Classifying the Breast Cancer tumors through Machine Learning Classification algorithms

Sanjanashree S [1], Shilpa G S [1], Supriya N Shetty [1], Trapthi R Shetty [1], Avinash B[2]

[1]Student, ISE/ The National Institute of Engineering, Mysuru, Karnataka, India

[2]Assistant Professor, ISE/ The National Institute of Engineering, Mysuru, Karnataka, India

(E-mail: trapthi.shetty31@gmail.com)

***Abstract*—Changes caused to the DNA within the cells results in Cancer. The individual genes present in the cells have the set of instructions to be performed by the cells. The errors in these cells cause the abnormal growth and allow the cells to become cancerous. The first stage of cancer diagnosis is to diagnose whether tumors are benign or malignant. Malignant tumors are cancer leading tumors where as benign tumors are harmless tumors. Breast cancer is the second leading cancer that causes death in cancerous women. The Breast cancer is a malignant tumor. This paper deals with classifying the tumors caused in breast as benign and malignant using classification algorithm and also comparing the accuracy of results of algorithms. It evaluates the accuracy of classification algorithm such as Logistic Regression, Nearest Neighbour Intuition, Naive Bayes Classifier, Random Forest, Support Vector Machines using K-fold cross validation. Cross validation improves the performance and validates the data for testing the model. Evaluation of algorithm is done using the University of California, Irvine(UCI) machine learning repository dataset, was curved to predict the existence of Breast Cancer.**

Keywords*— Breast Cancer, Logistic Regression, K-Nearest Neighbour Intuition, Naive Bayes Classifier, Random Forest, Support Vector Machines, k-fold cross validation.*

I. INTRODUCTION

Cancer is caused by the changes to the DNA within cells. The DNA inside the cell is packed into a large number of individual genes. Each gene contains set of instructions to be performed by the cells as well as how to grow and divide. Tobacco consumption, obesity, poor diet, lack of physical activity, certain infection, exposure to ionizing radiations, environment pollutants causes abnormalities in genome which results in errors in these instructions. This causes abnormal mutation which leads to the lumps like structure called tumors. These tumors are classified as Benign Tumors and Malignant Tumors. Benign tumors do not spread and are harmless. Malignant tumors are cancerous tumors which have abnormal cell growth with ability to spread to other parts of the body. Most common types of cancers include Breast cancer, Leukemia, Prostate cancer, Lung cancer, Non-Melanoma Skin cancer, Kidney cancer etc.

Breast Cancer is the second leading cancer that causes death in cancerous women. Cancer cells growing in breast tissue is breast cancer. It most often begins with milk-producing ducts. Along with lumps other signs of breast cancer may include change in breast shape, a newly inverted nipple, red or patches in skin, fluid coming from the nipple. One in eight women will be diagnosed with breast cancer in her life time. We can avoid breast cancer by following healthy activities. Women with certain risk factors like drinking alcohol, obesity, menopause, beginning of menstruation etc are more likely to have breast cancer. As per 2015, 2.1 million are affected from breast cancer and among them 533600 deaths are found. Survival rate is approximately 85%.

II. CLASSIFICATION ALGORITHM OF MACHINE LEARNING

Classification Algorithms of Machine Learning are used to predict the category of new data set it belongs based on the basis of training set categories which are already known.

This paper includes the comparison of accuracies of different classification algorithms such as Logistic Regression, K-Nearest Neighbours, Support Vector Machine, Naive Bayes, Random Forest classification.

A. Logistic Regression:

Logistic Regression is one of the Classification algorithms of machine learning. It is the special case of Linear Regression. It is applied to categorical data and the linear kind of model. It is used to narrate the relationship between the dependent variable and one or more independent variables. This paper includes the comparison of accuracies of different Classification algorithms such as Logistic independent variables whose values are between 0 and 1.

It is used when there are several independent variables influencing one dependent variable and its outcome.

Disadvantages: It assumes that all the predictors are independent to each other.

B. K- Nearest Neighbour:

KNN is a Supervised Classification algorithm. It is applied to both classification and regression problems. It classifies the new test case to the categories by storing all the available categorical data and considering the majority votes of the k nearest points to the new test case. The items having least Euclidean distance from the new test case are considered as nearest neighbours. It is non parametric and lazy algorithm. Because it does not make any assumptions from data distribution and do not learn anything from the model i.e it does not use training data set to do generalization.

Disadvantages: It requires high memory space to store all data and is computational expensive as it calculates Euclidean distance from each point to the new test point.

C. Support Vector Machine:

Support Vector Machine (SVM) is a supervised classification algorithm used in both classification and regression problems. It constructs the Hyper plane that separates the largest distance between the margins drawn to the nearest points of the categorical training data points. Least errors are achieved by considering the larger distance between the margins. It gives high accuracy in classifying the new test case to the categorical variables. It is used in classification of images, hand-written characters etc.

Disadvantages: It does not provide any probability estimates.

D. Naive Bayes:

Naive bayes is a supervised classification algorithm which is based on Bayes theorem and assumes that all predictors are independent i.e presence of one feature is independent of any other feature of predictor. It is useful for large data sets and textual data analysis like Natural Language Processing.

Disadvantages: It requires predictors to be independent of each other. In most of real world problems predictors are dependent on each other.

E. Random Forest Classification:

Random Forest is a supervised algorithm which is used for both classification and regression problems. It creates a set of decision trees from randomly selected train set. Instead of using only one classifier it uses multiple classifiers to predict the target. Here, each decision tree is a single classifier. Based on the maximum votes obtained by each classifier the target is predicted. Since it combines several classification models obtained by each classifier, it overall increases the performance of the model. Instead of searching the most important feature while splitting a node it searches for best feature. This gives rise to wide diversity which results in better model.

Disadvantages: Sometimes it over fits the model.

III. MODEL SELECTION TECHNIQUES

The model selection methods such as k-fold Cross Validation, Grid search are used to select the model which gives high accuracy by comparing the scores obtained from different classification algorithm model. In this paper, we are using cross validation algorithm.

Cross validation is used to predict the accuracy of data models by splitting the dataset into train set and test set randomly. Cross validation is a statistical method. Understanding and implementation of K-fold cross validation is easier. The accuracy scores obtained from k-fold cross validation is estimated which are lower biased result. Hence we have used k-fold cross validation. As specified in the name of the algorithm, k indicates the number of splits performed on the dataset. Hence, when we choose k=10, it will be 10-fold cross validation.

Advantages of using K-fold cross validation are:

- Reduction in computation time
- All the data sets will be tested exactly once and it will be used k-1 times in the training
- There will be reduction in biasing

Here we are taking breast cancer dataset from University of California, Irvine (UCI) Machine Learning repository and it was created by William H Wolberg .The dataset has total 569 instances, out of which 357 are benign and 212 are malignant. The dataset contains the features which are taken from a digitized image of a Fine Needle Aspirate (FNA) of a breast mass and it describes the characteristics of the cell nuclei present in the image. We are taking ten cell dimension variables such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, fractal dimension and then mean value, standard error and worst or largest of these features are computed for each image which results in total of 30 features. This analysis of data helps us to distinguish how worst the dimensions of malignant nuclei are greater than the worst dimensions of healthier benign ones. The visualization of data helps us to understand the data. In order to find the data distribution of features, we are using pandas visualization matplotlib which is shown in figure 1.

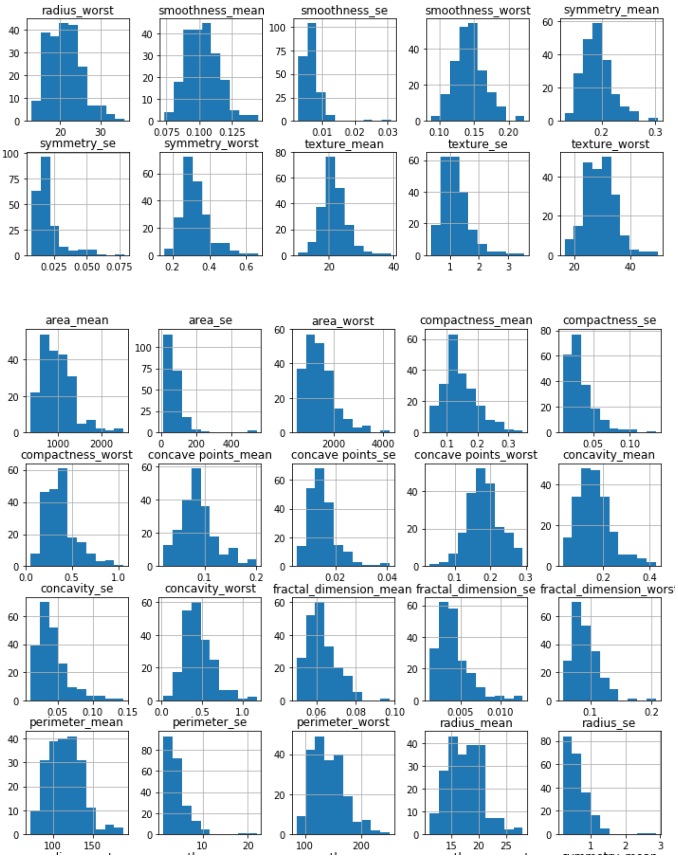

FIGURE 1

IV. OUR APPROACH

Logistic Classification Algorithm for Breast Cancer Classification:
Algorithm:

- The slope drawn by the logistic function give the best fitting line that can fit the binary classification problem dataset(here benign or malignant).
- We have to select line based on knowledge of our dataset(usually 50% is taken).Anything that falls below this line tells the probability of predicted benign tumor.
- Anything that falls above the line tells the probability of the predicted malignant tumor.

Nearest Neighbour Algorithm for Breast Cancer Classification for k=5:
Algorithm:

- Choose the number of neighbours to be considered(k).
- Take the k nearest neighbour for the new test set instance by calculating the Euclidean distance.
- Among the k neighbour, count the number of data points in each category.
- Assign the new data instance to the category where you counted the most neighbour.

We consider nearest neighbours parameters k=5, Distance metric parameter m='minkowiski' and power parameter for minkowiski p=2 which is to use standard Euclidean distance.

Support Vector Algorithm for Breast Cancer Classification:
Algorithm:

- Collect the training set {x,y}
- Choose the kernel and its parameters as well as any regulation needed
- Form the correlation matrix k
- Train your machine, exactly or approximately, to get contraction coefficient
- Use those coefficients, create your estimators

We have considered Linear Kernel for Support Vector Machine for the given data set.

Naive Bayes Algorithm for Breast Cancer Classification :
Algorithm:

- Data is divided into 2 classes D and sets of features T.
- Mean and Standard deviation is calculated for features and classes that are considered
- Probability of each feature is calculated using density of normal distribution
- The prediction of the class is made from the test set by calculating the probability of each class

Random Forest Algorithm for Breast Cancer Classification for N=10:
Algorithm:

- Pick the random k data points from the set.
- Build the Decision Tree Associated to these k data points
- Choose the number N tree of trees you want to build and repeat the above 2 steps
- For a new data point, make each one of your N tree trees. Predict the category to which the data points belongs, and assign the new data point to the category that wins the majority vote.

We have considered number of trees N=10, and criteria as entropy for the given information gain.

K Fold Cross Validation:

- Choose the value for k ie number of splits to be done
- The data sets divided into k-1 splits for training the model and the last split is used for test case. The split is done randomly.

- Fit the model by applying the above classification algorithm for each data set.
- Array of scores is obtained for each split of each classification model. The mean() of these scores is taken.
- Mean values obtained from all classification algorithm models are compared

The scores of all the classification algorithms is given below in the table:

algorithm	Logistic regression	SVM	KNN	NB	RF
accuracy	0.98	0.97	0.96	0.93	0.94

The maximum of scores of these models is computed. The model which has maximum score is used to predict the tumor type of the given test instance. This model is used to predict whether the tumor is benign or malignant

V. CONCLUSION

In this paper, we have applied classification algorithms on Wisconsin diagnostic dataset of breast cancer. We used Cross Validation Algorithm to select the models which gives high accuracy for the prediction of tumors type. The dataset that we have considered is low biased and has high variance. In order to avoid the offerfitting in prediction of tumors and to have maximum prediction accuracy k-fold cross validation is considered. As per our approach on the wisconsin data set, Logistic Regression, Naive Bayes, KNN has given high accuracy in predicting the tumor type. The predicted results may vary on the features of dataset considered.

REFERENCES

[1] Evaluating machine learning algorithms for applications with humans in the loop. Aravind Kota Gopalakrishna ; Tanir Ozcelebi ; Johan J. Lukkien ; Antonio Liotta, 2017 IEEE 14th International Conference on Networking, Sensing and Control (ICNSC), 03 August 2017

[2] Machine Learning and Learning Analytics: Integrating Data with Learning, Filippo Sciarrone, 2018 17th International Conference on Information Technology Based Higher Education and Training (ITHET) : 06 August 2018

[3] A Study of Machine Learning in Healthcare, Rohan Bhardwaj ,Ankita R. Nambiar,Debojyoti Dutta, 2017 IEEE 41st Annual Computer Software and Applications Conference

[4] Interpretable Machine Learning in Healthcare, Muhammad Aurangzeb Ahmad ; Ankur Teredesai, 2018 IEEE, International Conference on Healthcare Informatics (ICHI) 26 July 2018

[5] L.A. Altonen, R. Saalovra., P. Kristo, F. Canzian, A. Hemminki, Peltomaki P, R. Chadwik, A. De La Chapelle, "Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of molecular screening for the disease", N Engl J Med, Vol. 337, pp. 1481–1487, 1998.

Trapthi R Shetty
Final year BE The National Institute of Engineering, Mysuru

Currently pursuing final year bachelor's degree in Information science and Engineering in The National Institute of Engineering, Mysuru. She has done projects in the area of machine learning, Web Development, DBMS.

Supriya N Shetty
Final year BE The National Institute of Engineering, Mysuru

Currently pursuing final year bachelor's degree in Information science and Engineering in The National Institute of Engineering, Mysuru. She has done projects in the area of machine learning, Web Development, DBMS.

Shilpa G S
Final year BE The National Institute of Engineering, Mysuru

Currently pursuing final year bachelor's degree in Information science and Engineering in The National Institute of Engineering, Mysuru. She has done projects in the area of machine learning, Web Development, DBMS.

Sanjanashree S
Final year BE The National Institute of Engineering, Mysuru

Currently pursuing final year bachelor's degree in Information science and Engineering in The National Institute of Engineering, Mysuru. She has done projects in the area of machine learning, Web Development, DBMS.