Machine Learning Algorithms: Diagnosing Breast Cancer

Sridevi N, Kulkarni Varsha, Maria Navin J R

Abstract*: Breast Cancer has become one of the common diseases not only in women but also in few men. According to research, the demise rate of females has increased mainly because of Breast Cancer tumor. One out of every eight women and one out of every thousand men are diagnosed with breast cancer. Breast cancer tumors are mainly classified into two types: Benign tumor which is a non-cancerous tumor and other one is malignant tumor which is a cancerous tumor. In order to know which type of tumor a patient has; the accurate and early diagnosis is a very crucial step. Machine Learning (ML) algorithms have been used to develop and train the model for classification of the type of tumor. For accurate and better classification several classification algorithms in ML have been trained and tested on the dataset that was collected. Already algorithms like Naïve Bayes, Random Forest, K-Nearest Neighbor and SVM showed better accuracy for classification of tumor. When we implemented Multilayer Perceptron (MLP) algorithm it gave us the best accuracy levels among all both during training as well as testing .i.e. 97%. So, the exact classification using this model will help the doctors to diagnose the type of tumor in patients quickly and accurately.

Keywords-Benign, Malignant, Naïve Bayes, Random Forest, K-Nearest Neighbor, SVM, MLP, Accuracy.

I. INTRODUCTION

Breast Cancer is one of the most common cancers seen in most of the women across the globe. So as a precaution regular checkup should be made mandatory in hospitals. Mammography is the technique which is used to detect whether any tumor is present or not. Later, Biopsy is the technique that is used to test or classify what type of tumor does the patient have. Breast cancer tumors are mainly classified into two types: One is Benign which is a non-cancerous tumor and has no harmful effect on the body where as another type is Malignant which is a cancerous tumor which causes severe damage to the surrounding tissues in the body. If the patient is diagnosed with malignant tumor the doctors will perform biopsy to know the severity of the tumor which is time consuming process. So we are using Machine Learning (ML) algorithms to develop and train the model which can easily classify the type of tumor with high accuracy levels. Several ML algorithms have been used for training and testing on the dataset that was collected. The results showed that the accuracy levels kept on increasing for different algorithms and the algorithm that showed the best accuracy was selected and trained well.

Revised Manuscript Received on July 22, 2019.

Sridevi N, Dept of CS&E, Sri Venkateshwara College of Engineering, Bangalore, Karnataka, INDIA

Kulkarni Varsha, Dept of CS&E, Sri Venkateshwara College of Engineering, Bangalore, Karnataka, INDIA

Maria Navin J R, Dept of IS&E, Sri Venkateshwara College of Engineering, Bangalore, Karnataka, INDIA

Email : n.sridevi5@gmail.com, varsha_kulkarni@yahoo.com, marianavin.jr@gmail.com

II. MACHINE LEARNING TECHNIQUES

Machine Learning algorithms build a mathematical model based on sample data or also termed as training data in order to make predictions or decisions without being explicitly programmed to perform the task.

Several Machine learning algorithms are used for classification of breast cancer tumor.

A. Naïve Bayes Classifier:

Naïve Bayes Classifier is the simple technique used for classification purpose. It is a probabilistic classifier and is mainly based on Bayes theorem. It is a classification algorithm for binary and multi-class classification problems.

Bayesian Rule:

$$P(A|B) = P(B|A)\,P(A)/P(B)$$

An advantage of naïve Bayes is that it only requires a small number of training data to estimate the parameters necessary for classification.

When our model was trained using Naïve Bayes classifier, the results showed 96% accuracy for training and 95% as the testing accuracy.

B. Random Forest:

Random Forest algorithm is a supervised classification algorithm. This algorithm creates the forest with number of trees. The higher number of trees in the forest gives the higher accuracy results as shown in figure 1.

When we used random forest algorithm for our classification, results it showed 99% as training accuracy which was very high but when we tested our model it could give only 94% accuracy for testing.

C. *K-Nearest Neighbour Algorithm:*

K-Nearest Neighbor is a non-parametric machine learning algorithm. The main feature and use of this algorithm is a database which is separated into several classes from the data points to predict the classification of new sample point.
When our model is trained it is giving 96% for training accuracy and 95% for testing accuracy.

D. *Support Vector Machine:*

Support vector machine is a supervised learning used for regression and classification challenges. This algorithm is used to get fast and better outcomes and results for different learning tasks.

When our model is trained by using this algorithm it is giving 97% for training accuracy and 96% for testing accuracy.

III. EXISTING SYSTEM

In our existing system different Machine Leaning algorithms were applied to find out the accuracy levels of the Breast Cancer dataset.
The results shown for Naïve Bayes, Random Forest, K-Nearest Neighbor, and SVM algorithms were 95%, 94%, 95% and 96% of testing accuracy respectively. But the major disadvantages of the existing system are that these algorithms will show good accuracy levels only for small datasets. If the datasets are large, the accuracy levels decrease. Another drawback of existing system is that it is very tedious work and is time consuming.
In order to overcome the drawbacks of existing system and also to improve the accuracy levels we have used another algorithm in Machine Learning is that Multiple Layer Perceptron (MLP) algorithm to train our model so that it can classify the type of Breast Cancer tumor quickly and accurately.

IV. PROPOSED SYSTEM

The proposed system is a machine learning model that can diagnose the type of breast cancer (malignant or benign) with greater accuracy than what has been mentioned in the existing system. To achieve this, we applied the Multilayer perceptron algorithm to train our dataset, which in turn gave us better accuracy than the algorithms used in the existing systems.

A. *Multilayer Perceptron Algorithm*

The Multilayer perceptron algorithm is a feed forward artificial neural network also known as the MLP algorithm. The MLP algorithm can have multiple neurons in one layer and in this way it can have multiple layers which perform computation which in turn results in the model having a better accuracy. This algorithm makes use of activation functions in order to make non-linear predictions. MLP consists of an input layer, hidden layers, and an output layer.

During the process, the data is sent to the input layer. This data is further multiplied with its respective weight which is calculated with respect to the comparison of its importance with other inputs and then it propagates forward through the network to the neuron of the first hidden layer. This neuron then makes use of an activation function to process the output. This is shown in equation (1).

$$\text{Output}=f(X1*W1+X2*X2\ldots\ldots Xn*Wn) \quad (1)$$

Here X represents the input, W represents the weight and f represents the activation function and n represents the total number of input features.

The very common activation function that is used is the hyperbolic tangent function. The activation function does non -linear classification which gives the model a better accuracy. MLP utilizes a backpropagation technique for training.

B. *Backpropogation Technique*

MLP algorithm trains the data set using the backpropagation technique. The output obtained after the computation is compared with the output that we have in the supervised data set. If there is an error, the output value is propagated backwards through the neural network and the weights are adjusted accordingly. This process continues until the output is in maximum close proximity with the expected output in the supervised data.

V. METHODOLOGY

Methodology is a process that gives complete flow of the process in step by step manner.
We collect the data after mammography; convert the data into .csv format then dataset is preprocessed. Preprocessing the dataset includes data transformation, data cleaning i.e. removing null values and data anonymization i.e. removing of sensitive data that reveals patients identity. Once pre-processing is done the data is split into 80% training and 20% testing. The machine is trained via algorithms on dataset and once its trained we can know the percentage of accuracy at which it has got trained. Based on the trained data it will predict the testing accuracy of 20% of data which it had not come across during the training. The last stage is validating, we can key in the input attributes of the patient in user interface and find out whether the patient has malignant tumor or benign tumor.

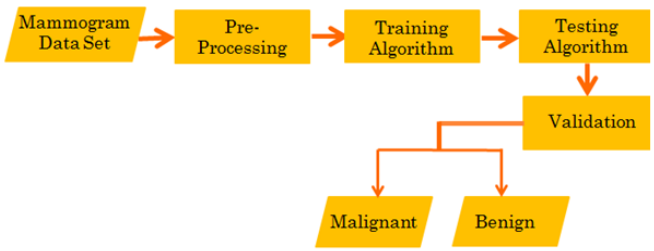

Figure 2. Methodology

VI. MODEL PERFORMANCE TESTING

One of the most important and critical factor that determines the performance of a machine learning model is testing. For this, we need to consider a confusion matrix. Based on the values obtained from the confusion matrix, the accuracy, precision, recall and f1score are calculated.

A. *Confusion matrix*

Confusion matrix gives information about the model that is classified as patients having

malignant or benign tumor and diagnosed correctly or visa-versa. The patients who have malignant tumor and are correctly diagnosed with malignant tumor are termed as true positive and patients who have malignant tumor but diagnosed as benign tumor termed as false negative. Similarly the patients who have benign tumor and are correctly diagnosed with benign tumor are termed as false positive and patients who have benign tumor but diagnosed as malignant tumor are termed as false negative.

TABLE I. CONFUSION MATRIX

	Diagnosed with Malignant Tumor (cancerous)	Diagnosed with Benign Tumor (non-cancerous)
Malignant Tumor	True positive (TP)	False Negative (TN)
Benign Tumor	False positive(FP)	True Negative(TN)

a. Classification is malignant and benign tumor. (Table footnote)

B. *Accuracy*

Accuracy determines the percentage of right predictions out of all the predictions made by the model. The equation with regard to accuracy is given below

$$\text{Accuracy} = \frac{TP+TN}{\text{Total Predictions}} \times 100\%$$

C. *Precision*

Precision determines the percentage of right malignant predictions out of all the malignant predictions made by the model and the percentage of right benign predictions out of all the benign predictions made by the model. The equation with regard to precision is shown in below

$$\text{Precision (Malignant)} = \frac{TP}{TP+FP} \times 100\%$$

$$\text{Precision (Benign)} = \frac{TN}{TN+FN} \times 100\%$$

D. *Recall*

Recall determines the right malignant predictions out of all the patients labeled as having malignant tumor in the dataset and the percentage of right benign predictions out of all the patients labeled as having benign tumor in the dataset. The equation is shown below.

$$\text{Recall (Malignant)} = \frac{TP}{TP + FN} \times 100\%$$

$$\text{Recall (Benign)} = \frac{TN}{TN + FP} \times 100\%$$

E. *F1 Score*

F1 score determines the accuracy of the model with respect to precision and recall.

$$\text{F1 Score} = \frac{2 * (\text{Recall} * \text{Precision})}{(\text{Recall} + \text{Precision})} \times 100\%$$

VII. CONCLUSION

As we know that the existing system, the PCA algorithm is used with Classification algorithms Naïve Bayes, Random Forest, K-Nearest neighbor and Support Vector Machine and found the accuracy to 95% for small dataset, but doesn't give a same/good accuracy for larger datasets.

Therefore by introducing neural networks and concepts of deep learning i.e. by proposing a system which uses Multilayer Perceptron algorithm we can find whether the breast cancer is benign or malignant accurately even for larger dataset with 97% accuracy.

REFERENCE

[1] Rafaqat Alam Khan, Taseer Suleman, Muhammad SajidbFarooq, Muhammad Hassan Rafiq and Muhammad Arslan Tariq. "Data Mining Algorithms for Classification of Diagonostic Cancer Using Genetic Optimization Algorithms",2017, Vol 17 No.12,December 2017.

[2] Dana Bazazeh and Raed Shubair. "Compartive Study of Machine Learning Algorithms for Breast Cancer Detectio and Diagonsis",2016.978-1-15090-5306-3/ 2016.

[3] Python Machine Learning – Sebastian Rashka & Vahid Mirjalili.