International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 9 Issue 6 ǁ 2021 ǁ PP. 69-73

Breast Cancer Analysis and Prediction by Using Machine Learning

Harjasdeep Singh
Assistant Professor
Dept. of CSE, MIMIT, Malout

Raushan Raj
Research Scholar
Dept. of CSE, MIMIT, Malout

Abstract: *Breast Cancer (BC) is the most common cancer in women after skin cancer and has become a serious health issue and it has the potential to cause death in women. As a result, it's important to diagnose BC properly.*

We all know that Machine Learning (ML) techniques have a unique benefit and that is why they're widely used to analyze advanced BC datasets and predict disease. An effective way to classify the data through classification or data mining. This becomes very practical, especially in the medical field where identification is done through these techniques.

The Wisconsin Breast cancer dataset(WBCD) is used to perform a comparison between SVM, Logistic Regression, Naïve Bayes(NB), and Random Forest. To estimate the correctness in classifying the data based on accuracy and time consumption which is used to determine the efficiency of those algorithms, which is the main objective. Based on this result after performing experiments, the Random Forest algorithm shows the highest accuracy rate (99.76%) and also the least error rate.

Keywords: *Breast Cancer, Accuracy, Algorithm, Random Forest, Decision Tree, Machine Learning, Wisconsin Breast Cancer Dataset(WBCD)*

Date of Submission: 25-05-2021 Date of acceptance: 07-06-2021

I. INTRODUCTION

Breast cancer (BC) is one of the highest common cancers among ladies worldwide, representing the bulk of recent cancer cases and cancer-related deaths per world statistics, creating it a major public pathological state in today's society.

Signs of BC could symbolize a lump within the breast, a modification in breast form, dimpling of the skin, fluid returning from the nipple, a newly inverted nipple, or a red or scaly patch of skin. In those people spreading this disease which may be causing the illness, there could also be bone pain, swollen lymph nodes, shortness of breath, or yellow skin. Regarding 5–10% of cases which measure the results of a genetic predisposition heritage from a human folk, as well as BRCA1 and BRCA2 among others. BC most ordinarily develops in cells from the lining of milk ducts and therefore the lobules that offer these ducts with milk. The identification of BC is confirmed by taking a diagnostic test for that regarding tissue.

Once the detection is possible, further tests are done to determine if cancer has spread far away from the breast and which treatments are most likely to be productive. There are mainly two types of classifications in breast cancer: Benign tumour and Malignant Tumour. Benign tumours are not part of cancer; they can be seen anywhere in the body and removed by proper medication and treatment, but Malignant tumours are cancerous, they grow abnormally uncontrolled and can spread to the other organs. If the possibilities of cancer can be predicted at the early stage, then the survivability chances of that patient could be increased. In another way to identify BC by using machine learning algorithms for the prediction of an abnormal tumour. Thus, the research paper is carried out for the proper diagnosis and categorization of patients into malignant and benign groups. In this paper, we have used machine learning techniques such as Naive Bayes, CART, Random Forest, KNN, Support Vector, Artificial Neural Network. Steps involved are pre-processing, data splitting, applying methods, and comparing the performance of the methods based on accuracy.

II. RELATED WORK

This paper mentioned about the study of a few important research papers.

[1]. B. Nithya applied the three categorizing methods such as Decision Tree, k-Nearest Neighbour, and Naïve Bayes for the different datasets. They are cancer & Iris datasets in the open source R tool environment. The authors also inspect the evaluation metrics like accuracy and error rate. The implementation was focused on a type of attribute of a dataset and its characteristics.

[2]. Md. Milan Islam processed the two inspected learning classifiers that are SVM and KNN. They forecasted in terms of accuracy, sensitivity, specificity, false discovery rate, false omission rate, and Mathew's correlation coefficient. They proposed a system with 10-fold cross-validation and achieved the accuracy of 98.57% by SVM and 97.14% by KNN on the dataset Wisconsin Breast Cancer Diagnosis.

[3]. Bayrak differentiates machine learning techniques SVM and ANN for breast cancer diagnosis. In this paper Wisconsin breast cancer dataset(WBCD)
is used. In differentiation to other methods, with the "WEKA" machine learning method SVM framework, both SVM and ANN are applied, providing the best classification

[4]. Turgut Machine learning procedure compared with SVM, KNN, DT, Logistic Regression, Random forest, ADA Boost, and Gradient Boost. In this paper feature selection (FS) techniques like Recursive Feature Elimination and Randomized Logistic Regression are done in pre-processing steps. After comparing these 8 algorithms, SVM has a maximum accuracy of 88.82% and the lowest error rate.

[5]. Fu, B proposed the XGBOOST algorithm to forecast breast cancer. In this paper, a model which is used is the combination of XGBOOST and stratified feature selection model is used for early breast cancer prediction. Other ML algorithms such as SVM, Random Forest, and ADA BOOST were compared to the XGBOOST. The highest accuracy was recorded for the XGBOOST algorithm.

[6]. Narasingarao.M presents a survey of the work conducted to predict or diagnose breast cancer using different machine learning techniques explaining their perspectives and limitations.

[7]. Junaid Ahmed achieved 84.21% accuracy by using Adaptive Reasoning Theory In this, the Wisconsin Breast Cancer data set was used, which was acquired from the Machine Learning Repository of UCI, that contains 569 rows of data, and also contains 32 attributes.

[8]. Breiman's Bagging Random Forest R.F extended from Random Forest based on the trees which group each tree based on a group of random variables. R.F is also the collection of multiple decision trees.

[9]. Fernandez-Delgado used 179 ML algorithms on around 121 UCI datasets in which Random Forest gives the best result in all.

[10]. Jacob collatesvarious classifier algorithms on the Wisconsin Breast Cancer diagnosis dataset. They found that Random Tree and SVM categorization algorithm which gives the best result i.e. 100% accuracy.

III. EXPERIMENT

To compare the behaviours of LR, NB, SVM, and Random Forest, the experiment conducted was focused on the evaluation of the algorithms.

Experimental Environment

We use Spyder to work on this dataset. Firstly go to import all the necessary libraries(NumPy, Pandas, Matplotlib, Seaborn) and import our dataset(dataset is obtained in a CSV format from their database) to Spyder. which is a development environment that supports python. It is a powerful IDE for python compared to others. It also has self-analysis features. Since our problem strength requires those features and also debugging is easier on this platform, so that it is preferred.

Dataset

The dataset used in this paper is publicly available(Dataset is available in CSV format from UCI Machine Learning Repository [www.ics.uci.edu]) and it was created by Dr. William H. Walberg, who is a physician at the University Of Wisconsin Hospital at Madison, Wisconsin, USA. To create this dataset Dr. Walberg uses fluid samples, taken from patients with solid breast masses and they are an easy-to-use graphical computer program called XCYT, which is more capable to perform the analysis of cytological features which is based on a digital scan. Dataset is used for this paper; it contains 569 rows and 32 columns. In the column of diagnosis where we are going to predict that if the cancer is M = malignant or B = benign. Where 1 means that cancer is malignant and 0 means that it is benign. Out of the 569 persons we identify, 357 are labeled as B (benign) and 212 as M (malignant).

The program which is used in this curve-fitting algorithm to compute the 10 features from each one of the cells in this sample, then it evaluates the mean value, extreme value, and standard error of each feature for the image, returning to the 30 real-valued vectors.

10 actual value characteristics are calculated for each of the cell nucleus.

i. Radius (mean distances from the canter to the points on the perimeter)
ii. Texture (standard variation of grey-scale values)

iii. Perimeter
iv. Area
v. Smoothness (local difference in radius lengths)
vi. Compactness (perimeter^2 / area - 1.0)
vii. Concavity (strength of concave parts of contour)
viii. Concave points (no of the concave part of contour)
ix. Symmetry
x. Fractal dimension (coastline approximation - 1)

Data Visualization

Data visualization is a key aspect of the data science. It helps to comprehend and also to carry the data to another person in a meaningful manner. Matplotlib & Seaborn both are the python data visualization libraries. It is essential for analysing large amounts of information and to make a decision. It engages the use of pictorial elements such as maps, plots, patterns, graph trends, etc. That provides an easy way for the user to comprehend the data.

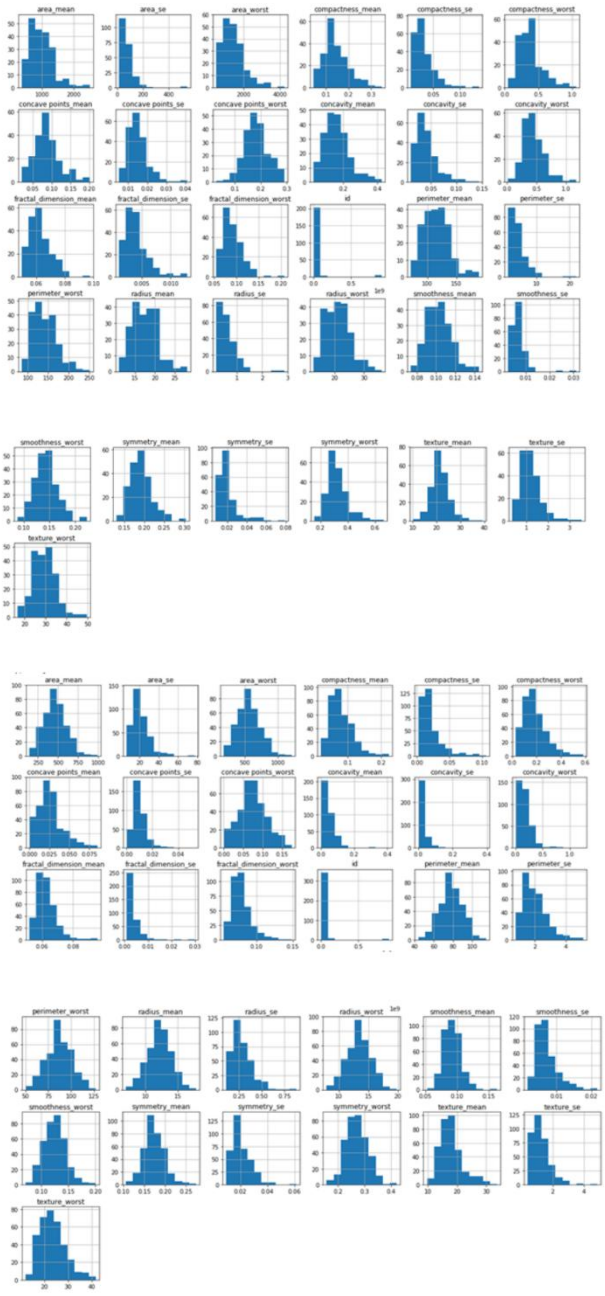
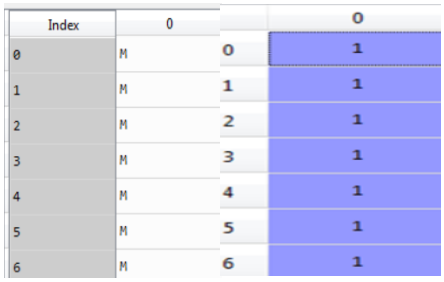


Fig : Visualization of Dataset

Categorical Data

Categorical data is the variables which contain label values rather than numeric values. The number of possible values is frequently limited for the fixed set. So, here we have a represented benign cells as value 0 and malignant cells as value 1.



Index	0		0
0	M	0	1
1	M	1	1
2	M	2	1
3	M	3	1
4	M	4	1
5	M	5	1
6	M	6	1

Fig : Data Conversion

Splitting Dataset

The data which is used is usually split into training data and test data. In this project, 75% of data is trained data and 25% of data is test data

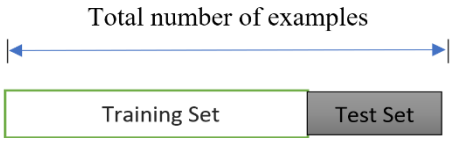


Fig : Splitting Dataset

Accuracy Method

We use the categorization Accuracy method to find the accuracy in our models. Categorization Accuracy is what we usually mean when we use this term accuracy. It is the ratio of the number of correct predictions to the total no of the predictions made.

$$Accuracy = \frac{No\ of\ correct\ prediction}{Total\ no\ of\ prediction\ made}$$

Fig : Accuracy

Confusion Matrix

To check the correct guess we have to check the confusion matrix object and add the guess results diagonally. They will be the number of correct predictions and then divide by the total number of predictions.

	0 (Normal)	1 (Abnormal)
0 (Normal)	87	3
1 (Abnormal)	3	50

Fig:Confusion Matrix

Experimental Results

After applying the different types of categorization models, we get this accuracy with different models

S. No	Algorithm	Accuracy %
1	Logistic Regression	95.8%
2	Nearest Neighbour	95.1%
3	Support Vector Machines	97.2%
4	Kernel SVM	96.5%
5	Naive Bayes	91.6%
6	Decision Tree Algorithm	95.8%
7	Random Forest Classification	98.6%

IV. CONCLUSIONS

In this paper, we used a Wisconsin Breast Cancer dataset for the prediction of breast cancer using ML techniques. So, finally, we have made our classification model and we can see the Random Forest Classification algorithm gives the best results for this dataset. But it is not always applicable for every dataset. To choose our model we always need to analyse our dataset and then apply it to the machine learning model.

REFERENCES

[1]. B. Nithya, V. Ilango, 2017, "Relative Analysis of categorization Methods in R Environment with two Different Datasets.", Intl J Scientific Research and Computer Science, Engineering and Information Technology (IJSRCSEIT), vol 2, Issue 6, ISSN: 2456-3307.

[2]. Mohd. Milon Islam. 2017.Performing the Breast Cancer prognosis by using Support Vector Machine & K-Nearest Neighbours.

[3]. Ahmad LG, Eshlaghy AT, Ebrahimi M, & Razavi AR. They implement Three Machine Learning Methods for the Forecasting of Breast Cancer Regularity. Journal of Health & Medical Informatics

[4]. RAZIA, S., & Narasinga Rao, M. R. (1943). Evolution & survey of Support Vector Machine Approach for Early prognosis of Breast Cancer and Thyroid.

[5]. J. A. Bhat, V. G, Cloud Computing with Machine Learning Could help Us in Early Detection of a Breast Cancer, (2015).

[6]. Breiman, L. (2001). Machine Learning 24(2). 123-140.

[7]. Breast Cancer Wisconsin Dataset. Available at: UCI Machine Learning Repository[www.ics.uci.edu].

[8]. Dataset Description. Available at: UCI Machine Learning Repository.

[9]. R. M. Mohana, R. Delshi Howsalya Devi, Anita Bai, "Performing the Lung Cancer Diagnosis by using the Nearest Neighbour Classifier", (IJRTE), September 2019

[10]. Ch. Shravya, K. Pravalika, Shaik Subhani, Breast Cancer by using Supervised Machine Learning System(IJITEE), Issue-6, April 2019.

[11]. Haifeng Wang ang Sang Won Yoon, "Breast Cancer Prognosis by using Data Mining procedure", Proceedings of the 2015 Industrial & Systems Engineering & Research Conference,

[12]. Abdelghani Bellaachia, Erhan Guven, "Foretell Breast Cancer endurance Using Data Mining Techniques.

[13]. Y. Khourdifi and M. Bahaj, ''Put in best machine learning algorithms for breast cancer prognosis and categorization,'' in Proc. Int. Conf. Electron., Control, Optimize. Computer. Science. (ICECOCS), Dec. 2018, pp. 1–5.

[14]. SA Medjahed, TA Saadi, A Benyettou "Performing Breast cancer prognosis by using k-nearest neighbor with different distances categorization rules" International Journal of a Computer Applications 62 (1), 2013.

[15]. M. Amrane, S. Oukid, I. Gagaoua, and T. Ensarİ, "Breast cancer categorization is performed by using machine learning technique," Electric & Electronics, Computer Science, Biomedical Engineering's Meeting (EBBT), Istanbul, 2018, pp. 1-4.

[16]. Prakash, K.B., Ananthan, T.V. & Raja Varman, V.N. 2014, "Neural network framework for code-switching web documents", Proceedings of 2014 International Conference on Contemporary Computing & Informatics, IC3I 2014, pp. 392.

[17]. kibeom Jang, Minsoon Kim, Candace A Gilbert, Fiona Simpkins, Tan A Ince, Joyce M Slingerland "WEGFA activates an inherited pathway to synchronize an ovarian cancer-initiating cells" Embo Molecular Medicines Volume 9 Issue 3 (2017).

[18]. Joseph A. Cruz & David S. Wishart "Implements Machine Learning technology in cancer prognosis Cancer informatics" 2(3):59-77 · February 2000.

[19]. F. K. Ahmad and N. Yusuf, ''Analyse breast cancer types based on fine needle desire biopsy data using random forest classifier,'' in Proc. 13th Int. Conf. Intelligent Syst. Design Appl., Dec. 2013, pp. 121–125.

[20]. M. Shahbaz, S. Faruq, M. Shahen, and S. A. Masood, ''Cancer detection using data mining technology,'' Life Sci. J., vol. 9, no. 1, pp. 308–313, 2012.

[21]. Pranay Shah, Rahul Deshpande, Nikhil Rao, Breast Cancer Detection System, (IRJET), Volume: 07 Issue: 05 | May 2020.

[22]. Ajay Kumar, R. Sushil, A. K. Tiwari, Comparative Study of Classification Techniques for Breast Cancer Diagnosis, Vol.-7, Issue-1, Jan 2019.

[23]. Vinoothna Manohar Botcha, Bhanu Prakash Kolla, Predicting Breast Cancer using Modern Data Science Methodology, ISSN: 2278-3075, Volume-8 Issue-10, August 2019.

[24]. Sivapriya J, Aravind Kumar V, Siddarth Sai S, Sriram S, Breast Cancer Prediction using Machine Learning, ISSN: 2277-3878, Volume-8 Issue-4, November 2019.

[25]. Maria Mohammad Yousef, Big data analytics in healthcare: A review paper, International Journal of Computer Science & Information Technology (IJCSIT) Vol 13, No 2, April 2021.