Prediction of Benign and Malignant Tumor Cells in Human Breast Using Machine Learning Techniques

Angela Makolo and Olanrewaju Sule-Balogun

Computer Science Department, University of Ibadan, Nigeria
a.makolo@ui.edu.ng, suleabimbola@gmail.com

Abstract

Among women, breast cancer is the most common cancer and cause of cancer deaths [1]. It is, therefore, necessary to minimize this subjectivity with digitized images of the fine needle aspirates and machine learning techniques. The adoption of the machine learning techniques will help to further classify tumor cells as being malignant or benign. The presence of malignant tumor cells in breasts was predicted for using Wisconsin diagnostic breast cancer (WDBC) data. Several machine learning methods, such as support vector machine (SVM), naive bayes (NB), and artificial neural network (ANN) were used for the model prediction. The performance of these models were tested using standard metrics and after optimizing the individual models, the final accuracy results were 0.9960, 0.9920 and 0.9890 for SVM, NB and ANN models respectively.

Keywords: *Breast cancer, Malignant, Benign, Machine Learning, Support Vector Machine, Naïve Bayes, Artificial Neural Network.*

1. Introduction

Cancer begins in cells; it is an abnormal growth of cells which are the building blocks that forms the tissues. These tissues are found in elements of the physical body, as well as breasts. Once traditional cells age, they shrink to die, then, new cells are formed. Sometimes, this method doesn't follow the traditional means. Some new cells are a unit shaped after they don't seem to be required, and recent cells don't die to permit new cells to exchange them. This unusual creation of the cells forms a mass of tissue, also called, a lump or tumor. Cancer that forms within the tissues of breast, sometimes within the ducts (tubes that carry milk to the nipple) and within the lobules (glands that build milk) is named Breast Cancer [2]. Early detection followed by acceptable treatment can scale back the deadly risk. Technology like machine learning will considerably improve the diagnosing accuracy. In this study, we will adopt of various supervised machine learning algorithms and techniques to build and train models that will help in the classification and differentiation of the two types of breast cancer scenarios. The aim of the study is to develop a functioning supervised machine learning algorithm (SVM, NB, ANN) for classifying breast cancer as either malignant or benign. In order to achieve this, the objectives are: (i) formulate a predictive model to determine the likelihood of having breast cancer, (ii) compare the performance of trained models using supervised machine learning algorithms such as NB, ANN and SVM to correctly predict the classes of tumor cells as either malignant or benign, (iii) optimize the efficiency of the models using proper evaluation techniques, (iv) reducing the false positive likelihood within the breast cancer

Diagnosed process.

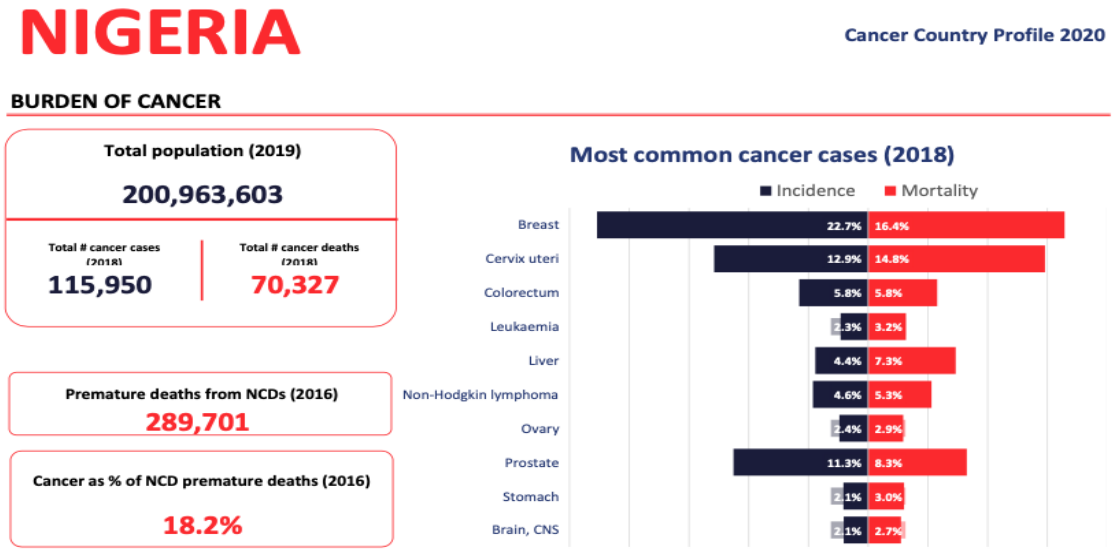

Figure 1: Most Common cancer cases in Nigeria 2018 [1]

2. Related Work

Several researchers have earlier studied various techniques for predicting breast cancer. The learning systems often used for these studies are data mining techniques, machine learning techniques and the hybrid form of data mining and machine learning systems [3]. A lot of research, however, has used one of the three different Wisconsin breast cancer datasets (Wisconsin Breast Cancer (WBC), Wisconsin Diagnosis Breast Cancer (WDBC), and Wisconsin Prognosis Breast Cancer (WPBC)).[4], discussed comparative study of six machine learning algorithms for breast cancer prediction namely as SVM, Artificial Neural Network, K-nearest Neighbor, Decision tree and Random Forest using Anaconda Python tool. They made comparisons and, on the performance, basis concluded that SVM and Random Forest have high accuracy, whereas Naives Bayes classifier have highest precision.

Comparative analysis has been made between Decision Tree J48 algorithm and Bayesian classification to determine breast cancer among the women showing the result that J48 have75.875% accuracy and 75.17% of Bayesian using WEKA tool. The data set includes 32 attributes and 286 instances [5].

Data mining classifiers namely NB, Decision Trees, KNN using different parameters to predict the cancer and as a result shows that NB is more superior to the other two [6]. Survey of breast cancer prediction using data mining technology shows that classification algorithms perform better than clustering algorithms in predicting breast cancer. SVM and C5.0 have the same accuracy [7]. Gayathri [8] used various machine learning algorithms (Supervised Learning, Unsupervised Learning, Semi-supervised Learning, Transduction, and Learning to learn) and methods to improve the accuracy of predicting breast cancer. [9], adopted the use of genetic algorithms and artificial neural networks to improve the diagnosis of breast cancer and they present an attempt to diagnose cancer by processing the quantitative and qualitative information obtained from medical infrared imaging. The best diagnosis parameters amongst the available parameters are selected and its precision in cancer diagnosis is done by utilizing genetic algorithm and artificial neural network. [11]

It is crystal clear from prior research works that a lot of studies still has to be done to improve the accuracy of prediction for the early detection of breast cancer.

3. Methodology

3.1.1 Naïve Bayes

Naïve Thomas Bayes algorithm is a classification algorithm belonging to the family of probabilistic classifiers adopted for classification issues based on Bayes' theorem. Each feature of the attribute is independent to every different attribute and a classification technique that was designed to classify the high-dimensional datasets. It is not only known for clarity and its high scalability but also as a functional algorithm. Naïve Thomas Bayes algorithm requires a number of parameters linear in the number of variables (i.e. features or predictors) in a learning problem. Naïve Thomas Bayes formula provides us a technique to calculate the probability. Formally, Bayes' theorem equation is given below:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \quad (1)$$

- P (A|B): Probability of event A, given that event B is true.
- P (A) and P (B): Possibility of the occurrence of event A and B, respectively.
- P (B|A): Possibility of the incidence of event B given the event A is true

3.1.2 Support Vector Machine

Support vector machine (SVM) is a supervised learning model used for analyzing data for classification problems. For this particular problem, the data needs to be classified into two groups, 'benign' and 'malignant'. The SVM accomplishes this classification by constructing a maximum margin line that separates the data space into two areas, called classes. Depending on the number of input features, this maximum margin line is constructed in a certain dimensional space. The extension from the linear to the nonlinear case is achieved by the use of the kernel trick. All benign samples are on one side of the maximum margin line and all malignant samples are on the other side. Both maximal margin hyperplanes are optimized to be as far away from each other as possible and still separate the classes. The data points touching the maximal margin hyperplanes are called the support vectors.

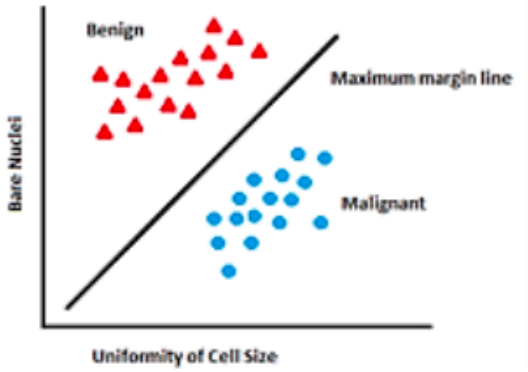

Figure 2. Class Distribution in WBC [4]

Transformation is done using kernels, polynomial and exponential kernels calculate separation line in higher dimension [12]. In the figure 2 below shows a clear representation of how kernels have played a significant role in obtaining distinction in the dataset.

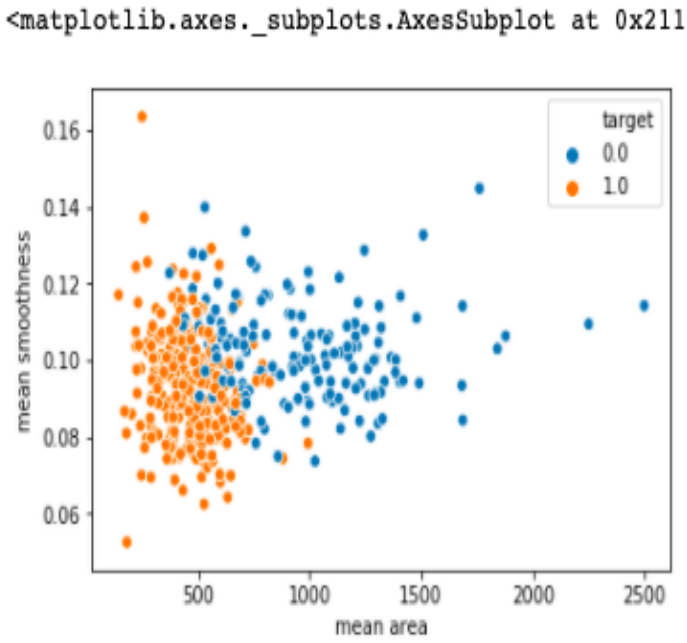

Figure 3. Classification using Kernel

3.1.3 Artificial Neural Network

Artificial neural network (ANN) has its name due to the fact that it has certain performance characteristics in common with biological neuron networks. Information get processed in the neurons thus making it a vital element. The needed signals are transported between the connecting neurons, where each connection has an associated weight. This weight will be multiplied with the signal being transmitted. Once the signal reaches the neuron, an activation function is applied to the input to determine the output.The activation function used in this research is sigmoid, which gives a value between [0,1]. In mathematical form, this transformation in each neuron is as follows

$$o = s\left(\sum_j \mathrm{w_j i_j} + \mathrm{b}\right) \qquad (2)$$

where o is the output of the neuron, wj are the weights of the neuron, ij are the inputs of the neuron, and b is a possible bias term. The sigmoid activation function S is as follows.

$$S(x) = \frac{e^x}{e^x} + 1 \qquad (3)$$

The network is trained by the input data and will produce an output corresponding to the size of the input data. For the input layer of the neural network, 8 neurons are selected (the number of neurons in the input layer is selected based on the fault percentage of the network output).For the middle layer of the neural network 6 neurons are selected to form the best structure [15].Choosing the correct number of neurons in the middle layer of the network is

very important, because it will reduce the time of the neural network's training process and keep the network in learning system [9]. The result of the classification problem will produce an appropriate output for the neural network.

3.2 Dataset

The dataset consists of traits obtained from a digitized image of a fine needle aspirate of a breast mass which was obtained from the Wiscoin Breast Cancer Data. The dataset is made readily available as an open source repository UCI Machine Learning [13].

Description of dataset

- Number of instances: 569
- Class Distribution: 212 – Malignant, 357 - Benign
- Number of attributes: 30 numeric, predictive attributes and the class
- Missing Attribute: None
- Attribute Information:
  - radius (mean of distances from center to points on the perimeter)
  - texture (standard deviation of gray-scale values)
  - perimeter
  - area
  - smoothness (local variation in radius lengths)
  - compactness (perimeter^2 / area - 1.0)
  - concavity (severity of concave portions of the contour)
  - concave points (number of concave portions of the contour)
  - symmetry
  - fractal dimension ("coastline approximation" - 1)

The mean, standard error, and "worst" or largest (mean of the three worst/largest values) of these features were computed for each image, resulting in 30 features. For instance, field 0 is Mean Radius, field 10 is Radius SE, field 20 is Worst Radius.

```
In [9]: print(cancer['target_names'])

        ['malignant' 'benign']

In [10]: print(cancer['feature_names'])

         ['mean radius' 'mean texture' 'mean perimeter' 'mean area'
          'mean smoothness' 'mean compactness' 'mean concavity'
          'mean concave points' 'mean symmetry' 'mean fractal dimension'
          'radius error' 'texture error' 'perimeter error' 'area error'
          'smoothness error' 'compactness error' 'concavity error'
          'concave points error' 'symmetry error' 'fractal dimension error'
          'worst radius' 'worst texture' 'worst perimeter' 'worst area'
          'worst smoothness' 'worst compactness' 'worst concavity'
          'worst concave points' 'worst symmetry' 'worst fractal dimension']

In [11]: cancer['data'].shape

Out[11]: (569, 30)
```

Figure 4. Description of dataset classes, features and shape

3.3 Simulation Software

Anaconda conveniently installs Python, the Jupyter Notebook, and other commonly used packages for the training process. The dataset was randomly divided into two; the training set and the testing set. The training set contains 455 cancer cells selected at random. The testing set contains the remaining 114 loss instances. Pandas was used for data manipulation, analysis and also used to import the dataset into the appropriate format.

Scikit-learn is used to feature the support vector machine and it also supports Python numerical and scientific libraries like NumPy (useful for working with data structures, multi-dimensional arrays) and SciPy (useful for manipulating and visualizing the data using a wide range of high-level Python commands). Matplotlib is used for data visualization and useful to generate high quality line plots, scatter plots, histograms, bar charts, and much more. Keras is a high-level interface and uses Tensorflow for its backend. It is used to support all the models of neural network adopted in this work. Sigmoid activation function is used to predict the probability of the output. Since probability of any event exists only between the range of 0 (benign) and 1 (malignant), sigmoid is the right choice. A confusion matrix is a technique for summarizing the performance of a classification algorithm.

3.4 Framework of Proposed System

3.4.1 Steps for the proposed system model

- Analyzing the Wisconsin dataset extracted from University of Irvine machine learning repository.
- Data preprocessing, data cleaning, handling and integration
- Data transformation and visualization
- Splitting of data into testing and another further split to get the training set
- Develop supervised machine learning algorithms (ANN, NB AND SVM) to help with the classification problem.
- Predict the probability of breast cancer cells to be malignant or benign
- Evaluate the performance of the models developed using metrics such as accuracy, sensitivity and specificity

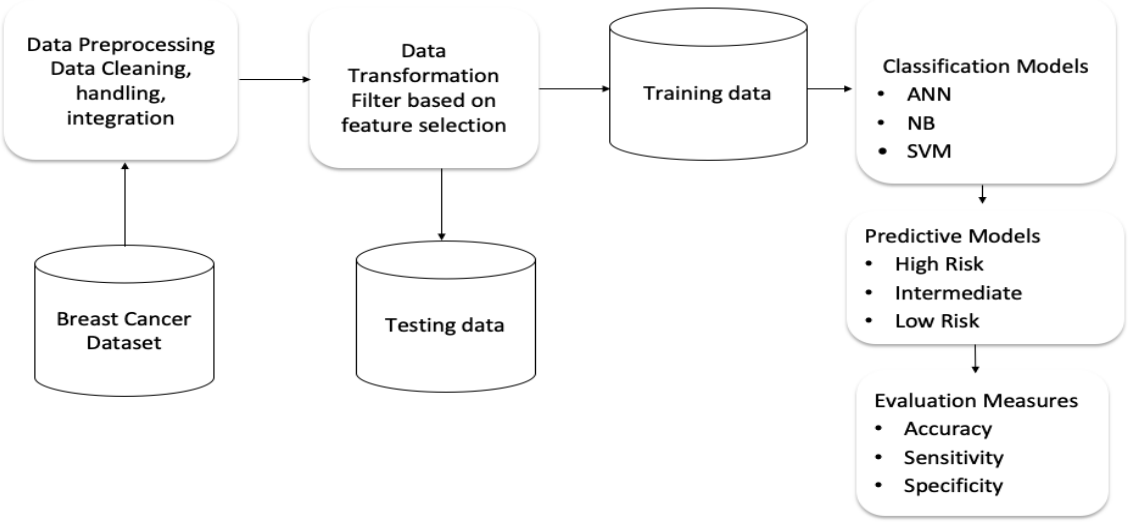

Figure 5. Architectural Framework of the Proposed System

4. Result and Analysis

Three models are evaluated on the test set using the selected features and the optimal parameter values. The validation set and test set performance accuracies are compared with the 10-fold cross validation accuracy to check the overall performance of the models. In addition, the performances between the models are compared based on the following metrics:

True Positive (TP): Group of positive instances that are correctly classified by the algorithm as positive. In this case this refers to the number of classes correctly classified as malignant [14]

True Negative (TN): Group of negative instances correctly classified by the algorithm as negative. In this case this refers to the number of classes correctly classified as benign [14]

False Positive (FP): False positives are the cases when the actual class of the data point was 0(False) and the algorithm predicted it to be 1(True). In this case this refers to the number of benign cells that are misclassified as malignant.

False Negatives (FN): False negatives are the number of instance when the actual class of the data point was 1(True) and the algorithm predicted it to be 0(False). In this case this refers to the number of malignant cells that are misclassified as benign.

Accuracy: Number of instances that was correctly labelled as either malignant or benign out of all the instance proved

Accuracy = (TP+TN)/(TP+FP+FN+TN)

Precision: Ratio of correctly true positive instance labelled by the algorithm to all positive labels. Total number of instances labelled as malignant are actually malignant

Precision = TP/(TP+FP)

Recall (Sensitivity): Ratio of correctly true positive instance labelled by the algorithm to all positives of people who have malignant cells in the real world. In this case, it is the number of instances of malignant cells that the algorithm correctly predicted

Recall = TP/(TP+FN)

Specificity: Ratio of the correctly negative instances labeled by the algorithm to all who are cancer free in the real world. In this case number of instances of benign cells that the algorithm correctly predicted

Specificity = TN/(TN+FP)

In this section, the obtained results of the used models are presented. These results will be evaluated in below sections

Table 1. Models and chosen features from Validation and Test set accuracy

Model	10-fold Cross Validation	Validation Set	Test Set
ANN	0.9717	0.9735	0.9735
NB	0.9601	0.950	0.975
SVM	0.9872	0.9745	0.9823

After performing several feature selection methods and trying several subsets of features, the final classifiers are constructed using the following subsets of features.

Naive Bayes: all original features are used

Neural network: texture + perimeter + concave Points + radius SE + perimeter SE + fractal Dimension SE + texture Worst + area Worst

Support Vector Machine: all original features are used

The Naive Bayes machine model and support vector machine does not need any prior feature selection. The performance of the model does not significantly improve when selected features are used.

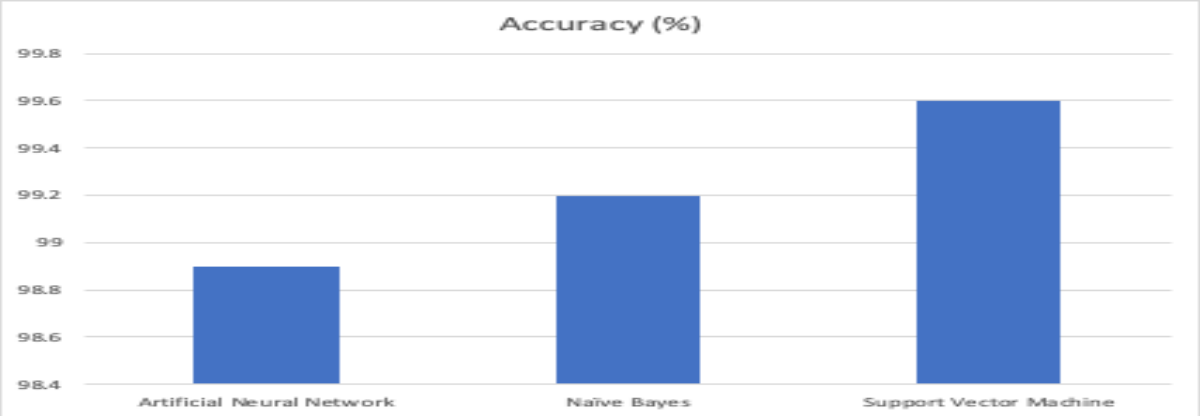

Figure 6. Graph for the accuracy of the algorithms

Figure 6 above shows the accuracy for each of the models which was used. The accuracy is percentage of all correct predictions over all predictions from the dataset. It is observed that the SVM has a better accuracy performance prediction score of 0.9960 due to the fact that there was a clear margin of separation between classes while the NB has a score of 0.9920 and 0.9890 for the ANN.

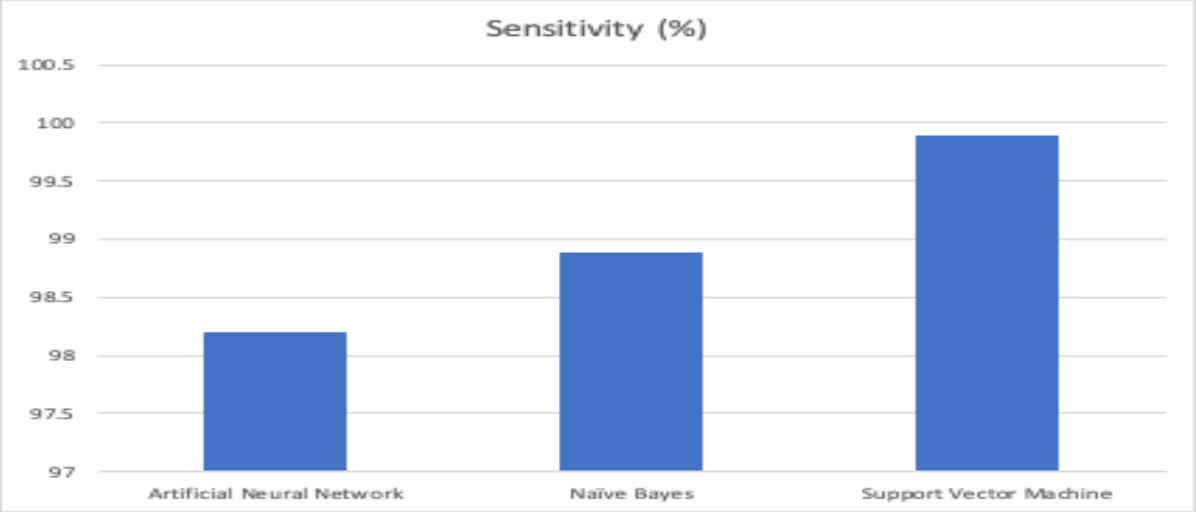

Figure 7. Graph for the sensitivity of the algorithms

Figure 7 above shows the recall (sensitivity) for each model that was used in the classification problem. The recall is the fraction of samples predicted to belong to a class with respect to examples that truly belongs to the class. In this case, it is the number of instances of malignant cells that the algorithm correctly predicts. It is observed that the SVM has a better recall prediction score of 0.9980 due to the fact that the number of dimensions is greater than the number of samples while the NB has a score of 0.9890 and 0.9830 for the ANN.

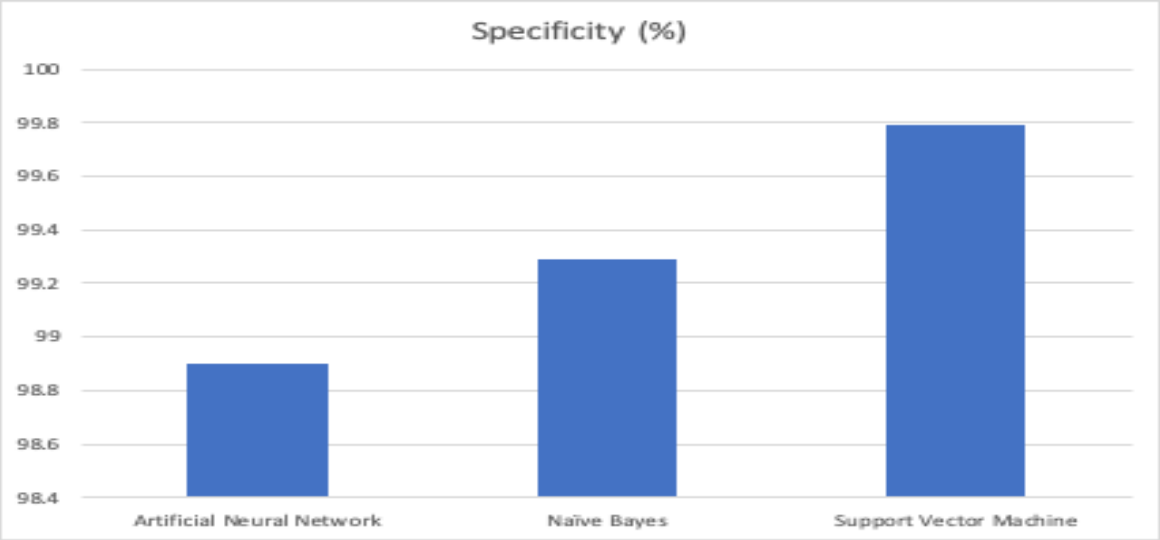

Figure 8. Graph for the specificity of the algorithms

Figure 8 above shows the graph of specificity for each model that was used in the classification problem. Specificity is the actual percentage of the correctly negative instances labeled by the algorithm to all who are cancer free in the real world. In this case, it is the number of instances of benign cells that the algorithm correctly predicts. It is observed that the SVM again has a better specificity performance prediction score of 0.9979 while the NB has a score of 0.9930 and 0.9890 for the ANN.

5. Conclusion and Future Work

From the results, it can be concluded that all three models obtain very promising performances in classifying the possible breast cancer. All models are optimized based on the accuracy. Selecting the best-suited model for this specific problem also depends on the sensitivity value, because it is important to have a low number of false positives. The tumor cell nuclei are best predicted by the SVM and NB. Both have the highest performance values for accuracy, sensitivity, and specificity. However, the support vector machine is the model which also has the highest value for the AUC. Therefore, the support vector machine model is recommended to use for this specific problem.

Future work will include experimenting with other programming language such as R , MATLAB and some other machine language algorithm that will be suitable for use with large dataset, able to perform well when the target classes are overlapping, be able to give a clue as to when there's a probing solution in order not to reduce trust in the network and also handling performance issues in situations where independent assumptions doesn't hold.

References

[1]. World Health Organization (2021, February 27) Cancer Country Profile 2020 https://www.who.int/cancer/country-profiles/NGA_2020.pdf?ua=1

[2]. Ponraj, N.Jenifer,M.E,.Poongodi,P.Manharan,J.(2011) ,"A survey on the processing Techniques of Mammogram for the Detection of Breast Cancer" Journal of Emerging Trends in Computing and Information Sciences(ISSN) 2079-8407 2(12).

[3]. Wei-Pin, Chang, Der-Ming, Liou (2008), "Comparison of Three Data Mining Techniques with Genetic Algorithm in the Analysis of Breast Cancer Data", Journal of Telemedicine and Telecare

[4]. Akshya Yadav, Imlikumla Jamir et al (2019), "Breast Cancer Prediction using SVM with PCA Feature Selection Method", International Journal of Scientific Research in Computer Science Engineering and Information Technology (IJSRCSEIT), ISSN: 2456-3307, Volume 5 Issue 2, pp. 969-978, March-April 2019.

[5]. Kashish Goyal et al, Comparative Analysis of Machine Learning Algorithms for Breast Cancer Prognosis, Springer Nature Singapore Pte Ltd., pp.727-734, (2019).

[6]. Chintan shah et al, Comparsion of data mining algorithms classification for breast Cancer prediction, pp.1-4, (2013)

[7]. Dono Sara Jacob et al, A Survey on Breast Cancer Prediction Using Data Mining Techniques, IEEE Conference on Emerging Devices and Smart System, pp.256-258 (2018).

[8]. Gayathri.B, M, Sumathi, C, P,Santhanam, T "Breast cancer diagnosis using machine learning algorithm a survey" . (2013). International Journal of Distributed and Parallel Systems 4(3).

[9]. Hossein GhayoumiZadeh, JavadHaddadnia, et al, "Diagnosis of Breast Cancer using a Combination of Genetic Algorithm and Artificial Neural Network in Medical Infrared Thermal Imaging", Iranian Journal of Medical Physics, Vol. 9, No. 4, Autumn 2012, 265-274

[10]. S. Padmapriya, M Devika, V Meena, S.B Dheebika and R. Vinodhini: "A survey on breast cancer analysis using data mining techniques", IEEE, vol-2, issue-4, pp.970-974, (2014).

[11]. Narang, S, Verma, H,K, Sachdev ,U "Review of breast cancer detection using ART model of neural network " . (2012). International Journal of Advanced Research in Computer Science and Software Engineering. ISSN 2277-128X, 2(10).

[12]. MeriemAmrane, SalihaOukid, "Breast CancerClassificationUsing Machine Learning,Proceedings", (2010), IEEE Student Conferenceon Research and Development (SCOReD 2010),13 - 14 Dec 2010,Malaysia.

[13]. M. Lichman, UCI Machine Learning Repositry,2013. Online]. Available:https://archive.ics.uci.edu/.

[14]. A. Makolo, "Support Vector Machine for improving Performance of TCP on Hybrid Network", African Journal of Computing & ICT, Vol.5, pp.107-112, (Dec, 2012).

[15]. Hirose Y, Yamashita K, Hijiya S. Back-propagation algorithm which varies the number of hidden units. Neural Networks. 1991;4(1):61-6.