ANALYSING THE PERFORMANCE OF MACHINE LEARNING ALGORITHMS FOR EFFECTIVE CLASSIFICATION OF BREAST CANCER

M.M.Achini Nisansala

Department of ICT, University of Vavuniya, Sri Lanka

nisansala491@gmail.com

ABSTRACT: *Breast cancer is the most common type of cancer diagnosed in women throughout the world. It can occur at any age in women's lives, but the risk increased with the age. In 2020 around 2.3millions of women are diagnosed with breast cancer and among them, around 0.68 million died globally. There are two types of breast cancer tumors: benign and malignant. Diagnosing breast cancer is kind of tough due to the compound nature of the breast cancer cells. However, the treatments for breast cancer are very effective when the disease is diagnosed at an early stage. In this study seven machine learning algorithms are used: Logistic Regression (LR), Linear Discriminant Analysis (LDA), K-Nearest Neighbor (KNN), Gaussian Naïve Bayes (GN), Decision Tree Classifier (C4.5), Support Vector Classifier (SVC) and Random Forest (RF) on Wisconsin Breast Cancer Dataset (WBCD) collected from UCI repository for classifying the tumors into benign and malignant. This analysis is carried out in two parts without removing the outliers from the dataset and after removing the outliers from the dataset. Based on the analysis without removing the outliers SVC outperforms other classifiers with 97.82% accuracy. After removing the outliers RF gives the highest accuracy of 96.18%.*

Keywords: Breast cancer, Classification algorithms, accuracy

1. INTRODUCTION

In previous years the number of patients suffering from cancer diseases has increased rapidly. This makes cancer, the second leading cause of death throughout the world Among them the most common type of cancer affecting women is breast cancer. In 2020 around 2.3 million women were diagnosed with breast cancer and 685 000 deaths have been reported (WHO | Breast Cancer, 2021). There are two types of breast cancers: Benign tumors and malignant tumors. Benign tumors are considered noncancerous tumors or less harmful tumors as they are growing very slowly and do not spread. But the malignant tumors enlarge very fast and they invade and damage other healthy tissues and expand throughout the body (Stanford Health Care, 2022). To escalate the survival rate of breast cancer, early detection is the most important thing. In order to detect breast cancer patients has to go through several medical examinations as these tumors are very hard to detect even by specialists in the field. Mammography, biopsy, and ultrasound are some examination types.

By taking the microscopic images numerical features like area, texture, perimeter, and radius of the cells and tissues are calculated. This paper mainly addresses the comparison of the performance and the accuracy level between seven machine learning algorithms: Logistic Regression (LR), Linear Discriminant Analysis (LDA), K-Nearest Neighbors (KNN), Gaussian Naïve Bayes (GN), Decision Tree (CART), Support Vector Classifier (SVC) and Random Forest (RF) for accurately determining the benign and malignant tumors. These algorithms are the most appropriate algorithms to solve categorical data-related classification problems. The performance of these seven algorithms on the Wisconsin Breast Cancer Dataset (WBCD) is taking in two procedures:

- All the data in the WBCD are taken.
- Remove the outliers of the dataset by taking the interquartile range and taking the remaining data.

Outliers are any observations that gives some abnormal values that do not fall within the expected distribution of particular data values. In most of the occasions these may be some errors happened while collecting the data while in some occasions these can be due to the varying body structures such as very high obesity and skinny. Besides both of mentioned facts, comparing the accuracy and finding best approach may help in finding best method for overall body structures and average level body structures separately. The accuracy and the performance of these two approaches will be compared and determine the best method to classify benign and malignant tumors

The remainder of this work is organized as follows. The literature review is presented in section II. The overall methodology is explained in Section III. Section IV discussed the results obtained. And at last Section V conclude the entire work.

2. LITERATURE REVIEW

Previous works have conducted several experiments and developed different models using Machine Learning (ML) and Deep Learning (DL) approaches on medical datasets of breast cancers. Ara,Das, & Dev (2021) uses a correlation bar plot and eliminates less correlated features for increasing accuracy. Six machine learning approaches: Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Naïve Bayes (GN), Decision Tree (C4.5), and K-Nearest Neighbor (KNN) have been used and the maximum accuracy of 96.5% has been achieved in Support Vector Machine (SVM) and Random Forest (RF) .Authors Nasien, Enjeslina, Adiya & Baharum (2022) used Artificial Neural Network(ANN) back propagation method using MATLAB R2016a software and achieved the best accuracy of 96.929% with 1000 epochs, and learning rate of 0.01, and a goal of 0.001 and hidden layer five. Naji et al. (2021b) use K-Nearest Neighbor (KNN), Naïve Bayes(NB), Decision Tree(C4.5), and simple logistic and ensemble methods like Majority Voting and Random Forest with 10 cross-field techniques on the Breast cancer dataset of the UCI repository. The majority ensemble technique reaches an accuracy of 98.1 % with the least error rate of 0.01% and surpasses all other algorithms. Ming et al.(2019) compared ML-based estimates and estimates from the Breast Cancer Risk Assessment Tool (BCRAT)model and Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm (BOADICEA) model using eight synthetic simulated datasets and two actual observational datasets with the same risk factor for all the Machine Learning algorithms as in BCRAT and BOADICEA models.

Naji et al.(2021a) applied five machine learning classifiers Support Vector Machine(SVM), Random Forest(RF), Logistic Regression(LR), Decision Tree(C4.5), and K-Nearest Neighbor(KNN) on the WBCD dataset. They evaluated and compare the models using the performance matrix: confusion matrix, accuracy, precision, sensitivity, F1 score, and AUC and found. Support Vector Machine (SVM) surpasses all other classifiers giving 556 correct predictions for the confusion matrix,0.98 precision 0.94 sensitivity, 0.96 F-measure, and 0.96 ROC. Ayyoubzadeh, Sohrabei, Esmaeiii, & Atashi (2022) used Synthetic Minority Oversampling Technique (SMOTE) for balancing the training data as the class records were not balanced. They used three learners Random Forest (RF), Gradient Boosting Tress(GBT),

ISBN: 978-955-627-013-6

and Multi-Layer Perceptron(MLP) for applying to the dataset, and 3-fold validation was used for getting the optimized hyperparameter for each model. Performance comparison was carried out using demographic features only and as a combination of demographic features and mammographic features. The optimized ROC of 0.974, the accuracy of 95%, the sensitivity of 96.14%, and the specificity of 93.94% were in RF when the model was optimized by genetic algorithm (GA). Aamir et al. (2022) used a hybrid of correlation-based feature elimination strategy and recursive feature elimination for the best selection of optimal features. Five ML methods Random Forest (RF), Artificial Neural Network (ANN), Gradient Boosting, Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP) were used and classification accuracy was taken for three train-test split sizes as 60-40 70-30 and 80-20. The best accuracy of 99.12% was achieved by the MLP model, and the best train-test split was 80-20.

Saleh, Abd-el ghany, Alyami, & Alosaimi(2022) followed two approaches: the regular Machine learning (ML) approach, and the deep learning approach (DL)for predicting breast cancer. They used two types of feature selection algorithms: univariant feature selection and recursive feature elimination (RFE). Deep RF achieved the best performance using univariant giving an accuracy of 99.89%, precision of 99.89%, recall of 96.74%, and F1 score of 99.89%. The same performance was recorded for correlation and RFE. Deep RNN also achieved the best performance using univariant with an accuracy of 96.74%, precision of 96.39%, recall of 96.74%, and F1 score of 96.8%. The same performance was recorded for both correlation and RFE. In recent work, Khourdifi (2018) experimented on a breast cancer dataset of Wisconsin with 569 data and 30 attributes using machine learning algorithms K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Random Forest (RF) and Naïve Bayes (NB). SVM marked the highest correct classifications (557), and the best accuracy rate of 97.9%. They evaluate the effectiveness of classifiers based on the time taken to build the model, the number of correct classifications, incorrect classifications, and accuracy by them.

Risk factors for breast cancer were identified using wrapper-48, wrapper-SVC, wrapper-NB, LR, correlation-based feature selection methods. The performance of five machine learning algorithms was compared before and after the feature section for predicting breast cancer and found that confidence-weighted voting method achieve the best result (Shanbehzaden, Kazemi-Arpanahi, Ghalibah, & Orooji, 2022).The breast cancer prediction done using Rapid miner 7.0 tool for data set assessment and using decision tree and deep learning methods were used to locate features to ensure the patients with malignant tumors from remaining patients. Deep learning algorithms recognized as the best for prediction (Saranya & Sasikala, 2020). Performance and efficiency measured using accuracy, sensitivity, and area under curve have concluded deep learning is the best algorithm for predicting breast cancer. Bayesian classifier is suitable for large scale predictions and classification tasks on complex and incomplete datasets rather than multilayer perceptron classification and C4.5 based on the results gain by doing the classification on WEKA software (Soria, Garibaldi, Biganzoli, & Ellis, 2008).

3. METHODOLOGY

Quantitative Approach

Dataset and Attributes

This research paper uses the publicly available dataset Wisconsin Breast Cancer Dataset (WBCD). This research paper uses the edited dataset collected from the UCI Machine Learning repository_(Wolberg, Street, & Mangasarian, n.d.). The dataset consists of 32 attributes including the id and the diagnosis. Each record consists of two output possibilities: benign or malignant which is included in the diagnosis column. The number of detailed data attributes related to breast cancer is thirty and all these attributes consist of numerical values. These 30 attributes are the average(mean), standard error (SE), and worst where each attribute consists of radius, texture, perimeter, area, smoothness, density, indentation, concave point, symmetry, and fractal dimensions (Nasien, Enjeslina, Adiya, & Baharum, 2022). There are a total of 569 records available 357 benign and 212 malignant cases as shown in Figure 4.

Experimental Environment

All the experiments carried out in this research run on the anaconda environment Jupiter notebook. Machine learning models were implemented using the sci-kit-learn package. Classification reports, confusion matrices, and accuracy score metrics have been used for the performance evaluation of each of the algorithms. Python 3.10.6 was used for full implementation and basic libraries like pandas, matplotlib, NumPy, and seaborn were used.

Model Optimization and Training

This research is carried out in two approaches: Approach_1 and Appriach_2 and compares the accuracy results of both approaches. Then get the best accuracy provided by all the machine learning algorithms in accurately predicting breast cancer. There are seven machine learning algorithms used here. LR,LDA,KNN,C4.5,NB,SVM and RF were used to compare and get the best accuracy. In both of these approaches, the dataset is split into two portions as 80:20 training and testing sets respectively. Models developed by these machine learning algorithms are trained using this training dataset which we split and take. Then, at last, the testing dataset which is a completely new dataset for the model is used to check the performance of the model with new data. Two approaches followed in our study are as follows:

- Approach_1: Without removing any data or the record from the dataset all the data are used in this approach. Scaling of data and hyperparameter tuning methods are used for increasing the accuracy of the model.
- Approach_2: Check for the outliers in the dataset by taking the interquartile range for all the numerical attributes in the dataset. Then these outliers are removed from the dataset. After removing the outliers from the dataset number of records available in the dataset was reduced to 519 total records as 349 benign results and 170 malignant as given in Figure 4.

Increase the accuracy of each model using scaling data and hyperparameter tuning are carried out. Grid search with Cross Validation is used for optimizing machine learning algorithms and for improving the performance of each model. Figure 6 contains the main steps we used in the proposed system of predicting breast cancer.

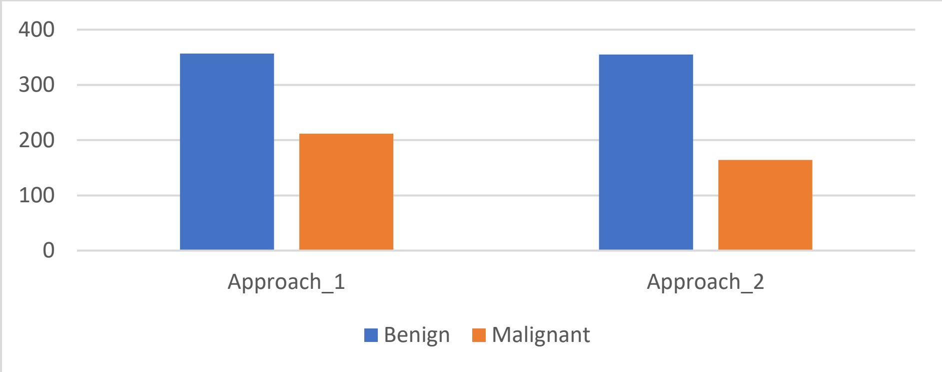

Figure 4.Dataset Distribution

Evaluation of models

These models are evaluated using Accuracy (AC), Precision (PR), Recall (RE), and F1-Score(F1). Equations for each of these performance evaluation methods are given in Table 7.

TP -True Positive (Model predicts as positive and actual value is also positive)
TN-True Negative (Model predicts as negative and the actual value is also negative)
FP-False Positive (Model predicts as positive but the actual value is negative)
FN-False Negative (Model predicts as negative but the actual value is positive)

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

$$PR = \frac{TP}{TP+FP} \quad (2)$$

$$RE = \frac{TP}{TP+FN} \quad (3)$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision+Recall} \quad (4)$$

4. DISCUSSION AND RESULTS

After concluding the implementation of machine learning algorithms on the Wisconsin Breast Cancer Diagnostic dataset (WBCD) different performance metrics such as confusion matrix, accuracy, precision, recall, and F1-score are used for evaluating the performance of each of these models belonging to both approaches we followed in this research. Table 7 summarizes the accuracy obtained by each of the classification algorithms without removing the outliers from the dataset (Approach_1) and after removing outliers from the dataset(Approach_2).

Table 7. Accuracy Comparison in Approach_1 and Approach_2

	Approach_1		Approach_2	
Algorithm	Training accuracy	Testing accuracy	Training accuracy	Testing accuracy
LR	96.94%	95.61%	97.83%	93.55%
LDA	95.61%	87.88%	95.42%	92.45%
KNN	97.15%	93.86%	96.86%	95.36%
GN	94.29%	93.86%	92.79%	95.15%

C4.5	91.43%	92.95%	92.06%	95.18%
SVC	98.02%	97.82%	97.83%	95.36%
RF	95.16%	93.03%	95.91%	96.18%

Based on Table 2 we can find that the best accuracy of 97.82% on the test set is achieved by the Support Vector Classifier (SVC). As well the best accuracy on the training set is also achieved by the Support Vector Classifier and the value is 98.02%. Based on the results we obtained in Approach_1 in our research study we can determine that SVC outperforms all other classification algorithms and shows the maximum accuracy in the first approach. Table 8 gives the classification performance of the Support Vector classifier in Approach_1.

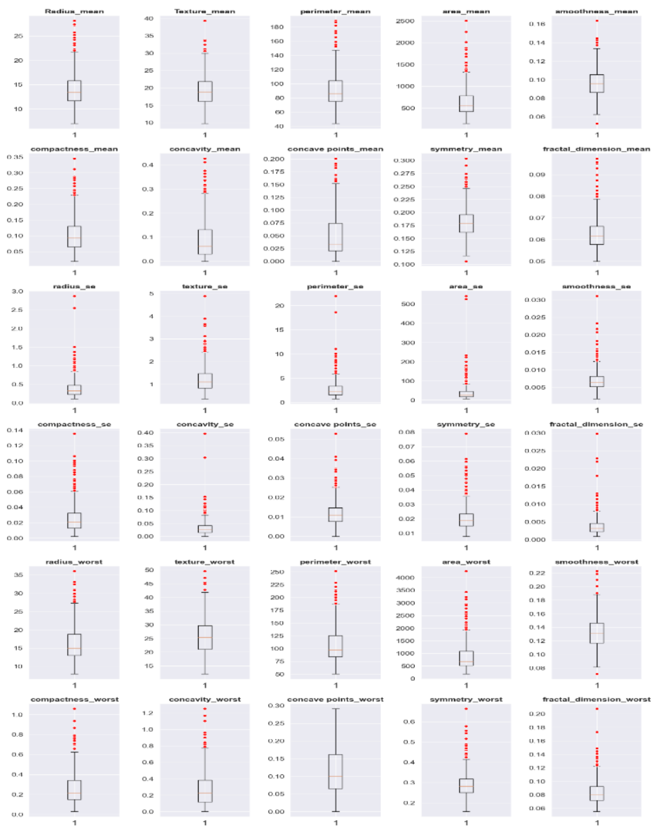

Figure 5.Outlier Detection in Different Attributes

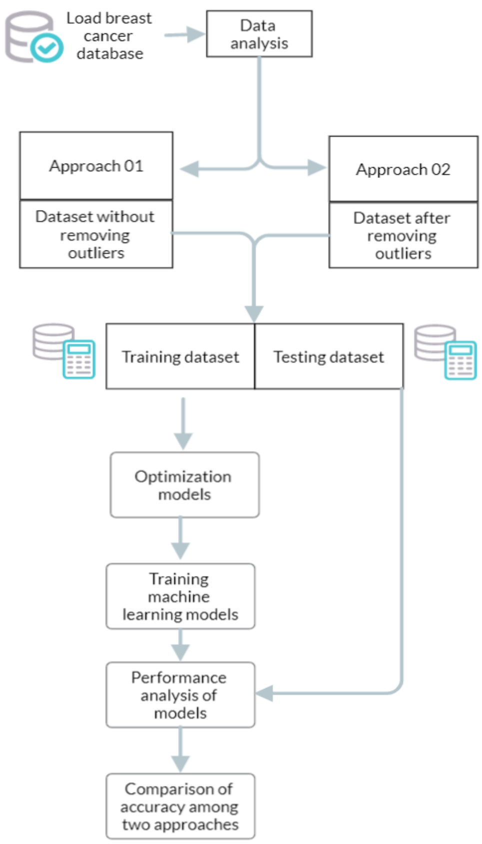

Figure 6.The proposed system for Predicting Breast Cancer

Table 8: Performance classifier of SVC

	Precision	Recall	F1-score	Class
SVC	0.97	1.00	0.99	Benign
	1.00	0.95	0.98	Malignant

From the results of Table 1, we can identify that all different classifiers in Approach_2 have various training and testing accuracies. Among all the classifiers RF gives the maximum accuracy of 96.18% on the testing dataset. When comparing the training set LR and the SVC gives the highest accuracy of 97.83% but the testing accuracy of these two algorithms has been reduced rather than the Random Forest classifier. Table 9 present the calculated performance measures of the Random Forest (RF) which outperform all other algorithms in Approach_2.

Table 9.Performance classifier of Random Forest

	Precision	Recall	F1-score	Class
RF	0.98	0.96	0.97	Benign
	0.92	0.97	0.95	Malignant

Based on the achieved results of the accuracies in all seven classifiers comparative graph of different classifiers is given in Figure 7. Support Vector classifier marks the maximum accuracy of 97.82% in Approach_1. At the same time LR, KNN, GN, RF, C4.5, and LDA show accuracies of 95.61%, 93.86%, 93.86%, 93.03%, 92.95%, and 87.88% respectively. RF classifier marks

ISBN: 978-955-627-013-6

the highest accuracy of 96.18% in Approach_2. When comparing the accuracy between the best classifiers in both approaches, we can determine that the SVC provides more accuracy when compared to the RF.

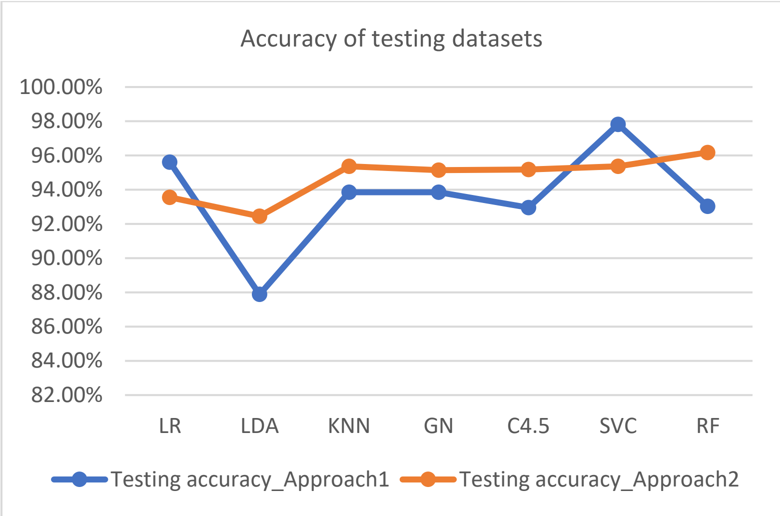

Figure 7.Accuracy of Testing Datasets

5. CONCLUSION AND FUTURE WORK

The application of machine learning algorithms on medical datasets and finding the best methods of diagnosing different types of diseases, predicting the severity range, predicting mere reasons for the diseases, predicting the condition of the disease are very significant works. They can help in increasing the lifetime of a patient and early detection of the disease can help in gaining the proper treatments on time before making it into the critical stage.

In this work, we have used the Wisconsin Breast Cancer Diagnostic Dataset (WBCD) for applying 7 machine learning algorithms in two different approaches. These two different approaches: without removing the outliers and removing the outliers were carried out to compare and evaluate different results obtained based on confusion matrix, accuracy, precision, recall, and F1-score. Based on these results we can determine whether keeping the dataset with these outlier data will affect the performance of the algorithms in a positive way or in a negative way. After comparing these results in both approaches, we found that the Support Vector Classifier in the Approach_1 gives the best accuracy and precision from all the classifier models in both approaches.

In future work, we can apply more outlier-removing mechanisms in more than one dataset related to one disease and can compare and increase the performance of the models. As well we intend to combine machine learning algorithms with deep learning approaches and test with more disease types.

REFERENCES

Aamir, S., Rahim, A., Aamir, Z., Abbasi, S. F., Khan, M. S., & Alhaisoni, M. (2022). Predicting breast cancer Leveraging Supervised Machine Learning Techniques.

Computational and Mathematical Methods in Medicine, 13. Retrieved from https://doi.org/10.1155/2022/5869529

Ming, C., Viassolo, V., Probst-Henshe, N., Chappuis, P. O., Dinov, D. I., & Katapodi, C. M. (2019). Machine learning techniques for personalized breast cancer risk prediction: comparison with the BCRAT and BOADICEA models. *Breast Cancer Res 21*. doi:https://doi.org/10.1186/s13058-019-1158-4

Nasien, D., Enjeslina, V., Adiya, H. M., & Baharum, Z. (2022). Breast Cancer Prediction Using Artificial Neural Networks Back Propagation Method. *Journal of Physics:Conference Series, 2319*. doi:1EU1jBxj8nKfvCaAzdeq1yafPEGrimcg8k

Ara, S., Das, A., & Dey, A. (2021). Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms. *2021 International Conference on Artificial Intelligence (ICAI)*, 97-101. doi:10.1109/ICAI52203.2021.9445249.

Khourdifi, Y., & Bahaj, M. (2018). Applying Best Machine Learning Algorithms for Breast Cancer Predition and Classification. *018 International Conference on Electronics, Control, Optimization and Computer Science*, 1-5. doi:10.1109/ICECOCS.2018.8610632.

Naji, M. A., Filali, S. E., Aarika, K., Benlahmar, E. H., Abdelouhahid, R. A., & Debauche, O. (2021b). Machine Learning Algorithms For Breast Cancer Prediction And Diagnosis. *Procedia Computer Science, 191*, 481-486. Retrieved from https://www.sciencedirect.com/science/article/pii/S1877050921014617

Naji, M. A., Filali, S. E., Bouhlal, M., Benlahmar, E. H., Abdelouhahid, R. A., & Debauche, O. (2021a). Breast Cancer Prediction and Diagnosis through a New Approach based on Majority Voting Ensemble Classifier. *Procedia Computer Science, 191*(1877-0509), 481-486. doi://doi.org/10.1016/j.procs.2021.07.062

Rabiei, R., Ayyoubzadeh, S. M., Sohrabei, S., Esmaeiii, M., & Atashi, A. (2022). Prediction of Breast Cancer using Machine Learning Approaches. *J Biomed Phys Eng*, 297-308. doi:10.31661/jbpe.v0i0.2109-1403

Saleh, H., Abd-el ghany, S. F., Alyami, H., & Alosaimi, W. (2022). Predicting Breast Cancer Based on Optimized Deep Learning Approach. *Computational Intelligence and Neuroscience, 2022*, 11. doi:10.1155/2022/1820777

Saranya, S., & Sasikala, S. (2020). Diagnosis Using Data Mining Algorithms for Malignant Breast Cancer Cell Detection. *2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, 1062-1067.

Shanbehzaden, M., Kazemi-Arpanahi, H., Ghalibah, M. B., & Orooji, A. (2022). Performance evalution of machine learning for breast cancer diagnosis:A case study. *Informatics in Medicine Unlocked, Volume 31*.

Soria, D., Garibaldi, J. M., Biganzoli, E. M., & Ellis, I. O. (2008). A comparison of three different methods for classification of breast cancer data. *Machine Learning and Applications 2008 (ICMLA'08) Seventh International Conference on Seventh International Conference on Machine Learning and Applications*.

Stanford Health Care. (2022). (Stanford Medicine) Retrieved from https://stanfordhealthcare.org/medical-conditions/cancer/cancer.html

WHO | Breast Cancer. (2021, March 26). (WHO) Retrieved from https://www.who.int/news-room/fact-sheets/detail/breast-cancer

Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (n.d.). *breast-cancer-wisconsis-data*. Retrieved from UCI.