

**R & D Project Report**

**Academic Year- 2021-22**

On

**Machine Learning Based Classification of Breast Cancer**

Submitted by

1. Sourabh Asharma BT19GBT163 B.Tech Biotechnology
2. Siddhesh Shinde BT19GCS125 B.Tech Computer Science
3. Aditya Choudhari BT19GCS158 B.Tech Computer Science

Dr. Utkarsh Raj

Logo, company name

Description automatically generated

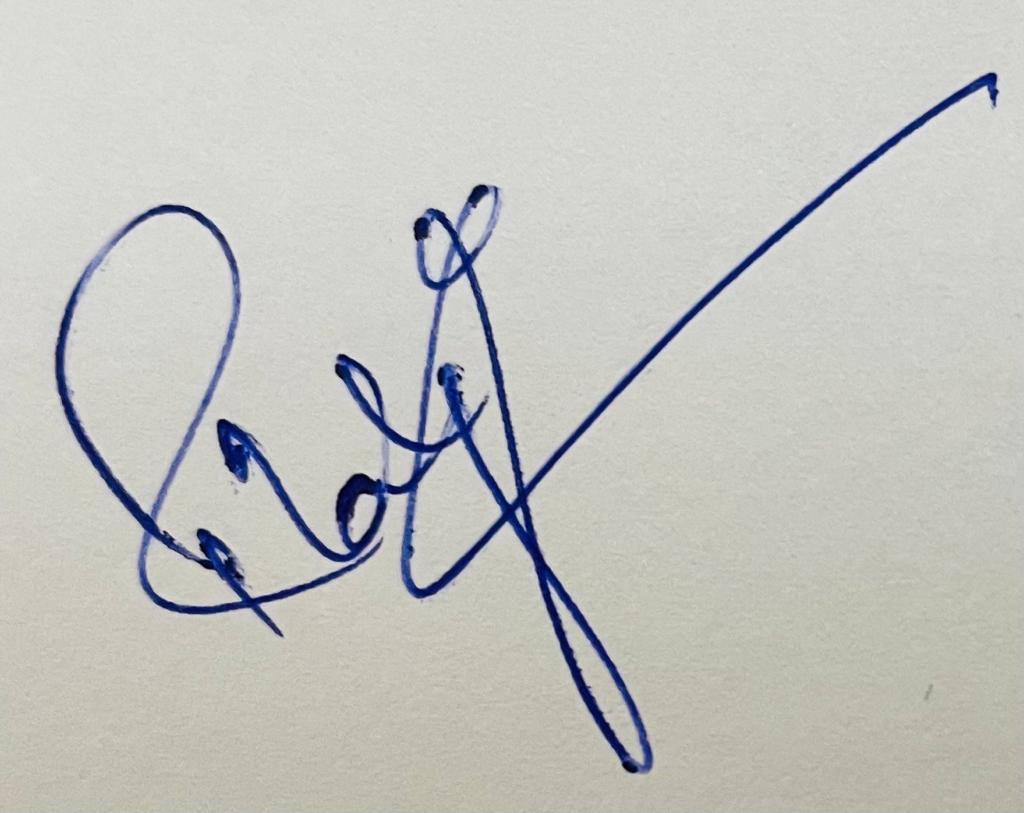
=================================================================

**CERTIFICATE BY SUPERVISOR(S)**

This is to certify that the present R&D project entitled **Machine Learning Based Classification of Breast Cancer** is being submitted to NIIT University, Neemrana, in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology, in the area of BT/CSE/ECE/GIS, embodies faithful record of original research carried out by **Siddhesh Shinde, Sourabh Asharma, Aditya Choudhari**. She / He / They has / have worked under my/our guidance and supervision and that this work has not been submitted, in part or full, for any other degree or diploma of NIIT or any other University.

Place: NIIT University

Signature:



Name of the Supervisor(s): Dr Utkarsh Raj

Date: 28th May, 2022

Logo, company name

Description automatically generated

=================================================================

**DECLARATION BY STUDENT(S)**

I/We hereby declare that the project report entitled **Machine Learning Based Classification of Breast Cancer** which is being submitted for the partial fulfillment of the Degree of Bachelor of Technology, at NIIT University, Neemrana, is an authentic record of my/our original work under the guidance of **Dr. Utkarsh Raj**. Due acknowledgements have been given in the project report to all other related work used. This has previously not formed the basis for the award of any degree, diploma, associate/fellowship or any other similar title or recognition in NIIT University or elsewhere.

Place: NIIT University

Date: 28th May, 2022

1. Sourabh Asharma BT19GBT163 B.Tech Biotechnology
2. Siddhesh Shinde BT19GCS125 B.Tech Computer Science
3. Aditya Choudhari BT19GCS158 B.Tech Computer Science

**TABLE OF CONTENTS**

| **S.No.** | **Topic** | **Page Number** |
| --- | --- | --- |
| 1 | Certificate By Supervisor | 2 |
| 2 | Declaration By students | 3 |
| 3 | Content Page | 4 |
| 4 | List of Tables and Figures | 5 |
| 5 | Introduction | 6 |
| 6 | Problem Statement | 7 |
| 7 | Literature Review | 7 |
| 8 | Proposed Methodology - Workflow and Techniques | 9 |
| 9 | Result and Analysis | 14 |
| 10 | Comparison with other work and Analysis |  |
| 11 | Conclusion |  |
| 12 | Future Scope |  |
| 13 | References |  |

**LIST OF TABLES**

| **S.No.** | **Table Number** | **Description of Table** |
| --- | --- | --- |
| 1. | Table 1 | Table of correlation between features |
| 2. | Table 2 | Summary of Predicted values of the implemented ML Classifiers |

**LIST OF FIGURES**

| **S.No.** | **Figure Number** | **Description of Figure** |
| --- | --- | --- |
| 1. | Figure 1 | Flowchart showing basic methodology |
| 2. | Figure 2 | Flowchart illustrating Machine Learning Classifiers implemented on dataset |
| 3. | Figure 3 | Correlation depiction between different features using pairplot |
| 4. | Figure 4 | Heatmap of correlation between features |
| 5. | Figure 5 | Accuracy of various ML classifiers in % |
| 6. | Figure 6 | Precision of various ML classifiers |
| 7. | Figure 7 | Recall of various ML classifiers |
| 8. | Figure 8 | F1 Score of various ML classifiers |
| 9. | Figure 9 | ROC plot of ML Classifiers |

**INTRODUCTION**

Breast cancer has several causes, including hormonal changes, radiation therapy, excessive amounts of body fat, gene-level mutations, and also reproductive variables. According to a report by WHO, around 1,000,000 women every year are newly detected with breast cancer, and around 40% of them die because the cancer is most of the times detected at a later stage. Breast cancer is caused generally due to variation in our DNA structure or damage to it, this further leads to mutation in gene structure, and because the instructions in the DNA are messed up, the altered genes don't act properly, resulting in cancer. If the problem is not fixed within a few months, millions of cells with erroneous instructions, known as cancer cells, will form. A tumour develops and grows gradually as cancer cells divide. Tumors can be benign (noncancerous) or malignant (cancerous). Malignant tumours (cancerous tumours) can spread to neighbouring cells, causing metastasis and starts expanding to other body organs and parts, but benign tumors cannot spread to other tissues and can only grow within the benign mass. Because there are no symptoms in the early stages of breast cancer, identification might be challenging. However, a proper diagnosis is done to distinguish between benign and malignant tumours, following a set of clinical testing. False positives and false negatives are rare with effective detection.

Machine learning can be regarded as a collection of tools that may be used to create and test algorithms that help with prediction, pattern recognition, and classification. Collecting data, finalizing the ML model, training the ML model on the dataset, and finally testing the model on the dataset are the four major processes of any Machine Learning model. The research on the relationship between Breast Cancer and Machine Learning is not a new field, many researchers has been working on it for decades to classify breast cancer tumors and other malignancies, to forecast cancer-causing gene sequences, and to define prognostic factors.

Random Forest, AdaBoost, XGBoost, Support Vector Machine are being utilized in this work as machine learning classifiers, to classify out whether a patient has either a malignant or benign tumor, based on the 10 real valued features computed for each cell nuclei, using fine needle aspiration biopsy. In this research, we analyze four ML classifiers for breast cancer categorization. The Breast Cancer Wisconsin Dataset from the University of Wisconsin is being used for the study. The aim is to implement the above mentioned four classifiers in Breast Cancer Wisconsin Dataset data set, and hyper-tuning the parameters of each classifier to develop efficient and optimal machine learning classifiers for the categorization of breast cancer. Precision, sensitivity, specificity, recall, accuracy, and F1 score are all performance metrics used to calculate the effectiveness of each machine learning classifier.

**PROBLEM STATEMENT**

Breast cancer is the most frequent and prevalent malignancy among Indian women at the moment. In general, if a patient's breast symptoms or the findings of an imaging test (such as a mammography) indicate that the patient might have breast cancer, the patient is advised to undergo a breast biopsy. We are focusing on one of the most common breast biopsies which is Fine Needle Aspiration here. In Fine Needle Aspiration biopsy, a little amount of breast tissue or fluid is taken from a suspicious location with a thin, hollow needle and examined for cancer cells. The problem statement is to use Breast Cancer Wisconsin Dataset which contains the numerical values of features of each cell nucleus calculated after Fine Needle Aspiration biopsy, implement various machine learning classifiers on the dataset and estimate the type of breast cancer being, benign or malignant. It is also expected that the accuracy of the predictions made by each classifier is increased either by hyper-tuning the parameters of each classifier or optimizing the classifiers so that a higher accuracy of the predictions can be achieved, and maximum lives can be saved by the predictions made.

**LITERATURE REVIEW**

The dataset used in this research paper includes both numeric and image data. They have used mainly 3 techniques for classification which are Random forest classifier, KNN classifier and Decision trees. The random forest classifier has shown the result as F1-score of **92%** which is highest among all, followed by KNN classifier. In the numeric dataset it is observed that Random Forest Classifier and Decision trees gave almost similar results. The **Resitine** and **BMI** are found to be two important biomarkers[2].

The dataset used in this study has a digitized image set of a fine needle aspirate. Various ML classifiers are used to train the dataset and give precise results. The methodology used in this focuses on the geometrical and textural features of biopsy that were accurate were taken for the study. To separate between benign and malignant cancer cells features like perimeter, radius,texture,compactness and concavity. They have also used concave points of the cell image for differentiation. The results of this study showed that around 569 people were considered out of which the study showed that around 63 percent of them were treated with benign tumors and 37 percent of them treated with malignant tumors. Out of all the features 3 features were selected primarily like area,perimeter and radius. Other features did not have strong results to show any strong preference to a particular treatment over another. One of the most significant results was that the mean value of benign cells were found to be lower than malignant tumor cells. This result strongly suggests that the malignant tumor spreads in the body.[3]

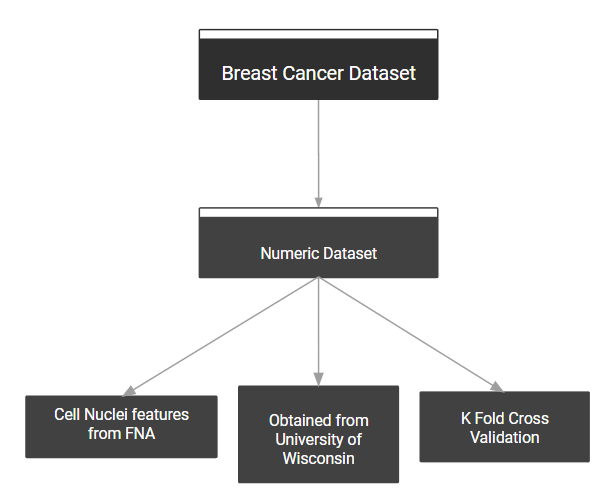
The nature of this study was quantitative and compared accuracies of primarily three machine learning classifiers. The classifiers used in this study were KNN, Support vector machine and Naive Bayes. The results showed that with k=3 the model had the highest accuracy,around 96.85 percent. Two SVM classifiers were considered C-SVM and nu-SVM.It says the most accurate results came with **C = and ɣ =** that gave around 96.85 percent accuracy. One more methodology used was Naïve Bayes which gave the result with accuracy 95.99%.[4]

In this another research paper we found that the dataset used was in the form of Image. The key points discussed in this paper are how to choose the right dimensions and how to deal with annotation problems in the training dataset. They also illustrated how one should identify a perfect training dataset. There were two main drawbacks in the study they have found: (a) Development of task specific techniques that require long research time. (b) Searching the optimal parameters for algorithms. The results are as follows: They found F-score for nuclei segmentation as **0.83**, for epithelium segmentation as **0.84**, for tubule segmentation as **0.83**, for lymphocyte detection as **0.90**, for mitosis detection as **0.53**, for invasive ductal carcinoma detection as **0.7648**, for lymphoma classification they evaluated out the classification accuracy as **0.97.**[5]

**PROPOSED METHODOLOGY**

**WORKFLOW:**

**Data gathering and pre-processing:** Breast Cancer Wisconsin dataset is available on both Kaggle and UCI Machine Learning Repository, and is obtained from the same. The Wisconsin dataset is public, and various studies have already been conducted using the dataset. The dataset was converted into CSV format in order to examine the data in the Jupyter notebook. The dataset is required to be pre-processed in order to eliminate the id column, the last column with a null value, and any rows with some null value. In the end, 32 characteristics were employed in the detecting procedure.

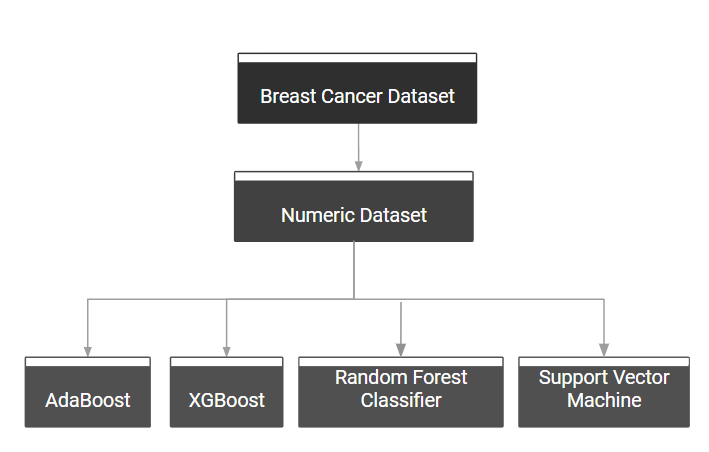


**Figure 1.** Flowchart showing basic methodology

**Analysis:** The analysis was carried out utilizing a python jupyter notebook to display the correlations between the data and their relationships.

**Classification Process:**

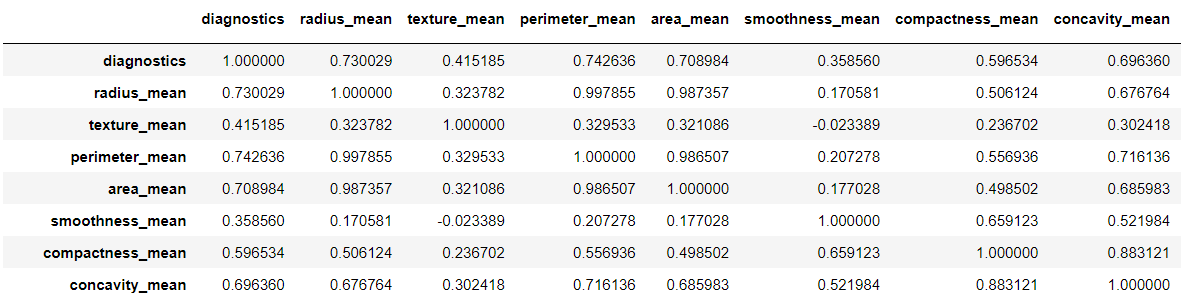
In the classification, five methods were used: Random Forest Classifier, AdaBoost, XGBoost, Support Vector Machine, and Sequential Minimal Optimization. To achieve ideal accuracy, the algorithms were tuned at various hyper-parameters. Modifications were made to the model in order for the classifier-generated model to properly categorize the random data.



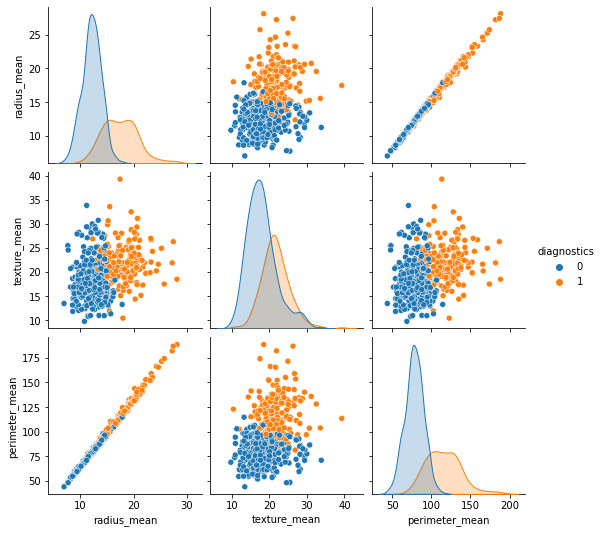
**Figure 2.** Flowchart illustrating Machine Learning Classifiers implemented on dataset

**Correlation:**

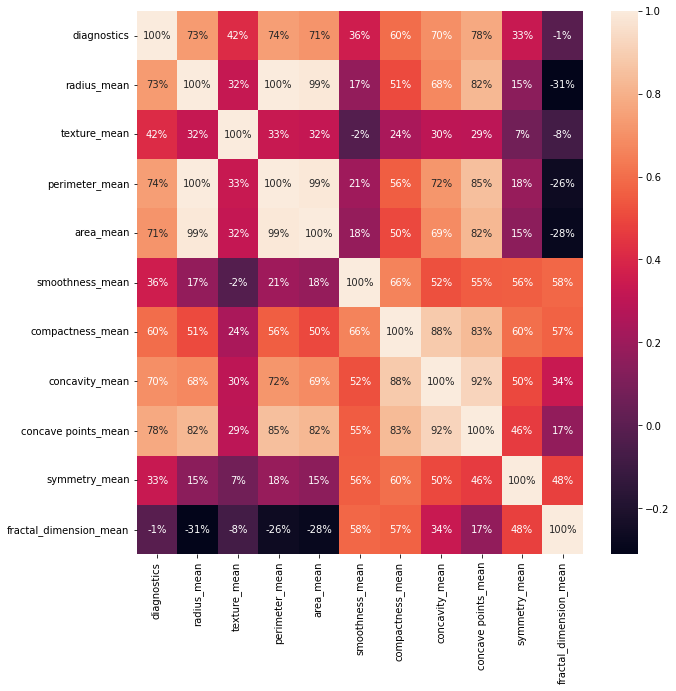
To have a better understanding of the data, we created a correlation table between all of the elements and visualized it with pair plotting and a heatmap. Since the size of the correlation table and plot was huge we have shown a part of their result.



**Table 1.** Table of correlation between features



**Figure 3.** Correlation depiction between 3 features of dataset using pairplot



**Figure 4.** Heatmap of correlation between features

**TECHNIQUES / ALGORITHMS:**

Below we have mentioned the basic working of the algorithms which are being implemented. Along with that, the hyperparameters of each algorithm are mentioned. Hyperparameters are the specific parameters which increase the performance of the respective algorithms.

1. **Random Forest Classifier:**

Working:

1. Selection of random samples from the dataset having n number of records.
2. Construction of individual decision trees for each record.
3. Output generated from each individual decision tree.
4. Majority Voting or Averaging is being done to select the final result.

Important Hyperparameters involved: n\_estimators, max\_features, min\_samples\_leaf

1. **AdaBoost:**

Working:

1. Equal weights are assigned to each record in the dataset.
2. On the subset of data, a model is trained. Predictions are made on the whole dataset using the same model.
3. Comparison of prediction and actual values to calculate the errors.
4. Higher weights assigned to the records which were predicted incorrectly, while constructing the next model.
5. In general, higher the error, more the weight assigned to that record, similarly lower the error, less weight is assigned.
6. Process repeated until the number of estimators doesn’t reach the maximum limit or error function doesn’t change.

Important Hyperparameters involved: n\_estimators, random\_state

1. **XGBoost:**

Working:

1. Predictions from many models are merged into a single forecast.
2. Then iteratively model each prediction depending on the mistake of its predecessor.
3. The predictors that perform better are given more weights.
4. A gradient descent approach is used in XGBoost to minimize the loss function.
5. Decision trees as weak predictors are being used.

Important Hyperparameters involved: booster, reg\_alpha and reg\_lambda, max\_depth, subsample, num\_estimators

1. **Support Vector Machine:**

Working:

1. Support Vectors are used.
2. Say in a dataset we have two classes, red and black. We want to classify the new datapoint as either red or black. We try to find decision boundaries for the same.
3. The best hyperplane is the one that has the greatest distance between both classes.
4. We find different hyperplanes which classify the labels in the best way.
5. Select the one which has a maximum margin or is farthest from the data points.

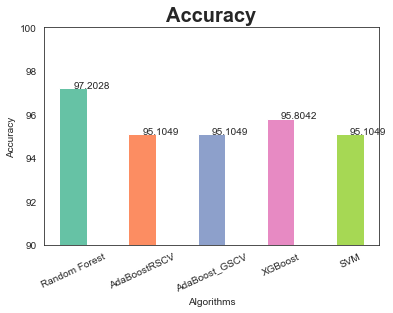
Important Hyperparameters involved: c, gamma, kernel

**RESULT**

**Accuracy:**

In statistical analysis accuracy is a parameter to measure the degree of closeness to the true value. In another way it is the ratio of true values.

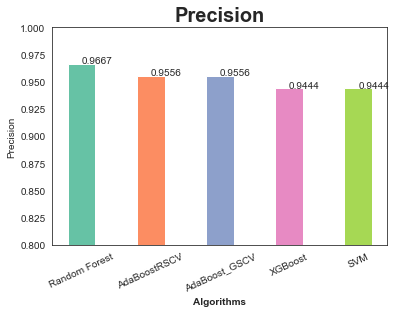
In this study accuracy of the all mentioned approaches are presented in **Figure 5** and summarized in **Table 2**



**Figure 5.** Accuracy of various ML classifiers in %

**Precision:**

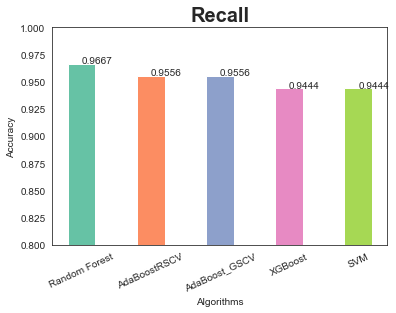
Here we are asking how well we can find all the positive results actually in your data. So you want to simply take the ratio of the actual positives your algorithm can find (TP), divided by all the ones present in your data set - which would be the ones you correctly found along with the ones you mistakenly predicted to be negative instead of positive. Result of the precision mentioned in **Figure 6** and summarized in **Table 2**

****

**Figure 6.** Precision of various ML classifiers

**Recall:**

Think of it as “uncovering” (which is a little like ‘recalling’) the positive cases that are hiding there in your data set. Recall is also sometimes called “sensitivity”, which is a bit better — you are measuring how well your algorithm can ‘sense’ or find all those positive results, (even though it may also be sensing some negative ones in the process). Result of the recall is mentioned in **Figure 7** and summarized in **Table 2**

****

**Figure 7.** Recall of various ML classifiers

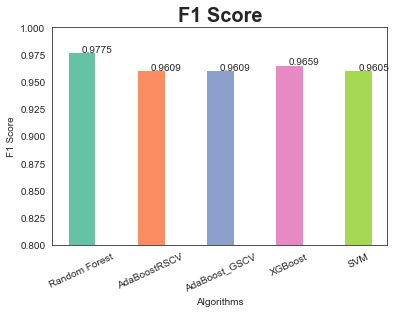
**F1-Score:**

F1 score is simply calculated as .

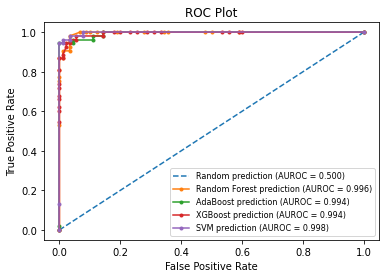
Result of the f1 score is mentioned in **Figure 8** and summarized in **Table 2**

**ROC-Curve:**

It is a graphical curve which is used to measures the performance of binary classifiers. It can be defined as the ratio of true positive vs. total actual positive. Value of the ROC curve of various classifiers is plotted in **Figure 9** and summarized in **Table 2.**



**Figure 8.** F1 Score of various ML classifiers



**Figure 9.** ROC plot of ML Classifiers

| **S.No** | **Algorithms** | **Accuracy** | **Precision** | **Recall** | **F1-Score** | **ROC-Curve** |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | Random Forest | 97.202 | 0.967 | 0.967 | 0.978 | 0.996 |
| 2 | AdaBoost | 95.804 | 0.956 | 0.956 | 0.961 | 0.994 |
| 4 | XGBoost | 95.804 | 0.945 | 0.945 | 0.966 | 0.994 |
| 5 | SVM | 95.105 | 0.9445 | 0.945 | 0.960 | 0.998 |

**Table 2.** Summary of Predicted values of the implemented ML Classifiers

**COMPARISON AND ANALYSIS**

In training a Machine Learning model there are two important things by which we evaluate whether our model is good enough or not, they are precision and F1-score. The precision tells us that how precise our model is i.e. how of total observed positive result, how many of them was correct. F1 score is more useful than accuracy especially if we are having uneven class distribution. So In different-different research papers we have found different methodology for extraction of tumor dara for detection of cancer, different parameters which they focused on and different algorithms they have used to analyze the data and classify the cell in benign and malignant. With the difference in methodology the results were different, so our aim was to choose the best algorithm in terms of precision and F1-score. The findings of different research papers are: 92% of F1-score[2], 96.85% of accuracy[4] etc. In our model we not only compared all algorithms and chosen the best one, but also we tried to increase the accuracy of all those algorithms with help of different techniques such as boosting, bagging and SVM classifier. Hence we concluded our research with highest accuracy 97.202% as of Random Forest with F1-score 0.978.

**CONCLUSION**

With 99% of the cases among women, Breast Cancer is the second leading cause of cancer death among women. Every year millions of womens are diagnosed with cancer. Risk of developing cancer increases with increase in age. Being genetically inherited the risk of getting breast cancer increases vastly with the family history of breast cancer. Even being cancer, the symptoms may vary among the women, and in some cases the symptoms are not visible at all. Thus, early detection is key as it greatly increases the survival rates. With early detection, the five- year survival rate for breast cancer is nearly 100%.

Thus, we utilized the ensemble learning algorithm which greatly boosted the performance of the algorithm. High accuracy of the algorithm helps in correct diagnosis of the patient which greatly reduces the time of biological analysis, machinery involved and financial investment of the patient.

Because this is an algorithm-driven procedure, a diagnosis may be made with only one click and no professional assistance is necessary. This could be beneficial in countries in southern Africa where medical support is scarce. With a maximum accuracy of 97.2 percent, our model can prevent 97 out of every 100 women from breast cancer, and can give more life to live with their families.

**FUTURE SCOPE**

Till now we had done our preliminary research on the numerical dataset part and in future we will be focusing on the Image dataset. The numerical model we have implemented provides good performance results in terms of performance metrics, however the model trained via image dataset will unlock the potential benefits. For example we are planning to train the model on an image dataset using CNN model and will deploy it on an application, using which one can determine whether the cell in the image is benign or malignant. To do this we won't be requiring a professional specialist laboratorist, because the classification which was performed by them, now can be done easily and quickly via this model, hence will reduce the cost, the time taken during the process and human errors.

**REFERENCES**

1. Malvia S, Bagadi SA, Dubey US, Saxena S. Epidemiology of breast cancer in Indian women. Asia‐Pacific Journal of Clinical Oncology. 2017 Aug;13(4):289-95.
2. Ray R, Abdullah AA, Mallick DK, Dash SR. Classification of benign and malignant breast cancer using supervised machine learning algorithms based on image and numeric datasets. In Journal of Physics: Conference Series 2019 Nov 1 (Vol. 1372, No. 1, p. 012062). IOP Publishing.
3. Yedjou CG, Tchounwou SS, Aló RA, Elhag R, Mochona B, Latinwo L. Application of Machine Learning Algorithms in Breast Cancer Diagnosis and Classification. International journal of science academic research. 2021 Jan;2(1):3081.
4. Akbugday B. Classification of breast cancer data using machine learning algorithms. In2019 Medical technologies congress (TIPTEKNO) 2019 Oct 3 (pp. 1-4). IEEE.
5. Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. Journal of pathology informatics. 2016;7.
6. Asri H, Mousannif H, Al Moatassime H, Noel T. Using machine learning algorithms for breast cancer risk prediction and diagnosis. Procedia Computer Science. 2016 Jan 1;83:1064-9.
7. Cruz JA, Wishart DS. Applications of machine learning in cancer prediction and prognosis. Cancer informatics. 2006 Jan;2:117693510600200030.
8. Wolberg WH, Mangasarian O, Aha DW. Breast cancer wisconsin dataset.

**Source code/hardware description/equipment description (As Annexure)**

Attached along with the Final Report submission