2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N) | 978-1-6654-3811-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICAC3N53548.2021.9725688

Breast Cancer Prediction using Different Machine Learning Algorithms

Udit Pratap,
C-DAC, Noida, India,
uditpratap619@gmail.com

Saurabh Chhabra,
C-DAC, Noida, India,
sau.chhabra@gmail.com

Abstract*—Due to Breast Cancer, many women die all across the globe. Breast Cancer has two stages Benign and Malignant. Several factors are responsible for causing Breast Cancer. Machine Learning is used to find out the optimal parameters from all the parameters and to predict the stage of cancer based on these parameters. Different machine learning algorithms are utilized to identify highly important features, and the cancer stage is forecasted based on these features. To determine the best model for Breast Cancer prediction, the SVM, RF, and KNN are implemented. A variety of factors are used to evaluate the models. With Accuracy=98.25, Precision=98.26, Recall=98.25, and F1-Score=98.26, it *has been determined that SVM is the best model for Breast Cancer prediction.

Keywords— Support Vector Machine (SVM), K Nearest Neighbors (KNN), Random Forest (RF), Logistic Regression (LR), Convolutional Neural Network (CNN)

I. INTRODUCTION

As per W.H.O in 2015, 8,800,000 people died of cancer. The analysis of cancer is very important. The analysis of cancer and its prediction with machine learning techniques is being done. During the early stage of Breast cancer, the symptoms are not well presented, so it is hard to diagnose if there is Breast cancer or not.

The chances of occurrence of cancer are more in urban areas and this trend is going upward globally. Breast cancer can be confirmed by a variety of factors, including a biopsy and family history. These characteristics may also influence cancer risk. With the help of available data, the efficiency of cancer prediction has increased significantly.

Cancer has different stages Benign and Malignant. Benign is the early stage of cancer while Malignant is the later stage of cancer, it is easy to cure cancer in the Benign stage, while it is difficult to cure in the Malignant Stage. But it is seen that due to the variation in symptoms of cancer it is not easily identified in the early stages and it became too late to provide efficient treatment. For Breast Cancer prediction there are many parameters i.e. symptoms on which the probability of occurrence of cancer depends.

Symptom's dataset is fitted to the model to find the stage of Breast Cancer.Supervised learning algorithms are generally used for Breast cancer detection in which there is a dataset that consists of features as well as labels, defining whether Breast Cancer is Benign or Malignant. With this data set system is trained, and whenever there comes some new data it classifies whether that is in the Benign or Malignant stage.

II. RELATED WORK

Breast cancer is the subject of numerous studies. "A review was done on 932 males with breast pathology to identify cases of Antidiuretic Hormone. It was concluded that nineteen males were diagnosed with Antidiuretic Hormone. All had Gynecomastia. The surgical procedure was mastectomy in 8 patients and excision in 11" [1]."Two data sets were taken and different algorithms such as RF, Naïve Bayes, LR were applied to those data set. And the results were compared. In this study, it is seen that obtained only some features were relevant for Breast Cancer Prediction" [2]. "Ensemble Machine Learning models were proposed that predict breast cancer with high accuracy compared with without ensemble models. The ensemble model improves the system performance with un-biasing. Here 6 different machine learning algorithms such as decision tree, support vector machine, multilayer perceptron, K- nearest neighbors, logistics regression, and random forest were used and its prediction evaluation was compared with the ensemble and without ensemble techniques" [3]." RF was created with the goal of detecting and forecasting breast cancer. RF was shown to be the most effective for detection. A variety of datasets were utilized to train the RF model. As a result, when new data values are supplied, the results are being predicted. " [4] ." Naive Bayes and KNN were utilized since the purpose and challenge of breast cancer classification are to construct precise and trustworthy classifiers. Following a thorough comparison, it was determined that K-Nearest Neighbor had a higher efficiency of 97.51 percent, while NB had a respectable accuracy of 96.19 percent. However, in the case of a large dataset KNN's running time would rise. [5]." The various Breast Cancer detection machine learning methods were compared. The Wisconsin Diagnosis Breast Cancer data set was used to examine the effectiveness of machine learning approaches. Each method showed a 94 % accuracy rate in detecting whether a tumor was benign or malignant.." [6]." Breast cancer was predicted using machine learning techniques such as KNN, SVM, RF, and Naive Bayes. To improve classifier performance measurement, simulation error is used. Because it had the lowest error rate and the quickest return time, SVM was picked as the best model.." [7]."CNN was used. After the implementation of this method, 99.67% accuracy was achieved. In this model, 12 features were used for the Beast Cancer Prediction" [8]. "Breast cancer is a remarkably risky disease that causes a lot of death for numerous ladies all over the world. So, early detection of this cancer can save a lot of valuable life. Proposed a

ISBN: 978-1-6654-3811-7/21/$31.00 ©2021 IEEE

model that predicts breast cancer based on Support Vector Machine and K-Nearest Neighbors. The Support Vector Machine implemented in Python was found to be the most effective in classifying the diagnostic data set into two classes. They used five learning algorithms: Support Vector Machine, Random Forest, Naive Bayes, and K-Nearest Neighbors. These were applied to the breast cancer dataset and tried to compare the algorithms. Support-Vector Machine has proved its performance on several levels in front of others, especially by the lowest error rate, and shortest turnaround time." [9].

III. PROPOSED METHODOLOGY

1.Dataset

The Wisconsin Breast Cancer (Diagnostic) Data Set is used. The data set was created by Dr. William H Wolberg, a physician at the University of Wisconsin Hospital in the United States. The following is a list of qualities. Radius. Size, shape, and texture are all factors to consider. There are 32 qualities in total, with a total of 569 instances.

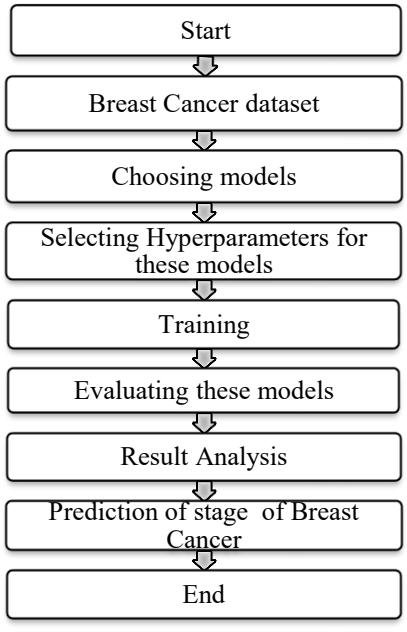

Fig. 1. Proposed Methodology

A target variable diagnostic is divided into two categories: benign and malignant. The non-cancerous stage is known as benign. Cells stop growing after a specific size in the Benign stage. The tumor grows slowly in this case. The cancerous stage is called malignant. Cells stop growing after a specific size in the Benign stage. The tumor grows slowly in this case.

2.Optimal feature Selection

Correlation is being found to get relevant features.These features add to the amount of information available for determining the stage of cancer By using the correlated features model's computation cost reduced.

$$\text{Correlation Cofficent} = \frac{n(\sum pq)-(\sum p)(\sum q)}{\sqrt{[n\sum p^2-(\sum p)^2][n\sum q^2-(\sum q)^2]}}$$

p = data value in the first set, q = data value in the second set, n = Total number of values.

By using a good correlation for optimal parameters, attributes decreased from 32 to 20.

1. perimeter_se
2. radius_mean
3. texture_mean
4. perimeter_mean
5. area_mean
6. compactness_mean
7. concavity_mean
8. concave points_mean
9. radius_se
10. area_se
11. concave points_se
12. radius_worst
13. texture_worst
14. perimeter_worst
15. area_worst
16. smoothness_worst
17. compactness_worst
18. concavity_worst
19. concave points_worst
20. symmetry_worst

Fig. 2. Optimal Parameters

3.Hyperparameter Tuning

The attributes that regulate the entire training process are known as hyper-parameters. When a model is being trained, hyper-parameters have a significant impact on its performance. For selecting optimal hyper-parameters Grid Search is used in this model. A Grid search considers all the possible values given in the hyper-parameters and finds the combination of hyper-parameters that are best for given models. Hence the grid search ensures the optimal set of hyperparameters for the given model.

4. Algorithms

A. SVM is utilized for classification since it can handle large amounts of contiguous and categorical data. The data points closest to the hyperplane are known as Support Vectors. A decision plane that splits a collection of items into classes is known as a hyperplane. The margin is the perpendicular distance between the decision surface and the closest point. C tells SVM optimization how much you don't want each training example to be misclassified. The smaller margin hyperplane with the greater value of C is picked. A hyperplane with a narrower margin is chosen. Kernal values: The kernel function is a mathematical function that accepts data and changes it into the desired format.

- The linear kernel is the most basic type of kernel, and it performs best when there are a lot of characteristics. Other types of kernels are slower than linear kernels. functions. f(x,xj)=sum(x,xj) where x, xj represent the data to be classified.
- The RBF Kernel is one of the most used SVM kernels. It's typically used with nonlinear data. When there is no prior knowledge of data, it aids in proper separation. $f(x,xj)=\exp^{-(\text{gamma}(x-xj)2)}$
- Polynomial Kernal is a more generalized representation of the linear kernel f(x,xj)= (x.xj+1)degree ,f(x,xj)= indicate a decision line that divides the supplied class.

B. Random Forest creates decision trees, and find prediction from each of them and then select best from them. Hyper Parameters for Random forest are :

- Max_features are the maximum attributes taken into account for determining the optimum node split.
- The criterion function is used to assess a split's quality.
- Min_sample_split is the attributes required to produce a split in an intermediate node.
- Min_sample_leaf is the attributes that can be stored in a tree leaf before it is referred to as a leaf.

C. K-Nearest Neighbors is a classification method that calculates the distance between new data points using neighboring data points.

- n_neighbors: Number of neighbors taken into consideration.
- Weights: These are the magnitudes assign to the neighbors' data point.
- Algorithm: These are the algorithms to compute the nearest neighbors.
- Leaf_size: the size of the leaf passed to the BallTree or KDTree

5.Implementation

A comparison study is conducted using several models such as Support Vector Machine, Random Forest, and K-Nearest Neighbors. Pandas, NumPy, Matplot, Seaborn, and Sklearn are some of the Python Machine libraries utilised in the implementation. There is a link between all of the characteristics and the target variable diagnostic from a total of 32 attributes.The feature that is correlated is taken for further consideration resulting in a total of 20 features. The data is separated into two categories: training and testing.30 percent of data used for testing and 70 percent used for training resulting in 171,20 samples for testing and 398,20 samples for training. Different values of Hyperparameters are set for Random Forest "n estimators": [9, 10, 15], "max features": ['log2','sqrt','auto'], "criterion": ['entropy','gini'], "max depth": [5,10,30]. , "min samples leaf": [1, 5, 8], "min samples split": [2, 3, 5]. For Support Vector Machine different Hyperparameter values are : "kernel": ['linear'], "C": [1, 10, 100]. For k-Nearest Neighbors, different Hyperparameter values are: "n_neighbors": [3, 4, 5, 10], "weights": ['uniform', 'distance'], "algorithm": ['auto', 'ball tree', 'kd tree', 'brute'].

IV.RESULT ANALYSIS

Different performance matrices, for example, Accuracy, Precision, Recall, F1 score are calculated for each model.

Support Vector Machine

Fig. 3. Confusion Matrix for SVM

Table No. I Performance Measure Indices for SVM

	precision	recall	f1-score
1	.9907	.9815	.9860
0	.9688	.9841	.9764
Avg	.9826	.9825	.9825

Random Forest

Fig. 4. Confusion Matrix for RF

Table No. II Performance Measure Indices for RF

	precision	recall	f1-score
1	.9677	.9524	.9600
0	.9725	.9815	.9770
Avg	.9707	.9708	.9707

K Nearest Neighbors

Table No. III Performance Measure Indices for KNN

	precision	recall	f1-score
1	.9720	.9630	.9674
0	.9375	.9524	.9449
Avg	.9593	.9591	.9591

V. CONCLUSION/FUTURE WORK

A comparison of Support Vector Machine, Random Forest, and K-Nearest Neighbors is presented in this paper. The best model for Breast Cancer prediction is found by performing a performance analysis on the UCI Breast Cancer Wisconsin (Diagnostic) dataset. With an accuracy of 98.25 percent, precision of 98.26 percent, recall of 98.25 percent, and F1-Score of 98.26 percent, Support Vector Machine was shown to be the most successful in predicting whether malignancy is benign or malignant. As a result, the models presented in this work can be used to predict whether a person is healthy or has cancer

REFERENCES

[1] S. B. Coopey, K. Kartal, C. Li, A. Yala, R. Barzilay, H. R. Faulkner, T. A. King, F. Acevedo, J. E. Garber, A. J. Guidi and K. S. Hughes, "Atypical ductal hyperplasia in men with gynecomastia: what is their," Breast Cancer Research and Treatment, vol. 175, no. 1, 2019.

[2] D. R, I. R. Paul, S. S. Akula, M. Sivakumar and J. J. Nair, "A Comparative Study for Breast Cancer Prediction using Machine Learning and Feature Selection," in International Conference on Intelligent Computing and Control Systems *(ICCS)*, 2019.

[3] N. R. S. K and A. R. Nair, "Efficient Breast Cancer Prediction Using Ensemble Machine Learning Models," in 4th International Conference on Recent Trends on Electronics, Information, Communication & Technology (RTEICT), 2019.

[4] M. S. Yarabarla, L. K. Ravi and A. Sivasangari, "Breast Cancer Prediction via Machine Learning," in 3rd International Conference on Trends in Electronics and Informatics (ICOEI), 2019ontrol Systems (ICCS), 2019.

[5] M. Amrane, S. Oukid, I. Gagaoua and T. Ensari, "Breast Cancer Classification Using Machine Learning," in Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), 2018

[6] S. Sharma, A. Aggarwal and T. Choudhury, "Breast Cancer Detection Using Machine Learning Algorithms," in 2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), 2

[7] 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), 2018. 018.

[8] N. Khuriwal and N. Mishra, "Breast Cancer Diagnosis Using Deep Learning Algorithm," in 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2018.

[9] M. M. Islam, H. Iqbal, M. H. Rezwanul and M. K. Hasan, "Prediction of breast cancer using support vector machine and K-Nearest neighbors," in 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), 2017.