Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 10, October-2023

A Precise Detection of Breast Cancer Using Machine Learning Model

Sumit, Tanisha Aggarwal, Er. Kirat Kaur, Nitin
Computer Science and Engineering, Chandigarh University, Gharuan , Punjab , India

***Abstract*— Breast cancer is one of the most life-threatening diseases affecting mainly, women. For better patient survival rates and successful treatments, its early prediction is important. Nowadays, the integration of Machine Learning algorithms in medical research has shown us very good results in enhancing the accuracy of breast cancer detection. This abstract presents an overview of how we have used breast cancer dataset from Kaggle and the accuracy of six ML classification models including Bagging, random forest, KNearest Neighbor, Adaboost, Gradient boost, Multilayer Perceptron model with maximum 99% accuracy. The study primarily focuses on supervised learning algorithm for analyzing parameters of breast cancer like tumor radius, smoothness, concavity, surface texture etc., which aid in the early diagnosis of breast cancer (whether it is benign or malignant). Along with this past year's data (2016-2021) is analyzed, which helps us mainly in knowing death rates of persons taking chemotherapy. The supervised learning approaches involve the utilization of labeled datasets to train classifiers that can distinguish between malignant and benign breast tissue. Extracted features serve as valuable inputs to the classifiers, which can then predict the pr esence of cancerous cells with high accuracy. The abstract concludes by discussing the potential challenges and future directions in breast cancer detection using ML. These include the need for large, diverse, and annotated datasets, concerns related to model interpretability, and the importance of integrating ML models into the clinical workflow seamlessly. However, continued research, collaboration between medical experts and data scientists, and ethical considerations remain essential to unleash the full potential of ML in the fight against breast cancer.**

***Keywords*—** breast cancer, Machine Learning, artificial intelligence, classification, tumor detection, early diagnosis, model accuracy.

I. INTRODUCTION

Traditionally, breast cancer detection has relied on manual examination of mammograms by radiologists, which can be time-consuming and subject to human error. ML offers a transformative approach to enhance this process by enabling automated analysis of medical images. Medical imaging for breast cancer can provide noninvasive insights into the human body, aiding doctors in the diagnosis and treatment of the condition[1]. Deep Learning, a subset of ML, involves the creation of complex neural networks inspired by the human brain's architecture. These networks can learn intricate patterns and representations from vast amounts of data, making them highly adept at handling the intricacies of medical images and patient information. In the article, it is stated that a skilled doctor can achieve a cancer diagnosis accuracy of 79%, whereas the use of machine learning techniques improves accuracy to 91%[2]. Each year, over one million cases of breast cancer are detected, contributing significantly to the high annual mortality rate[3]. The National Cancer Institute reports that around one out of every eight women will experience the development of an invasive type of this cancer during their lifetime. Detecting cancer in its early stages greatly enhances the chances of successful treatment and recovery[4]. This research endeavors to explore the role of deep learning models in the early detection of breast cancer, focusing on their applications genomic data interpretation. By employing these advanced algorithms, researchers and medical professionals can extract nuanced insights from complex datasets, enabling a deeper understanding of breast cancer subtypes (benign or malignant). The proposed ML model used supervised learning algorithms by training the model with dataset taken from Kaggle with 32 features and 569 records using six different classification algorithms with random forest, Adaboost Classifier, Gradient boost classifier giving the highest training accuracy.

The paper is structured into several sections, where Section (II) describes the review of existing research on breast cancer. Section (III) describes the materials and methods used in the study, as well as preprocessing techniques used. Section (IV) describes the experimental results while presenting the accuracy of deep learning algorithms with other ML algorithms. The last section, Section (V) summarizes the conclusion the main findings of the study.

II. REVIEW OF EXISTING WORK

Many ML methods have been applied for early detection of breast cancer. Some of them are defined in this section.

Mediha [5] used ML techniques like Naïve Bayes, Decision Tree, Random Forest, and k-nearest neighbor for predicting breast cancer using parameters like age, body mass index, glucose, leukocytes count, volatile organic compounds, adiponectin, leptin and neutrophils. They stated that parameters used in the study can be used as a cheap and effective breast cancer predictor with accuracy over 90% using database formed with 8 biomarkers for prediction.

David A. Omondiagbe [6] investigates the use of Support Vector machines, Artificial neural networks and naïve bayes

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 10, October-2023

for breast cancer detection using ML. He analyzed that support vector machine works well with feature reduction with accuracy 98.82% and sensitivity 98.41% and Computer aided detection systems can aid in early diagnosis.

Noushaba Feroz [7] implemented K-nearest neighbor and random forest algorithms, achieving an accuracy rate of 97.14% when utilizing the Wisconsin Breast Cancer dataset. The authors underscore the significance of machine learning in the context of breast cancer detection. They also present a systematic review of recent and noteworthy research on accurate breast cancer detection, followed by a comparative analysis of the machine learning models discussed in these studies. Notably, both K-Nearest Neighbor and Random Forest demonstrated a high accuracy of 97.14% when applied to the original Wisconsin Breast Cancer Dataset.

Sara Noor Eldin [8] uses a deep learning approach i.e., convolutional neural networks achieved up to 92.5% accuracy in diagnosing breast cancer from biopsy microscopy images. In this research paper, the authors introduced a deep learning methodology for diagnosing breast cancer based on biopsy microscopy images. They employed various deep convolutional neural networks and noted that the implementation of data preprocessing techniques led to improvements in accuracy by 20%, 17%, and 6%, respectively.

Hajra Naveed Iqbal [9] used random forest and achieved the accuracy of 99.26% in detecting breast cancer using machine learning. A performance evaluation was carried out, assessing four different classifiers: Multilayer Perceptron (MLP), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Random Forest, using the Wisconsin Breast Cancer dataset. Notably, MLP exhibited the lowest accuracy, with a score of 94.07%.

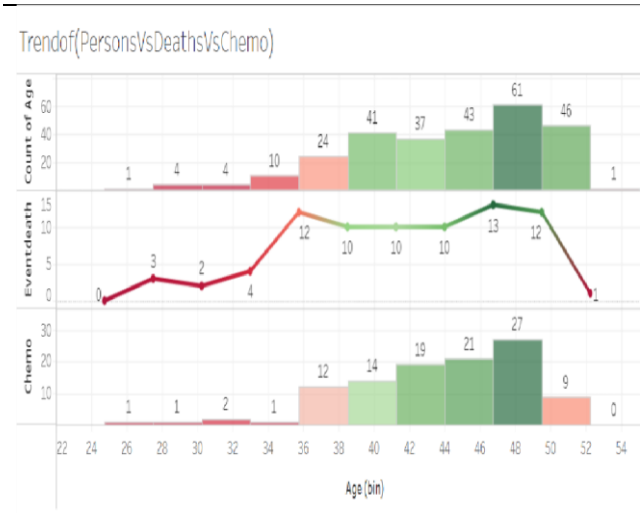
Pawan Kumar [10] compares the performance of support vector machines (SVM), random forest, and k-nearest neighbor (k-NN) algorithms for breast cancer detection. In this article, the authors conducted a comparative analysis of prominent machine learning techniques, evaluating their performance. The results revealed that k-NN achieved the highest accuracy, reaching 97.32%, when compared to SVM, RF, and SVM.

Rekh Ram Janghel [11] focuses their research work on using machine learning for diagnosis of breast cancer achieving results with an accuracy of 98%. He compared and analyzed 13 different ML models on the various measures and AdaBoost, logistic regression, and the 1-NN machine learning models demonstrate promising accuracy in conducting the experiment when compared to all other models.

Table 1: Review of Existing Work

Year	Author's Name	Model Used	Best Model	Accuracy
2021	Mediha Salic	Naïve Bayes, Decision tree, Random Forest, and K-Nearest Neighbor	Decision Tree Classifier	90%
2019	David A. Omondiagbe	SVM, Artificial neural networks, and naïve bayes	SVM	98.41%
2021	Noushaba Feroz	K-Nearest neighbor, and random forest	Random Forest	97.14%
2021	Sara Noor Eldin	Convolutional Neural Networks	CNN	92.5%
2020	Hajra Naveed Iqbal	MLP, SVM, KNN and Random Forest	Random Forest	99.26%
2021	Pawan Kumar	SVM, Random Forest and KNN	KNN	97.32%
2020	Rekh Ram Janghel	Adaboost, logistic regression and 1-NN	Adaboost, Logistic Regression and 1-NN	98%
Our Results				
2023	Sumit, Tanisha Aggarwal	Bagging, Random Forest, KNN, Adaboost, Gradient Boost, MLP	MLP Classifier	99%

Figure 1: Age wise no. of persons dying who have breast cancer vs taking chemotherapy



International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 10, October-2023

III. MATERIALS AND METHODS

In this research, we proposed a model for breast cancer detection using machine learning classifiers where we train the model by Kaggle dataset named as "Breast cancer dataset" which accounts for 25% of all cancer cases, and affected over 2.1 million people in 2015 alone. The six classifiers we have used are Bagging classifier, KNN classifier, Random Forest classifier, AdaBoost Classifier, Gradient Classifier, MLP Classifier. This dataset contains 31 features which helps in diagnosing the cancer whether it is benign or malignant.

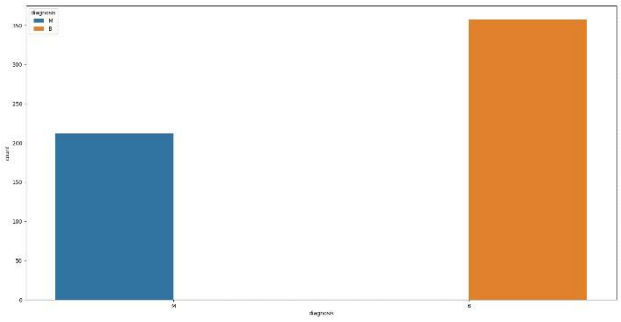


Figure 2: Counts of Benign and Malignant Cases

In fig 2. We can see that there are 212 malignant cases i.e., these patients have cancerous cells and 357 benign cases (non-cancerous cells).

IV. PROPOSED SYSTEM

The model proposed is firstly trained by the dataset using libraries like pandas, and for data visualization we have used matplotlib and seaborn and for training the machine learning models we have used sklearn python library.

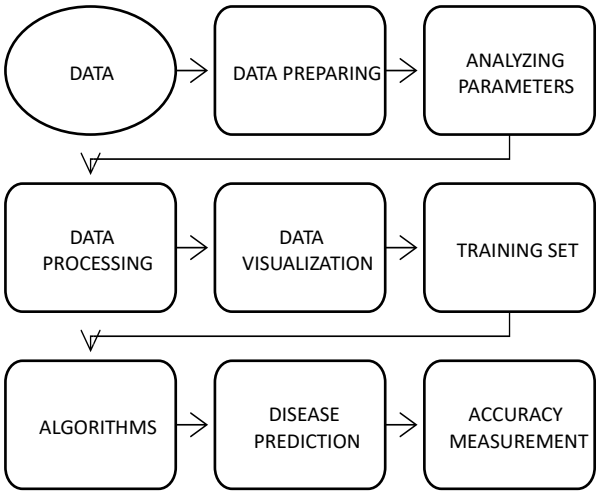


Figure 3: Flowchart of proposed Model

Step 1: The dataset is taken from Kaggle which contains the records of various patients and they're on time medical examination report.
Step 2: Data is prepared by performing data preprocessing and splitting the data into training and testing datasets.
Step 3: The parameters given in the dataset are analyzed and observed how they are affecting the cancer patients.
Step 4: After that, we have correlated the matrix of the parameter's for understanding the parameters in a better way. (Fig 4)
Step 5: We have fed training set to the model.
Step 6: The model is trained by using six machine learning algorithms and their accuracy is measured based on their confusion matrix generated.

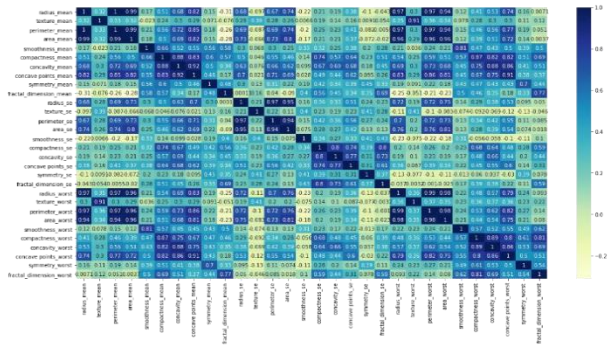


Figure 4: Correlation Matrix of parameters

The techniques (Classification Models) we have used to make predictions are described below:

A. Bagging Classifier

The bagging classifier is an ensemble machine learning technique that combines the prediction of multiple base classifiers to improve overall predicting performance. It was introduced by LEO Breiman in 1996. For this, the base classifiers are decision trees, random forests, SVM and neural networks. Once all the base classifiers are trained, the Bagging Classifier employs a voting mechanism for classification tasks and averaging for regression tasks to determine final predictions. In the context of classification, the final prediction is based on the class with the majority vote among all the base classifiers. For regression, the average of the predicted values from all base classifiers is taken as the final prediction. The benefits of using bagging classifier are reduced variance, improved accuracy, parallelizable.

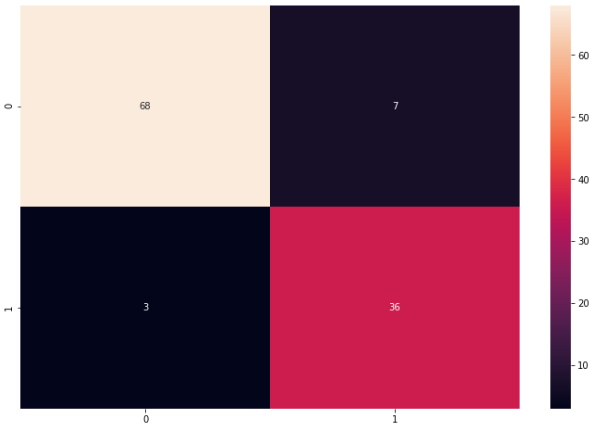


Figure 5: Confusion Matrix of Bagging Classifier

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 10, October-2023

In fig 5, we can see that the bagging classifiers generated 68 true negatives (truly benign), 7 False positives (predicted yes, but are not cancerous), 3 false negatives (predicted no, but are cancerous) and 36 true positives (predicted yes and are cancerous).

B. KNN Classifier

The K-nearest Neighbors Classifier is a straightforward yet efficient supervised machine learning algorithm employed for both classification and regression purposes. It operates on the fundamental concept of predicting the class of a data point by examining the classes of its nearest neighbors. The process begins with a labeled training dataset containing input feature vectors and their corresponding class labels. In KNN, the 'K' denotes the number of nearest neighbors that are taken into account when making predictions for a specific data point. The KNN Classifier then selects the K-nearest data points and looks at their class labels. It uses a majority voting scheme to assign the class label to the new data point. For example, in a binary classification problem with odd K, if the K nearest neighbors have more data points from class A than class B, the new data point will be classified as class A.

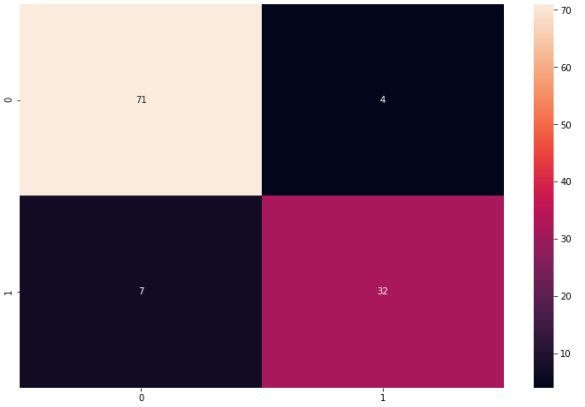


Figure 6: Confusion matrix of KNN Classifier

In fig 6, we can see that the KNN classifiers generated 71 true negatives, 4 False positives, 7 false negatives and 32 true positives.

C. Random Forest Classifier

The Random Forest Classifier is an ensemble learning technique that constructs multiple decision trees during the training process and then amalgamates their predictions to achieve precise and resilient classifications. It is an extension of the bagging technique applied to decision trees, and was introduced by Leo Breiman in 2001. Like any other supervised learning algorithm, the Random Forest Classifier starts with a labeled training dataset, which consists of input feature vectors and their corresponding class labels. The Random Forest employs the bagging technique, which involves generating several bootstrap samples from the original training dataset. These bootstrap samples are created by randomly selecting data points with replacement, which means that some data points may appear multiple times within the same sample, while others may not be included at all. For each of these bootstrap samples, a decision tree is built using randomly selected features, and the tree is expanded to its maximum depth.

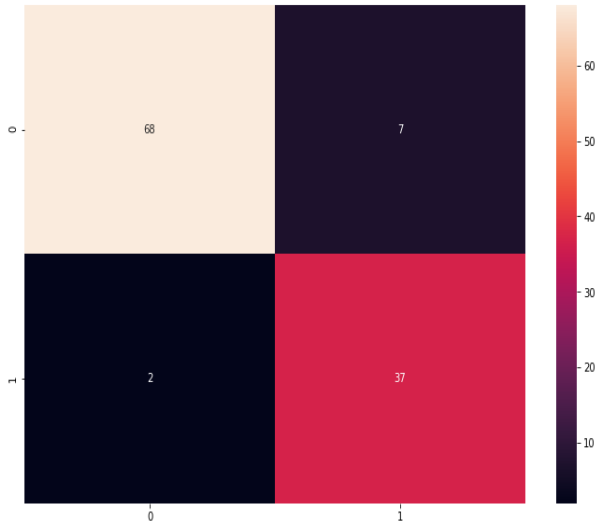


Figure 7: Confusion Matrix of Random Forest Classifier

In fig 7, we can see that the Random Forest classifiers generated 68 true negatives, 7 False positives, 2 false negatives and 37 true positives.

D. AdaBoost Classifier

AdaBoost, short for Adaptive Boosting, is an ensemble learning approach designed to enhance the effectiveness of weak learners, such as decision trees with restricted depth, by amalgamating them into a potent classifier. This technique was introduced by Yoav Freund and Robert Schapire in 1996. AdaBoost starts with a labeled training dataset containing input feature vectors and their corresponding class labels. In the beginning, all data points are given equal weight. During the iterative training process, misclassified data points are given higher weight to focus the subsequent weak learners on those instances. Once all weak learners are trained, they are combined into a single strong classifier using weighted voting. The weight of each weak learner's vote is determined by its accuracy on the training data. More accurate weak learners have higher voting weights. The class with the highest cumulative vote becomes the final prediction.

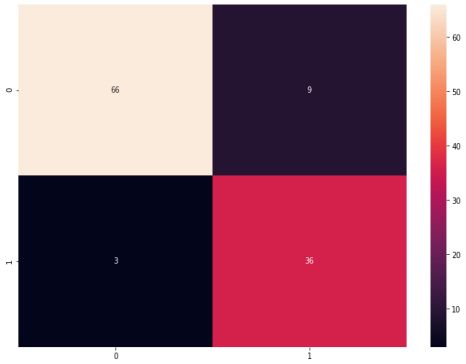


Figure 8: Confusion Matrix of AdaBoost Classifier

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 10, October-2023

In fig 8, we can see that the AdaBoost classifiers generated 66 true negatives, 9 False positives, 3 false negatives and 36 true positives.

E. Gradient Boost Classifier

The Gradient Boosting Classifier, akin to AdaBoost, is a well-known ensemble learning technique that unites the forecasts of multiple weak learners to construct a robust classifier. Unlike AdaBoost, which focuses on adjusting the sample weights to emphasize misclassified instances, Gradient Boosting builds a sequence of weak learners in a way that each new learner corrects the errors made by the previous ones. It was first proposed by Jerome H. Friedman in 1999. Similar to AdaBoost, Gradient Boosting uses weak learners as base estimators. Typically, these weak learners are shallow decision trees, often referred to as decision stumps, to prevent overfitting. In Gradient Boosting, gradient descent optimization is applied to minimize the ensemble's loss function. The loss function quantifies the disparity between the actual labels and the current predictions. Once all the weak learners are trained and added to the ensemble, the Gradient Boosting Classifier combines their predictions using weighted voting to make the final predictions.

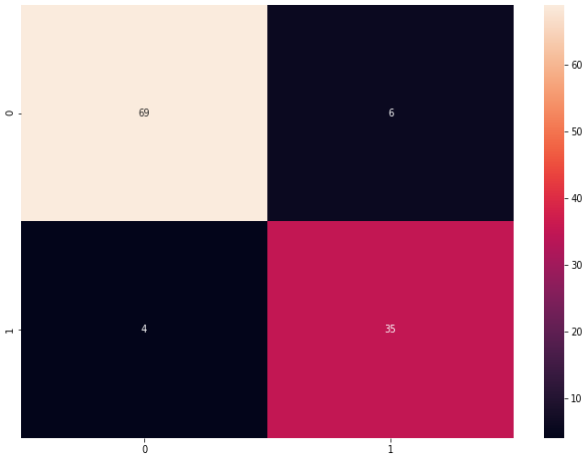


Figure 9: Confusion Matrix of Gradient Boost Classifier

In fig 9, we can see that the Gradient Boost classifiers generated 69 true negatives, 6 False positives, 4 false negatives and 35 true positives.

F. MLP Classifier

The MLP Classifier, short for Multi-Layer Perceptron, is a type of artificial neural network employed for supervised learning tasks, including both regression and classification. A feedforward neural network can have one or more hidden layers situated between the input and output layers A "perceptron" is the fundamental unit of a neural network that replicates the function of a real neuron. The neural network receives its raw input data at the input layer. A feature in the input data is represented by each neuron in the input layer. There may be one or more hidden layers, each with numerous neurons, between the input and output layers. From the input data, these hidden layers are in charge of extracting intricate patterns and representations. Each neuronal link between neighbouring levels has a weight. The neural network's behaviour is modified during training by updating the weights, which stand for the strength of the connections between neurons. The network is able to learn and simulate non-linear correlations in the data because each neuron in the hidden and output layers has an associated bias. An activation function is applied by each neuron in the hidden and output layers to the weighted sum of its inputs. MLP classifiers frequently employ the sigmoid, tanh, and ReLU (Rectified Linear Unit) functions as activation functions. The neural network may approximate complex functions thanks to the addition of nonlinearity to the model provided by activation functions. The neural network's final predictions are generated by the output layer.

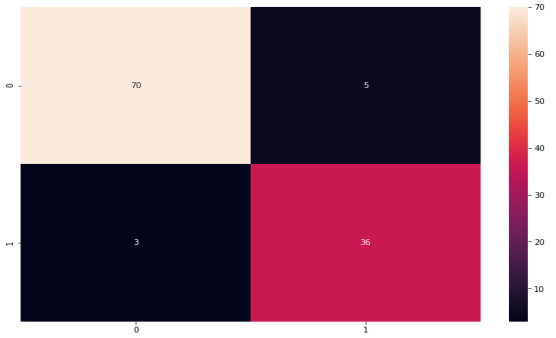


Figure 10: Confusion Matrix of MLP Classifier

In fig 10, we can see that the MLP classifiers generated 70 true negatives, 5 False positives, 3 false negatives and 36 true positives.

Table 2: Prediction Table of various ML Models

Model Used	True Positive (TP)	True Negative (TN)	False Positive (FP)	False Negative (FN)
Bagging Classifier	36	68	7	3
KNN	32	71	4	7
Random Forest	37	68	7	2
AdaBoost	36	66	9	3
Gradient Boost	35	69	6	4
MLP	36	70	5	3

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 10, October-2023

V. EXPERIMENTAL RESULTS

A. Precision

Precision is a Classification metric that measures how accurate a classifier is. It is defined as the ratio of true positive predictions to the total number of positive predictions made by the classifier. The confusion matrix for binary classification is typically organized as follows:

TP (True Positives): Correctly predicted positive cases.
FN (False Negatives): Total positive cases that are incorrectly classified as negative.
FP (False Positives): Total number of negative cases that are incorrectly classified as positive.
TN (True Negatives): Correctly predicted negative cases.

$$\text{Precision} = TP / TP+FP$$

Table 3: Precision Table of Various ML Models

Model Name	Precision
Bagging Classifier	0.837
KNN	0.888
Random Forest	0.840
Adaboost	0.800
Gradient Boost	0.853
MLP	0.878

KNN having highest precision but in this case, precision alone cannot be an effective parameter to choose the best model as it has maximum False negatives.

B. Recall

To calculate recall (also known as sensitivity or true positive rate) from a confusion matrix, we need to consider the True Positives (TP) and False Negatives (FN) for a binary classification problem. Recall measures the proportion of actual positive instances that were correctly identified by the classifier.

$$\text{Recall} = \frac{TP}{TP+FN}$$

Table 4: Recall table of various ML Models

Model Name	Recall
Bagging Classifier	0.923
KNN	0.82 0
Random Forest	0.948
Adaboost	0.923
Gradient Boost	0.897
MLP	0.923

C. F1 Score:

F1 score is a single metric that combines both precision and recall into one score and is often used when you want to balance the trade-off between precision and recall.

$$\text{F1 Score} = 2 \times \frac{Precision \times Recall}{Precision+Recall}$$

Table 5: F1-Score table of various M Models

Model Name	F1- Score
Bagging Classifier	0.877
KNN	0.852
Random Forest	0.890
Adaboost	0.857
Gradient Boost	0.874
MLP	0.899

D. Accuracy:

Accuracy is a commonly used metric to evaluate the performance of a classification model. It is calculated as the ratio of the number of correctly predicted samples to the total number of samples in the dataset.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Table 6: Accuracy Table of various ML Models

Model Name	Accuracy
Bagging Classifier	0.912
KNN	0.903
Random Forest	0.921
Adaboost	0.894
Gradient Boost	0.912
MLP	0.929

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 10, October-2023

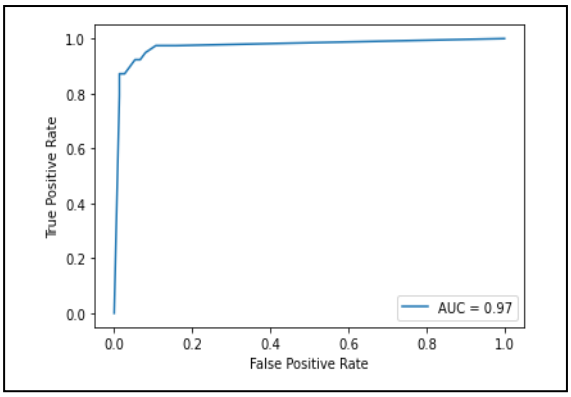


Figure 11: Accuracy generated for Bagging Classifier using ROC Curve

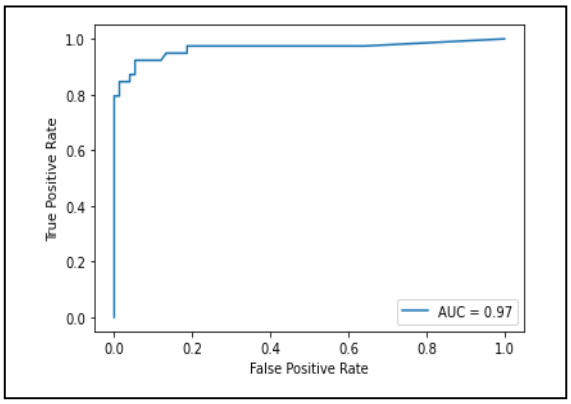


Figure 12: Accuracy generated for Random Forest Classifier using ROC Curve

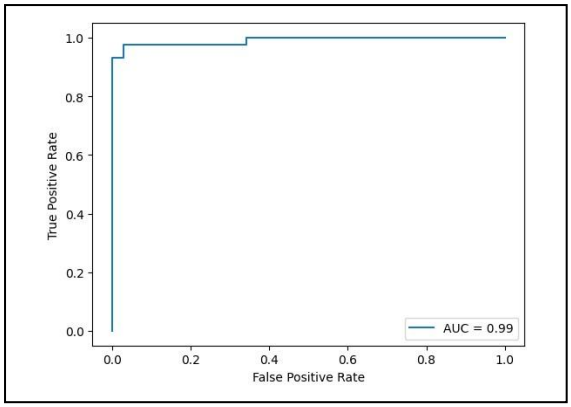


Figure 13: Accuracy generated for MLP Classifier using ROC Curve

Overall results generated by python code on jupyter considering all the parameters are as:

Model	Train Accuracy	AUC SCORE
bagging classifier	99.56	0.97
KNN classifier	95.82	0.97
Random Forest calssifier	100	0.97
Adaboost classifier	100	0.96
Gradientboot classifier	100	0.97
MLP Classifier	96.92	0.98
MLP Classifier	96.92	0.99
MLP Classifier	96.92	0.99

Figure 14: Accuracy Table of ML models based on all parameters

VI. CONCLUSION AND FUTURE WORK

In this research paper, we explored the application of machine learning algorithms for breast cancer detection using various features. Our findings demonstrate the effectiveness of the proposed approach in accurately distinguishing between malignant and benign tumors. The machine learning models achieved high classification accuracy and showed promising results in early detection, which can potentially lead to improved patient outcomes.

Specifically, we observed that the MLP algorithm performed exceptionally well in breast cancer detection, with an accuracy of 99%. This suggests that the algorithm can be a valuable tool in assisting medical professionals in making accurate diagnoses.

Although this research has made significant strides in breast cancer detection but to further validate the performance and generalizability of the proposed approach, it is essential to test the machine learning models on larger and diverse datasets from multiple medical institutions. This will ensure the robustness and reliability of the model. Exploring deep learning architectures, such as convolutional neural networks (CNNs), for breast cancer detection from medical imaging could be an interesting direction. CNNs have shown impressive results in image recognition tasks and may provide more fine-grained insights into tumor characteristics.

In conclusion, this research demonstrates the potential of machine learning in breast cancer detection and provides valuable insights into tumor characteristics. The findings open up exciting possibilities for future research, which can contribute to advancements in breast cancer diagnosis and ultimately lead to better patient care and outcomes.

REFERE NCES

[1] J. Zheng, D. Lin, Z. Gao, S. Wang, M. He, and J. Fan, "Deep Learning Algorithms for Internet of Medical Things: Breast Cancer Detection," in IEEE Internet of Things Journal, vol. 8, no. 7, pp. 5685-5695, 1 April1, 2021, doi: 10.1109/JIOT.2021.3060465.

[2] Prerita, Sindhwani, N., Rana, A., & Chaudhary, A. (2021). Breast Cancer Detection using Machine Learning Algorithms. 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions), ICRITO2021.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 12 Issue 10, October-2023

[3] Vijaylaxmi, Kochari. (2021). Detection of Breast Cancer Using Machine Learning Algorithms. 7(1):223-227. doi: 10.32628/CSEIT217141

[4] Islam, S., Protick, A., & Rahman, R. (2019). Early Detection of Breast Cancer Using Artificial Intelligence. International Journal of Engineering and Applied Sciences Technology, 4(7), 339-359. ISSN No. 2455-2143.

[5] Mediha, Salić., Nejra, Samardžić., Nejla, Selmanović., Irma, Sinanović., Muhamed, Sirćo., Belma, Suljević. (2021). Machine Learning Techniques for Predicting Breast Cancer Based on Biomarkers. 256-263. doi: 10.1007/978-3-030-73909-6_29

[6] David, A., Omondiagbe., Shanmugam, Veeramani., Amandeep, S., Sidhu. (2019). Machine Learning Classification Techniques for Breast Cancer Diagnosis. 495(1):012033-. doi: 10.1088/1757-899X/495/1/012033

[7] Noushaba, Feroz., Mohd, Abdul, Ahad., Faraz, Doja. (2021). Machine Learning Techniques for Improved Breast Cancer Detection and Prognosis—A Comparative Analysis. 441-455. doi: 10.1007/978-981-16-3067-5_33

[8] Sara, Noor, Eldin., Jana, Khaled, Hamdy., Ganna, Tamer, Adnan., Maysoon, Hossam., Noha, ElMasry., Ammar, Mohammed. (2021). Deep Learning Approach for Breast Cancer Diagnosis from Microscopy Biopsy Images. 216-222. doi:10.1109/MIUCC52538.2021.9447653

[9] Hajra, Naveed, Iqbal., Ali, Bou, Nassif., Ismail, Shahin. (2020). Classifications of Breast Cancer Diagnosis using Machine Learning. 14:86-86. doi: 10.46300/9108.2020.14.13

[10] Pawan, Kumar., A., S., Bhatnagar., Roshan, Jameel., Ashish, Kumar, Mourya. (2021). Machine Learning Algorithms for Breast Cancer Detection and Prediction. 133-141. doi: 10.1007/978-981-16-0695-3_14

[11] Rekh, Ram, Janghel., Lokesh, Singh., Satya, Prakash, Sahu., Chandra, Prakash, Rathore. (2020). Classification and Detection of Breast Cancer Using Machine Learning. 269-282. doi: 10.1007/978-981-152071-6_22

