International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 03 | Mar 2022 www.irjet.net p-ISSN: 2395-0072

Machine Learning Aided Breast Cancer Classification

Sohom Sen[1], Ankit Ghosh[2]

[1]Student, Department of Computer Science and Engineering, Heritage Institute of Technology, Kolkata, India
[2]School of Nuclear Studies and Application, Jadavpur University, Kolkata, India

---

Abstract - *One of the most common disease a women faces is Breast Cancer. Majority of the breast lumps or abnormal growth are benign i.e. not malignant. Non cancer (benign) breast lumps are not life-threatening and does not spread outside the breast area. But these benign lumps can agitate the growth of malignant cells in the long run, leading to Breast Cancer. The malignant cells can start spreading from different parts of the breast, such as Lobules (causing Lobular Cancer), Ducts (causing Ductal Cancer), Nipple (causing Paget disease of the breast), Stroma (causing Phyllodes tumour) and Lymph Vessels (causing Angiosarcoma). Detection of Breast Cancer thus is a vital responsibility of Doctors. Machine Learning (ML) and Artificial Intelligence (AI) can contribute significantly in this field by speeding up the diagnosing time and reducing human-errors. In this research-paper we have identified the problem as a Binary Classification Model and implemented 8ML Algorithms namely, Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Logistic Regression (LR), Decision Tree Classifier (DTC), Random Forest Classifier (RFC), XGBoost Classifier (XGBC), Gradient Boosting Classifier (GBC) and Naive Bayes (NB) to automate the detection of Breast Cancer. The Dataset consists of all the necessary parameters listed by American Medical Association (AMA) required for detection. Finally a performance comparative studyis conducted to rank the ML Algorithms based upon the following metrics: accuracy, precision, recall-value, F1-score, AUC-ROC Curve and AUC-PR Curve. After conducting the performance analysis, XGBoost Classifier and Gradient Boosting Classifier surpassed other implemented ML Algorithms with an accuracy of 98.84%, Precision of 0.9688, Recall Value of 1.00 and F1-Score of 0.9841.*

Key Words: **American Medical Association, Artificial Intelligence, Binary Classification Model, Breast Cancer, Machine Learning, Performance Comparative Analysis**

1. INTRODUCTION

On close inspection, it is revealed, the cardinal cause leading to false diagnosis of an ailment is human-error. A Doctor's decision making process is impacted by patient's family history, overlapping symptoms, incorrect medical reports and other influencing factors. The best solution to reduce the risk of false diagnosis is Algorithm based Automation [1]. An Algorithm is unbiased unlike us, Humans.

Over the past decade, Artificial Intelligence (AI) and Machine Learning (ML) have played game-changing roles in the Medical Field [2], [3]. Not only they have helped in near accurate diagnosing of an ailment, but also play a vital role in R&D Department of Drug Manufacturing. Based on a person's health condition, AI and ML are able to suggest personalised medicine routine with an accuracy of over 95%. With the development of algorithms, we can minimise the inaccuracy to a great extent.

The paper focuses on implementation of 8ML Algorithms to automate the process of detecting Breast Cancer [4], based on given medical attributes listed by American Medical Association (AMA).

2. RELATED WORK

Bayrak et al. [5] presented two of the most admired ML Algorithms for the classification of Breast Cancer from Wisconsin Breast Cancer dataset. Support Vector Machine came up with the highest accuracy when evaluated against selected parameters.

Teixeira et al. [6] in his research paper deployed 6ML Algorithms including Deep Neural Network (DNN) to classify Breast tumour into benign and malignant. His best results happen to emerge from DNN, with an accuracy of 92%.
Kolay and Erdoğmuş [7] worked with the breast cancer data obtained from University of California-Irvine (UCI), using the WEKA data mining software to deploy ML Algorithms. They came up with accuracy ranging from 45% to 79% across all the deployed Algorithms.

Bektas and Babur [8] deployed multiple ML Algorithms for the diagnosis of Breast Cancer. They identified the active genes present in Breast Cancer by utilising the attribute selection method. 90.72% is the highest accuracy they were able to achieve from the deployed models.

3. DATASET

The Dataset constitutes of 32 columns. The key attributes required for detection as stated by AMA are as follows:

- Radius : it is the mean of distances from the centre to the points on the circumference
- Texture : the standard deviation of the grey-scale values
- Perimeter : Circumference of Tumour
- Area : Area of the Tumour

- Smoothness : it is the local deviation in radius
- Compactness : defined as [(perimeter^2)/area - 1]
- Concavity : the gravity of concave portions on the silhouette
- Concave points : number of concave portions on the silhouette
- Symmetry
- Fractal dimension : its a characteristic parameter used to describe the irregular extent of coastline

The arithmetic mean, standard deviation and the maximum values of each key attribute was calculated resulting in a total of 30 working attributes.
The dataset that has been used comprises of 569 observations. There are 357 records of those patients who have been diagnosed with Benign Breast Tumour while the remaining 212 observations are those of patients with Malignant Breast Tumour.
In future, with the influx of more data, our implemented models can be tested for a better accuracy.
Despite using a relatively small dataset, the implemented models have generated appreciable results. However, their performances are likely to improve while using a larger dataset

4. STATE-OF-THE-ART ML ALGORITHMS

Without complicating the classifying parameters, we can broadly categorise ML Algorithms into three classes namely, Supervised Learning, Unsupervised Learning and Reinforcement Learning. In this paper, we are majorly dealing with Supervised Learning Algorithms. In case of Supervised Learning, the algorithm is provided with labeled input data and its function is to map the input data with known input-output pairs.

Brief illustrations of the implemented ML Algorithms have been discussed below.

- Support Vector Machine (SVM) - Belongs to the class of Supervised Learning. To put forward in a palatable form, SVM creates a hyperplane (simply a divider) that segregates two (Binary Classification) or more (Multi-Class Classification) unique classes.

- K-Nearest Neighbour (KNN)- Belongs to the class of Supervised Learning. The algorithm can be used for both Binomial and Multi-Class Classification. KNN selects out a particular area from the classes near the testing value and uses a voting mechanism to predict the output of the testing value.

- Logistic Regression (LR) - Belongs to the class of Supervised Learning. Basically it is a statistical model that makes use of a Logistic function to evaluate the probability of a class. LR is quite efficient in case of Binary Classification.

- Decision Tree Classifier (DTC) - Belongs to the class of Supervised Learning. The model can be used for Binary Classification. It is a Tree-Structured Model, where every parent node depicts the attributes of the dataset. The model learns from basic decision rules deduced from the attributes and predicts the outcome.

- Random Forest Classifier (RFC) - Belongs the class of Ensemble Learning (sub-class of Supervised Learning). It is one of the Bagging Techniques that makes use of numerous Decision Tree Classifiers to predict the outcome of a testing data point.

- XGBoost Classifier (XGBC) - Belongs to the class of Boosting Techniques (sub-class of Ensemble Learning). The Boosting Technique is a cumulative process, where the low accuracy predictors are boosted into new models with higher accuracy and precision. It makes use of the Gradient Boosting method also.

- Gradient Boosting Classifier (GBC) - Belongs to the class of Boosting Techniques (sub-class of Ensemble Learning). This algorithm has an iterative learning approach. They combine every weak sub-model into a stronger unit, yielding a higher accuracy.

- Naive Bayes (NB) - Belongs to the class of Supervised Learning. The Algorithm is based upon Bayes Theorem and it is primarily used for Binary Classification.

5. METHODOLOGY

Steps involved for the implementation and evaluation of the ML Algorithms are displayed in the form a Flow-Chart in Fig-1 for better visualisation and understanding.

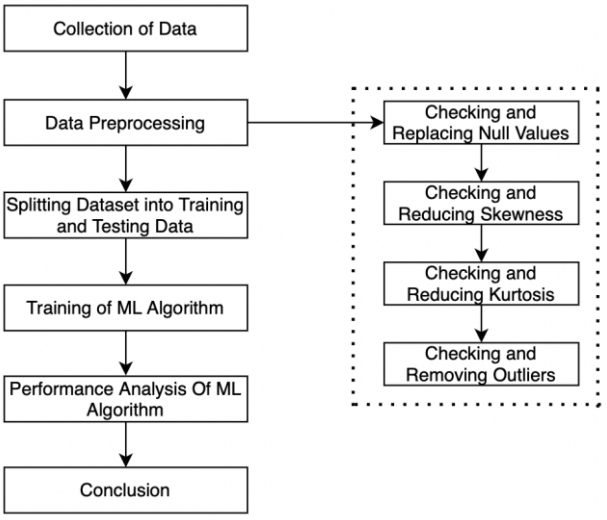

Fig - 1: Flowchart

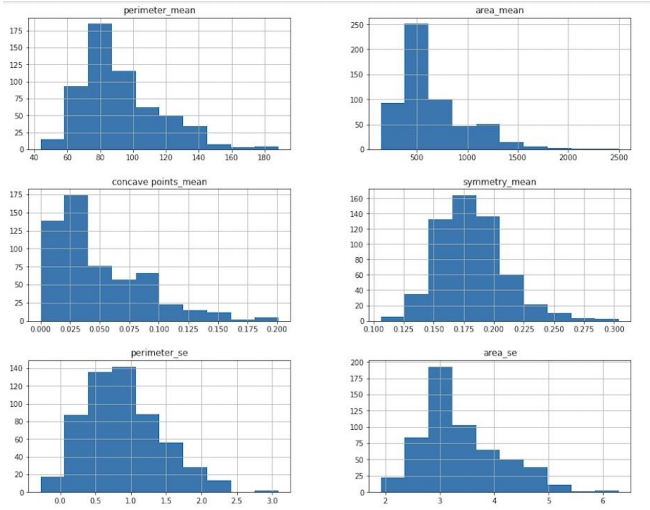

Fig - 2 : Distribution Plot of few attributes of the Dataset

From Fig-1, all the steps from Data Collection to Evaluation of the ML Model is shown chronologically. The steps followed in Data Preprocessing are also highlighted for clarity.

The steps adopted for data preprocessing are described in the next section.

5.1 Data Preprocessing

Being a Binary Classification problem, we have mapped Malignant Tumour (M) to 0 and Benign Tumour (B) to 1, for the sake of simplicity.
The major potholes in the journey of feeding the processed dataset into the ML Algorithms are Null Values. For sufficiently large datasets, all the Null Value entries are dropped, leading to a negligibly small dataset in comparison to the original one to be worked upon. For a relatively small dataset, the Null Values are either replaced by the arithmetic mean of that attribute (if feasible) or the median. In our case, we have replaced each Null Value with its arithmetic mean [9].

The Distribution Plot of each attribute can contribute significantly in the accuracy of the Generated Function. The attributes must be Normally Distributed (Gaussian Curve or Bell Curve). The Distribution Plots for few of the attributes is shown in Fig-2.

In Fig-2, it is noticeable, the majority of the attributes (perimeter_mean, perimeter_se, symmetry_mean and area_se) are Normally Distributed [10] and the remaining Concave points_mean and area_mean are not ideally what we call Normal Distribution.
Few of the attribute's distribution are asymmetric from their mean position or distorted in some form. These distributions are labeled as Skewed Distributions [11] and can be resolved using little tricks such as taking the logarithmic values (for all entries greater than 0) or taking the square-root values (for all non-negative entries) or simply taking the square of each value.

5.2 Splitting the Processed Dataset

After the Data Preprocessing stage, the Dataset is divided into Training Set and Testing Set. We have divided the dataset in the ratio 85:15 which portrays the Training Set to constitute of 483 cases and the Testing Set to constitute of 86 cases.

5.3 Fitting Training Set into ML Algorithms

The Training Set is fitted into an ML Algorithm or simply a Learning Algorithm for training. After the ML Algorithm is trained, it is prepared to take inputs and produce an Output from what it has learned after the learning/training process and this is shown in Fig-3 for a better understanding.

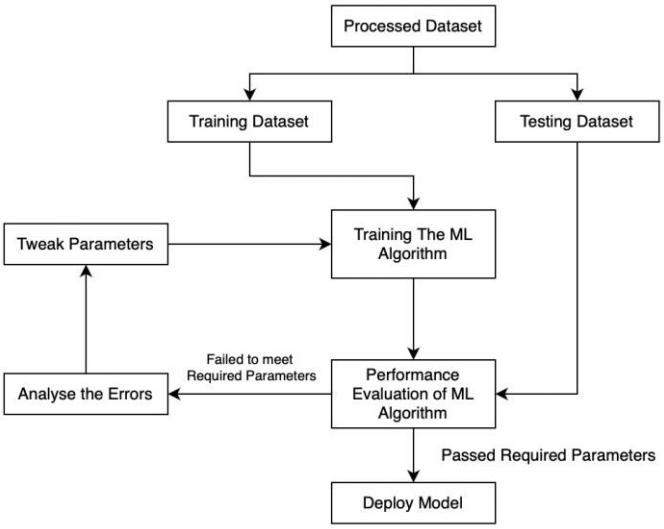

Fig - 3: ML Algorithm Working Principle

In Fig-3, the detailed mechanism is shown from the point of fitting the training set to the point of predicting the output and evaluation.

5.4 Evaluating an ML Algorithm

Each value from the Testing Dataset yields an output after the training process of the Algorithm. Every output is then weighed-up against the Testing input-output pair. The model with highest number of correct predictions is broadcasted as the best-suited model for disentangling that particular problem.

For a clear view, we need to first have an idea about Confusion Matrix [12].

A confusion matrix is 2x2 matrix, which constitutes of 4 unique values labelled as True Positive (TP), False Positive (FP), False Negative (TN) and True Negative (FN). The terms are self explanatory.

True Positive (TP)	False Positive (FP)
False Negative (FN)	True Negative (FN)

Fig - 4: Confusion Matrix

Fig-4 gives us a brief idea about how a confusion matrix looks like. We evaluate Accuracy, Precision, Recall and F1-Score from the confusion matrix.

6. EXPERIMENTAL RESULTS

We have implemented 8 ML Algorithms for this Binary Classification Model and the results are portrayed below in Table-1.

ML Algorithm	Accuracy	Precision	Recall	F1-Score
SVM	0.9767	0.9375	1.00	0.9677
KNN	0.9535	0.875	1.00	0.9333
LR	0.9651	0.9062	1.00	0.9508
DTC	0.9418	0.875	0.9655	0.9180
RFC	0.9538	0.9063	0.9667	0.9335
XGBC	0.9884	0.9688	1.00	0.9841
GBC	0.9884	0.9688	1.00	0.9841
NB	0.9186	0.8438	0.9310	0.8821

Table-1 : Experimental results obtained

Table-1 gives us all the parameters needed to evaluate a particular ML Algorithm. From a bird's eye view, we can point out, XGBoost Classifier and Gradient Boosting Classifier possess the highest accuracy of 98.94%.

ML Algorithm	TP	FP	FN	TN
SVM	31	1	1	53
KNN	28	4	0	54
LR	30	2	1	53
DTC	28	4	1	53
RFC	29	3	1	53
XGBC	31	1	0	54
GBC	31	1	0	54
NB	28	4	3	51

Table-2 : Confusion Matrix of Implemented Algorithms

From Table-2, Support Vector Machine, XGBoost Classifier and Gradient Boosting have the maximum True Positive Value of 31 and minimum False Positive Value of 1, while XGBoost Classifier and Gradient Boosting Classifier have a maximum True Negative Value of 54.

6.1 AUC-ROC Curves of Implemented ML Algorithms

Area Under Cover - Receiver Operating Characteristic (AUC-ROC) Curve [13] is basically a computing parameter which indicates how well an ML Algorithm can differentiate between two classes (in our context Class 0-Malignant and Class 1-Benign). The AUC-ROC Curves of the implemented ML Algorithms are shown in Fig-5, 6, 7, 8, 9, 10, 11 and 12.

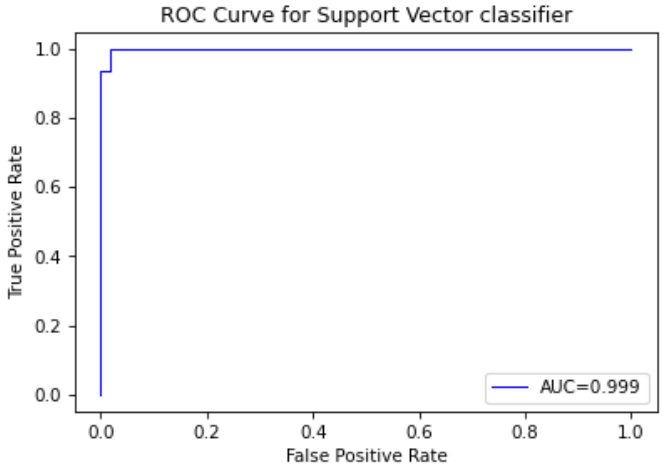

Fig - 5: AUC-ROC Curve of SVM

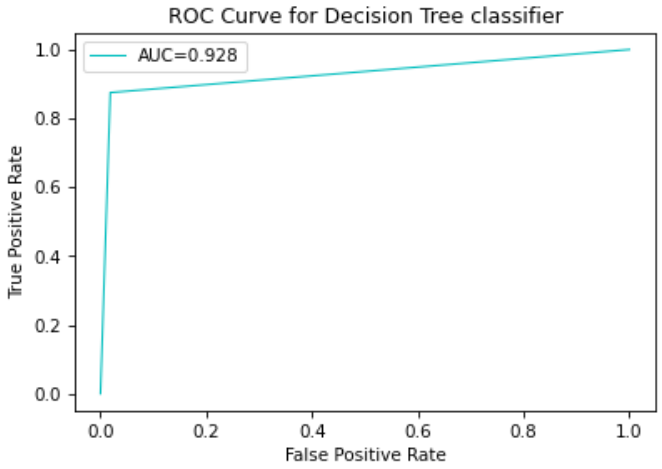

Fig - 8: AUC-ROC Curve of DTC

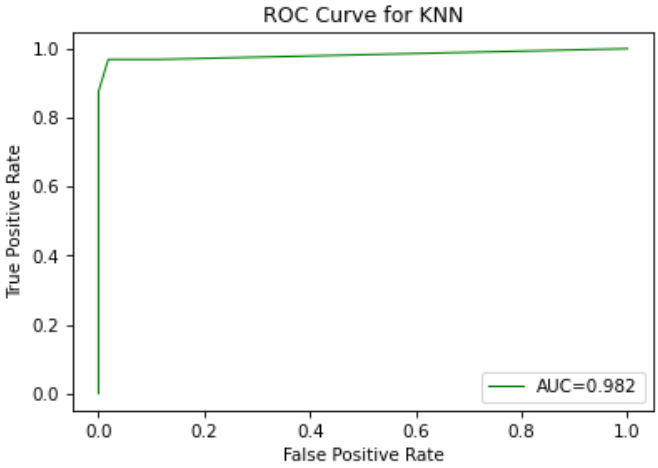

Fig - 6: AUC-ROC Curve of KNN

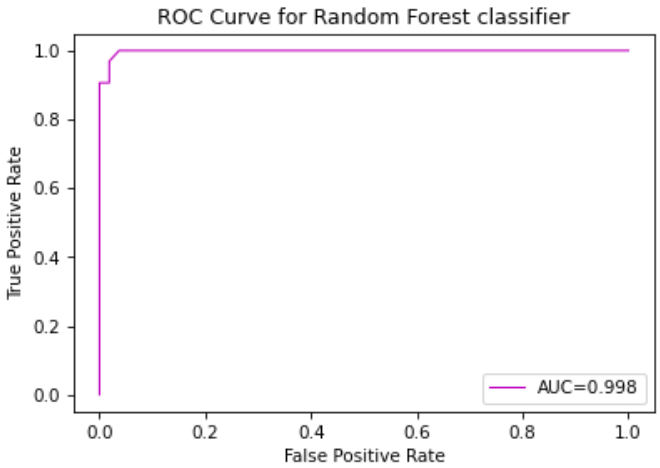

Fig - 9: AUC-ROC Curve of RFC

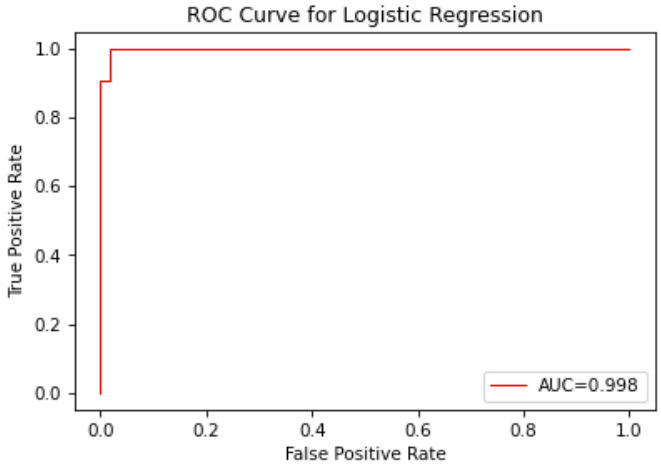

Fig - 7: AUC-ROC Curve of LR

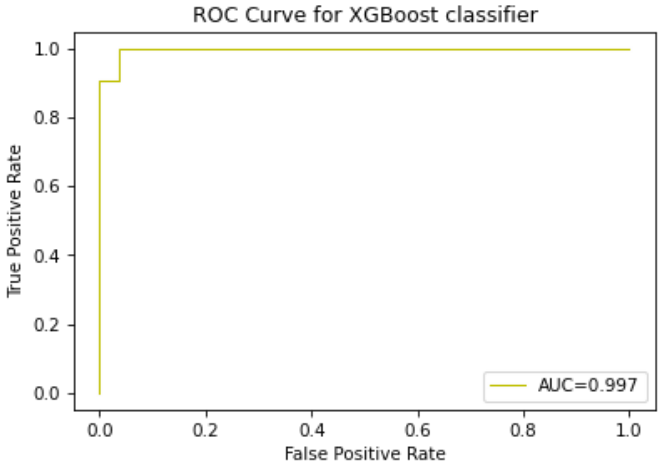

Fig - 10: AUC-ROC Curve of XGBC

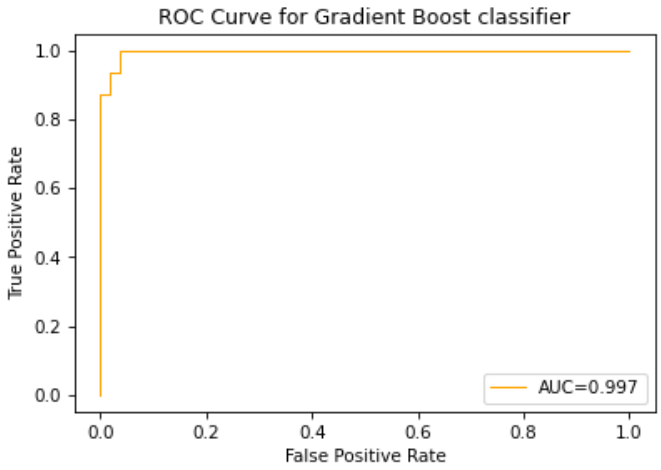

Fig - 11: AUC-ROC Curve of GBC

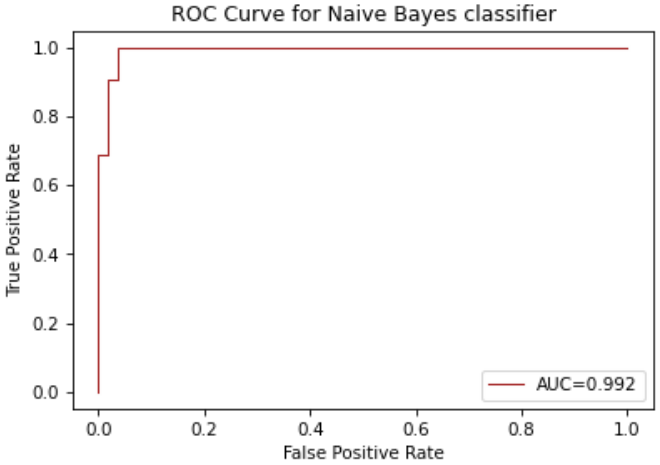

Fig - 12: AUC-ROC Curve of NB

The AUC-ROC values of each ML Algorithm is summarised in Table-3 for a better understanding.

ML Algorithm	AUC-ROC Value
SVM	0.999
KNN	0.982
LR	0.998
DTC	0.928
RFC	0.998
XGBC	0.999
GBC	0.997
NB	0.992

Table-3 : AUC-ROC values of implemented ML Algorithms

At a glance, XGBoost Classifier and Support Vector Machine came up with the highest value of 0.999, followed by Logistic Regression and Random Forest Classifier with a value of 0.998.

6.2 AUC-PR Curves of Implemented ML Algorithms

In simple words, AUC-PR signifies Area Under the Precision-Recall Curve. This parameter is used to differentiate the ML Algorithms based upon its performance in classifying a given duty. The AUC-PR Curves of the implemented ML Algorithms are shown in Fig-13, 14, 15, 16, 17, 18 , 19 and 20.

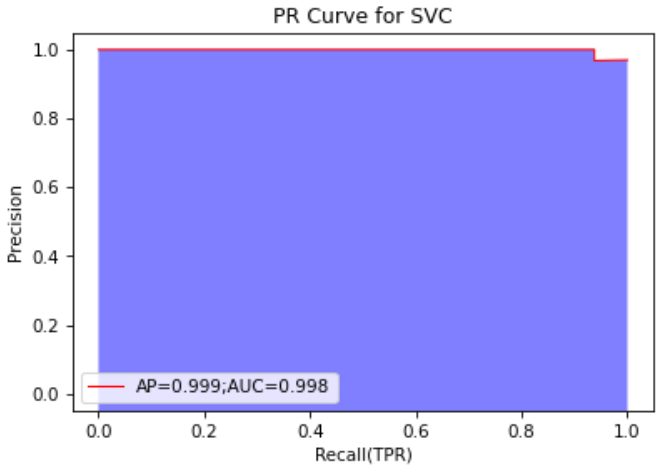

Fig - 13: AUC-PR Curve of SVC

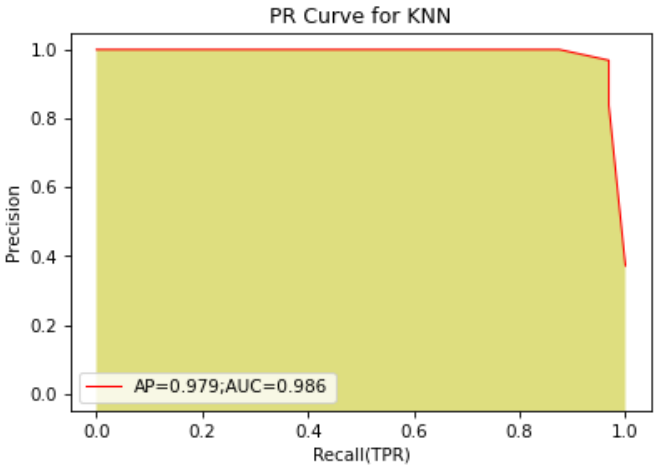

Fig - 14: AUC-PR Curve of KNN

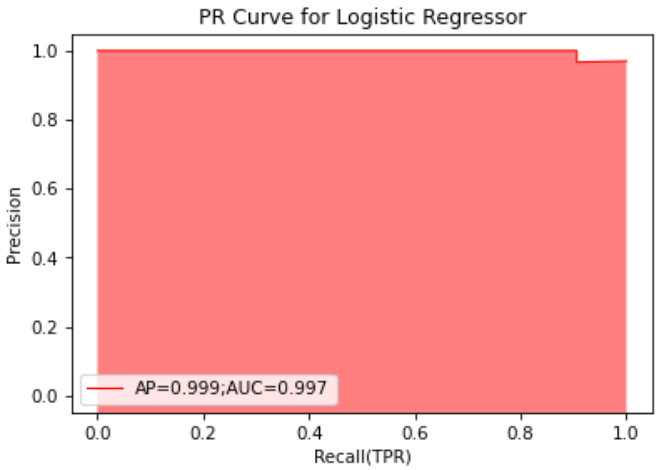

Fig - 15: AUC-PR Curve of LR

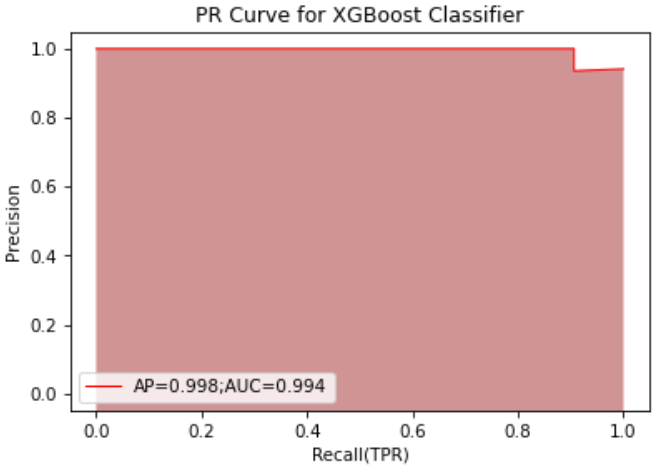

Fig - 18: AUC-PR Curve of XGBC

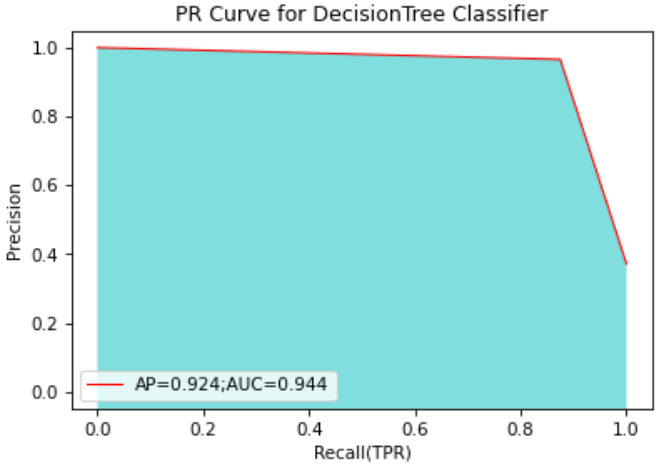

Fig - 16: AUC-PR Curve of DTC

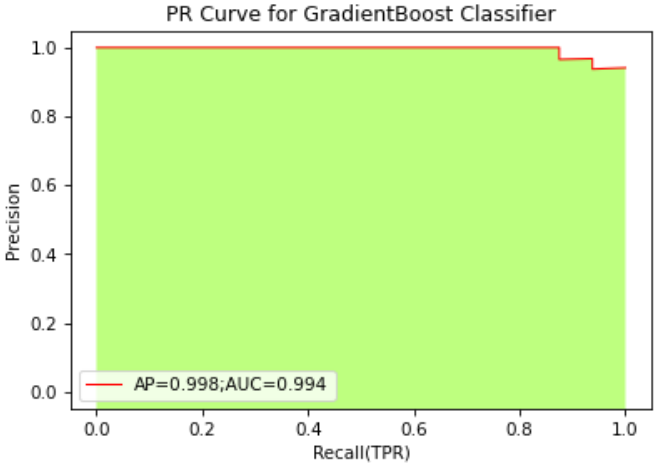

Fig - 19: AUC-PR Curve of GBC

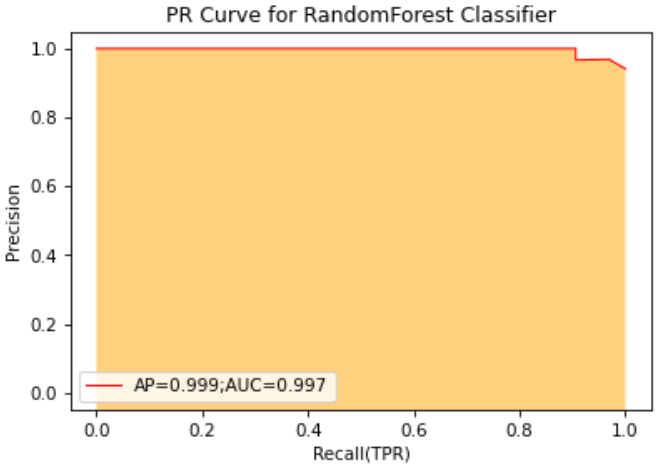

Fig - 17: AUC-PR Curve of RFC

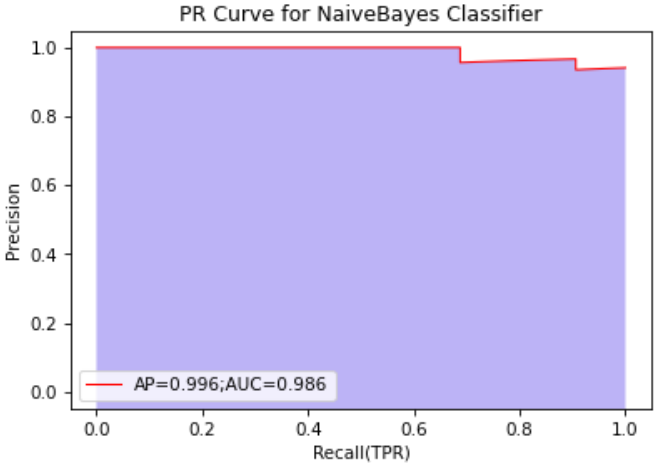

Fig - 20: AUC-PR Curve of NB

The AUC-PR values of each implemented ML Algorithms is summarised in Table-4 for a better understanding.

ML Algorithm	AUC-PR Value
SVM	0.999
KNN	0.979
LR	0.999
DTC	0.924
RFC	0.999
XGBC	0.999
GBC	0.998
NB	0.996

Table-4 : AUC-PR values of implemented ML Algorithms

From Table-4, it is evident 4 of the ML Algorithms namely, Support Vector Machine, Logistic Regression, Random Forest Classifier and XGBoost Classifier got values as high as 0.999 .

6.3 Performance Comparative Study of the different ML Algorithms

The four parameters that we have appointed for the ML Algorithms to be evaluated against are Accuracy, Precision, Recall, F1-Score, AUC-ROC Curve and AUC-PR Curve. To add a little more depth to these parameters, they are defined as follows [14], [15]:

- Accuracy - It is defined as the ratio of total number of correct outputs to the total number of outputs (including both correct and wrong outputs).

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

- Precision - It is defined as how palatable the measurements are even if they are not accurate.

Formula = (TP)/(TP+FP)

- Recall - It is defined as the measure of the Algorithm's ability to classify a Binary Classification or Multi-Class Classification problem.

Formula = (TP)/(TP+FN)

- F1-Score - It is defined as the weighted harmonic mean of Precision and Recall value.

Formula = (2*Recall*Precision)/(Recall+Precision)

- AUC-ROC - It is basically a computing parameter which indicates how well an ML Algorithm can differentiate between two classes
- AUC-PR - It signifies Area Under the Precision-Recall Curve.

ML Algorithm	Accuracy	Precision	Recall	F1-Score	AUC-ROC	AUC-PR
SVM	0.9767	0.9375	1.00	0.9677	0.999	0.999
KNN	0.9535	0.875	1.00	0.9333	0.982	0.979
LR	0.9651	0.9062	1.00	0.9508	0.998	0.999
DTC	0.9418	0.875	0.9655	0.9180	0.928	0.924
RFC	0.9538	0.9063	0.9667	0.9335	0.998	0.999
XGBC	0.9884	0.9688	1.00	0.9841	0.999	0.999
GBC	0.9884	0.9688	1.00	0.9841	0.997	0.998
NB	0.9186	0.8438	0.9310	0.8821	0.992	0.996

Table-5 : Performance Analysis of the Implemented Algorithms

From Table-5 it is noticeable that XGBoost Classifier and Gradient Boosting Classifier surpasses all the other Algorithms with an Accuracy of 98.84%, Precision value of 0.9688, Recall value of 1.00, F1-Score of 0.9841, AUC-ROC value of 0.999 and AUC-PR Value of 0.999 and 0.998 respectively.

7. CONCLUSIONS

Now we have reached a stage to summarise our entire presentation.

Automation is the key to future. Efficient and accurate diagnosis of an ailment can be achieved with the help of ML Algorithms as they are devoid of all human emotions and works solely for accuracy. Our Performance Analysis revealed XGBoost Classifier and Gradient Boosting Classifier to bear the highest accuracy with minimum False-Positive Cases.

Till now we are only dealing with numeric values or 2D image scans for the detection of Malignant Tumours using Automation. We can extend the field by using 3D scans of infected areas and feed the data directly into our ML Algorithms to obtain desired results. With each day passing, ML and AI are evolving to handle complex data inputs and yield accurate results in minimum possible real-time.

REFERENCES

[1] Ankit Ghosh, Purbita Kole and Alok Kole, "Automatic Identification ofCovid-19 from Chest X-ray Images using Enhanced Machine LearningTechniques", International Research Journal of Engineering andTechnology (IRJET),vol.8, issue.9, no.115, pp.765-772, 2021.

[2] M. Gupta and B. Gupta, "An Ensemble Model for Breast Cancer Prediction Using Sequential Least Squares Programming Method (SLSQP)," 2018 Eleventh International Conference on Contemporary Computing (IC3), 2018, pp. 1-3.

[3] GHOSH, ANKIT; KOLE, ALOK (2021): A Comparative Study of Enhanced Machine Learning Algorithms for Brain Tumor Detection and Classification. TechRxiv. Preprint.

[4] S. Ara, A. Das and A. Dey, "Malignant and Benign Breast Cancer Classification using Machine Learning Algorithms," 2021 International Conference on Artificial Intelligence (ICAI), 2021, pp. 97-101.

[5] E. A. Bayrak, P. Kırcı and T. Ensari, "Comparison of Machine Learning Methods for Breast Cancer Diagnosis," 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT), 2019, pp. 1-3.

[6] F. Teixeira, J. L. Z. Montenegro, C. A. da Costa and R. da Rosa Righi, "An Analysis of Machine Learning Classifiers in Breast Cancer Diagnosis," 2019 XLV Latin American Computing Conference (CLEI), 2019, pp. 1-10.

[7] N. Kolay and P. Erdoğmuş, "The classification of breast cancer with Machine Learning Techniques," 2016 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), 2016, pp. 1-4.

[8] B. Bektaş and S. Babur, "Machine learning based performance development for diagnosis of breast cancer," 2016 Medical Technologies National Congress (TIPTEKNO), 2016, pp. 1-4.

[9] S. Samsani, "An RST based efficient preprocessing technique for handling inconsistent data," 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 2016, pp. 1-8.

[10] A. Ahmad, I. Azmira, S. Ahmad and Nur Ilyana Anwar Apandi, "Statistical distributions of load profiling data," 2012 IEEE International Power Engineering and Optimization Conference Melaka, Malaysia, 2012, pp. 199-203.

[11] C. Ozansoy, "Performance Analysis of Skewness Methods for Asymmetry Detection in High Impedance Faults," in IEEE Transactions on Power Systems, vol. 35, no. 6, pp. 4952-4955, Nov. 2020.

[12] F. J. Ariza-Lopez, J. Rodriguez-Avi and M. V. Alba-Fernandez, "Complete Control of an Observed Confusion Matrix," IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium, 2018, pp. 1222-1225.

[13] S. A. Khan and Z. Ali Rana, "Evaluating Performance of Software Defect Prediction Models Using Area Under Precision-Recall Curve (AUC-PR)," 2019 2nd International Conference on Advancements in Computational Sciences (ICACS), 2019, pp. 1-6.

[14] M. Junker, R. Hoch and A. Dengel, "On the evaluation of document analysis components by recall, precision, and accuracy," Proceedings of the Fifth International Conference on Document Analysis and Recognition. ICDAR '99 (Cat. No.PR00318), 1999, pp. 713-716.

[15] D. Tanouz, R. R. Subramanian, D. Eswar, G. V. P. Reddy, A. R. Kumar and C. V. N. M. Praneeth, "Credit Card Fraud Detection Using Machine Learning," 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), 2021, pp. 967-972.