Breast Cancer Relapse Prognosis by Classic and Modern Structures of Machine Learning Algorithms

Farhana Kausar[1], Karan Chaudhary[2], Karthik GS[3], Chetan J[4]

[1]*Assistant Professor, Atria Institute of Technology, Visvesvaraya Technological University, Bangalore*
[2,3,4]*Student, Atria Institute of Technology, Visvesvaraya Technological University, Bangalore*

Abstract: *Breast Cancer is the principle cause of death from cancer among women globally and is the most common cancer in women worldwide. One of the efficient ways to reduce death due to breast cancer is to detect it earlier. Diagnosis of breast cancer requires the reliable procedure by the physicians to distinguished weather the cancer is benign or malignant. Machine learning algorithm helps them to detect the cancer using automated tools and by prediction. Breast cancer is the most diagnosed cancer and is the major cause of cancer death among population worldwide. In 2019, about 268,600 new cases of invasive breast cancer was diagnosed among women and approximately 2,670 cases were diagnosed in men. In addition, approximately 48,100 cases of DCIS were diagnosed among women. Breast Tumor can be categorized into three types: benign breast tumors, in situ cancers, and invasive cancers. Majority of breast tumors detected by mammography are benign. They are non-cancerous growths and cannot spread outside of the breast to other organs. In some cases, it is difficult to distinguish certain benign masses from malignant with mammography. Therefore, early detection of breast cancer is essential. In our study, we are focusing on the differentiation between benign and malignant tumors.*

Key Words: Correlation Testing, Logistic Regression, Machine learning, Methodology, Multi-layer Perceptron, Outlier Detection, Support Vector Machine.

1. INTRODUCTION

Breast cancer is major reason for woman death. It is one of the world biggest issue by which woman is dying day by day. In 2019, an estimated 268,600 new cases of invasive breast cancer will be diagnosed among women and approximately 2,670 cases will be diagnosed in men. In addition, an estimated 48,100 cases of DCIS will be diagnosed among women. In order to improve breast cancer outcomes and survival, early detection is critical. There are two early detection strategies for breast cancer: early diagnosis and screening. Limited resource settings with weak health systems where the majority of women are diagnosed in late stages should prioritize early diagnosis programs based on awareness of early signs and symptoms and prompt referral to diagnosis and treatment. Lump in the chest, discharge of blood from breast, breast pain are the major symptoms of breast cancer. Many women were suffering from breast cancer nowadays and they were unable to predict whether the cancer is benign or malignant because of which they have to lost their life soon. The major problem of breast cancer in the contest of India is uneducated behavior because the person who is uneducated did not understand the stage of cancer and so that they have to lost their life.

2. MACHINE LEARNING BASED ESTIMATION AND DETECTION

2.1 Machine Learning Methods

Machine learning algorithm has been successfully applied in a wide range of areas with excellent performance. With the help of machine learning we can train the model and test the model in the efficient manner and try to predict the output that we obtain from the machine.

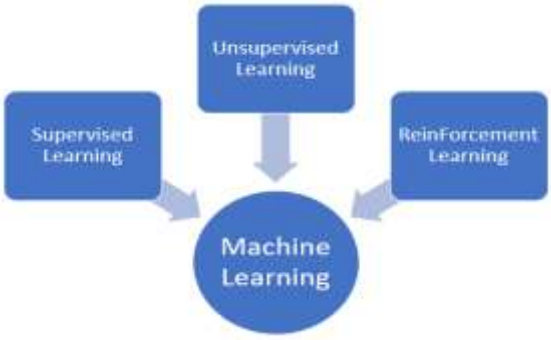

Fig 1. Machine Learning categories

Machine Learning mainly divided into three categories: Supervised Learning, Unsupervised Learning, Reinforcement Learning.

2.2. System Architecture

Machine Learning Architecture includes Data Acquisition, Data Processing, Data Modeling, Execution, Deployment. Machine Learning Architecture occupies the major industry interest now as every process is looking out for optimizing the available resources and output based on the historical data available, additionally, machine learning involves major advantages about data forecasting and predictive analytics when coupled with data science technology.

3. METHODOLOGY

We took the breast cancer dataset from UCI and used jupyter notebook as the platform for the purpose of coding. Our methodology involves use of classification techniques like Logistic Regression, Support Vector Machine (SVM), Random Forest Classifier, Multi-Layer Perception with Accuracy Visualization technique i.e. Visualize the accuracy using confusion matrix.

3.1. Data Exploration

Data Exploration is an approach where data was visualized and characteristics of data were explored. It consists calculation of statistics for numerical purpose, getting information about the dataset and its data types to detect null values and plotting the Histograms to Visualize Feature Distributions in the dataset (Detect Skewness)

3.2. Correlation Testing

Correlation is any statistical relationship, whether causal or not, between two random variables or bivariate data. In the broadest sense correlation is any statistical association, though it commonly refers to the degree to which a pair of variables are linearly related. It shows the relation between two entities/variable. With the help of correlation, it is possible to have a correct idea of the working data. With the help of it, it is also possible to have a knowledge of the various qualities of an entity.

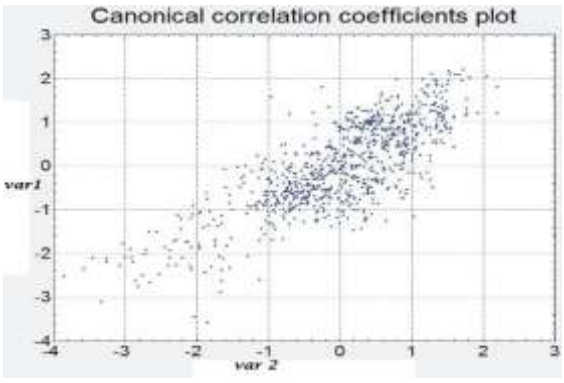

Fig 2. Correlation matrix plot

3.3. Outlier Detection

Outlier can be source of information in the dataset on the other hand it may be lead to bad results or biased result. So, we need to detect them and treat them in the Preprocessing stage.

The presence of outliers can often skew results. There are many techniques for how to detect and deal with the outliers in a dataset. Outlier step is calculated as factor multiplied the interquartile range (IQR). A data point with a feature that is beyond an outlier step outside of the IQR for that feature is considered abnormal.

4. MODEL SELECTION

Selection of algorithm plays an important role in machine learning model. We can use more than one kind of techniques to large datasets. But, at advanced level all those different algorithms can be classified in two groups: supervised learning and unsupervised learning. **Supervised learning** is the machine learning task of learning a function that maps an input to an output based on example input-output pairs. It infers a function from labeled training data consisting of a set of training example. Supervised learning classified into two categories of algorithms. A **classification problem** is when the output variable is a category, such as "Red" or "blue" or "disease" and "no disease". A **regression problem** is when the output variable is a real value, such as "dollars" or "weight". **Unsupervised learning** is the training of machine using information that is neither classified nor labeled and allowing the algorithm to act on that information without guidance. Here the task of machine is to group unsorted information according to similarities, patterns and differences without any prior training of data. Unsupervised learning classified into two categories of algorithms. A **clustering problem** is where you want to discover the inherent groupings in the data, such as grouping customers by purchasing behavior. An **association rule learning problem** is where you want to discover rules that describe large portions of your data, such as people that buy X also tend to buy Y. In our dataset we have the outcome variable having only two set of values, either B(Benign) or M (Malignant). So, supervised learning algorithm is applied on it. We have chosen four different types of classification algorithms in Machine Learning.

1. Support Vector Machine (SVM)

2. Logistic Regression

3. Random Forest Classifier

4. Multi-Layer Perceptron (MLP)

4.1. Support Vector Machine (SVM)

Support Vector Machines (SVM) is a data classification method that separates data using hyperplanes. The concept of SVM is very intuitive and easily understandable. If we have labeled data, SVM can be used to generate multiple separating hyperplanes such that the data space is divided into segments and each segment contains only one kind of data. SVM technique is generally useful for data which has non-regularity which means, data whose distribution is unknown. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the

examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on the side of the gap on which they fall. SVM is used to train a support vector machine. It can be used to carry out general regression and classification (of nu and epsilon-type), as well as density-estimation. SVM can be used as a classification machine, as a regression machine, or for novelty detection. Depending of whether y is a factor or not, the default setting for type is C-classification or eps-regression, respectively, but may be overwritten by setting an explicit value. Valid options are c-classification, nu-classification, one-classification, eps-regression, nu-regression.

4.2. Logistic Regression

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. Logistic regression is used when the response variable is categorical in nature. Predicting a defaulter in a bank using the transaction details in the past is an example of logistic regression. Logistic regression is a statistical method for analyzing a dataset in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes).In logistic regression, the dependent variable is binary or dichotomous, i.e. it only contains data coded as 1 (TRUE, success, pregnant, etc.) or 0 (FALSE, failure, non-pregnant, etc.).The goal of logistic regression is to find the best fitting (yet biologically reasonable) model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. Logistic regression generates the coefficients (and its standard errors and significance levels) of a formula to predict a *logit transformation* of the probability of presence of the characteristic of interest:

$$logit(p) = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_kX_k$$

where p is the probability of presence of the characteristic of interest. The logit transformation is defined as the logged odds:

$$odds = \frac{p}{1-p} = \frac{probability\ of\ presence\ of\ characteristic}{probability\ of\ absence\ of\ characteristic}$$

and

$$logit(p) = \ln\left(\frac{p}{1-p}\right)$$

Rather than choosing parameters that minimize the sum of squared errors (like in ordinary regression), estimation in logistic regression chooses parameters that maximize the likelihood of observing the sample values.

4.3. Random Forest Classifier

Random forest is a type of supervised machine learning algorithm based on ensemble learning. Ensemble learning is a type of learning where you join different types of algorithms or same algorithm multiple times to form a more powerful prediction model. The random forest algorithm combines multiple algorithm of the same type i.e. multiple decision trees, resulting in a forest of trees, hence the name "Random Forest". The random forest algorithm can be used for both regression and classification tasks. Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set. Random forest has nearly the same hyperparameters as a decision tree or a bagging classifier. Fortunately, there's no need to combine a decision tree with a bagging classifier because you can easily use the classifier-class of random forest. With random forest, you can also deal with regression tasks by using the algorithm's regressor. Random forest adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model. Therefore, in random forest, only a random subset of the features is taken into consideration by the algorithm for splitting a node. You can even make trees more random by additionally using random thresholds for each feature rather than searching for the best possible thresholds (like a normal decision tree does).

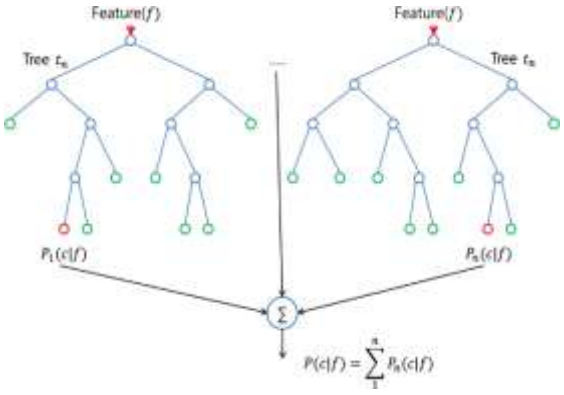

Fig 3. Random Forest Classifier

4.4. Multi-Layer Perceptron

Multi-Layer perceptron (MLP) is a feedforward neural network with one or more layers between input and output layer. Feedforward means that data flows in one direction from input to output layer (forward). This type of network is trained with the backpropagation learning algorithm. MLPs are widely used for pattern classification, recognition, prediction and approximation. Multi-Layer Perceptron can solve problems which are not linearly separable. In the Multilayer perceptron, there can more than one linear layer (combinations neurons). If we take the simple example the three-layer network, first layer will be the *input layer* and last will be *output layer* and middle layer will be called *hidden layer.* We feed our input data into the input layer and take the output from the output layer. We can increase the number of the hidden layer as much as we want, to make the model more complex according to our task. The output layer of MLP is typically Logistic regression classifier, if probabilistic outputs are desired for classification purposes in which case the activation function is the SoftMax regression function.

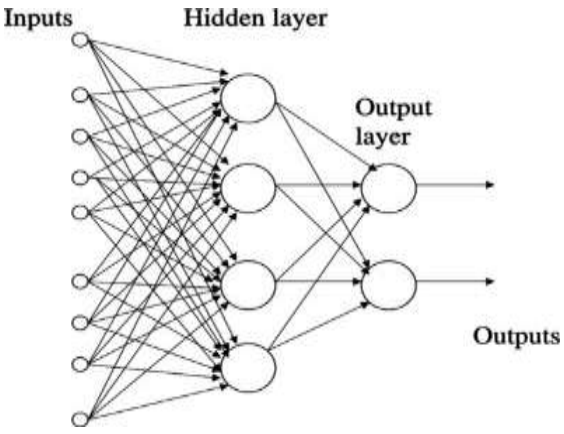

Fig 4. Multi-Layer Perceptron

5. RESULTS AND DISCUSSION

Our dataset contains 31 attributes dimensionality reduction contributes a lot in decreasing the multi-dimensional data to a few dimensions. Of all the four applied algorithms Logistic Regression, Support Vector Machine, Random Forest Classifier and Multi-Layer Perceptron. Using various technique such as Correlation, skewness and outlier deletion, SVM gives the highest accuracy of 98.3% when compared to other three algorithms. So, we conclude that SVM is the best suited algorithm for the prediction of Breast Cancer Occurrence with complex datasets. Accuracy Visualization plays an important for predicting and visualizing the correctness of the algorithm. Confusion matrix is used to detect the two types of errors False Positive and False Negative.

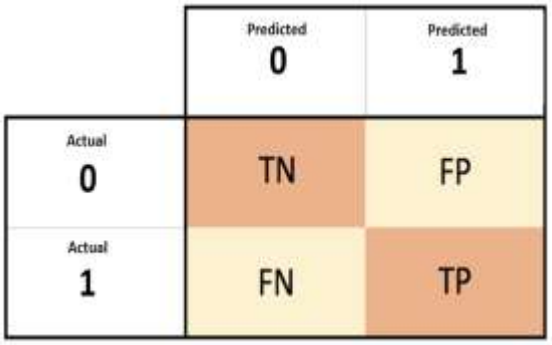

Fig 5. Confusion Matrix

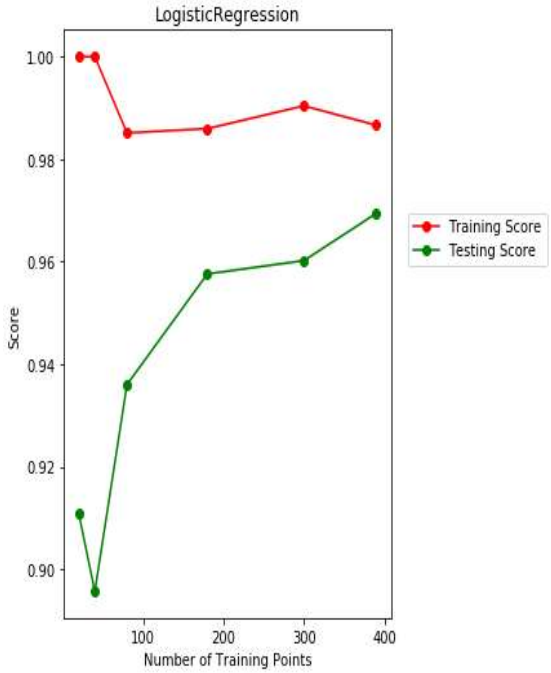

Fig 6. Learning Performance for Logistic Regression

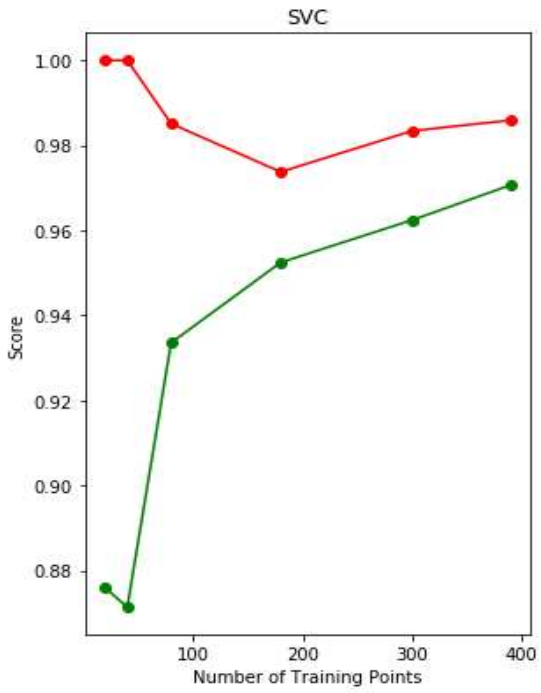

Fig 7. Learning Performance for SVC

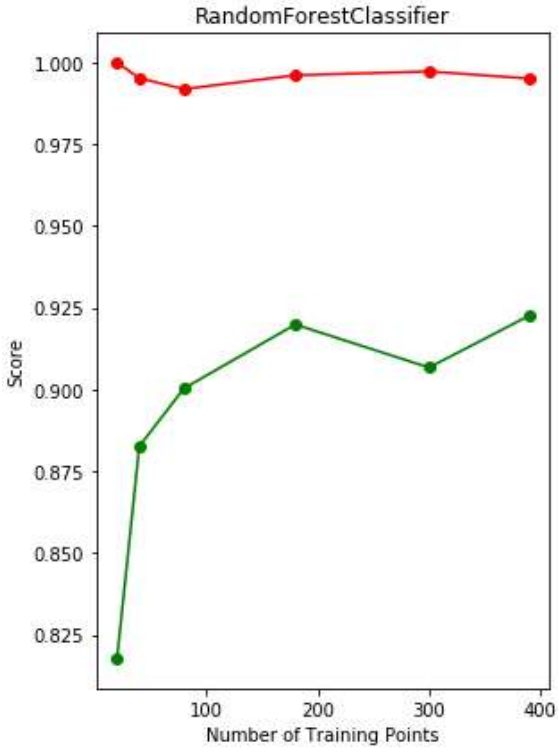

Fig 8. Learning Performance for Random Forest Classifier

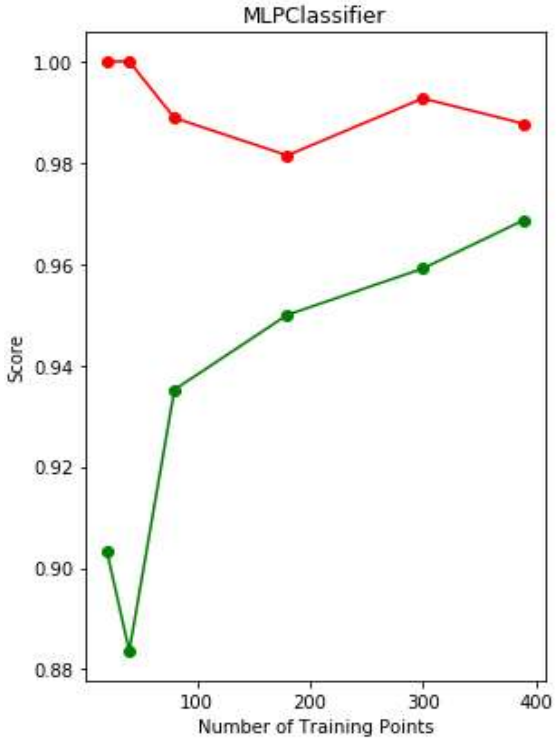

Fig 9. Learning Performance for Multi-Layer Perceptron

Table1 shows the comparison of Accuracy Score of Logistic Regression, Support Vector Machine, Random Forest Classifier and Multi-layer Perceptron.

Table1.Comparison of the accuracy score of various algorithms

Algorithm	Accuracy Score
1.Support Vector Machine	98.3
2.Linear Regression	97.2
3.Random Forest Classifier	93.2
4.Multi-Layer Perceptron	96.8

6. CONCLUSION

Our work mainly focused on comparing different types of machine learning algorithm and choosing the best model to achieve good accuracy in predicting valid disease outcomes. The analysis of the results signifies that the presence of various unwanted data will affect the accuracy of the predicting model. It is also clear that machine learning methods generally improve the performance or predictive accuracy of most prognoses, especially when compared to conventional statistical or expert-based systems. We believe that if the quality of

studies continues to improve, it is likely that the use of machine learning classifier will become much more commonplace in many clinical and hospital settings. Further research in this field should be carried out for the better performance of the classification techniques so that it can predict on more variables.

ACKNOWLEDGMENT

We would like to thank our Research Guide Prof. Farhana Kausar, and Prof. Srinivas Achar, Associate Professor in Computer Science Department, Atria Institute of Technology, Bangalore for their continuous support and guidance regarding this project work done by us. Authors are also thankful to the reviewer for going through the manuscript and giving valuable suggestions for the renovation of manuscript. We would also like to thank the Department of Computer Science, Atria Institute of Technology, Bangalore for providing us with the facility for carrying out the simulations. Last, but not the least we would like to thank our family, who has acted as a beacon of light throughout our life. Our sincere gratitude goes out to all our comrades and well-wishers who have supported us through all the ventures.

REFERENCES

[1] Ch. Shravya, K. Pravalika, Shaik Subhani," Prediction of Breast Cancer Using Supervised Machine Learning Techniques", in International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-8 Issue-6, April 2019.

[2] S. Kharya, D. Dubey, and S. Soni," Predictive Machine Learning Technique for Breast Cancer Detection", in (IJCSIT) Vol. 4 (6), 2013, 1023-1028

[3] Vikas Chaurasia, BB Tiwari and Saurabh Pal, "Prediction of benign and malignant breast cancer using data mining techniques", Journal of Algorithms and Computational Technology

[4] Haifeng Wang and Sang Won Yoon – Breast Cancer Prediction using Data Mining Method, IEEE Conference paper.

[5] Logistic Regression for Machine Learning - Machine LearningMasteryhttps://machinelearningmastery.com/logistic-regression-for-machine-learning/

[6] Comparative Study of Machine Learning Algorithms for Breast Cancer Detection and Diagnosis Dana Bazazeh and Raed Shubair [978-1-5090-5306-3/16/$31.00c 2017 IEEE]

[7]https://dataaspirant.wordpress.com/2014/09/19/supervised-and-unsupervised-learning/

[8] Pooja Mudgil, Mohit Garg, Vaibhav Chhabra, Parikshit Sehgal, Jyoti," Breast Cancer Prediction Algorithm Analysis", in International Journal of Advance Research, Ideas and Innovations in Technology (Volume 5, Issue 3)

[9] Vishabh Goel," Building a simple machine learning model on Breast Cancer Data".