Breast Cancer Prediction Using Statistical Features

Shruti Sharma[1] and Deval Verma[2]

[1]Chandigarh University
1Department of Mathematics, Chandigarh University, Punjab- 140413
srts1906@gmail.com
[2]Bennett University, Times Group
School of Computer Science Engineering and Technology
Greater Noida-201310
deval09msc@gmail.com

Abstract

As we all know breast cancer is a very devastating disease. By each passing day number of deaths in women are increasing because of breast cancer. The accuracy rate of each technique varies depending on the situation, tools, and datasets. Since their effectiveness has been established, machine learning techniques have gained popularity as a field of study. They can greatly help with the early diagnosis and prediction of breast cancer processes. In this work a logistic regression and random forest machine learning techniques were used to detect and predict breast cancer. In this work, breast cancer prediction is carried out using statistical properties and classified using Random Forest Classifier and Logistic Regression on Wisconsin Breast Cancer Diagnosis Dataset.

Keywords. Breast Cancer Prediction, Machine Learning, Ensemble Techniques, Data Mining, Deep Learning, Algorithms.

1. INTRODUCTION

The proportion of women who die from breast cancer worldwide has increased dramatically in the last few decades, making it the most lethal and heterogeneous disease of our time. Women die from this disease more often than any other disease [1]. Breast cancer is caused by abnormal growth of fatty and fibrous tissues in the breast. A combination of data mining and machine learning algorithms is being used to predict breast cancer [2]. One of the most important tasks is finding the most appropriate and suitable algorithm to predict breast cancer. Breast Cancer begin to spread when cellular growth becomes unchecked, leading to malignant tumours [3]. There are different stages of cancer caused by the cancer cells spreading throughout the tumours. A cancerous condition caused by the spread of cells and tissues throughout the body is breast cancer.

When we collect data of various types of breast cancer prevailing worldwide, we get to know that it's a huge raw dataset which we need to clean and analyse, using data mining

techniques, and algorithms. Any kind of disease can be discovered with the help of these functions. Statistics and machine learning are employed in the diagnosis of cancer disorders such as, lung cancer, prostate cancer and leukaemia [6], as well as databases, fuzzy sets, storage warehouses, and neural networks can be used. In traditional cancer detection, three tests are conducted: clinical examination, radiological imaging, pathology tests. This method is called "the gold standard [7]". The model is aimed at predicting unseen data and delivering good results in both the training and testing phases [8]. As far as machine learning is concerned, it is based on three main strategies-feature selection, pre-processing, classification.

There are three sections in this paper: the first discusses related work for prediction of benign and malignant classes, and the second discusses proposed methodology, the third section discusses experiments and results for breast cancer diagnosis.

2. SURVEY FOR BREAST CANCER PREDICTION

As we know, machine learning models are those algorithms which learn from the data of the past. Based on a machine learning model, we analyse many data and predict the future based on those data [9]. Based on regression and classification models, the decision tree is constructed. Subsets of the dataset are divided into smaller ones. It is possible to make predictions with the highest level of precision using smaller sets of data [12]. In this K-Nearest Neighbour (KNN) algorithm, more dependent variables are included in the learning process. Using this algorithm, a binary response is generated. Based on a particular set of data, logistic regression [11] can provide a continuous outcome. A statistical model with binary variables is used in this algorithm [10]. In Naive Bayes Algorithm (NB) an assumption is made that the training dataset will be large in this model. By using the Bayesian method, the probability is calculated [13]. In support vector machine (SVM) classification and regression problems are solved using this supervised learning algorithm [14]. It consists of large datasets that can be predicted with the highest accuracy rate using this method. In addition to using 3D and 2D modelling, it is an effective machine learning method [11], [15]. In K-Mean Algorithm, using a clustering algorithm, the K-mean algorithm divides data into small clusters. Data is compared using an algorithm to determine their similarity. Data containing at least one cluster can be used to evaluate a large dataset [17]. In this work we have used two machine learning algorithms which are used to predict breast cancer are as below:

- Random Forest (RF): A Random Forest algorithm is an efficient way to solve problems of supervised learning both for classification and regression. In machine learning, this is a basic building block that is used to predict new data based on previous datasets [11].

- Logistics Regression (LR): In this algorithm, more dependent variables are included as part of the supervised learning process. Responses from this algorithm are binary in nature. It is possible to obtain a continuous outcome of specific data using logistics regression [11]. The algorithm is based on the use of a binary variable statistical model [10].

3. PROPOSED METHODOLOGY

The seven phases of the suggested framework are as follows:

In this paper, the goal is to create a method for predicting the benign and malignant classes using regression and classification of breast cancer.

3.1. Dataset Description

On the UCI machine learning repository, a dataset is accessible. 569 samples in all are included in this collection. Our samples have a malignant (M) or benign (B) classification (B). These are medical terminology that describe the two types of tumour cells that we discussed earlier: benign and malignant. The properties have all their values. The distribution of our samples shows that 357 are benign and 212 are malignant. Figure 3.1 and Figure 3.2 shows the images of benign and malignant cancer.

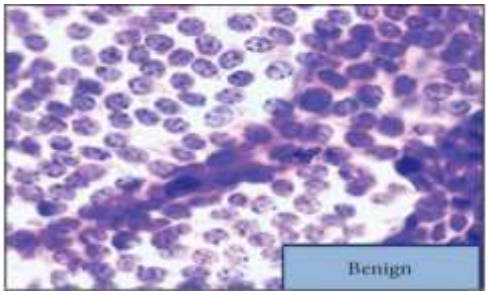

Figure 3.1: Benign Class

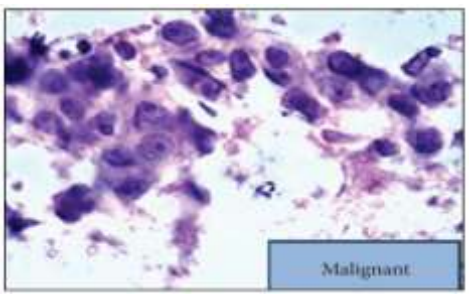

Figure 3.2 Malignant Class

Figure 3.1 shows benign cancer cells which are abnormal in nature but non-cancerous collection of cells. It grows very slowly and doesn't spread to additional bodily parts whereas malignant cancer cells are shown in Figure 3.2. These cells grow rapidly and invade other body organs very soon, it is cancerous and metastatic in nature. Figure 3.3 represents the number of malignant and benign cells present in our dataset.

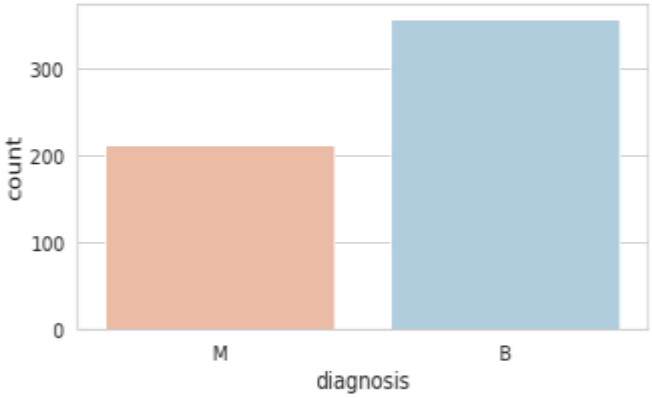

Figure 3.3. Wisconsin Breast Cancer Diagnostic Dataset

3.2. Feature Extraction

There are a total of 31 features namely, mean of radius, texture mean, mean value of perimeter, mean value of area, smoothness mean, mean of compactness, mean of concavity, mean of concave points, mean of symmetry, mean of fractal dimension, worst texture, worst perimeter, worst area, worst smoothness, worst compactness,worst part of concavity, worst part of concave points, worst part of symmetry, and worst part of fractal dimension.

4. EXPERIMENTS AND RESULTS

All experiments are performed using Sci-Kit Learn library available in Python programming language. The dataset is splitted into various ratios for training and testing samples. Using data inputs, our predictive algorithm will determine whether a cancer is benign or malignant in nature. In this study, we have applied two machine learning techniques to know which algorithm is more accurate in prediction. Firstly, we imported the required libraries such as NumPy and Pandas in our python then download the dataset from Kaggle. Then we got the information about our dataset and checked for the missing values. In this dataset, there is no missing value. So, we have spitted data into training and testing for both classifiers.

4.1. Performance Analysis

After splitting the data, we have evaluated accuracy score of training data which was 99% in Logistic regression and 99.5% in Random Forest classifier. After this, we tested our model accuracy on test data on confusion matrix. The important parameters are recall, F1 score, precision, accuracy on which we tested our model and compared with each other. In logistic regression model, the confusion matrix was [TP=86, TN=50, FP=4, FN=3] and in Random Forest Classifier the confusion matrix was [TP=87, TN=51, FP=3, FN=2] after applying the accuracy formula in both we got 95.10% and 96.5% testing accuracy respectively. The results are as below-

	Algorithm	Training Accuracy	Testing Accuracy
1.	Logistic Regression	99%	95.10%
2.	Random Forest Classifier	99.5%	96.50%

4.2. Comparative Analysis

The comparative analysis of both performances is shown in Figure 4.1:

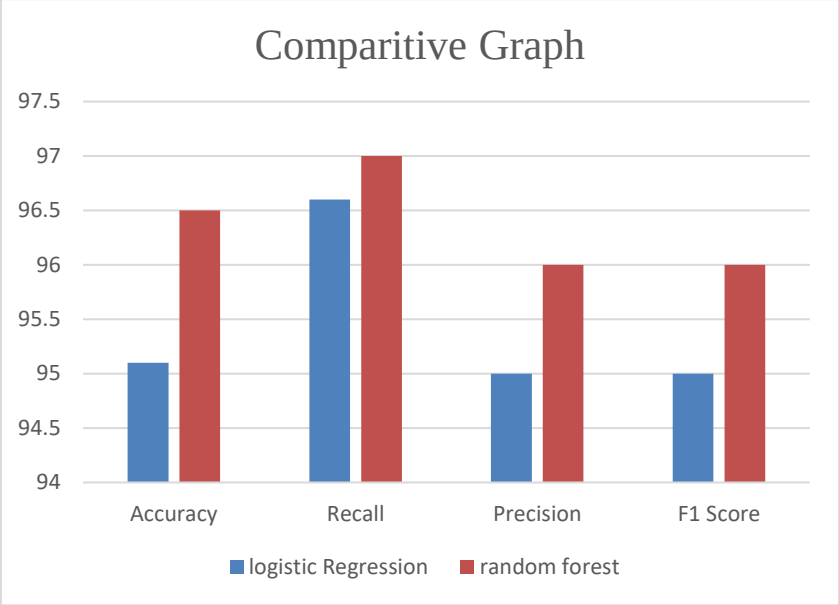

Figure 4.1. Comparative Graph

5. CONCLUSIONS

In this work, breast cancer prediction is carried out using statistical properties and classified using Random Forest Classifier and Logistic Regression. The results are evaluated and compared, based on accuracy's a result, the Random Classifier model can be used to predict breast cancer because it performed well in terms of accuracy of prediction (96.5%), which will greatly aid doctors in making an accurate prediction. The limitation we have seen that is after being accurate Random Forest classifier model can predict wrong as it is not 100% accurate Several issues still need to be addressed in future research. As we all know data is everything in every domain. Hence, it can be said that data availability is the most significant challenge for deep learning and machine learning in predicting breast cancer. Most researchers are now searching for medical images of patients with cancer, which contain sensitive information, and are publicly available as raw images. To overcome the problem of limited patient data, many researchers are now using data augmentation schemes, such as cropping, filtering, rotating, and cleaning. Data from more patients can be obtained using this technique.

6. REFERENCES

[1] Y. S. Sun et al., 'Risk factors and preventions of breast cancer', International journal of biological sciences, vol.13, no.11, pp.1387, 2017.

[2] Y. Khourdifi and M. Bahaj, "Applying best machine learning algorithms for breast cancer prediction and classification," in 2018 International Conference on

Electronics, Control, Optimization and Computer Science (ICECOCS), pp. 1–5, IEEE.

[3] Y. Lu, J. Y. Li, Y. T. Su, and A. A. Liu, 'A review of breast cancer detection in medical images', IEEE Visual Communications, and Image Processing (VCIP), pp. 1–4, IEEE in 2018.

[4] F. K. Ahmad and N. Yusoff, 'Classifying breast cancer types based on fine needle aspiration biopsy data using random forest classifier', 13th International Conference on Intelligent Systems Design and Applications, pp. 121–125, IEEE, 2013

[5] R. Hou et al., 'Prediction of upstaged ductal carcinoma in situ using forced labelling and domain adaptation', IEEE Transactions on Biomedical Engineering, Vol. 67, no. 6, pp. 1565-1572, 2019.

[6] D. Delen, 'Analysis of cancer data: a data mining approach', Expert Systems, vol. 26, no. 1, pp.100–112, 2009.

[7] A. Reddy, B. Soni, and S. Reddy, 'Breast cancer detection by leveraging machine learning', ICT Express, vol. 6, no. 4, pp. 320-324, 2020.

[8] Z. Salod and Y. Singh, 'Comparison of the performance of machine learning algorithms in breast cancer screening and detection: A protocol', Journal of Public Health Research, vol.8, no. 3, pp. 1677-1686, 2019

[9] C.M. Bishop, 'Pattern recognition and machine learning' Springer, vol. 4, no. 4, pp. 738). New York: springer. 2006.

[10] H. Tran, 'A survey of machine learning and data mining techniques used in multimedia system', no. 113, pp.13-21, 2019.

[11] C.Y.J .Peng ,K.L.Lee, and G.M.Ingersoll, 'An introduction to logistic regression analysis and reporting', The journal of educational research, vol. 96, no. 1, pp. 3–14, 2002.

[12] H. Sharma and S. Kumar, 'A survey on decision tree algorithms of classification in data mining', International Journal of Science and Research (IJSR), vol. 5,no. 4, pp.2094–2097, 2016.

[13] A. A. Ibrahim, A. I. Harshad and N.E.M. Shawky, 'A Comparison of Open-Source Data Mining Tools for Breast Cancer Classification', pp. 636–651.IGI Global, 2017

[14] T. Evgeniou and M. Pontil, 'Support vector machines: Theory and applications', in Advanced Course on Artificial Intelligence, pp. 249–257, Springer, 2005.

[15] Y. Yang, J. Li, and Y. Yang, 'The research of the fast SVM classifier method', in 12th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), pp. 121–124, IEEE,2015.

[16] L. Breiman, 'Random forests', Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[17] Y. Li and H. Wu, 'A clustering method based on k-means algorithm', Physics Procedia, vol. 25, pp. 1104–1109, 2012.