Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 11, November-2021

Breast Cancer Prediction using Relative Analysis of Machine Learning and Deep Learning Techniques

Manas Bhole
Department of Information Technology
Ramrao Adik Institute of Technology

Yash Bharambay
Department of Electronics Engineering
Ramrao Adik institute of Technology

Pranav Chavan
Department of Information Technology
Ramrao Adik Institute of Technology

Anish Nair
Department of Computer Science
Don Bosco Institute of Technology

***Abstract*—Breast cancer is one of the deadliest disease caused to the women in this world. Women are the ones who are more likely to be diagnosed with it. A major cause of increased mortality in women. A breast cancer diagnosis takes time, and because systems are limited, it is vital to design a system that can automatically diagnose breast cancer in its early stages. According to the statistics, 7 billion people and out of which 3.4 billion are women and that 1 out of every 22 women is diagnosed with breast cancer. Though this method cannot definitively detect cancer, it can assist clinicians in determining whether a biopsy is necessary by giving information on whether the patient has breast cancer. Confusion matrix and ROC analyses were used to evaluate the definite diagnosis for each patient. The dataset has been taken from the alcrase dataset, which contains approx 16000 datasets and 30 features that would be used in detecting the results from the algorithm applied. The main idea for the paper is to do a comparative research of the machine learning and deep learning methods and to show which is the best performing algorithm amongst all of them.**

Keywords*—Comparative analysis, machine learning, deep learning, random forest, breast cancer prediction

I. INTRODUCTION

Breast cancer is one of the worst cancers that arises in the breast cells and is a very common disease. Breast cancer, like lung cancer, is a life-threatening disease for women. Breast cancer is divided into different types depending on the appearance of the cells under a microscope. Invasive ductal carcinoma (IDC) and ductal carcinoma in situ (DCIS), the latter of which advances slowly and has minimal influence on patients' daily life, are the two most frequent types of breast cancer. The DCIS type accounts for a modest percentage of all cases (between 20% and 53%); the IDC type is more dangerous, enclosing the whole breast tissue.For the vast majority of breast cancer patients, this is the case (about 80 percent). The most lethal cancer is lung cancer, which is followed by breast cancer. Breast cancer accounts for around 11% of all new cancer cases, with women accounting for almost 24%. People seek the opinion of an oncologist if they see any cancer signs or symptoms. The methods used in the the papers are support vector machine , logistic regression , random forest classifier , artificial neural networks, decision tree classifier and the ensmeble method XGBoost.

II. LITERATURE OVERVIEW

Breast cancer is classified using immunohistochemistry (IHC), histopathologic features, and molecular characterization. The two most prevalent histologic subtypes of invasive breast cancer are invasive ductal carcinoma and invasive lobular carcinoma (80 percent to 85 percent and 10 percent to 15 percent of all cases, respectively). Several histologic cancer subtypes occur in the remaining 1% of invasive breast tumours. Breast cancer HC characterization is crucial for assessing treatment options and predicting prognosis. Biomarkers such as the oestrogen receptor (ER), the progesterone receptor (PR), and the human epidermal growth factor receptor (HEGFR) must be expressed. ER and PR expression is less than 1% in roughly 75% of persons with hormone receptor (HR)–positive breast cancer. 12 Furthermore, according to IHC, 15 to 30 percent of breast cancer patients have HER2 that has been amplified or overexpressed. 13 Breast cancer that is triple-negative lacks ER and PR expression as well as HER2 overexpression (TNBC). The TNM model has traditionally been used to classify breast cancer into stages 0, 1, 2, and 3. Prognostic indicators (such as histologic tumour grade, ER, PR, HER2, and multigene test-b) were included to breast cancer staging in the 8th edition of the American Joint Committee on Cancer's Cancer Staging Manual in 2017.

III. METHODOLOGY

A. DATA PREPARATION

Data was acquired from Wisconsin breast cancer diagnostic data, which has been used in a number of studies. Because the data is difficult to come by, the only way to run the model and make a forecast was to use data from a reliable source.. To compute features, a digitized image of a fine needle aspirate (FNA) of a breast mass is employed. The attributes of the image's cell nuclei are defined by them. Two attributes are the ID number and the diagnosis (M = malignant, B = benign). Smoothness, compactness, concavity, radius, texture, perimeter, area, radius, texture, perimeter, area and radius. The below figure shows the two types of tumors that occur from the dataset.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 11, November-2021

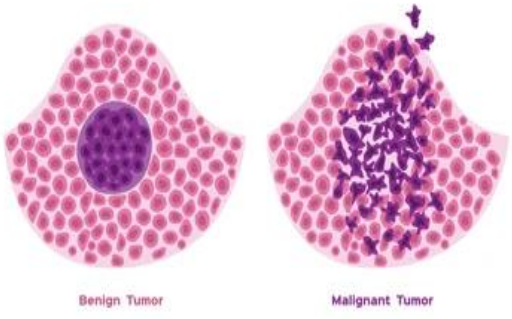

Figure 1: Samples of Benign and Malign Tumor

B. DATA PREPROCESSING

Data preparation is critical to the algorithm's or model's performance. The speed of the method is determined on whether or not the data has been preprocessed. If done correctly, it can speed up the method and make it perform well when applied to a huge dataset. As a result, preprocessing is needed to finish the algorithm or model, which will serve as the foundation for the next step.

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280

Figure 2: Dataset after the preprocessing is performed

C. FEATURE SELECTION

In the real world, all the variables in a dataset are unlikely to be significant for building a machine learning model. Duplicate variables limit the model's generalization capacity and may lower the overall accuracy of a classifier. Adding extra variables to a model also increases the model's overall complexity. As a result, feature selection is becoming increasingly important in the development of machine learning models. So the samples are classified as malignant or benign based on two primary characteristics. We'd also want to show off some additional features.

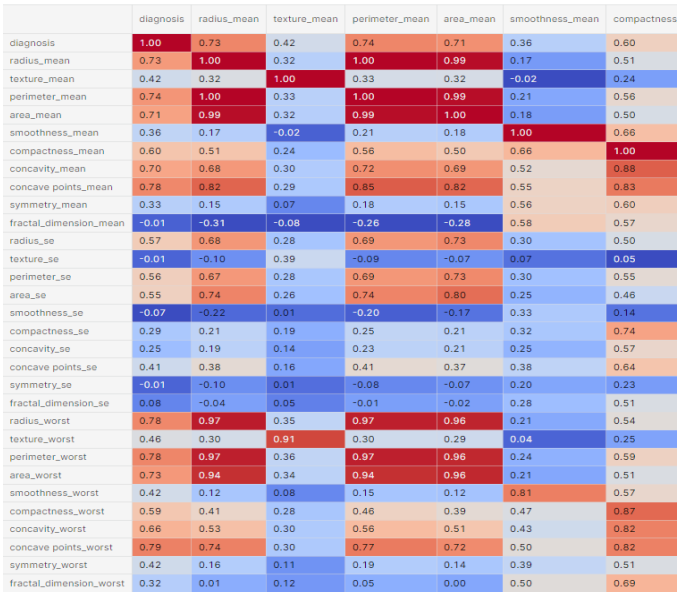

Figure 3: Correlation map between the features

D. MODEL ARCHITECTURE

1) RANDOM FOREST

It's a technique which belongs on the ensemble model category. It may be used to develop a good prediction model by combining classification and regression techniques. In this work, decision trees are used as the foundation estimators. On their own, decision trees are a poor predictor, but when combined with other decision trees, they improve. Decision trees vote on how to categorize a specific instance of input data in classification tasks, and they output the class that is the mode of the classes or the mean of predictions in regression tasks. In this manner, we may prevent parameter tinkering and reduce overfitting.

2) SUPPORT VECTOR MACHINE

To implement nonlinear class borders, Support Vector Machines use a linear model. To separate the target classes, support vectors (lines or hyperplanes) are created. To handle a nonlinear problem, the model uses a mapping function to apply numerous transformations to the data and then trains a linear SVM model to classify the data in a higher-dimensional feature space.

3) LOGISTIC REGRESSION

The method of modelling the probability of a discrete result given an input variable is known as logistic regression. The most frequent logistic regression models have a binary outcome, which might be true or false, yes or no, and so forth. Multinomial logistic regression can be used to m

4) XGBOOST

XGBoost mainly known as the decision tree ensemble machine learning algorithm especially designed for speed and performance. It uses the gradient boosting framework, when it comes to prediction of problems involving unstructured data the artificial neural networks tend to outperform all the other designed algorithms and frameworks. But, when it comes to small and medium

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 10 Issue 11, November-2021

structured data, the tree based algorithms are considered best.

5) ARTIFICIAL NEURAL NETWORK

ANN stands for Artificial Neural Network which is composed of the structure and role of the biological neural network which is mainly the reason for the designing of the Artificial neural network architecture. It is generally consisted by neurons which are in layers and similar neurons which are present in the human brain. A processing unit absorbs input signal, whilst the output layer generates network output. The input layer and the output layer are always present in these connected arrangements, as they are in all network topologies.

6) DECISION TREE

A decision tree is a decision-making aid that employs a tree-like model of decisions and their potential results, such as chance event outcomes, resource costs, and utility. It's one approach to show an algorithm made up entirely of conditional control statements.

7) KNN

firstly we would select the number of clusters k , then we assume the centroid of the cluster , Any random item can be used as the initial centroid, or the first k objects in a sequence can also be used. So the algorithm works into three steps , firstly we determine the coordinate of the centroid after that We'll figure out how far each object is from the centroid ad finally grouping begins based on the minimum distance. Following the method, we are able to obtain a centroid

IV. EXPRIMENTAL RESULTS

The process of selecting the dataset to preprocessing the dataset for making a proper one to improve the accuracies of the models selected. The models selected show great accuracies, which has been listed in the table below. Starting with the logistic regression which provides us the accuracy of ~97 percent then svm with a ~96 percent then decision tree with ~91 percent, random forest with ~96 percent, neural network with ~95 percent, knn with ~94 percent and finally the XGBoost with ~96%. Which makes one observe that the Logistic Regression method gives out the accuracy better than the selected methods.

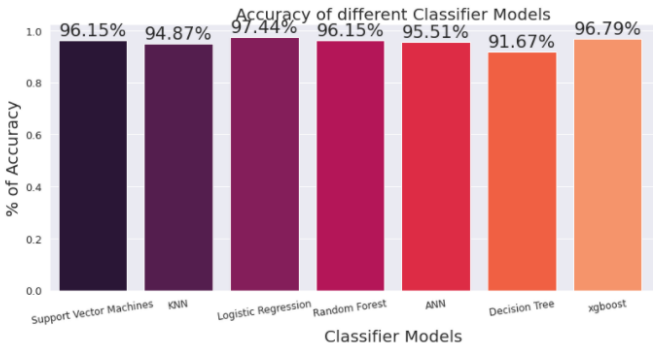

Figure 4: Result Comparison

V. CONCLUSION

Finally after performing all the models described starting from support vector machine , knn, logistic regression, random forest, ANN, decision tree and XGBoost the author observes that the accuracy received by the Logistic Regression which is a regression method used in machine learning and is the highest amongst the entire selected models. The accuracy would be used by the other researchers too while choosing the best model out amongst the other models and would definitely help in treating the breast cancer most of the times.

VI. REFRENCES

[1] Akay MF. Support vector machines combined with feature selection for breast cancer diagnosis. Expert systems with applications. 2009 Mar 1;36(2):3240-7

[2] Marcano-Cede˜no A, Quintanilla-Dom´ınguez J, Andina D. WBCD breast cancer database classification applying artificial metaplasticity neural network. Expert Systems with Applications. 2011 Aug 1;38(8):9573-9.

[3] Polat K, G¨une¸s S. Breast cancer diagnosis using least square support vector machine. Digital Signal Processing. 2007 Jul 1;17(4):694- 701.

[4] Kaya Y, Uyar M. A hybrid decision support system based on rough set and extreme learning machine for diagnosis of hepatitis disease. Applied Soft Computing. 2013 Aug 1;13(8):3429-38.

[5] Jemal A, Murray T, Ward E, Samuels A, Tiwari RC, Ghafoor A, Feuer EJ, Thun MJ. Cancer statistics, 2005. CA: a cancer journal for clinicians. 2005 Jan 1;55(1):10- 30.

[6] Yeh WC, Chang WW, Chung YY. A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method. Expert Systems with Applications. 2009 May 1;36(4):8204-11.

[7] Nahato KB, Harichandran KN, Arputharaj K. Knowledge mining from clinical datasets using rough sets and backpropagation neural network. Computational and mathematical methods in medicine. 2015;2015.

[8] Liu L, Deng M. An evolutionary artificial neural network approach for breast cancer diagnosis. InKnowledge Discovery and Data Mining, 2010. WKDD'10. Third International Conference on 2010 Jan 9 (pp. 593-596). IEEE.

[9] Chen HL, Yang B, Liu J, Liu DY. A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. Expert Systems with Applications. 2011 Jul 1;38(7):9014-22.

[10] Overview of Breast Cancer and Implications of Overtreatment of Early-Stage Breast Cancer: An Indian Perspective Gouri Shankar Bhattacharyya, Dinesh C. Doval, Chirag J. Desai, Harit Chaturvedi, Sanjay Sharma, and S.P. Somashekhar JCO Global Oncology 2020 :6, 789-798

[11] S. Modi and M. H. Bohara, "Facial Emotion Recognition using Convolution Neural Network," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 2021, pp. 1339-1344, doi: 10.1109/ICICCS51141.2021.9432156

[12] X. Zhang and Y. Sun, "Breast cancer risk prediction model based on C5.0 algorithm for postmenopausal women," 2018 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC), 2018, pp. 321-325, doi: 10.1109/SPAC46244.2018.8965528.

[13] . Singhal and S. Pareek, "Artificial Neural Network for Prediction of Breast Cancer," 2018 2nd International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), 2018 2nd International Conference on, 2018, pp. 464-468, doi: 10.1109/ISMAC.2018.8653700

[14] A. Bharat, N. Pooja and R. A. Reddy, "Using Machine Learning algorithms for breast cancer risk prediction and diagnosis," 2018 3rd International Conference on Circuits, Control, Communication and Computing (I4C), 2018, pp. 1-4, doi: 10.1109/CIMCA.2018.8739696.

[15] Lin, YL., Xu, DZ., Li, XB. et al. Consensuses and controversies on pseudomyxoma peritonei: a review of the published consensus statements and guidelines. Orphanet J Rare Dis 16, 85 (2021). https://doi.org/10.1186/s13023-021-01723-6