Cancer: Prediction & Analysis

MUDIT SHARMA
Student, Maharaja Agrasen Institute of Technology, Rohini, Delhi

Abstract- Breast cancer is one of the fundamental drivers for the disease spread and the leading cause of death to most women across the globe. Early diagnostics builds the odds of right treatment and endurance, however this cycle is dreary and regularly prompts a contradiction between pathologists. Although many individuals who suffer breast cancer have no family history but women who have blood relatives suffering from the same disease are at higher risk. Besides, a high risk of developing breast cancer includes aging, genes, thick breast tissues, obesity, and radiation exposure. Malignant and benign are two different types of tumors and to distinguish between these two, physicians need a reliable diagnostic procedure. The mammography method is used to detect breast cancer but radiologists exhibit significant variation in interpretation. In any case, early recognition and anticipation can altogether reduce the fatality. Henceforth, it is foremost to detect breast cancer as early as possible. In this paper, we present a prediction of breast cancer with different machine learning algorithms compare their prediction accuracy, area under the receiver operating characteristic curve (AUC) and performance parameters, wherein the model gets trained by considering the parameters such as: radius, texture, perimeter, area, smoothness, concavity, concaveness, and compactness. Here, all these parameters are taken in mean and overall values are considered. Further, these algorithms can be modified with their mathematical models to increase the prediction of breast cancer.

Indexed Terms- Neighbouring tissues, Breast Cancer, Machine Learning, Concave Point mean, Malignant.

I. INTRODUCTION

Breast Cancer has now surpassed lung cancer as the leading cause of global cancer incidence in 2020, with an estimated 2.3 million new cases, representing 11.7% of all cancer cases. Epidemiological studies have shown that the global burden of BC is expected to cross almost 2 million by the year 2030. The risk of breast cancer begins in the breast cells of the human body and these modified cells can easily invade its neighbor tissues [7]. The illness effect may be subject to cancer, risk level, and age of patients. The classification of breast cancer is either by observing a lump in the breast or through mammogram screening by using categorical data.

There are many types of categorical data which needs categorical variable for various categories. The lump is classified as either benign or malignant tumors, which is the abnormal growth of cells. People's live will be protected if the disease is diagnosed early and is more conveniently stopped from spreading[12]. The X-ray was the only method that was used to detect breast cancer. FNAC is also widely adopted in the diagnosis of breast cancer, but the average correct identification rate is only 90%[8]. However, many methods have been generated and proposed for detecting a process that is more efficient than X-ray procedures such as, artificial intelligence and data mining.

Breast cancer (BC) is the commonest malignancy among women globally. In India, the incidence has increased significantly, almost by 50%, between 1965 and 1985. The estimated number of incident cases in India in 2016 was 118000 (95% uncertainty interval, 107000 to 130000), 98.1% of which were females, and the prevalent cases were 526000 (474000 to 574000)[5]. Over the last 26 years, the age-standardised incidence rate of BC in females increased by 39.1% (95% uncertainty interval, 5.1 to 85.5) from 1990 to 2016, with the increase observed in every state of the country. As per the Globocan data 2020, in India, BC accounted for 13.5% (178361) of all cancer cases and 10.6% (90408) of all deaths with a cumulative risk of 2.81[9].

Machine Learning (ML) is a branch of Artificial Intelligence (AI) is a scientific discipline concerned with the design and development of algorithms based on data. It is more feasible to properly assess the available data automatically to determine breast cancer survival rate and risk-specific factors[3]. These automated models that helps doctors make reliable predictions of a patient, based on their recorded collection of parameters. Machine Learning which is widely used to predict, prognosis, and treat important frequent diseases such as cancers, hepatitis, and heart diseases.

We classify ML algorithms as supervised and unsupervised. Supervised learning is a function that maps an input to output, input-output pairs. It is a labeled training data set which analyzes the training data and produces an inferred function, which can be used for mapping new examples. Unsupervised learning is a type of ML that looks for previously undetected patterns in data with no pre-defined labels and with a minimum of human intervention [13]. Methods used in unsupervised learning are principal components and cluster analysis.

The rest of the paper is organized as: II - Aim of Research. III - Related work. IV - Deals with different kinds of ML algorithms used in the analysis process. V - Represents an experimental setup and results obtained from ML algorithms. VI - Represents the conclusion and future work.

II. AIM OF RESEARCH

Aim of the research is to build several models that use the data's numerical values, which describe the geometric shape of a breast tumor, to predict if the subject's tumor is malignant [4]. There are two classifications of a breast tumor that are malignant, which I mentioned before which means cancerous, and benign, which means the tumor is non-cancerous.

It's important to accurately classify whether a breast tumor is cancerous or not because malignant tumors can become extremely dangerous if it's gone untreated. Malignant tumors have the potential to invade their surrounding tissue or spread around one's body [14]. Once the cancerous tissues invade neighbouring tissues and spreads, it becomes difficult to contain/control the spread resulting in devastation.

The main metric I will use to justify my models are accuracy. Accuracy is defined as the number of correctly predicted subjects over the total number of subjects. It can also be described as the number of how many people are told they don't have breast cancer when in reality they do. There is a higher importance on the proposed recall, when compared to accuracy and precision because it can become deadly if a person is told they don't have breast cancer and they move on without seeking necessary treatment [2]. This prediction can help doctors prescribe different medical examinations for the patients based on the cancer type. This helps save a lot of time as well as money for the patient.

I decided to split the data into training and test data sets. The training data set is used to train the models and the test data set is used to validate the performance of the models. The Breast Cancer Wisconsin (Diagnostic) Data Set collected from UCI machine learning repository must be read. It consists of 569 rows and 32 columns, 212 malignant, and 357 begin[15]. Also, the data does not consist of any missing values which help with setting up the environment

III. RELATED WORK

Random Forest [RF] uses several decision trees to achieve classification accuracies. The random forest was first proposed by Ho in 1995. The advantage of the increased accuracy of this algorithm is it has “randomness” i.e. as more decision tree have developed from the system accuracy increases further. The developed system of random forests, decision trees have a great deal of flexibility from each other[1]. Each decision tree has several leaf nodes and tree depth makes sure the results produced by each decision tree are independent. Train each decision tree to yield results, and determine to provide the best solution from the results. Results were obtained from this model with an accuracy of 97.88%.

The proposed KNN algorithm, considers the output as a target class. For the input attributes, feature selection is based on Performance Component Analysis (PCA) so the data set reduced with significant features data[6]. Data is then regarded as a reduced and truly valued positive integer by the qualified majority of its neighbors (K). The k value is computed based on the input, if the value of K is reasonably large then the influence of noise on the output class is minimized. KNN refers to as lazy-learning algorithm whose function is only calculated locally and the whole learning cycle is delayed until the test step is finished. The proposed model KNN yields an accuracy of 97.36%.

The proposed Logistic Regression [LR] model for classification which can be used for mammography images adopting techniques of image processing for feature extraction. The proposed model has hypothesis, cost function and after repeated number iterations the functions return optimal value. The proposed model has LR has predicted the highest accuracy of 99.12%.

The proposed Gradient Boosting algorithm is one of the reinforcement gradient algorithms with a very good performance in classification and performs the best classification for each of the data. In this method, the trees are trained one after another; each subset tree is taught primarily with data erroneously predicted by the previous tree. This process continuously reduces the model error since each model is sequentially improved against the weaknesses of the previous model. The proposed model has predicted the accuracy of 98.83%.

The proposed SVC relay on boundaries between different classes. The SVC model proposed use grid search to optimize parameters. When the subset is generated by grid search using 10 fold, it is trained using these subset values. The procedure is continued till all features appear in the subset. The proposed model yielded the accuracy of 97.66%%.

The proposed AdaBoost algorithm is initiated by setting the weight of the training set. The training set (u1, v1),…(un, vn) where each ui belongs to instance space U, and each labelvi is in the label set V, which is equal to the set of {-1,+1}. It assigns the weight on the training example i on round t as Dt(i). The same weight will be set at the starting point (Dt(i)=1/N, i=1,…,N). Then, the weight of the misclassified example from base learning algorithm (called a weak hypothesis) is increased to concentrate the hard examples in the training set in each round. The proposed AdaBoosting algorithm yields an accuracy of 97.66%.

The proposed Decision Tree algorithm applied a top-down approach to data so that given a knowledge set, they struggle to group and label observations that are similar between them, and appearance for the simplest rules that split the observations that are not the same between them until they reach a certain degree of similarity. They use a layered splitting process, where at each layer they struggle to separate the info into two or more groups, in order that data that fall under an equivalent group is most similar to every other (homogeneity), and groups are as die-rent as possible from one another. The proposed Decision Tree model has predicted the accuracy of 97.18%.

IV. MACHINE LEARNING ALGORITHMS FOR BREAST CANCER DETECTION

```
Algorithm 1Logistic Regression (LR) Algorithm:
Begin
Step 1: Input data set with real-valued input values. The output of this function $f(z) = 1/ (1 + e^{-z} )$ is converted into probability.
Step 2: If the probability is greater than 0.5 then belongs to class 0 (Malignant) or class 1 (Benign)
Step 3: Repeat Step 2, each repetition is called an epoch (increase in epoch increases accuracy). Update the training data set until the model is accurate enough.
Step 4: Use that training model for prediction.
End
```

```
Algorithm 2 Random Forest (RF) Algorithm:
Begin
Step 1: Input dataset with S number of attributes, and the attributes s are chosen arbitrarily from S to form the nodes for decision tree.
Step 2: Choose a training set m for the above decision tree, which is error free.
Step 3: Split tree based on chosen m and carry out preparation on each decision tree.
Step 4: Poll to find the optimal solution.
End
```

```
Algorithm 3 K- Nearest Neighbor (KNN) Algorithm:
```

Begin

Step 1: Input Data set, KNN acquires all neighbourhood data points. Data points that have a large amount of variation are significant elements in distance determination.

Step 2: Applying Euclidean distances formula to find distances between attributes (P1, P2, ...,P_n), and place them in 2-dimensional planes as

$$D(p_1, p_2) = \sqrt{\sum_{i=1}^{n}(P_1 - P_2)^2}$$

$$D(p_1, p_2) = \sqrt{\sum_{i=1}^{n}(P_1 - P_2)^2}$$

Step 3: Assume K as a positive integer and the first K distances are taken from the above arrangement. Using such distances K to measure the points K. For K greater than 0, K_i is the number of points corresponding to the i^{th} category.

Step 4: The condition of $k_i > k_j$, only when i≠j, then keep x(data point) in the category i.

End

Algorithm 4Decision Tree Algorithm:

Begin

Step 1: Begin the tree with the root node, says S, which contains the complete dataset.

Step 2: Find the best attribute in the dataset using Attribute Selection Measure (ASM).

Step 3: Divide the S into subsets that contains possible values for the best attributes.

Step 4: Generate the decision tree node, which contains the best attribute.

Step 5: Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

End

Algorithm 5 Support Vector Machine (SVM) Algorithm:

Begin

Step 1: Input a labelled data set (u_1, v_1), ...,(u_n, v_n), $v_i \in R_d$, and $v_i \in$ (1, +1), u_i is a vector for feature and v_i is a class label.

Step 2: The optimal hyperplane is defined as M * u + c = 0 to achieve vector for feature selection. And the binary classification can then be expressed as a function F(x) = sign (M * u + c). M is the weight vector; u is input feature and c is the bias

Step 3: All the elements with dissimilarity from the training data set must satisfy if $M * u_i + c \geq +1$ if $v_i = +1$ and $M * u_i + c \leq -1$ if $v_i = -1$. Repeat Step 3 until all the elements.

Step 4: Find M and c for the hyperplane to divide the data as Malignant or Benign.

End

Algorithm 6 Gradient Boosting Algorithm:

Begin

Step 1: build a base model to predict the observations in the training dataset.

Step 2: calculate the pseudo residuals which are (observed value – predicted value).

Step 3: we will build a model on these pseudo residuals and make predictions. Because we want to minimize these residuals and minimizing the residuals will eventually improve our model accuracy and prediction power.

Step 4: find the output values for each leaf of our decision tree. That means there might be a case where 1 leaf gets more than 1 residual, hence we need to find the final output of all the leaves.

Step 5: Finally the last step where we have to update the predictions of the previous model. It can be updated as: $F_m(x) = F_{m-1}(x) + v_m h_m(x)$, where m is the number of decision trees made.

End

Algorithm 7 AdaBoost Classifier Algorithm:

Begin

Step 1: Initially, all observations are given equal weights.

Step 2: A model is built on a subset of data.

Step 3: Using this model, predictions are made on the whole dataset.

Step 4: Errors are calculated by comparing the predictions and actual values.

Step 5: While creating the next model, higher weights are given to the data points which were predicted incorrectly.

Step 6: Weights can be determined using the error value. For instance,the higher the error the more is the weight assigned to the observation.

Step 7: This process is repeated until the error function does not change, or the maximum limit of the number of estimators is reached.

End

V. EXPERIMENTAL SETUP AND RESULT

The dataset contains 569 samples of malignant and benign tumor cells. The first two columns in the dataset store the unique ID numbers of the samples and the corresponding diagnosis (M=malignant, B=benign), respectively.

The columns 3-32 contain 30 real-value features that have been computed from digitized images of the cell nuclei, which can be used to build a model to predict whether a tumor is benign or malignant.
1= Malignant (Cancerous) - Present (M)
0= Benign (Not Cancerous) -Absent (B)

Then, visualize the data attributes using histograms, box plot, etc. now, to work with continuous variables the categorical variables in our data set are converted into continuous variables by performing one-hot encoding. Next the data set is split into the training data set and testing data set. Then feature scaling is performed by using Standard Scaler. Standard Scaler will normalize the features such that each feature will be having mean as 0 and standard deviation as 1.

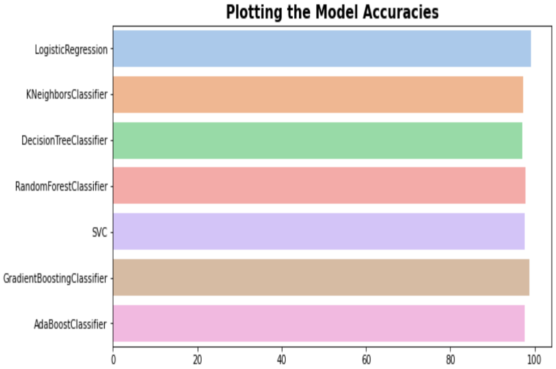
Then the training data set is given to learning algorithms like logistic regression, random forest, decision tree, etc. to generate models. Now to predict whether the cancer cells are malignant or benign, the test data should be given to the model which was generated by using the training data set.

Random Forest creates many decision trees while predicting the output and training of those decision trees require more computations. Logistic regression performs when attributes are linear data with independent attributes but in our case, we have dependent attributes. The model needs only a few relevant features otherwise it leads to over-fitting of data and also training data set to need more independent variables. The Decision tree supports automatic feature interaction and is is faster due to KNN's expensive real time execution. The SVC model performs better with non-linear data and classifying data by using kernel functions either linear or nonlinear. We have used a linear kernel function for prediction. The KNN is easy to implement even though new features were added but feature scaling is necessary before applying algorithm otherwise prediction accuracy will decrease. The algorithms are evaluated based on accuracy. Accuracy is the measure of the prediction made by algorithms and is given as –

Accuracy = (TP + TN)/(TN + FP + FN + TP)

Where TP true positive, TN true negative, FP false positive, FN false negative.

The ML Algorithms with Accuracy as shown below –



For the Breast Cancer Wisconsin (Diagnostic) data set Logistic Regression has achieved 99.12% accuracy, K-Nearest Neighbor has achieved 97.36% accuracy, Decision Tree Classifier has achieved 97.18% accuracy, Random Forest has achieved 97.88% accuracy, SVC has achieved 97.66% accuracy, Gradient Boost has achieved 98.83% accuracy and AdaBoost has achieved 97.66% accuracy. In the model we have calculated accuracy score of each model and tries to improves it. It compares the true value of diagnosis with the corresponding predicted value of diagnosis for test dataset. After creating predictive model, efficiency can be checked. For this, the models can be compared based on accuracy and time consumed.

It was really hard to choose the algorithm which has higher performance, greater accuracy and efficiency, since all of them ended very close in accuracy. The accuracy value of the algorithms from machine learning is shown in below table.

MODEL	ACCURACY
Logistic Regression	99.12%
K-Nearest Neighbor	97.36%
Decision Tree	97.18%
Random Forest	97.88%
SVC	97.66%
Gradient Boost	98.83%
AdaBoost	97.66%

CONCLUSION AND FUTURE SCOPE

When the small cells in breast progressively grow and go uncontrollable way breast cancer emerges which is now prevalent disease in women of all ages. Young women typically experience more hostile effects and lower levels of resilience as opposed to older women. Breast cancer forecast is very critical in the area of Medicare and Biomedical. We present in this paper a comparative study of different machine learning algorithms, for the detection of breast cancer. We considered Accuracy and performance parameters for comparison of these ML algorithms, carried out with the data set. Seven classifiers are observed and presented to all subsets resulting in qualified models.

It has been observed that each of the algorithms had an accuracy of more than 97% and apart the Logistic Regression has the highest of 99% accuracy after improving all the algorithms using different .

Supervised machine learning algorithms are suitable for early detection of breast cancer. The comparative study has given the knowledge that SVM model can be used in BC detection. In addition, these models can be utilized for extracting features and selection within this specific situation. We aspire to carry out detailed research of ML algorithms using deep learning models to adapt more complex deep learning techniques to enhance the performance. We can also use a dataset to predict the re-occurrence of breast cancer after a surgery or chemotherapy session. Artificial Neural Networks can be applied to make the prediction better and smarter. Accuracy can be increased by selecting better features. We will work towards optimizing outcomes and raising the number of patients identified late.

REFERENCES

[1] Akbugday B., "Classification of Breast Cancer Data Using Machine Learning Algorithms," 2019 Medical Technologies Congress (TIPTEKNO), Izmir, Turkey, 2019, pp. 1-4. 2019

[2] Ammu P K and Preeja V. Article: Review on Feature Selection Techniques of DNA Microarray Data. International Journal of Computer Applications 61(12):39-44, January 20. 2020.

[3] Aruna S., Rajagopalan S., Nandakishore L. Knowledge based analysis of various statistical tools in detecting breast cancer. Comput. Sci. Inf. Technol. 2021;2:37–45.2021.

[4] B. N. Dontchos, A. Yala, R. Barzilay, J. Xiang, C. D. Lehman, External validation of a deep learning model for predicting mammographic breast density in routine clinical practice, Acad. Radiol., 28 (2020), 475-480. 2020.

[5] Bing Li N., Chui C.K., Chang S., Integrating spatial fuzzy clustering with level set methods for automated medical image segmentation, Computers in Biology and Medicine, Volume 41, Issue 1, 2019, Pages 1-10, ISSN 0010-4825. 2019.

[6] Fatima N, Liu L, Hong S, Ahmed H. Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis. IEEE Access 2020;8:150360–76. https://doi.org/10.1109/ACCESS.2020.3016715 . 2020.

[7] Islam, M.M., Haque, M.R., Iqbal, H. et al. Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques. SN COMPUT. SCI. 1, 290. 2020.

[8] Nallamala S.H., Mishra P., Koneru S.V. (2019), ‘Pedagogy and reduction of K-NN algorithm for filtering samples in the breast cancer treatment’, International Journal of Scientific and Technology Research, 8(11), PP.2168-2173. 2019.

[9] Qu Z. Predicting diabetes mellitus with machine learning techniques. Front. Genet. 2019;9:515. 2019.

[10] Shen, L., Margolies, L.R., Rothstein, J.H. et al. Deep Learning to Improve Breast Cancer Detection on Screening Mammography. Sci Rep 9, 12495. 2019.

[11] Sutanto D., Ghani M.A., A Benchmark Of Classification Framework For Non-Communicable Disease Prediction : A Review 2018.

[12] Tumuluru P., Lakshmi C.P., Sahaja T., Prazna R. (2019), 'A Review of Machine Learning Techniques for Breast Cancer Diagnosis in Medical Applications', Proceedings of the 3rd International Conference on I-SMAC IoT in Social, Mobile, Analytics and Cloud, I-SMAC 2019, (), PP.618-623. 2019.

[13] Varalatchoumy M. and Ravishankar, "Comparative study of four novel approaches developed for early detection of breast cancer and its stages," 2021 International Conference on Inventive Computing and Informatics (ICICI), Coimbatore, 2021.

[14] Wang H., Yoon W.S. Breast cancer prediction using data mining method; Proceedings of the 2019 Industrial and Systems Engineering Research Conference; Nashville, TN, USA. 30 May–2 June 2019. 2019.

[15] Wolberg W.H. Wisconsin Breast Cancer Database. University of Wisconsin Hospitals; Madison, WI, USA: 1991.

[16] Zhou X, Li C, Rahaman MM, Yao Y, Ai S, Sun C, et al. A Comprehensive Review for Breast Histopathology Image Analysis Using Classical and Deep Neural Networks. IEEE Access 2020;8:90931–56. https://doi.org/10.1109/ACCESS.2020.2993788 . 2020.