use a Machine Learning algorithm named Random forest for the classification of breast cancer in the chosen Wisconsin (Diagnostic) dataset.

Machine Learning for Classification

Machine Learning algorithms are widely used for classification problems in several domains [5]. Learning algorithms can be grouped into: Supervised, Unsupervised and Semi Supervised [6]. Several studies have adopted these machine learning approaches in the detection of breast cancer. Breast cancer is one of the most dangerous and common reproductive cancers that affect mostly women [7]. Greene et al. in [8] pointed out that in machine learning, data plays an indispensable role, and the learning algorithm is used to discover and learn knowledge or properties from the data. There are different complexities that are taking into consideration in a Breast cancer research. For instance, taken the size of the breast cancer into consideration is very important. The size of a breast cancer indicates how large across the tumor is at its widest point. Doctors measure cancer in millimeters (1 mm = 0.04 inch). In most cases, size is used to help determine the stage of the breast cancer [9].

Random forest algorithm works by combining multiple algorithms of the same type while making classification or regression [5]. The algorithm applies the technique of bagging to decision tree learners to do this. In this work, automatic feature selection was arrived at as a result of the Random Forest algorithm chosen for the classification [4]. A random forest classifier combines the results from different tree-based models to achieve the classification. The proposed Random forest-based model focuses on classifying breast data into malignant or benign by using the selected features for the training and testing.

II. RELATED WORKS

Shamy et al. in [10] used K-Means and Convolutional Neural Network algorithms for the detection of Breast cancer. They argued that the proposed method was evaluated using Mammographic Image Analysis Society MIAS dataset. The technique achieved 95.8% accuracy. Similarly, Alom et al. in [11] classified breast cancer from histopathological images by using Recurrent Residual Convolutional Neural Network architecture. The emphasis of the study was on comparing the selected algorithms. The authors were silent on the issue of feature selection.

Nahid et al. in [1] carried out a survey of different Machine Learning techniques that have been used for classification of breast cancer in the past. Moreover, Dharhi in [12] used genetic programming and machine learning algorithms classifying benign and malignant breast tumors. The genetic programming technique was used to select the best features and perfect parameter values of the machine learning algorithms. The authors argued that the proposed method was based on sensitivity, specificity, precision, accuracy, and the roc curves. The study argued that genetic programming can automatically find the best model by combining feature preprocessing methods with the chosen classifier algorithms.

Eleyan in [13] presented two different Machine Learning classifiers for Breast cancer classification. The algorithms used in the study include Naive Bayes (NB) classifier and K-nearest neighbor (KNN). The focus of the study is comparison between the two new implementations and evaluates their accuracy using cross validation. Results show that KNN gives the highest accuracy (97.51%) with lowest error rate then NB classifier (96.19 %). Moreover, Tike-Thein et al. in [14] proposed an approach for breast cancer distinguishing between different classes of breast cancer. This approach is based on the Wisconsin Diagnostic and Prognostic Breast Cancer and the classification of different types of breast cancer datasets. The proposed system implements the island-based training method to be better accuracy and less training time by using and analyzing between two different migration topologies. However, the study used limited number of performance metrics for the evaluation.

Sivapriya et al. in [15] carried out a study for the classification of Breast cancer. The study made use of four Machine Learning Algorithms. However, the study is limited to investigation of the performances of these algorithms by using only Accuracy and running time as the performance metrics. Jabar in [16] proposed a decision support system by using the ensemble model. The ensemble model was built with Bayesian network and Radial Basis Function. The author argued that his approach is effective in cancer classification. In their work, the researchers in [17] used Deep Neural Networks to classify the presence of breast cancerous growth.

III. METHODOLOGY

Random Forest algorithm was chosen for building the predictive model in this study. As a supervised machine learning algorithm, the algorithm was employed to correctly classify breast cancer evidence in the dataset [1, 18]. The trained algorithm classifies the labels into malignant (cancerous) or benign (non-cancerous) using the selected features in the dataset. The experimentation was carried out using Python 3.7 and its set of libraries. Windows 10 operating system was used as the platform.

i) Description of the Dataset

The dataset chosen for this study is a Wisconsin breast cancer (Diagnostic) dataset that was obtained through UCI repository as released by [19]. The breast cancer dataset was originally released by the University of Wisconsin Hospitals, Madison. The dataset consists of features which were computed from digitized images of FNA tests on a breast mass [19]. The discriminative breast cancer features are used for training and testing machine learning-based models. The attributes give a description of characteristics of the cell nuclei present in the images. A dataframe showing a preview of the instances in the dataset as obtained in Python language environment is as shown in figure 1. As can be seen in the figure, the dataset contains numerical input features and a categorical target feature.

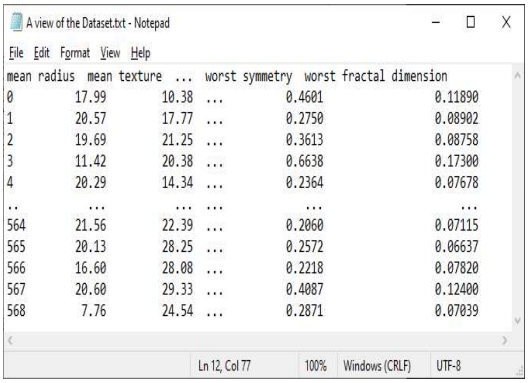

Figure 1: The Data frame in the dataset (Source: Exploratory Data Analysis)

ii) Methodological Flow

The steps followed in the classification are as shown in figure 2.

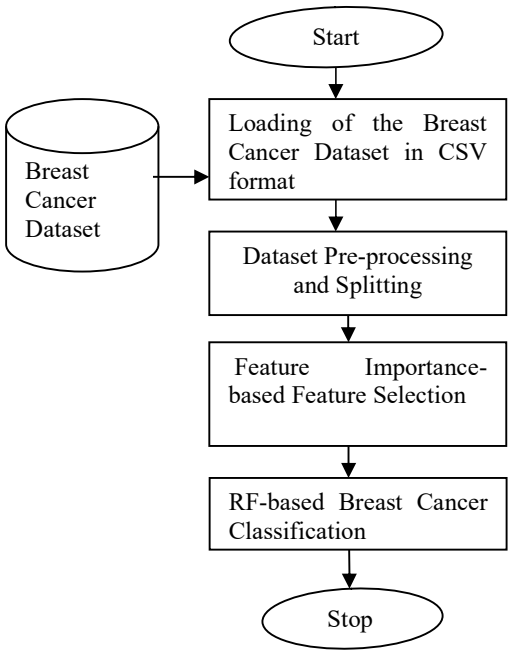

Figure 2: Flow of the processes in the work

iii) Dataset Preprocessing and Splitting

Exploratory analysis of the dataset revealed that the dataset contains no missing values. However, there are a few rows with a special character named "Question mark". The data pre-processing approach used involves deletion of few rows that has the special character. This was done so to put the dataset in usable format for the machine learning classifier. In every classification problem, the purpose of the training phase is to create detection models that can identify the class of detected botnet. Similarly, testing data is the subset of our data that the model has not been seen before. For this reason, the dataset was split in the ratio 75:25 as training and test sets respectively, during the experimentation.

iv) About Random Forest Algorithm

Random forest is a type of supervised bagged ensemble algorithm that is used for classification and regression problems. It builds multiple decision trees and merges them together to get a more accurate and stable prediction. Random forest classifier has in-built feature selection capability called feature importance. The summary of the algorithm is as shown in algorithm 1.

Algorithm 1: Random Forest Algorithm for Breast Cancer Classification

Start
Input: A set of random samples from dataset D
Output: Binary classification (of breast cancer)
Step 1: Pick N random samples from the breast cancer dataset.
Step 2: Build a decision tree based on these N samples (instances).
Step 3: Choose the number of trees required for the problem and repeat steps 1 and 2.
Step 4: In case of this classification problem, each tree in the forest predicts the category to which the new sample belongs.
Step 5: Assign the new sample to the category that wins the majority vote.
Stop

(Source: Author algorithm description of the classifier)

v) Feature Selection Approach

Feature selection was carried out using a feature selection technique named feature importance. The feature selection method is embedded in Random forest algorithm. The selection algorithm computes the importance of each feature. These features are ranked in their order of importance. The selected features were then used in the RF-based breast cancer classification process.

IV. RESULTS AND DISCUSSION

i) Results-Experimental Evaluation

(a) Confusion Matrix

A confusion matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. The table provides insights into the errors being made by a classifier. It also gives the types of errors that are being made. The values in the general table are taken as True Positives, True Negatives, False Positives and False Negatives.

True positives (TP): These are cases in which it is predicted breast cancer is present and the patients do have the disease.

True negatives (TN): These are situations in which the model predicted no, and the patients don't have the disease.

False positives (FP): These are situations in which the proposed model predicted yes, but the patients don't actually have the disease.

False negatives (FN): These are situations in which the model predicted no, but the patients actually do have the disease.

TABLE 1: COMPUTING CONFUSION MATRIX FORMULA

N=No of Predictions	**Predicted Value**	
Actual Value	**Non-cancerous**	**Cancerous**
Non-Cancerous	TN	FP
Cancerous	FN	TP

TABLE 2: CONFUSION MATRIX FOR THE CLASSIFICATION

N=143	**Predicted Value**	
Actual Value	**Non-cancerous**	**Cancerous**
Non-cancerous	51	2
Cancerous	4	86

The cancer classification is a binary classification problem. This is as a result of the nature of the chosen dataset.Table2 shows the proportion of the samples that fall into the categories of 'Cancerous' or 'Non-Cancerous' for both predicted and actual values. Table 3 was obtained from the experimental analyses in this study. The predictive values of the algorithm using five different metrics are as shown in table 3.

ii) Model Performance Metrics

The metrics used to measure the performance of the proposed model include: Accuracy, Precision, Recall, F1-Score and Cohen's Kappa Statistics [6]. Mathematically, each of the metrics used in this study can be computed using the formulae shown below:

$$\text{Accuracy}=(TP+TN)/(TP+TN+FP+FN) \quad (1)$$

$$\text{Precision}=TP/(TP+FP) \quad (2)$$

$$\text{Recall}=TP/(TP+FN) \quad (3)$$

$$\text{F1 score}=2\times(\text{Precision X Recall})/(\text{Precision} + \text{Recall}) \quad (4)$$

$$\text{Cohen's Kappa}= (Pr(a)- Pr(e))/N-Pr(e)) \quad (5)$$

The values of True Negative (TN), False Positive (FP), False Negative (FN) and True Positive (TP) are as shown in table 2.

The classification results of the breast cancer based on the selected metrics are shown in Table3.

TABLE 3: SUMMARY OF THE CLASSIFICATION RESULTS

S/N	**Performance Metrics**	**Results**
1	Accuracy	0.97
2	Precision	0.98
3	Recall	0.96
4	F1-Score	0.97
5	Cohen's Kappa Statistics	0.91

In table 3, all the values obtained are taken to the nearest two decimal places. The values were calculated from equations 1 to 5 respectively.

iii) Discussions of results

The values the for True Negative, False Positive, False Negative and True Positive are computed as shown in table 2. The results obtained from the experimental analysis showed that a predictive accuracy of 0.97 was arrived at using Random Forest Algorithm. 0.98 was obtained for precision. During experimentation, the value obtained for recall was 0.96. 0.97 was obtained for F1 score while the value obtained for Cohen Kappas's statistics was 0.91. The predictive ability of the selected machine learning algorithm is marginally better when compared with some similar studies reviewed. The study laid emphasis on testing the built classifier across five different metrics with a view to having a more comprehensive performance measures compared to some similar studies.

V. CONCLUSION

The work demonstrated the use of a machine learning technique for the classification of breast cancer. The problem is a two-class problem. The improvement achieved in the classification is at data-preprocessing and feature selection stages. The Random Forest-based model was trained on the seventy-five percent training set while the remaining twenty-five percent was used as test set. The study made use of five different performance metrics and their various values were computed and reported. From the results obtained, this work gave further insights on the detection of breast cancer using Random Forest Machine Learning algorithm. The results obtained are better when compared to similar studies reviewed.

REFERENCES

[1] A. Nahid & Y. Kong, "Involvement of Machine Learning for Breast Cancer Image Classification: A Survey", 2017. Available: *Computational and Mathematical Methods in Medicine*, *2017*. https://doi.org/10.1155/2017/3781951, accessed October, 2020

[2] R. Chtihrakkannan, P. Kavitha, T. Mangayarkarasi, & R. Karthikeyan," Breast cancer detection using machine learning", *International Journal of Innovative Technology and Exploring Engineering*, 2019, *8*(11), 3123–3126. https://doi.org/10.35940/ijitee.K2498.0981119

[3] American Cancer Society, "About Breast Cancer. *Breast Cancer Facts and Figures"*, 2017, 1–19. Available: https://www.cancer.org/content/dam/CRC/PDF/Public/8577.00.pdf%0Ahttp://www.breastcancer.org/symptoms/understand_bc/what_is_bc, retrieved on 7th September, 2020

[4] S. Chopra, "An Introduction to Building a Classification Model Using Random Forests in Python", Learn Data Sciences", 2019 Available: https://blogs.oracle.com/datascience/an-introduction-to-building-a-classification-model-using-random-forests-in-python, Accessed on 2. nd August, 2020

[5] M. Swamynathan, "Mastering Machine Learning with Python in Six steps", 2019, Available: *DOI:10.1007/978-1-4842-2866-1_3, Accessed on 7th September, 2020*

[6] M. A. Hall, "Correlation-based Feature Selection for Machine Learning", *a PhD Thesis at University of Waikato, 1999*

[7] S. Shahnorbanun, Q. Ashwaq, O. Khairuddin, A. Dheeb, A. Afzan, N. H. S. Siti, Azizi Abdullah, I. H. Rizuana, I. Fuad, A. Norlia, H. M. Suria, M. P. Hayati and A. Nurdashima, "Machine Learning Methods for Breast Cancer Diagnostic", 2018, Avialable: http://dx.doi.org/10.5772/intechopen.79446, retrieved on 13th September, 2020

[8] L. R. Greene & D. Wilkinson, "The role of general nuclear medicine in breast cancer", Journal of Medical Radiation Sciences, 62(1): 54–65. Available:https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4364807/

[9] Weiss Marisa, "Size of the Breast Cancer", Available: Breast Cancer.org, https://www.breastcancer.org/symptoms/diagnosis/size, Accessed 4th October, 2020

[10] S. Shamy, & J. Dheeba, "A research on detection and classification of breast cancer using k-means gmm & CNN algorithms", *International Journal of Engineering and Advanced Technology*,2019, *8*(6 Special Issue), 501–505. https://doi.org/10.35940/ijeat.F1102.0886S19

[11] M. Z. Alom., C.Yakopcic, M. S. Nasrin, T. M. Taha, & V. K. Asari, "Breast Cancer Classification from Histopathological Images with Inception Recurrent Residual Convolutional Neural Network. *Journal of Digital Imaging"*, *32*(4), 605–617.Available: https://doi.org/10.1007/s10278-019-00182-7, retrieved on 5th October, 2020

[12] H. Dhahri, E. Al Maghayreh, A. Mahmood, W. Elkilani, & Nagi M. Faisal, "Automated Breast Cancer Diagnosis Based on Machine Learning Algorithms", 2019, *Journal of Healthcare Engineering*, *2019*. Available: https://doi.org/10.1155/2019/4253641, accessed on 12th November, 2020

[13] A. Eleyan. "Breast cancer classification using moments, in the proceedings of Signal Processing and Communications Applications Conference (SIU), 2012, Available: DOI: 10.1109/SIU.2012.6204778:

[14] H.T.Tike Thein, & K. M. Mo Tun, "An Approach for Breast Cancer Diagnosis Classification Using Neural Network". *Advanced Computing: An International Journal*, 2015 *6*(1), 1–11. Available: https://doi.org/10.5121/acij.2015.6101, Accessed on 4th November, 2020

[15] J. Sivapriya, Kumar V. Aravind, Sai S. Siddarth & S. Sriram, "*Breast Cancer Prediction using Machine Learning*", International Journal of Recent Technology and Engineering (IJRTE), 2019,8(4), 4879–4881. https://doi.org/10.35940/ijrte.D8292.118419

[16] M. A. Jabbar, "Breast Cancer Data Classification Using Ensemble Machine Learning", *Engineering and Applied Science Research*, *48*(1), 65-72. Available: https://ph01.tci-thaijo.org/index.php/easr/article/view/234959, retrieved on 21 Feb 2021

[17] S. Karthik, P. R. Srinivasa. and M. P. Chandra Mouli, "Breast Cancer Classification Using Deep Neural Networks", In S. Margret Anouncia and U. K. Wiil (eds.), Knowledge Computing and Its Applications,2019 *Volume 1*, 1–293 Available:https://doi.org/10.1007/978-981-10-6680-1_12, accessed 3rd July 2020

[18] Yue Wenbin, Wang Zidong, Chen Hongwei, Payne Annette & Liu Xiaohui, "Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis", 2018 1–17. https://doi.org/10.3390/designs2020013

[19] W. H. Wolberg, W. Nick Street, & Mangasarian L. Olvi, "UCI Machine Learning Repository" 1995, Available:https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic), accessed on 3rd September, 2020

