Grenze International Journal of Engineering and Technology, Jan Issue

Building a Breast Cancer Prediction Model using Machine Learning Algorithms

Dr.G.Vijaya
Associate Professor, Department of Computer Science & Engineering, Bapatla Women's Engineering College
Bapatla, Andhra Pradesh, India
Email: viji.pooshan@gmail.com

***Abstract*—When it comes to cancer in Indian women, breast cancer is one of the most common and frequently occurring cancer types. In most of the developing countries like India, the overhead charges for breast cancer are very high which leads to a poor mortality rate when compared to western countries. As of now, large numbers of researches have been carried out based on deep learning and machine learning prediction of Breast cancer. This paper reviews some of the existing machine learning algorithms like Decision Tree (DT), Multi-Layer Perceptron (MLP), Random Forest (RF) and Support Vector Machines (SVMs) in linear (SVM-L), polynomial (SVM-P), and radial basis function (SVM-RBF) on the Breast Cancer Wisconsin (BCW) data set using Python. An experimental result shows that SVM-L outperforms all the above algorithms with 96.5% accuracy.**

***Index Terms*— Decision Tree, Multi-Layer Perceptron, Random Forest, Support Vector Machine – Linear, - Polynomial, - Radial Basis Function, Breast Cancer Wisconsin dataset, Python.**

I. INTRODUCTION

Breast cancer (BC) is one of the common deadliest diseases worldwide. Normally like other cancer, BC also forms in the tissues of the breast. BC is found in both men and women, but in men BC is uncommon. In India [1], BC is typical cancer in women; particularly the women in cities are the ones who are affected very often than in rural India. In India, for every 4 minutes one woman is diagnosed and for every 13 minutes, one woman dies due to this cancer. When compared with European countries the survival rate of at least five years after the treatment is very low, because of lack of awareness and absence of screening. Nearly half of the existing patients are in Stage III and IV, where the probabilities of continued existence are exceptionally low. According to National Health Mission [2], BC touches its peak in the age group of the early 30s to late 50s in India, and the count increases rapidly. The appropriate screening technique to detect BC in the earlier 30s age group is the self-breast examination and in the late 50s, this can be accomplished by yearly mammograms. Once the symptoms such as changes in the breast size, development of lumps or the wrinkling of skin arise, immediately consult the doctor as soon as possible. If there is any doubt of having BC, the patient should undergo a mammogram which is followed by a biopsy. Once the cancer is confined through autopsy, the stages of the cancer are confirmed from cross-sectional imaging. During the past two decades, machine learning techniques play a vital role in diagnosis and detection of all sorts of images particularly cancer images. The summary of the existing machine learning approaches for Breast Cancer are depicted in Table I.

Grenze ID: 01.GIJET.8.1.59

TABLE I: SUMMARY OF EXISTING MACHINE LEARNING APPROACHED FOR BREAST CANCER DATASET

Author	**Algorithm used**	**Dataset used**	**Accuracy in %**
[4]	SVM	Wisconsin Breast Cancer	97.13
[5]	ANN	Wisconsin Breast Cancer	98.57
[6]	SVM	Wisconsin Breast Cancer	97.07
[7]	DNN	Real-time X-Ray images	96.3
[8]	DL using Adam Gradient Descent Learning	Wisconsin Breast Cancer	98.24
[9]	SVM	Wisconsin Breast Cancer	97.89
[10]	ANN	Wisconsin Breast Cancer	99
[11]	SVM	Wisconsin Breast Cancer	92.78
[12]	Random Forest	Wisconsin Breast Cancer	99.76

II. DATA SET USED

The dataset used for this experiment is the Breast Cancer Wisconsin (original) Dataset [3], acquired from Kaggle. Originally this dataset was generated by Dr. William H. Wolberg, General Surgery Dept., University Of Wisconsin, and Madison, USA. The features of this dataset are extracted from the patients using fluid needle aspirate (FNA) with solid breast masses and a convenient GUI-based computer program called Xcyt, which is skilled to perform the analysis of the features of cells based on MRI scans. The program then extracts ten mean values, ten standard error values, and ten 'worst' real-valued features from each cell in the sample, returning a total of 30 real-valued vectors.

	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
0	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3001	0.14710	0.2419	0.07871	1.0950	0.9053	8.589	153.40	0.006399	0.04904	0.05373	0.01587	0.03003
1	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435	0.7339	3.398	74.08	0.005225	0.01308	0.01860	0.01340	0.01389
2	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1974	0.12790	0.2069	0.05999	0.7456	0.7869	4.585	94.03	0.006150	0.04006	0.03832	0.02058	0.02250
3	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2414	0.10520	0.2597	0.09744	0.4956	1.1560	3.445	27.23	0.009110	0.07458	0.05661	0.01867	0.05963
4	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1980	0.10430	0.1809	0.05883	0.7572	0.7813	5.438	94.44	0.011490	0.02461	0.05688	0.01885	0.01756

Figure 1. First five rows of the BCW Dataset

From the Figure 1, it was clearly showed that the ten real-valued features for each cell nucleus are: radius which is computed from the mean of distances from center to points on the perimeter, the texture is the standard deviation of gray-scale values, perimeter, area, smoothness is the local variation in radius lengths, compactness which is computed from perimeter^2 / area - 1.0, the concavity is the severity of concave portions of the contour, concave points represents the number of concave portions of the contour, symmetry, and fractal dimension.

```
Features:  ['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']
Labels:  ['malignant' 'benign']
```

Figure 2. Features of the BCW Dataset

Once the BCW dataset is loaded and the features are extracted using Python and showed in Figure 2. According to the data set 30 input parameters are dependent variables and the one independent variable is the output parameter which is mentioned as 'Label'. Also, this data set has no 'missing attribute' values. The main aim of this study is to classify the patient having breast cancer as malignant or benign.
The output labeled as 'malignant' is specified as '1' and 'benign' is specified as '0', which is depicted in the Figure 3. The total numbers of instances used for this experiment is 569, out of which 357 are benign and 212 are malignant.

```
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 1 0 0 0 0 0 0 0 0 1 0 1 1 1 1 1 0 0 1 0 0 1 1 1 1 0 1 0 0 1 1 1 1 0 1 0 0
 1 0 1 0 0 1 1 1 0 0 1 0 0 0 1 1 1 0 1 1 0 0 1 1 1 0 0 1 1 1 1 0 1 1 0 1 1
 1 1 1 1 1 1 0 0 0 1 0 0 1 1 1 0 0 1 0 1 0 0 1 0 0 1 1 0 1 1 0 1 1 1 1 0 1
 1 1 1 1 1 1 1 1 0 1 1 1 1 0 0 1 0 1 1 0 0 1 1 0 0 1 1 1 1 0 1 1 0 0 0 1 0
 1 0 1 1 1 0 1 1 0 0 1 0 0 0 0 1 0 0 0 1 0 1 0 1 1 0 1 0 0 0 0 1 1 0 0 1 1
 1 0 1 1 1 1 1 0 0 1 1 0 1 1 0 0 1 0 1 1 1 1 0 1 1 1 1 1 0 1 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 1 1 1 1 1 1 0 1 0 1 1 0 1 1 0 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1
 1 0 1 1 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 0 1 0 1 1 1 1 0 0 0 1 1
 1 1 0 1 0 1 0 1 1 1 0 1 1 1 1 1 1 1 0 0 0 1 1 1 1 1 1 1 1 1 1 1 0 0 1 0 0
 0 1 0 0 1 1 1 1 1 0 1 1 1 1 1 0 1 1 1 0 1 1 0 0 1 1 1 1 1 1 0 1 1 1 1 1 1
 1 0 1 1 1 1 1 0 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 0 1 0 0 1 0 1 1 1 1 1 0 1 1
 0 1 0 1 1 0 1 0 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 0 1
 1 1 1 1 1 1 0 1 0 1 1 0 1 1 1 1 1 0 0 1 0 1 0 1 1 1 1 1 0 1 1 0 1 0 1 0 0
 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 0 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 0 0 0 0 0 0 1]
```

Figure 3. Output for the target

III. METHODOLOGY

The purpose of this paper is to predict the BC from the UCI data set by means of state-of-the-art algorithms namely: DT, MLP, RF, SVM-L, -P, & -RBF. The general architecture for predicting the BC data set is depicted in Figure 4. From Figure 4, the first step is to feed the input from the BCW data set. The next step is to extract the features from the data set and divide it into training and test set. Then and there, by means of different algorithms like DT, MLP, RF & SVM, different models have been built and the best one is chosen by means of Accuracy.

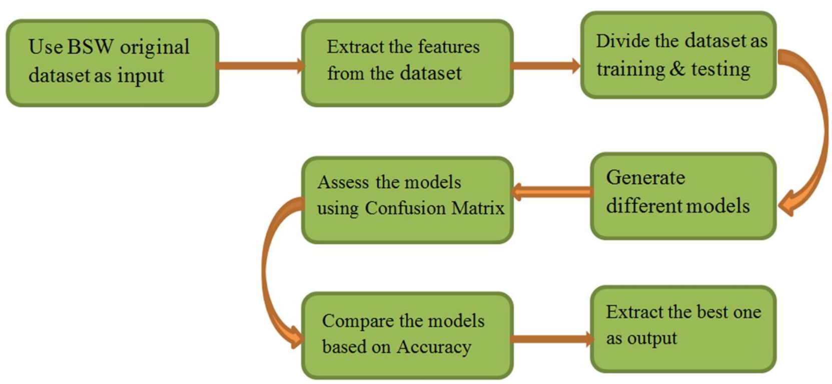


Figure 4. Basic architecture of Breast Cancer Detection model

A. Decision Tree (DT)

DT is a tree-like structure where the information is split into nodes and leaves based on diverse conditions. Mostly all the DTs are used for both classification and regression models. The root of the DT is the basic structure of the building block of the tree. The intermediate - term 'node' represents a test on the entities of the nodes and 'leaf' represents the outcome/result of the entities. The trackway from root to leaf represents the 'classification rules'.

B. Multi-layer Perceptron (MLP)

A perceptron is a one - dimensional (1D) classifier which classifies the input entities into two different groups by a straight line. As the name suggests, MLP contains more than one perceptron. MLP consists of 3 different layers as shown in Figure 5: input, hidden, and output layers. The input layer is the one that receives the input from the source, and the output layer is the one that produces the expected output. In between these two layers, a random number of hidden layers are presently based on the applications of MLP.

C. Random Forest (RF)

RF [13] is based on ensemble learning – a method to combine several classifiers to decode a difficult problem in order to upgrade the execution. Also, RF is a combination of different DTs to predict the class of a given data

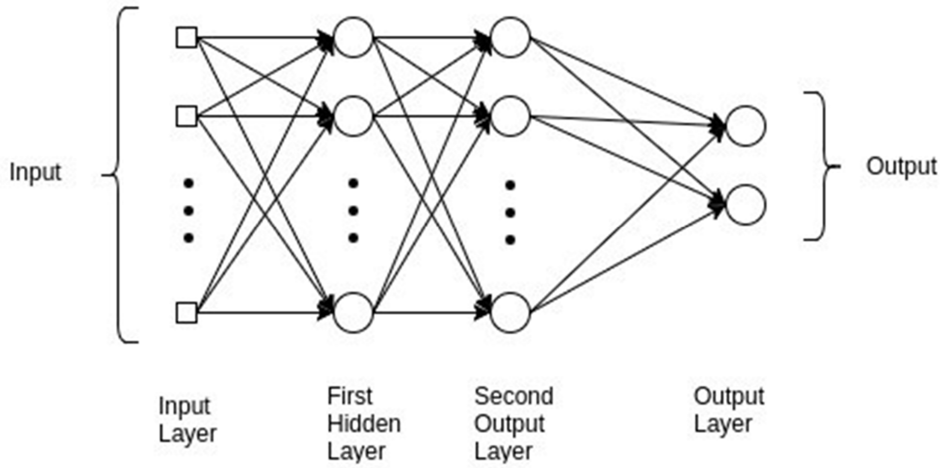


Figure 5. MLP Architecture

set. Figure 6 depicts the architecture of Random forest. The Working process can be explained in the below steps and diagram:

Step-1: Select random K data points from the training set.
Step-2: Build the decision trees associated with the selected data points (Subsets).
Step-3: Choose the number N for decision trees that you want to build.
Step-4: Repeat Step 1 & 2.
Step-5: For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

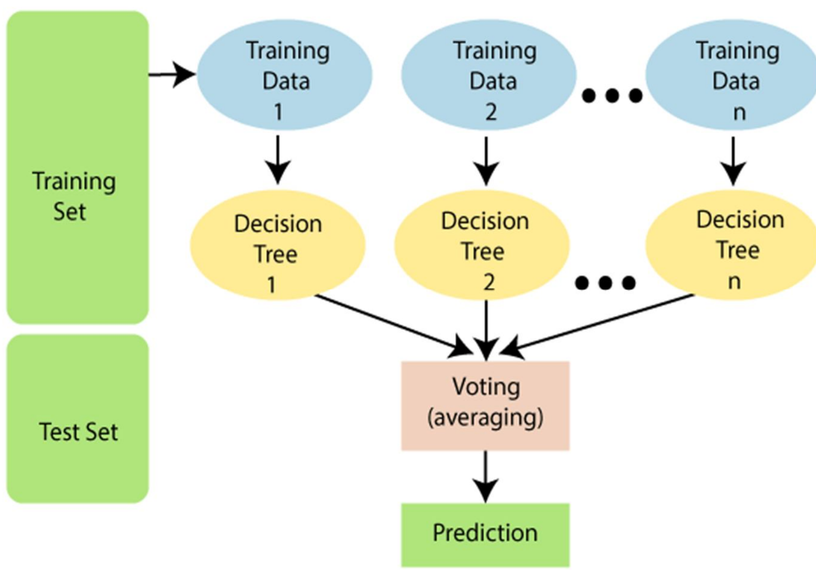
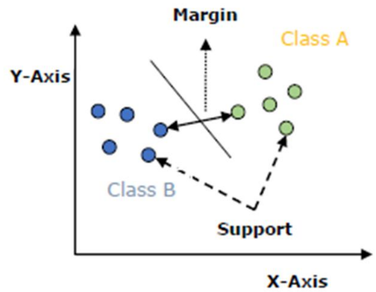


Figure 6. Decision tree architecture

D. Support Vector Machine (SVM)

Support vector machines (SVMs) [14] are flexible and commanding supervised machine learning algorithms primarily used for classification and regression. Almost always, they are used in classification problems. An SVM model is basically a representation of different classes in a hyperplane in a multidimensional space. The hyperplane will be generated in an iterative manner by SVM so that the error can be minimized. The goal of SVM is to divide the datasets into classes to find a maximum marginal hyperplane (MMH).



The followings are important concepts in SVM –

- **Support Vectors** − Datum that is nearest to the hyperplane is named as support vectors. Hyperplane has to be defined with the assistance of this datum.
- **Hyperplane** − is a sub-space that divides a set of objects of diverse classes.
- **Margin –** is defined as the gap between two lines on the nearest datum of diverse classes. A big margin is considered as a good one and a small margin is considered as a bad one.

The key objective of SVM is to split the datasets into modules to discover a maximum marginal hyperplane (MMH) by means of: −

- Producing hyperplanes through repetition, which segregates the modules in a preeminent way.
- Pick the hyperplane that splits the modules appropriately.

IV. Results & Discussion

To predict whether the patients having benign or malignant cancer, which is collected from the BCW dataset by means of the machine learning algorithms such as DT, MLP, RF, SVM – L, -P & - RBF. The accuracy of the above-said algorithms is illustrated in Table II. For experimentation purposes, Scikit – learn a non-proprietary software library for machine learning in Python programming language. Google Colaboratory, simply “Colab”, is a non-proprietary web application that permits everyone to note down and implement Python code through the browser.
The model should be evaluated by means of different metrics such as confusion matrix, area under curve (AUC), cross-validation etc. Here we are using confusion matrix as our performance measure. Confusion matrix is nxn matrix, where n is the number of predicted classes.

Confusion Matrix		Target			
		Positive	Negative		
Model	Positive	a	b	*Positive Predictive Value*	a/(a+b)
	Negative	c	d	*Negative Predictive Value*	d/(c+d)
		Sensitivity	*Specificity*	**Accuracy** = (a+d)/(a+b+c+d)	
		a/(a+c)	d/(b+d)		

Figure 7. Confusion matrix

From figure 7, we can define the following terms:
Sensitivity: is the ratio between Target Positive value and the Positive value of Model
Specificity: is the ratio between Target Negative value and the Negative value of Model
Positive Predictive Value: is the ratio between Positive values that were identified correctly
Negative Predictive Value: is the ratio between Negative values that were identified correctly
Accuracy: the ratio between the total numbers of predictions those were correctly identified
The calculated performance measures of the models are illustrated in Figure 8. Based on the performance models, SVM-L beats all machine learning algorithms we have studied so far with the peak accuracy of 96.49% whereas SVM – RBF & SVM – P holds the second and third most accuracy of 92.39% & 91.81% respectively. In the overall assessment, the SVM model masters the other machine learning algorithm in our study.

TABLE II. ACCURACY OF ALL THE MODELS

S.No.	Model	Accuracy
1.	Decision Tree	92.397
2.	Multi-Layer Perceptron (MLP)	90.058
3.	Random Forest	90.643
4.	**Support Vector Machine – Linear (SVM – L)**	**96.491**
5.	Support Vector Machine – Polynomial (SVM – P)	91.812
6.	Support Vector Machine – Radial Basis Function (SVM – RBF)	92.397

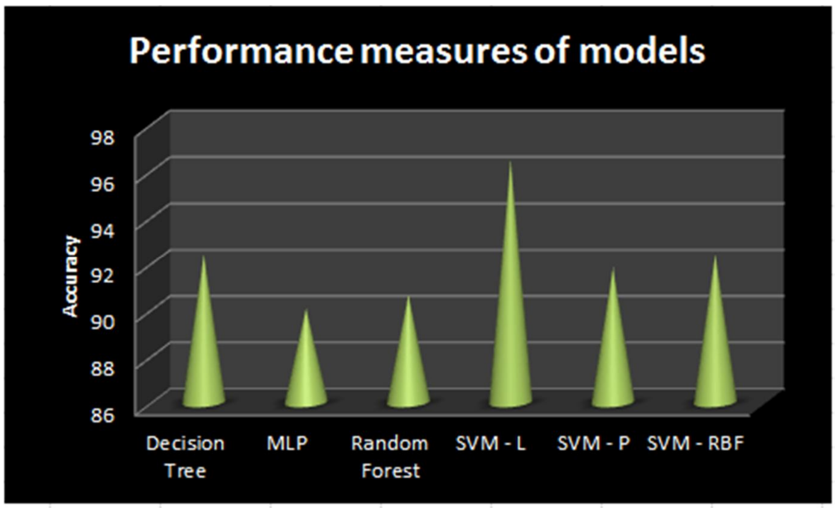


Figure 8. Performance measures of models

V. CONCLUSION

The diagnosis process in the medical field is much more complicated in terms of time and cost. Due to the advent of computer-aided diagnosis (CAD) system, the complications were resolved and even CAD acts as a suggesting tool after it is integrated with the machine learning techniques. This paper proposes one such complication named Breast cancer along with the six machine learning algorithms namely Decision Tree (DT), Multi-Layer Perceptron (MLP), Random Forest (RF) and Support Vector Machines (SVMs) in linear (SVM-L), polynomial (SVM-P), and radial basis function (SVM-RBF) on the Breast Cancer Wisconsin (BCW) data set and the performance measures have been predicted. Based on the prediction, SVM – L outbreathes the existing algorithms with the highest accuracy of 96.49%.

REFERENCES

[1] https://cytecare.com/blog/statistics-of-breast-cancer/
[2] https://darpg.gov.in/sites/default/files/National%20Health%20Mission.pdf
[3] https://data.world/health/breast-cancer-wisconsin
[4] Ramik Rawal, “Breast Cancer Prediction using Machine Learning”, Journal of Emerging Technologies and Innovative Research, May 2020; Vol.7, Issue 5, pp: 13-24
[5] Md.Milon Islam, Md. Rezwanul Haque, Hasib Iqbal, Md. Munirul Hasan, Mahmudul Hasan, Muhammad Nomani Kabir. Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques. SN Computer Science (2020), 1:290
[6] F.M. Javed Mehedi Shamrat, Md. Abu Raihan, A.K.M. Sazzadur Rahman, Imran Mahmud, Rozina Akter. An Analysis on Breast Disease Prediction Using Machine Learning Approaches. International Journal of Scientific & Technology Research. Feb. 2020; Vol.8 Issue 9; pp:2450-2455
[7] R. Chtihrakkannan, P. Kavitha, T.Mangayarkarasi, R.Karthikeyan. Breast Cancer Detection Using Machine Learning. International Journal of Innovative Technology and Exploring Engineering. Sep. 2019, Vol. 8, Issue 11; pp: 3123-3126

[8] Puja Gupta, Shruti Garg. Breast Cancer Prediction Using varying parameters of Machine Learning Models. Procedia Computer Science. Third International Conference on Computing and Network Communications (CoCoNet'19). 2020; 171, pp:593-601
[9] Somil Jain, Puneet Kumar. Prediction of Breat Cancer Using Machine Learning. Recent Advances in Computer Science and Communications. 2020; Vol. 13, Issue 5; pp: 901-908
[10] M.Divyavani, Dr.G.Kalpana. An Analysis of SVM & ANN Using Breast Cancer Dataset. AEGAEUM Journal. 2020; Vol. 8, Issue 12; pp: 369-379
[11] Ch. Shravya, K. Pravalika, Shaik Subhani, "Prediction of Breast Cancer Using Supervised Machine Learning Techniques", International Journal of Innovative Technology and Exploring Engineering (IJITEE). 2019; Vol.8, Issue 6; pp: 1106-1110.
[12] J. Siva Priya, V. Arvind Kumar, S. Siddarth Sai, S. Sriram. Breast Cancer Prediction using Machine Learning. International Journal of Recent Technology and Engineering. 2019; Vol.8, Issue 4; pp: 4879-4881
[13] https://medium.com/nerd-for-tech/digit-recognition-using-ml-algorithm-b8171caff87f
[14] https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_classification_algorithms_support_vector_machine.htm