Implementation of Machine Learning Algorithms for Breast Cancer Detection

Machine Learning for Breast Cancer Detection

Puneet, Ramakrishna Prasad Are, Anju R Babu*

Department of Biotechnology and Medical Engineering, National Institute of Technology Rourkela, Odisha, India

E-mail ID*: babua@nitrkl.ac.in

ORCID ID:0000-0003-2259-5896

***Abstract*- Breast cancer ranks as the second most prevalent cancer in women. Detection at an early stage can save many lives. Different machine learning (ML) algorithms can be extremely useful for predicting breast cancer. The Light Gradient Boosting Machine (LightGBM), Extreme Gradient Boosting (XGBoost), and Gaussian Nave Bayes (GaussianNB) ML algorithms were employed to predict breast cancer using the Wisconsin breast cancer dataset. The accuracy, precision, F1 score, and area under the curve (AUC score) of the receiver operating characteristics (ROC) curve were used to evaluate and compare the performance of different ML classifiers. GaussianNB had the lowest accuracy, at 95.74 percentage, while LightGBM had the most accuracy, at 98.40 percentage.**

Keywords— Accuracy; GaussianNB ; LightGBM; ROC curve ; XGBoost

I. INTRODUCTION

Breast cancer ranks second as the most prevalent kind of cancer in women. Furthermore, the World Health Organization wants to reduce 25% of breast cancer deaths by 2030 and 40% by 2040 by reducing breast cancer mortality by 2.5% annually [1]. Conventional treatment procedures, such as surgery, chemotherapy, and radiotherapy, have been utilised but have several disadvantages, like hair loss, joint discomfort, itching, peeling skin [2]. Recent breakthroughs include stem cell therapy, targeted therapy, ablation therapy, and therapy based on ferroptosis. Artificial intelligence (AI) can be used to analyse the images and find breast masses, segment breast masses, determine breast density, and determine cancer risk [3]. In mammograms, AI can detect cancer up to two years sooner than human experts, increasing the likelihood of saving lives. However, the use of AI also calls for inclusive and diverse datasets for training. The dataset's source population should be reflected in the patient population used to implement and apply these models. In this study, malignant and benign tissues were identified using machine learning classifiers from various features available in the dataset.

II. METHODOLOGY

A. Data

The data was collected from the UCI machine learning repository in .csv format [4]. To create features that described the characteristics of the cell nuclei apparent in the image, the fine needle aspirate (FAN) of the breast mass image was enhanced digitally. In a study of 569 subjects, 32 features were present in total. Of these 569 subjects, 357 were benign features, whereas 212 were malignant features (62% benign and 37.26 % malignant). The first feature consisted of the patient's identification number. Hence it was omitted during model training. The 'diagnosis' was the second feature, which serves as the model's target or output feature. The remaining attributes constitute the model's inputs. 30 input features are used to determine ten real-valued features for each cell nucleus.

B. Methods

After data acquisition, the dataset preprocessing was done in the following steps: data cleaning, attribute selection, output selection and feature extraction. The preprocessed data were then used for the building of the ML classifiers. Three ML classifiers—GaussianNB, LightGBM, and XGBoost—were used to predict the formation of benign and malignant tissue using various dataset attributes. After applying ML classifiers to the given breast cancer dataset, accuracy, precision, sensitivity, F1_score, and AUC_score were used to evaluate and compare the performances of these ML classifiers. The ML classifiers were trained and tested on 7:3.

III. RESULTS

The ROC curve for the used ML classifiers is shown in Figure I. The AUC score for XGBoost is 0.94, while it is 0.99 for GaussianNB and LightGBM. This shows that, based on the AUC score, GaussianNB and LightGBM outperformed the XGBoost classifier. The predictability of ML classifiers increases with the classifier AUC score.

TABLE I. PERFORMANCE EVALUATION OF DIFFERENT MACHINE LEARNING CLASSIFIERS

ML Classifier	Class	Sensitivity	Precision	F1_Score
XG Boost	Benign	0.98	0.99	0.99
	Malignant	0.98	0.97	0.98
Light GBM	Benign	0.98	1.00	0.99
	Malignant	1.00	0.96	0.98
Gaussian NB	Benign	0.94	0.94	0.94
	Malignant	0.90	0.89	0.90

Comparing the effectiveness of several ML classifiers for both malignant and benign tumours using sensitivity,

precision, and F1 score is shown in Table-I. LightGBM fared best for these assessment parameters, generating 98% sensitivity, 100% precision, and 0.99 F1 score for benign, while GaussianNB produced the worst results, 94% sensitivity, 94% precision, and 0.94 F1 score for benign.

TABLE II. PERFORMANCE EVALUATION OF DIFFERENT MACHINE LEARNING CLASSIFIERS

ML Classifier	Accuracy	AUC Score
GaussianNB	95.74%	0.99
Lightgbm	98.40%	0.99
XGBoost	98.40%	0.94

Table-II. Displayed the effectiveness of the various ML classifiers utilised. GaussianNB demonstrated an accuracy of 95.74%, whereas LightGBM and XGBoost showed an accuracy of 98.40%.

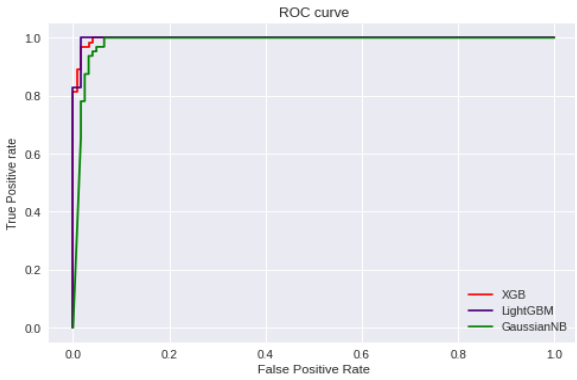

Figure-I ROC curve for different ML classifiers

Based on the evaluation criteria, it was evident that LightGBM outperforms the competition, whereas GaussianNB underperforms. With the help of this assessment, the feature importance ranking for the LightGBM classifier was determined, and the feature importance ranking graph is shown as seen in Figure II. The radius mean attribute is the most significant, while the fractal dimension worst attribute is the least important, as seen in Figure II.

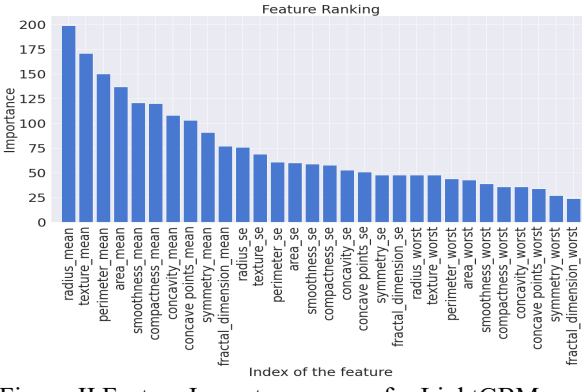

Figure-II Feature Importance curve for LightGBM

IV. Discussion

Mohammad et al.'s usage of a support vector machine (SVM) yielded the best accuracy of 97.7% and AUC_score of 0.97 on the provided dataset. In contrast, Naïve Bayes (NB) yielded the lowest accuracy of 92.9% and an AUC score of 0.98 [5]. This indicates that all classifiers employed in our work produce superior accuracy except for the Gaussian NB. Using different fold cross-validation (5,10, and 15) and other training splitting ratios (66.6% and 85.5%). Ahmed et al. produced better results than we did, achieving a classification accuracy of 99.01% with NB [6]. Different cross-validation and training splitting ratios may improve the outcomes. LazyIBK had the maximum accuracy of 99.14%, and NB had the lowest accuracy, according to research by Kumar et al., which was consistent with our study's finding that the NB classifier had the lowest accuracy [7]. Ribeiro et al. use of neural networks for diagnosis yielded subpar results compared to our work, with accuracy rates of 94.69% and 96.19%, respectively, using Sklearn and Keras (python libraries) [8].

V. Conclusion

Three different ML classifiers were applied to the given dataset. The LightGBM predicted breast cancer with the highest accuracy, whereas the XGBoost had the lowest accuracy. The classifiers can also be optimized to improve dataset analysis and predictability. More machine learning classifiers and more feature selections may be added to enhance performances and obtain higher predictability.

VI. References

[1] "Breast cancer," *Who.int*. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/breast-cancer. [Accessed: 31-Jan-2023].

[2] M. I. Nounou, F. ElAmrawy, N. Ahmed, K. Abdelraouf, S. Goda, and H. Syed-Sha-Qhattal, "Breast cancer: Conventional diagnosis and treatment modalities and recent patents and technologies," *Breast Cancer (Auckl.)*, vol. 9, no. Suppl 2, pp. 17–34, 2015.

[3] G. Dileep and S. G. Gianchandani Gyani, "Artificial intelligence in breast cancer screening and diagnosis," *Cureus*, vol. 14, no. 10, p. e30318, 2022.

[4] "UCI machine learning repository: Breast cancer Wisconsin (diagnostic) data set," *Uci.edu*. [Online]. Available: https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic). [Accessed: 31-Jan-2023].

[5] W. T. Mohammad, R. Teete, H. Al-Aaraj, Y. S. Y. Rubbai, and M. M. Arabyat, "Diagnosis of breast cancer pathology on the Wisconsin dataset with the help of data mining classification and clustering techniques," *Appl. Bionics Biomech.*, vol. 2022, p. 6187275, 2022.

[6] M. T. Ahmed, M. N. Imtiaz, and A. Karmakar, "Analysis of Wisconsin Breast Cancer original dataset using data mining and machine learning algorithms for breast cancer prediction," *J. Sci. Technol. Environ. Inform.*, vol. 9, no. 2, pp. 665–672, 2020.

[7] V. Kumar, B. K. Mishra, M. Mazzara, D. N. H. Thanh, and A. Verma, "Prediction of malignant and benign breast cancer: A data mining approach in healthcare applications," in *Advances in Data Science and Management*, Singapore: Springer Singapore, 2020, pp. 435–442.

[8] V. Ribeiro, E. J. Solteiro Pires, and P. B. de Moura Oliveira, "Breast Cancer Diagnosis using a Neural Network," in *2019 IEEE 6th Portuguese Meeting on Bioengineering (ENBENG)*, 2019, pp. 1–4.