# Breast Cancer Diagnostics using Machine Learning

Hitesh Thawani[1]
Computer Science Department
University of Massachusetts
Lowell, MA, USA
hitesh_thawani@student.uml.edu

Shraddha Kharche[2]
Computer Science Department
University of Massachusetts
Lowell, MA, USA
shraddha_kharche@student.uml.edu

***Abstract- The aim for the project is to take the Breast Cancer Wisconsin (Diagnostic) Data Set available at UCI Machine Learning Repository and apply machine learning models such as Logistic Regression, Naïve Bayes, Support Vector Machine, Decision Tree and Multi-Layer Perceptron to extract features from the data set which could be the most suitable for predicting the nature of the cancer. The purpose is to classify breast cancer whether it is benign or malignant. Accuracy of the models based on their predictions is determined to analyze and compare the generated models to each other and select the best out of the model. Multi-Layer Perceptron is the most accurate model out of the models tested with 97.2% accuracy.***

***Keywords- Breast Cancer(BC), Machine Learning(ML), Naïve Bayes(NB), Logistic Regression(LR), Support Vector Machine(SVM), Decision Tree(DT), Multi-Layer Perceptron(MLP)***

## I. INTRODUCTION

Breast Cancer is one of the most common cancers affecting women all over the globe. It is a significant public health problem today according to global statistics as it a cause of most new cancer cases and cancer-related deaths.

Diagnosing breast cancer in its early stages improves the prognosis and significantly increases the chance of survival because it results in timely clinical treatment to patients. Moreover, improved diagnostics will facilitate accurate classification of benign and malignant tumors which would prevent patients undergoing unnecessary treatment and save valuable resources for the ones in need. Thus, the correct diagnosis of BC and classification of patients into malignant or benign groups is the subject of much research. Because of its unique advantages in critical features detection from complex BC datasets, machine learning is widely recognized as the methodology of choice in BC pattern classification and forecast modelling.

Classification and data mining methods are an effective way to classify data. Especially in medical field, where those methods are widely used in diagnosis and analysis to make decisions.

In this paper, four machine learning models are established and tested for classifying tumors as benign or malignant. The initial project proposal consisted of only three proposed models– Logistic Regression, SVM and Naïve Bayes, but Decision Tree and Multi-Layer Perceptron models were added in the later stages of the project. Logistic Regression model is used as a baseline for testing accuracies of other models. The expected performance of MLP is supposed to be high as in previous research which are presented in the background section of this paper, MLP has performed quite well, with accuracies in the range of 95-99%.

The paper uses the Wisconsin Diagnostic Breast Cancer (WDBC) dataset on which all the models are applied which is publicly available at UCI Machine Learning Repository [1]. The data set provides data for 569 patients on 30 features of the cell nuclei obtained from a digitized image of a fine needle aspirate (FNA) of a breast mass. For each patient, the cancer was diagnosed as malignant or benign.

## II. BACKGROUND

Machine Learning methods are being heavily employed in medical field to make the diagnostics part of the field much more streamlined and robust and less prone to errors [2],[3]. However, in this paper, machine learning models to accurately predict the severity of breast cancer masses are taken into consideration, and there are many recent papers on the subject. MLP is a model that generally performs the best [4] [5] [6] [7] [8] with the highest accuracy achieved being 99.03% for MLP [8] using the Wisconsin Diagnostic Breast Cancer dataset, and

there are several other papers that are able to reach 95% or more with at least one algorithm. In particular, Khourdifi *et al.*[7] compared Random Forest, Naïve Bayes, Support Vector Machines, K-Nearest Neighbors, and Multilayer Perception, they were able to show that MLP performs better in terms of accuracy on the given dataset; which is in line with our research presented in this paper i.e. MLP model comes out to be the best out of the models in consideration. In general, most models can achieve accuracies in the 85-95% range.

## III. APPROACH

The problem presented in the paper is a classification problem and involves studying all suitable classification machine learning models. AS discussed in the background section of this paper, a conclusion can be drawn that SVM and MLP are the majorly used and popular classification models. In this paper, Logistic Regression, Naïve Bayes, Support Vector Machine, Decision Tree and Multi-Layer Perceptron algorithms are implemented. However, instead of applying these models directly onto the data, correlation analysis [10] considering the 30 attributes with respect to the diagnosis status is done to avoid bias in the model outcome. This helps in removal of significantly correlated attributes by keeping only one out of them while discarding the others, resulting in a better model as now there are fewer attributes and problem of over-fitting the data can be avoided.

### *A. Logistic Regression*

Logistic Regression [11] is a classification algorithm used where the response variable is categorical. The idea of logistic regression is to find out a relationship between features and probability of particular outcome. Unlike linear regression which outputs continuous number values, logistic regression transforms its output using the logistic sigmoid function to return a probability value which can then be mapped to two or more discrete classes. Binary, Multi and Ordinal are the three types of the logistic regression. In this case since the output desired is of a binary sort – benign or malignant – binary logistic regression is used. The sigmoid function shown below is used to transform the output of the model into a range between 0 and 1.

$$Sigmoid(x) = \frac{1}{1 + e^{-x}}$$

### *B. Naïve Bayes*

Naïve Bayes [12] classifier is a probabilistic machine learning model used for classification task. It is a simple but very popular machine learning classifier that functions by assigning most likely class to a given example determined by its feature vector. Naïve Bayes has three types, 1. Gaussian used for classification, assuming the features follow normal distribution 2. Multinomial used for discrete counts, e.g. word occurring in document. 3.Bernoulli used for binary feature vectors. In this paper, Gaussian Naïve Bayes classifier is used to solve the problem at hand.

### *C. Support Vector Machines*

SVM [13] is a supervised machine learning algorithm which can be used for both classification and regression challenges. In SVM algorithm, each data item as a point in n dimensional space is plotted with the value of each feature. The classification is performed by finding optimal hyper plane that predominantly differentiate the classes. The SVM algorithms uses set of mathematical functions (for example, linear, nonlinear, polynomial, radial basis function etc.) that are passed as a kernel parameter to SVM model function. However, for the classification problem discussed in this paper SVM is implemented using a linear mathematical function.

### *D. Decision Tree*

Decision Trees [14] can be used for both classification and regression problems. Implementing decision tree models is a two-step process which involves Induction and Pruning. Induction deals with generation of hierarchical decision boundaries based on the dataset provided to the model. Pruning is done after the decision tree has been constructed; it is a process of removal of unnecessary structures of the generated tree in order to avoid overfitting. Since, decision trees can be evaluated at each node, it makes them easy to understand and interpret. However, there is one

problem associated with decision trees, that is overfitting of data due to the way the model is trained.

*E. Multi-Layer Perceptron*

A Multi-Layer Perceptron [15] is a class of feedforward Artificial Neural Network (ANN) consisting of at least three layers viz. Input layer, Hidden layers (at least one) and Output layer. MLP takes advantage of Supervised Learning technique called backpropagation for training which is a procedure to repeatedly adjust the weighted and the threshold values accordingly to minimize the difference between the desire/targeted output and the obtained output. The biggest advantage of MLP over other models is the ability to be scaled to work with bigger datasets efficiently. Many times, it is hard to improve the accuracy of other models after training certain amount of data which is not the case with MLP.

## IV. RESULTS

The Wisconsin Diagnostic Breast Cancer Dataset has 357 benign and 212 malignant instances, a total of 569; and each of these instances are representations of FNA test measurements per diagnosis case. There are 32 attributes in a single instance, with the initial two attributes corresponding to a unique ID number and telling of the diagnosis status, whether it is benign or malignant. The remaining 30 features enable us to compute 10 real valued features; this includes their mean, standard error and the mean of the three largest values or "worst" value for each cell nucleus. A digitized image of a fine needle aspirate (FNA) of breast tumor enables us to compute 10 real values. These 10 real values are computed with four significant digits and describe characteristics of the cell nuclei present in the image

In order to compare the models with each other, different parameters such as accuracy, precision, recall and specificity are required for all the models in consideration. These calculations require confusion matrices for all the models to be determined beforehand.

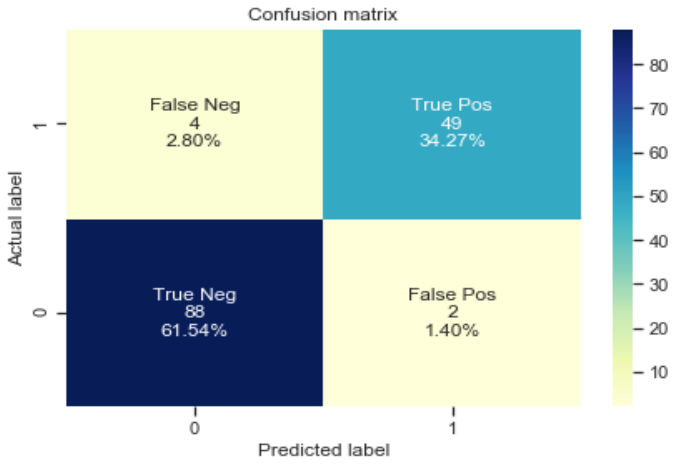


Fig. 1a: Confusion Matrix for Logistic Regression

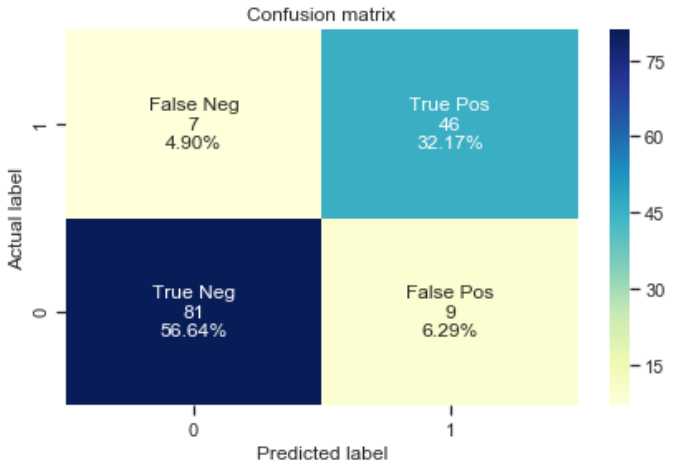


Fig. 1b: Confusion Matrix for Naïve Bayes

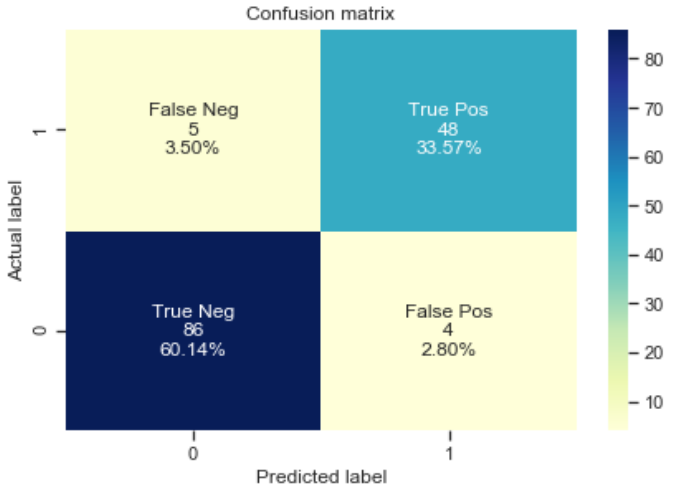


Fig. 1c: Confusion Matrix for SVM

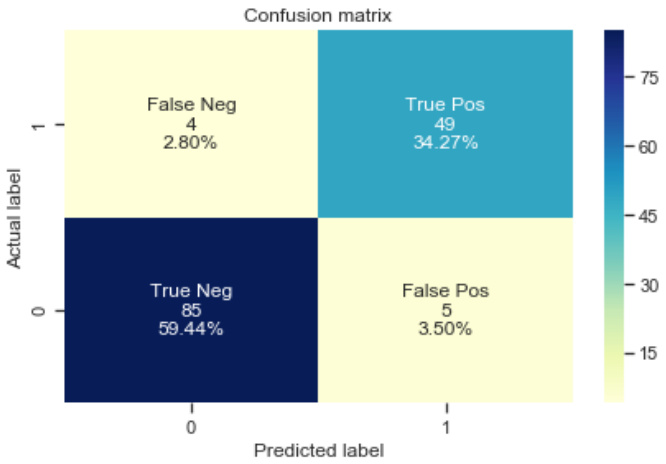


Fig. 1d: Confusion Matrix for Decision Tree

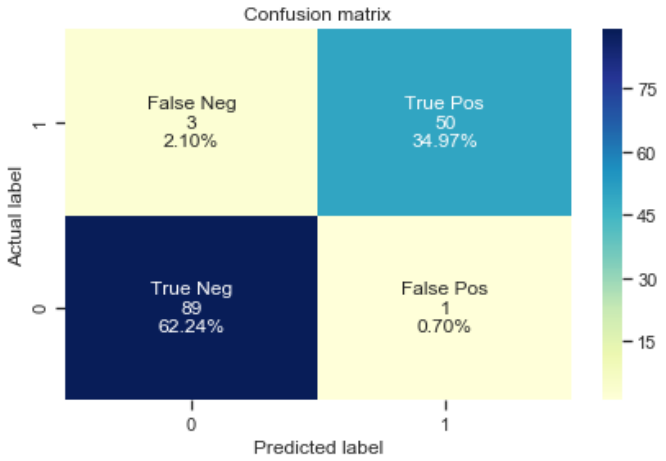


Fig. 1e: Confusion Matrix for MLP

Now, using the confusion matrices for the different models, we can calculate the different parameters for the models using the following definitions-

Accuracy is the percent of correct predictions in the set of all predictions made by the model.

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

Precision is the percent of positives that are true positives.

$$Precision = TP/(TP + FP)$$

Recall is the percent of positives in the data set that were detected by the model.

$$Recall = TP/(TP + FN)$$

Specificity is the percent of positives in the data set that were detected by the model.

$$Specificity = TN/(FP + TN)$$

| | Logistic Regression | Naïve Bayes | SVM | Decision Tree | MLP |
|---|---|---|---|---|---|
| Accuracy | 0.9580 | 0.8881 | 0.9370 | 0.9370 | 0.9720 |
| Precision | 0.9607 | 0.8363 | 0.9230 | 0.9074 | 0.9803 |
| Recall | 0.9245 | 0.8679 | 0.9056 | 0.9245 | 0.9433 |
| Specificity | 0.9777 | 0.9000 | 0.9555 | 0.9444 | 0.9888 |

Table 1: Comparison Parameters for Different Models

Table 1 shows these different parameters calculated using the confusion matrices for each of the models analyzed in this paper.

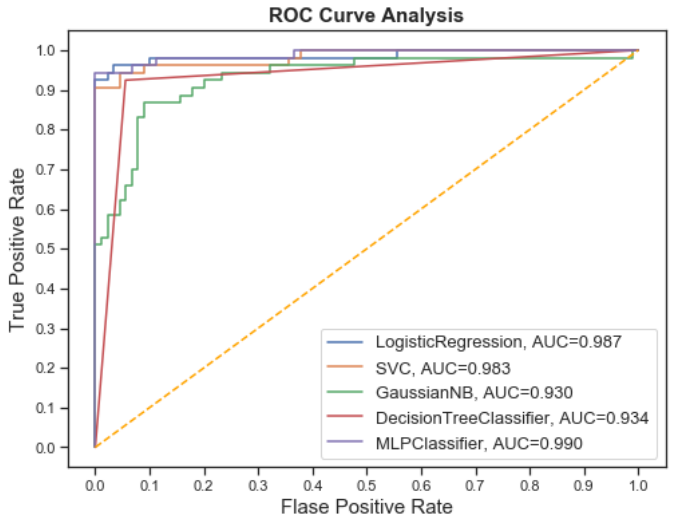


Fig. 2: Receiver Operating Characteristic (ROC) for Different Models

Fig. 2 shows the plots of ROC for each learning model. ROC is an important statistic of a learning model. It plots the rate of false negatives against the rate of false positives. A good ROC curve is the one that hugs the upper-left corner of the graph, as this indicates that a high rate of false positives can be achieved while simultaneously keeping false negatives low.

One important characteristic of a ROC graph is the area under the ROC curve (AUC). Areas close to 1 indicate a model that distinguishes between positive and negative samples well, and areas close to 0.5 indicate pure guessing.

The experimental results show that MLP model comes out to be on top of other models considered in this paper with an accuracy of 97.2% and has AUC as 0.990 which is closest to 1 in ROC curve analysis.

## V. CONCLUSION

In this paper, the implementation of five machine learning models for classifying breast tumors as benign or malignant was presented. The models were analyzed according a suite of performance measures, and MLP was found to be the best of the five. The MLP had the highest accuracy of the five at 97.2% and the highest recall at 94.3%. It also has the highest AUC (=0.990) out of all the models under consideration. MLP is the "best model" since one of the most important tasks of these models is identifying as many malignant masses as possible. One other point to be noted is that while the accuracy of the models was significantly high, the dataset on which the models trained on was not big enough in order to reach a definite conclusion. Future work could consider this handicap and improve upon the current research by seeking a dataset with more datapoints as compared to the one used here as well as possible modifications to MLP to make it more effective.


### *Acknowledgement*

A special gratitude to professor, Jerome Braun, Ph.D., whose contribution in stimulating suggestions and encouragement, helped in coordinating this project. His full effort in guiding the team in achieving the goal is highly appreciated.

Furthermore, the authors acknowledge with much appreciation the crucial role of the teammates, who helped each other to assemble the parts and gave suggestion about the task.


### *References*


[1]. William H Wolberg, W Nick Street, and Olvi L Mangasarian. 1992. "Breast cancer Wisconsin (diagnostic) data set". *UCI Machine Learning Repository*, 1992

[2]. Rahul C. Deo, "Machine Learning in Medicine." *Circulation* vol. 132,20, 2015.

[3]. Shruti Agarwal, Hitesh Thawani and Narina Thakur, "A Machine Learning Model for Arterial Blockage Risk Prediction", *International Journal of Information Systems & Management Science*, vol. 1, no. 2, 2018.

[4]. A. F. Seddik and D. M. Shawky, "Logistic regression model for breast cancer automatic diagnosis," *2015 SAI Intelligent Systems Conference (IntelliSys)*, 2015

[5]. M. Sewak, P. Vaidya, C. Chan and Zhong-Hui Duan, "SVM Approach to Breast Cancer Classification," *Second International Multi-Symposiums on Computer and Computational Sciences (IMSCCS 2007)*, 2007.

[6]. Puneet Yadav, Rajat Varshney and Vishan Kumar Gupta, "Diagnosis of Breast Cancer using Decision Tree Models and SVM", *International Research Journal of Engineering and Technology,* vol. 05, issue 03, 2018.

[7]. Y. Khourdifi and M. Bahaj, "Feature Selection with Fast Correlation-Based Filter for Breast Cancer Prediction and Classification Using Machine Learning Algorithms," *2018 International Symposium on Advanced Electrical and Communication Technologies (ISAECT)*, 2018.

[8]. Abien Fred M. Agarap, "On Breast Cancer Detection." *Proceedings of the 2nd International Conference on Machine Learning and Soft Computing – ICMLSC,* 2018.

[9]. Mochin Li and Raji Sundararajan, "Application of Machine Learning Algorithms on Breast Cancer Dataset", *Proceedings of 2018 Electrostatics Joint Conference,* 2018.

[10]. C.F. Dormann, S.J. Schymanski, J. Cabral, I. Chuine, C. Graham, F. Hartig, M. Kearney, X. Morin, C. Römermann, B. Schröder and A. Singer, "Correlation and process in species distribution models: bridging a dichotomy." *Journal of Biogeography*, 2012.

[11]. A. El-Koka, K. Cha and D. Kang, "Regularization parameter tuning optimization approach in logistic regression," *15th International Conference on Advanced Communications Technology (ICACT)*, 2013.

[12]. Harry Zhang, "The Optimality of Naive Bayes." *FLAIRS Conference,* 2004.

[13]. V. Vapnik, *The Nature of Statistical Learning Theory*. SpringerVerlag, New York, 1995.

[14]. J. R. Quinlan. "Induction of Decision Trees." *Machine Learning*, 1986.

[15]. J. Singh and R. Banerjee, "A Study on Single and Multi-layer Perceptron Neural Network," *International Conference on Computing Methodologies and Communication (ICCMC),* 2019.