International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 05 Issue: 05 | May - 2021 ISSN: 2582-3930

Early Detection of Breast Cancer

Snehal Mhatre[1], Dr. Kamal Shah[2]

[1]Information Technology, Thakur College of Engineering and Technology, Kandivali(E), Mumbai
[2]Information Technology ,Thakur College of Engineering and Technology, Kandivali(E), Mumbai

---

Abstract -Breast Cancer is more common hence early detection of breast cancer is necessary.Breast cancer are of two types the benign (non-cancerous) and malignant(cancerous).Benign breast cancer are abnormal growth, but they do no spread outside of the breast and they are not life threatening.Malignant breast cancer starts in the cells of the breast and is life threatening. This paper presents a comparison of several machine learning (ML) algorithms: Support Vector Machine (SVM), K-Nearest Neighbor (KNN), NaïveBayes (NB),Decision Tree (DT) ,Random Forest (RF), Adaptive Boosting (AdaBoost), eXtreme Gradient Boosting(XGBoost) on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset. The algorithm with the best results is used, it will then classify the cancer as benign or malignant.

Key Words: cancer, benign, malignant, detection, cancer in womens

1.INTRODUCTION

Early detection is the most effective way to reduce breast cancer deaths. Early diagnosis requires an accurate and reliable procedure to distinguish between benign breast tumors from malignant ones .Breast Cancer Types - three types of breast tumors: Benign breast tumors, In-situ cancers, and Invasive cancers. The majority of breast tumors detected by mammography are benign. They are non-cancerous growths and cannot spread outside of the breast to other organs. In some cases, it is difficult to distinguish certain benign masses from malignant lesions with mammography.If the malignant cells have not gone through the basal membrane but is completely contained in the lobule or the ducts, the cancer is called in-situ or noninvasive.If the cancer has broken through the basal membrane and spread into the surrounding tissue, it is called invasive.This analysis assists in differentiating between benign and malignant tumors.Breast cancer develops from breast tissues with abnormal cells growing, changing and multiplying out of control. It is the most common type of cancer among women in both developed and less developed nations with an estimated death of 508,000 women in the year 2011 alone [1] and accounted for 25% of all cancer cases and 15% of all cancer deaths among females in the estimated cancer case in 2012 [2].Most women who get breast cancer do not have a family history of the disease but women who have close blood relatives with breast cancer have a higher risk. Cancer constitutes an enormous burden on society in more and less economically developed countries alike. Cancer cases are becoming more common due to the growth and aging of the population, as well as a widespread rise of established risk factors such as smoking, overweight, physical inactivity. Early detection of cancer significantly increases the probability of recovering through successful treatment. Delays in diagnosis results in late-stage presentation with consequences of lower likelihood of survival, higher costs of treatments and even death. The most common techniques used for cancer detection are X-ray mammography and magnetic resonance imaging (MRI). However, these present innovations have a few downsides as they are very costly, extensive in size and are only affordable in large hospital facilities. The mentioned methods also may have some side effects and false positives. However, with the volume of data generating extremely fast in the field of biomedical and advancement of technology, machine learning techniques offer promising results. Machine learning helpsto extract information and knowledge from the basis of past experiences and detect hard-to perceive pattern from large and noisy dataset to give accurate results within a short period of time. Application of machine learning in the medical domain is growing rapidly due to the effectiveness of its approach in prediction and classification, especially in medical diagnosis to predict breast cancer, now it is widely applied to biomedical research. Researchers use machine learning for cancer prediction and prognosis.Machine learning allows inferences or decisions that otherwise cannot be made using conventional statistical methodologies.With a robustly validated machine learning model, chances of right diagnosis improve.It specially helps in interpretation of results for borderline cases.

2. BLOCK DIAGRAM

In the below block diagram, is the proposed system in which the WDBC dataset is used for the breast cancer detection.The set is divided into training and testing set.The same data is used for visualization so that the content is clear . Further seven algorithms are used to classify the dataset and the algorithm that gives the heighest accuracy will be used for deployment in healthcare project on Heroku Platform.Refer Fig-1.

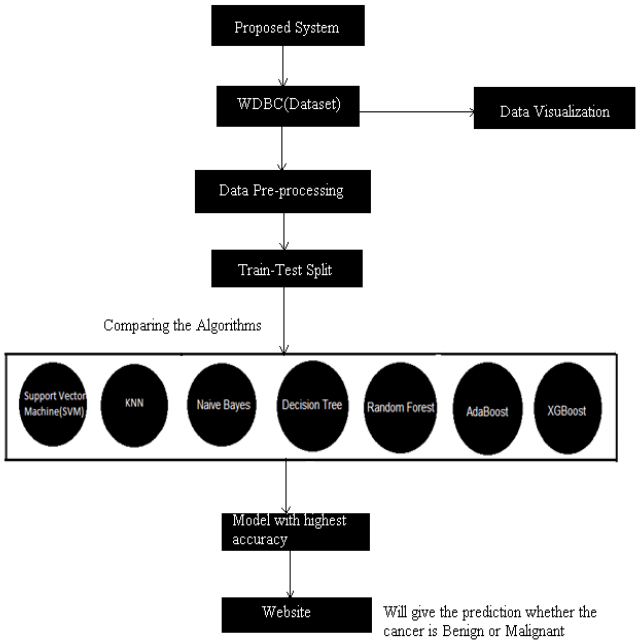

Fig -1: Block Diagram

3. BACKGROUND

1.1 Data Visualization:

Data visualization is the representation of data or information in a graph, chart, or other visual format. It communicates relationships of the data with images. This is important because it allows trends and patterns to be more easily seen.

1.2 Classification:

The Classification algorithm is a Supervised Learning technique that is used to identify the category of new observations on the basis of training data. In Classification, a program learns from the given dataset or observations and then classifies new observation into a number of classes or groups. Such as, Yes or No, 0 or 1, Spam or Not Spam, cat or dog, etc. Classes can be called as targets/labels or categories.Unlike regression, the output variable of Classification is a category, not a value, such as "Green or Blue", "fruit or animal", etc. Since the Classification algorithm is a Supervised learning technique, hence it takes labeled input data, which means it contains input with the corresponding output.

1.3 Data Pre-processing:

Train-Test Split:The data is normally split into two subsets: training data and testing data (and sometimes to three: train, validate and test). The training dataset is the actual dataset that is used to train the model. The model sees and learns from this data. The test dataset is the sample of data used to provide an unbiased evaluation of a final model fit on the training dataset. The splitting ratio depends on the total number of samples and the actual training model. The train/test split was implemented using the train_test_split class of scikit-learn's model_selection package.

Feature Scaling: The range of values of the attributes in the dataset varies widely and feature scaling is used to bring it to a standardized range. It is also known as data normalization. This is done because some algorithms will not function properly without it and data should be standardized. In this paper, scikit-learn module sklearn.preprocessing.StandardScaler is used to implement standardization[3].

1.4 Machine Learning Algorithms:

Machine learning (ML) is one of the fields of Artificial Intelligence (AI) where statistical techniques are used to provide the computer systems with the capability to "learn" and improve by itself progressively without being explicitly programmed. Machine learning explores the study and construction of algorithms that can learn from and make predictions on data and datasets[4]. The term 'machine learning' was initially coined by Arthur Samuel in 1959 [5]. Three important categories of machine learning can be described as follows:

Supervised Learning:In this form of learning, the machine uses data that is labelled and some of the data is already tagged with the correct answers to learn the mapping function from the input to the output. The goal is to approximate the mapping function so well that when you have new input data (x) that you can predict the output variables (Y) for that data.

Unsupervised Learning:The machine is trained using data that is neither classified nor labeled. This allows the algorithm to perform calculations on it's own accord without guidance. The task of the machine is to group unsorted data according to patterns and differences in order to find the hidden structure by itself.

Reinforcement Learning:The machine or the agent learn how to behave in an environment by performing actions and seeing the results based on the action allowing it to dictate the ideal action in a specific circumstance.

In the modern times, the vast amount of data available is not feasible for human being to keep up with and analyze them. Machine learning, which is a subset of computer science and an important branch of artificial intelligence, primarily focuses on the development and building of algorithms to over this problem. The very recent advancement in this field has opened up vast and almost limitless applications in fields ranging from financial industries, data security to medical fields. But there is still much space to make progress by means of using Machine Learning in social media services, disease prediction and identification, virtual assistants, search

engine refining, fraud detection, manufacturing, etc. It is only going to improve and integrate in our daily lives making it easier and more convenient in the future.Algorithms used are:

Support Vector Machine[6],K-Nearest Neighbor[7],Naïve Bayes[8][9],Decision Tree[10],Random Forest[11][12][13],Adaptive Boosting[14],eXtreme Gradient Boosting[15].

1.5 Result Analysis:

The performance of the algorithms differed with and without principal component analysis implementation on the dataset. Analysis and comparison of the performance of different models implemented on the test portion of the dataset was evaluated.

(a) Performance Metrics: This paper deals with classification problem and therefore the chosen performance metrics primarily focus on classification. For the detection of breast cancer, if the target variable is 1 then it is a positive instance, meaning the patient has a malignant tumor and therefore cancer. And if the target variable is 1, then it a negative instance, meaning the tumor is benign and the patient does not have cancer.

Parameter Tuning: Hyperparameter tuning is the process of determining the right combination of hyperparameters that allows the model to maximize model performance. Setting the correct combination of hyperparameters is the only way to extract the maximum performance out of models.

Confusion matrix: The Confusion matrix is one of the most intuitive and easiest metrics used for finding the correctness and accuracy of the model. It is used for classification problems where the output can be of two or more types of classes which makes it perfect for this paper. The table layout or the matrix layout helps to visualize the performance of an algorithm. Each row of the matrix in Table 1 represents the instances in an actual class while each column represents the instances in a predicted class or vice versa[16].

Table -1: Confusion Matrix

		Predictive Negative (0)	Predictive Positive (1)
Actual Negative(0)		True Negative (TN)	False Positive (FP)
Actual Positive (1)		False Negative (FN)	True Positive (TP)

Terms associated with Confusion matrix:

True Positives (TP): True positives are the cases when the actual class of the data point was True(1) and the predicted is also True(1) Ex: The case where a person is actually having malignant (0) tumor and the model classifying his case as malignant (0) comes under True Positive.

True Negatives (TN): True negatives are the cases when the actual class of the data point was False (0) and the predicted is also False (0). Ex: The case where a person having benign (1) tumor and the model classifying his case as benign (1) comes under True Negatives.

False Positives (FP): False positives are the cases when the actual class of the data point was False (0) and the predicted is True (1). False is because the model has predicted incorrectly and positive because the class predicted was a positive one (1). Ex: A person having a benign (1) tumor and the model classifying his case as malignant (0) comes under False Positives.

False Negatives (FN): False negatives are the cases when the actual class of the data point was True (1) and the predicted is False (0). False is because the model has predicted incorrectly and negative because the class predicted was a negative one (0). Ex: A person having malignant (0) tumor and the model classifying his case as benign (1) tumor comes under False Negatives. The ideal scenario for the model would be when it gives 0 False Positives and 0 False Negatives.Refer Fig-2.

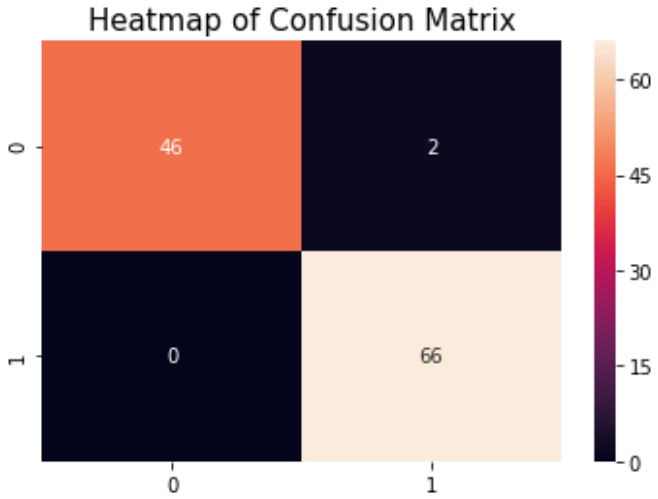

Fig -2: Confusion Matrix

Classification Report of Model:

Precision: Precision is the ratio of True Positives to the summation of True Positives and False Positives. Ex: Precision is a measure of proportion of patients that has been diagnosed as having malignant tumor, actually had malignant tumor. The predicted positives (People predicted as having

malignant tumor are TP and FP) and the people actually having a malignant tumor are TP. Precision =TP/TP+FP.

Recall: Recall is a measure that shows the proportion of patients that actually had malignant tumor was diagnosed by the algorithm as having malignant tumor. The actual positives (People having malignant tumor are TP and FN) and the people diagnosed by the model having a malignant tumor are TP. Therefore, if we want to focus more on minimizing False Negatives, we would want our Recall to be as close to 100% as possible. Recall =TP/TP+FN.

F1 Score: F1 Score is the Harmonic Mean between precision and recall. The range for F1 Score is [0,1]. It shows how precise the classifier is and how robust it is at the same time. F1 =2* Precision* Recall/Precision+Recall.

Support: Support is the number of actual occurrences of the class in the specified dataset. Refer Table -2.

Table -2 : Classification Report Model

	Precision	Recall	F1-Score	support
0.0	1.00	0.96	0.98	48
0.1	0.97	1.00	0.99	66
micro avg	0.98	0.98	0.98	114
macro avg	0.99	0.98	0.98	114
Weighted avg	0.98	0.98	0.98	114

Cross-validation of the model : Cross-validation, sometimes called rotation estimation or out-of-sample testing, is any of various similar model validation techniques for assessing how the results of a statistical analysis will generalize to an independent data set.

(b) Model Performance: Evaluating the performance of a model is one of the core stages in the data science process. It indicates how successful the scoring (predictions) of a dataset has been by a trained model.

Support Vector Machine: The Accuracy of Support Vector Machine model classifier is 57.89% and when the same is trained with the standard scaled data the accuracy is 96.49%.

K-Nearest Neighbor: The Accuracy of model K-Nearest Neighbor classifier is 93.85% and when the same is trained with the standard scaled data the accuracy is 57.89%.

Naïve Bayes: The Accuracy of model Naïve Bayes classifier is 94.73% and when the same is trained with the standard scaled data the accuracy is 93.85%.

Decision Tree: The Accuracy of model Decision Tree classifier is 94.73% and when the same is trained with the standard scaled data the accuracy is 75.43%.

Random Forest: The Accuracy of model Random Forest classifier is 97.36% and when the same is trained with the standard scaled data the accuracy is 75.43%.

Adaptive Boosting: The Accuracy of model Adaptive Boosting classifier is 94.73% and when the same is trained with the standard scaled data the accuracy is 94.73%.

eXtreme Gradient Boosting: The Accuracy of model eXtreme Gradient Boosting classifier is 98.24% and when the same is trained with the standard scaled data the accuracy is 98.24%.

The highest accuracy is given by the eXtreme Gradient Boosting Model that is about 98.24% .After the cross Validation of the model the mean accuracy value is 96.24% and XGBoost model accuracy is 98.24%.It is showing XGBoost is slightly overfitted.

4. DISCUSSION

After completing the implementation of all seven algorithms for detecting breast cancer from the dataset, the results can be compared.Refer Table -3.

Table -3: Algorithm Results

	Classifier	Train with Standard Scaled Data
Support Vector Machine	57.89%	96.49%
K-Nearest Neighbor	93.85%	57.89%
Naive Bayes	94.73%	93.85%
Decision Tree	94.73%	75.43%
Random Forest	97.36%	75.43%
Adaptive Boosting	94.73%	94.73%
eXtreme Gradient Boosting	98.24%	98.24%

After taking all the results into account the highest accuracy is given by the eXtreme Gradient Boosting Model i.e about 98.24%. The same model is saved and we have completed the project successfully with 98.24% accuracy which is great. Now we are ready to deploy our project model in the healthcare project using the Heroku Platform. Refer Fig -3.

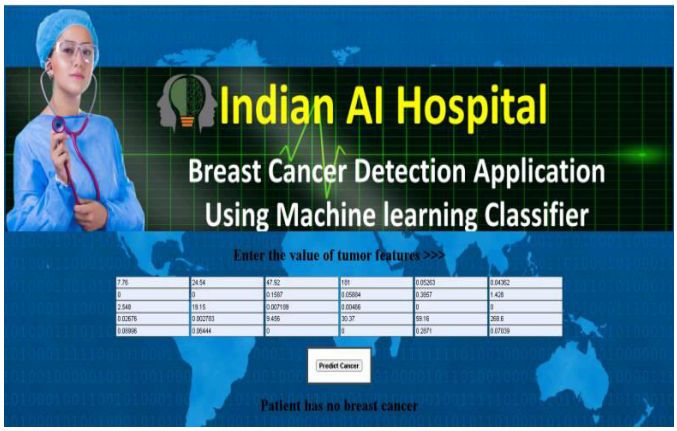

Fig -3: Result

5.CONCLUSION

The important aspects in breast cancer are early detection and risk reduction. Screening may identify early noninvasive cancers and allow treatment before they become invasive or identify invasive cancers at an early treatable stage. But screening does not, per se, prevent cancer. By using machine learning algorithms we will be able to classify and predict the cancer into being or malignant. Machine learning algorithms can be used for medical oriented research, it advances the system, reduces human errors and lowers manual mistakes. This project shows the comparative analysis of different machine learning algorithms in detecting breast cancer from a digitized image of a fine needle aspirate (FNA) of a breast mass. The simple, safe, accurate, and inexpensive procedure of FNA combined with the predictive model in this paper can be used for prognosis, diagnosis and assist doctors in making the final decision more accurately in shorter time span with less human and monetary resource.

REFERENCES

1. WHO (2016). Breast cancer: prevention and control.

2. Torre, L. A., Bray, F., Siegel, R. L., Ferlay, J., Lortet-Tieulent, J., and Jemal, A. (2015). Global cancer statistics, 2012. CA: a cancer journal for clinicians, 65(2):87–108.

3.Gholami, V., Chau, K., Fadaee, F., Torkaman, J., and Ghaffari, A. (2015). Modeling of groundwater level fluctuations using dendrochronology in alluvial aquifers. Journal ofhydrology, 529:1060–1069.

4.Kohavi, R. (1998). Glossary of terms. Special issue on applications of machine learningand the knowledge discovery process, 30(271):127–132.

5. Kohavi, R. (1998). Glossary of terms. Special issue on applications of machine learning and the knowledge discovery process, 30(271):127–132.

6.Hussain, M., Wajid, S. K., Elzaart, A., and Berbar, M. (2011). A comparison of svm kernel functions for breast cancer detection. In 2011 Eighth International Conference Computer Graphics, Imaging and Visualization, pages 145–150. IEEE.

7. Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, 46(3):175–185.

8. Kharya, S. and Soni, S. (2016). Weighted naive bayes classifier: A predictive model for breast cancer detection. International Journal of Computer Applications, 133(9):32–7.

9. Rennie, J. D., Shih, L., Teevan, J., and Karger, D. R. (2003). Tackling the poor assumptions of naive bayes text classifiers. In Proceedings of the 20th international conference on machine learning (icml-03), pages 616–623.

10. E. Venkatesan and T. Velmurugan, "Performance analysis of decision tree algorithms for breast cancer classification", *Indian Journal of Science and Technology*, vol. 8, no. 29, 2015.

11. Ho, T. K. (1995). Random decision forests. In Document analysis and recognition, 1995., proceedings of the third international conference on, volume 1, pages 278–282. IEEE.

12. Barandiaran, I. (1998). The random subspace method for constructing decision forests. IEEE transactions on pattern analysis and machine intelligence, 20(8).

13. Friedman, J., Hastie, T., and Tibshirani, R. (2001). The elements of statistical learning, volume 1. Springer series in statistics New York, NY, USA:.

14. T. Wang, W. Li, H. Shi and Z. Liu, *Software Defect Prediction Based on Classifiers Ensemble*, vol. 16, pp. 4241-4254, December 2011.

15. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system", *in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, pp. 785-794, 2016.

16. Powers, D. M. (2011). Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation.

.