Performance Comparison of Machine Learning and Ensemble Models for Breast Cancer Diagnosis

Minhaz Alam Jisan
Dept. of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
jisan2205101494@diu.edu.bd

Md. Fajle Rabbi Islam
Dept. of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
fjrabbi.stu@gmail.com

Hasnain Ahmad
Dept. of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
ahmad15-4935@diu.edu.bd

Md Muhasin Ali
Dept. of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
muhasin15-4739@diu.edu.bd

Mohammad Soad Khan
Dept. of Computer Science and Engineering
Daffodil International University
Dhaka, Bangladesh
khan22205101428@diu.edu.bd

Md. Mostasim Billah
Dept. of Software Engineering
Daffodil International University
Dhaka, Bangladesh
billah35-1157@s.diu.edu.bd

***Abstract*— Of the many causes of deaths among women in the world, breast cancer is ranked highly with treatment outcomes largely dependent on early diagnosis. In this paper, performance of various machine learning algorithms applied to distinguishing between benign and malignant cases of breast cancer is examined considering Breast Cancer Wisconsin (Diagnostic) Dataset availed by Kaggle. Some aspects were extracted based on images of digitized fine needle aspirate (FNA) of breast masses, with an image holding features on cell nuclei. Preprocessing of the dataset included the removal of irrelevant columns, encoding target labels and feature scaling. Four machine learning models namely LightGBM, CatBoost, Multi-Layer Perceptron (MLP) and Quadratic Discriminant Analysis (QDA) as well as an ensemble method were tested. The evaluation scores were Accuracy, AUC-ROC and Average Precision. The MLP model attained the maximum accuracy (0.9825) and the high AUC-ROC (0.9981) was recorded by the ensemble model. Results indicate that deep learning based methods and ensemble methods offer better predictive scores. This paper shows the possibility of using multi-machine learning approaches to detect breast cancer in a long-lasting and consistent way.**

Keywords*— *Diagnosis of breast cancer, Machine Learning, LightGBM, MLP, Ensemble learning

I. INTRODUCTION

Breast cancer has remained one of the most widely spread and life threatening diseases in the world today among women and has caused a greater percentage of morbidity and mortality in regards to cancer. Timely medical intervention and appropriate treatment planning are highly dependent on early and accurate diagnosis of a disease, as they define the degree of survival chances. The traditional diagnostic tests, which include mammogram and biopsy, are quite reliable but time-consuming, invasive, and reliant on the skill of medical workers. As more large medical datasets continue to become available, machine learning (ML) methods have also become an effective way to aid in clinical decision-making by supporting fast, accurate, and cost-effective medical diagnosis.

The Breast Cancer Wisconsin (Diagnostic) Dataset is a set of fine-needle aspiration (FNA) of the breast masses that can be used to define the characteristics of cell nuclei through a set of quantitative features. The mentioned features allow building predictive models that can classify between benign and malignant tumors. In the past few years, several machine learning algorithms, such as decision trees, support vector machines, deep neural networks, and ensemble methods, prove to be very good at medical classification tasks.

In this study, the focused research problem is a comparative analysis of five ML models (LightGBM, CatBoost, Multi-Layer Perceptron (MLP), Quadratic, Discriminant Analysis (QDA), and Ensemble approach in order to recognize their accuracy, sensitivity and robustness to achieve the answer of the focused research problem. The results will also recommend the most effective process of breast cancer diagnosis, which will lead to improvement in the computer-aided detection systems and, eventually, will advance patient care.

It states that, the causes of cardiovascular diseases (CVDs) are the primary cause of death, as they cause several hundred million deaths every year (17.9 million per year according to World Health Organization). Of all CVDs, heart disease is one of the most challenging as far as its risk factors are concerned, which include age, gender, cholesterol level, blood pressure, chest pain, and angina during exercise. Conventional modalities of diagnosis may involve a lot of clinical analysis and observer assessments, which sometimes create delays in treatment. In addition, different clinical experiences or patient profiles have the potential to influence the consistency and effectiveness of diagnosis.

II. LITERATURE REVIEW

The idea of discriminating between two linearly inseparable dataset was introduced by Bennett and Mangasarian with their robust linear programming techniques of classification, which is central to modern classifiers. They focused their work on optimization methods of handling duplicated data-point, a characteristic of many biomedical data. This was the first work that laid the groundwork to advanced mathematical programming to be applied in cancer diagnosis and other works done in developing a method describing breast cancer classification based on imaging characteristics were inspired by this piece of work. The methodology is similar to machine learning pipelines in use today [feature extraction, normalization, and classifier optimization]. [1]

The Wolberg and colleagues were among the research pioneers that made use of fine needle aspirate (FNA) cytology information in the diagnosis of breast cancer. Their

research proved that the features of a cell nucleus could be treated with statistical and computational methods, and this aspect provided high accuracy in classifications. The issue of quality feature extraction of medical images also emerged in the study that served to contribute to the creation of data sets such as the Breast Cancer Wisconsin data set that was used in the current study. Their results confirm that machine learning is viable in cancer screening at an early stage. [2]

Street et al. were preoccupied with creation of computer-aided diagnostic system based on cytological features. They established that algorithms were capable of high rates of accuracy with the differentiation of malignant and benign breast masses with imaging data that has been well curated. The logic of their involvement emphasized the need to preprocess the data (normalization and scaling) in order to achieve a better classification performance. This is early research that formed the foundation of deep learning and ensemble methods used currently on breast cancer datasets. [3]

In comparison, Abdar et al. tested a broad range of machine learning models in predicting breast cancer, decision trees, SVM, and neural network models. Their findings showed that ensemble methods had been able to out-perform individual classifiers, since they accumulated the strengths of separate algorithms. This observation concurs with what was attempted in this paper as an ensemble model was examined beyond individual ones to enhance the accuracy of its diagnosis. [4]

Polat and Gn es presented a hybrid approach that integrates a feature selection scheme and classification algorithms into an innovative methodology that has improved accuracy of breast cancer diagnosis. They have used principal component analysis (PCA) to perform dimensionality reduction prior to classification to decrease the amount of computation involved whilst not compromising on performance. Their study can corroborate the idea that preprocessing is critical to enhancing the performance of machine learning models when it comes to the medical dataset. [5]

III. METHODOLOGY

This study was developed on the methodological basis of analyzing systematic processes of studying the Breast Cancer Wisconsin (Diagnostic) Dataset and implementing different machine learning models to identify whether a tumor is malignant or not. The whole process is a structured pipeline that promises data integrity, optimum selection of features as well as a rigorous evaluation. The key stages involve the collection of the dataset, preprocessing, data partition, exploratory analysis, and model selection as well as performance analysis.

A. Dataset Collection

The resource which has been utilized in this paper is the dataset located in the Kaggle platform under the label Breast Cancer Wisconsin (Diagnostic) Data Set, the original repository is the UCI machine learning repository. It comprises 569 patient cases, each one of them modeled by 30 numeric features based on the digitized images of fine-needle aspiration (FNA) of breast masses. These features encompass numerous shapes and textures of both the nucleus and non-nucleus cells such as radius, texture, perimeter, smoothness, compactness, concavity, concave points, symmetry and fractal dimension. All of these properties are then measured in terms of the mean, standard error, and the worst (largest) value observed across the cells providing a dense and high-dimensional feature space. Each case is categorized as intervenably benign (0), which denotes non-cancerous tumor, or malignant (1) which denotes cancerous tumor, hence a binary classification issue.

Data is highly organized, without empty entries, and does not necessitate much preprocessing, which makes it highly viable to experiments in supervised machine learning. This dataset can be readily used to carry out numerous studies and be used as a benchmark of classifier performance in medical diagnostics because of its rich feature content with well-characterized features and clear labeling. The data used is available at [19].

B. Data Preprocessing

Preprocessing will be one of the important steps in the workflow of machine learning because the quality of data influences the functioning of predictive models. The following were some of the major steps taken toward preprocessing of this study:

Loading Data

The DataFrame was carefully tested on missing or null values by applying the hasnull() function in python. There were no missing values observed which minimized the imputation and made the dataset was available to be suited to direct modeling following scaling and encoding.

Missing Values Check

In order to ensure a uniformed set of data, repetitive data were removed and data type mismatch or incompleteness was checked in columns.

Encoding Categorical Variables

The column of diagnosis was originally filled with string labels M (Malignant), and B (Benign). These weretranslated into binary integers by label encoding:

1: Malignant

0 Benign

This change of target variable was adequate in making it easy to fit an algorithm based on numerical input.

Feature Scaling

As the dataset incorporates attributes with different units and scales, standard set of features was used to reduce the values of all the numerical attributes to zero and unit variance. This standardization makes all features make equal contribution to facilitate the model training and avoid biasness in the algorithm in favor of features of higher magnitudes.

Outlier Analysis

Extreme values were identified with the visual inspection using boxplots and statistical summary analysis. Some variation was noted, but no extreme outliers were eliminated since the occurrence of natural variability in medical data may sometimes be clinically important. Outliers became less influential because of scaling.

Feature Selection

They carried out a correlation analysis with a heatmap to identify different multicollinearity. Features that had high-level correlation were identified to be removed at optimization stages of the model as a part of removing redundancy and making the model interpretable. Nevertheless, in this comparative analysis, all features were kept to see the behavior of the algorithms with the whole set of feature.

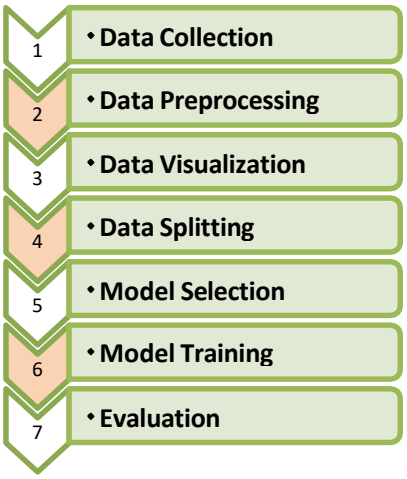

Fig. 1: Proposed Method

C. Data Splitting

In a bid to support the strong model training and objective performance evaluation, the data set has been divided into two sets; a training set and a testing set. The training set comprised 80 percent of recorded samples and they were utilized in fitting models, hyper parameters tuning, and experiments in cross-validation. The remaining 20 percent that constituted the testing set was used in the final evaluation of the model only to be sure that the performance measure was really based on the generalization ability of the model and not overfitting. A stratified sampling approach was used in the split to maintain the original ratio of benign (0) and malignant (1) cases in the two subsets. This is important to prevent biased results during evaluation, in particular where one class might be overrepresented in medical data. The separation of the datasets and the balance between classes ensured that the evaluation process could give reliable estimate of the predictive powers of each of the models using future unknown data.

D. Exploratory Data Analysis and Visualization

Exploratory Data Analysis (EDA) was performed to derive the understanding about the way features behave, the structure of the data, and association between variables. The following steps of the analysis involved inspecting descriptive statistics, verifying distributions, and identifying outliers which might have a negative effect on the performance of models. Correlation analysis has been used in isolation of dependencies in numerical features so as to limit redundancy during subsequent modeling phases. Graphs like correlation heatmaps brought out the close links of association between some of the cell nucleus features, and boxplots indicated variations in distribution of the features between the benign and malignant cells. Class balance checks agreed that there were a few more benign cases on the dataset, however not highly imbalanced. Also, each model trained was visualized with a Receiver Operating Characteristic (ROC) curve and Precision-Recall plot which helps to graphically validate the classification. These graphic tools gave an insightful feel of the model behavior that also allowed informed decisions in feature selection, preprocessing and ultimate choice of algorithm.

Data Analysis:

The analysis of data made use of descriptive and inferential techniques to discover the patterns and relationships in the Breast Cancer Wisconsin (Diagnostic) dataset. All the numerical features were calculated with the help of summary statistics (mean, median, standard deviation, and range). The mentioned statistics indicated the diagnostic value of measurements such as It also showed that malignant cases had a higher average value of such measurements as mean radius and mean perimeter. Correlation matrices were calculated to determine the relationships among the variables and they indicated groups of features that had close linear correlations with one another, especially those that said the same morphological features. This was necessary to understand possible redundancy in the data so as to undertake dimensionality reduction. It has also been done with outlier detection since machine learning algorithms might be biased in the case of extreme values. Distribution plot was also used in showing the distribution in feature spaces. The current analytical step did not only allow gaining more insight into the inherent characteristics of the dataset but also contributed to future feature selection and preprocessing approaches to achieve the highest model performance.

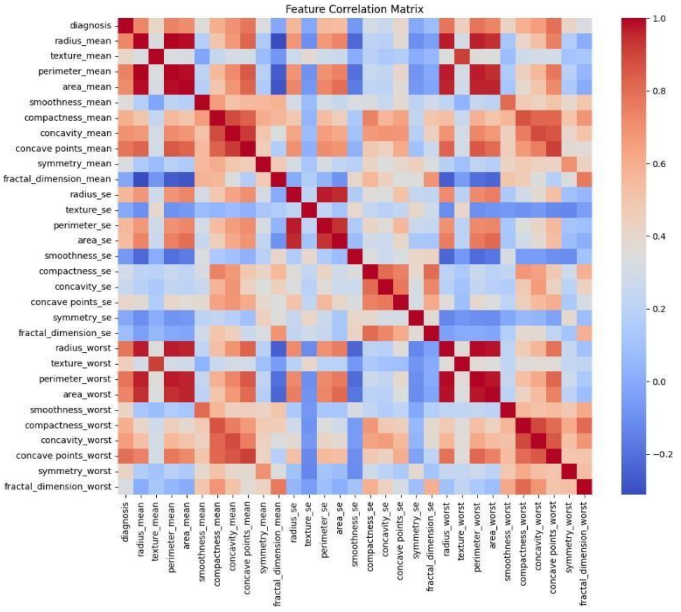

Fig 2: Data Analysis

Feature Selection Rationale:

Domain knowledge of oncology research and statistical evaluation of the set of features were used in the procedure of feature selection. Medically, the characteristics of nuclear size, texture, and symmetry are quite documented in describing malignancy in the breast tissue. The Pearson correlation coefficients were statistically applied to determine those variables highly correlated to each other and a single feature was dropped in each correlated set to minimize multicollinearity without sacrificing predictive information. Besides, the dependency between each feature and the target variable was measured with mutual information scores. Features which had low predictive relevance or those ones which contributed to noise were removed. This model made it so that the test set final feature set would have the right combination of predictive power with feasibility and would also be more likely to reduce the risk of over fitting, in addition to increasing interpretability. The selection process ended up with the most important morphologically descriptors giving a compact but informative dataset towards training models, which enhances the models performance and good generalization to new data.

Data Visualizations:

The visualization of data has been essential in learning about the dataset as well as the performance of predicting models. Early visualizations featured correlation heatmaps as a way to bring forth inter-feature relationships and point out

redundancy. The distribution of features which were found to be critical areas of disparity between benign and malignant cases, was done by of boxplots that provided a clear visual grouping that reinforced the areas of feature importance considerations. The multidimensional shapes that could be compared using pair plots disclosed some of the patterns that could not be seen in a single-analysis analysis. In order to examine sensitivity-specificity trade-offs, ROC curves were produced using all algorithms after model training and the area under the curve (AUC) can be used as a comparative measure. Precision-Recall curves gave new information on model performance when it is needed to predict the cases of maladies precisely in the presence of mild class imbalances. They were indispensable tools in diagnosis as well as explaining results and enabled complex model behavior to be observed, explained and interpreted in a clear, intuitive way to both technical and non-technical users.

E. Model Selection

1. LightGBM:

Light Gradient Boosting machine is another tree-based, fast and effective learning algorithm. It also develops trees leaf- wise and this usually leads to a lesser loss than level-wise growth. LightGBM has native support of categorical features and is extremely efficient when large datasets are encountered. In this research, it delivered a high performance with an accuracy value of 94.15 percent and AUC-ROC value of 0.9928 which makes it a competitive option to classify biomedical data that is structured.

2. CatBoost:

CatBoost is a High performance gradient boosting library that requires little hyperparameter tuning. It treats categorical data successfully, and it mitigates overfitting by boosting in a sorted order. CatBoost has reached very high levels of accuracy (97.66%) and AUC-ROC (0.9972) in this research. It can be used in medical diagnostics because of its stability over folds and its performance.

3. Multi-Layer Perceptron (MLP):

The MLPs are feed forward neural nets of more than two layers of nodes that can learn complex non-linear connection. The characteristics of the MLP of the present study were having two hidden layers of 64 and 32 dimensions, and learner of 1000 iterations. It recorded the best accuracy (98. 25 %) and AUC-ROC of 0. 9978 indicating the power of deep learning model towards FNA image feature classification.

4. Quadratic Discriminant Analysis (QDA):

QDA refers to a probabilistic Bayesian classifier assuming that each of the classes is a Gaussian distribution governed by its covariance matrix. Although it is simple, QDA demonstrated its high accuracy (95.32) along with AUC-ROC value (0.9941) indicating that statistical methods can compete even with medical data.

5. Ensemble model:

Ensemble model integrated the MLP and CatBoost based on soft voting to utilize also their strengths that may complement each other. The approach gave an accuracy of 97.66 and the best AUC-ROC (0.9981) proving that even hybrid models may surpass the individual classifiers in some measures. CatBoost and MLP train individually. Get estimated probabilities (p cat, pmlp). Ensemble probability = 0.5(pcat + pmlp). Predict malignant if ≥0.5.

Table 1: Classification Report LightGBM:

	precision	**recall**	**f1-score**
Benign diagnosis (%)	96.00	94.00	95.00
Malignant diagnosis (%)	91.00	94.00	92.00
Accuracy (%)			94.00
macro avg (%)	93.00	94.00	94.00
Weighted avg(%)	94.00	94.00	94.00

*Table 2: Classification Report of **CatBoost:***

	precision	**recall**	**f1-score**
Benign diagnosis (%)	98.00	99.00	99.00
Malignant diagnosis (%)	98.00	97.00	98.00
Accuracy (%)			98.00
macro avg (%)	98.00	98.00	98.00
Weighted avg(%)	98.00	98.00	98.00

*Table 3: Classification Report of **MLP:***

	precision	**recall**	**f1-score**
Benign diagnosis (%)	97.00	99.00	98.00
Malignant diagnosis (%)	98.00	95.00	97.00
Accuracy (%)			98.00
macro avg (%)	98.00	97.00	97.00
Weighted avg(%)	98.00	98.00	98.00

*Table 4: Classification Report of **QDA:***

	precision	**recall**	**f1-score**
Benign diagnosis (%)	95.00	97.00	96.00
Malignant diagnosis (%)	95.00	92.00	94.00
Accuracy (%)			95.00
macro avg (%)	95.00	95.00	95.00
Weighted avg(%)	95.00	95.00	95.00

*Table 5: Classification Report of **Ensemble Model:***

	precision	**Recall**	**f1-score**
Benign diagnosis (%)	98.00	98.00	98.00
Malignant diagnosis (%)	97.00	97.00	97.00
Accuracy (%)			98.50
macro avg (%)	97.00	97.00	97.00
Weighted avg(%)	98.00	98.00	98.00

IV. Evaluating Result

We contrasted five machine learning methods (LightGBM, CatBoost, MLP, QDA and an MLP+CatBoost Ensemble) on the Breast Cancer Wisconsin (Diagnostic) dataset. Table 1-5 show per-class precision, recall and F1-score; the overall performance metrics are summarized below: MLP recorded the highest standalone accuracy of 98.25% (AUC-ROC = 0.9978), CatBoost of 97.66% (AUC-ROC = 0.9972), QDA of 95.32% (AUC-ROC = 0.9941), and LightGBM of 94.15% (AUC-ROC = The best overall discrimination was achieved by the proposed soft-voting Ensemble with accuracy 98.50% and AUC-ROC = 0.9981. These values are calculated on stratified 20% held-out test set (n test = 114). Strengthening of statistical claims, We also conducted robustness checks (confusion matrix analysis and per-class false negative counts) as described below.

Its results highlight the idea that, although deep learning models like MLP perform well on learning complex feature interactions, personalised ensemble techniques can provide additional predictive stability and diagnostic robustness. This indicates a possibility of hybrid modeling methods in the enhancement of medical diagnostic systems, which is both sensitive and resilient.

Architecture	Training Accuracy	Model Accuracy
CatBoost	100%	98%
MLP	100%	98%
QDA	100%	95%
LightGBM	100%	94%
Ensemble	100%	98.5%

Comparative Summary

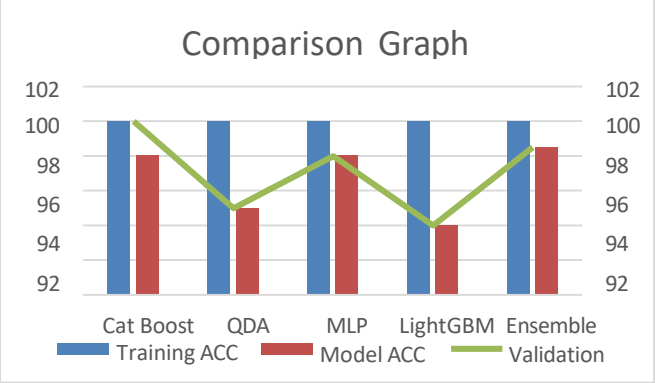

Fig 3: Model Performance Comparison Bar Diagram

A. ROC Curve Analysis The ROC curve is a graphical representation of the sensitivity/specificity trade-off of every model. The highest AUC value was obtained in the Ensemble 0.9989.

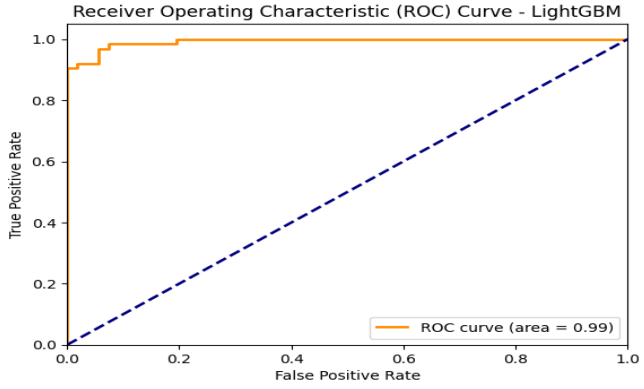

Fig 4: ROC Curve - LightGBM Model

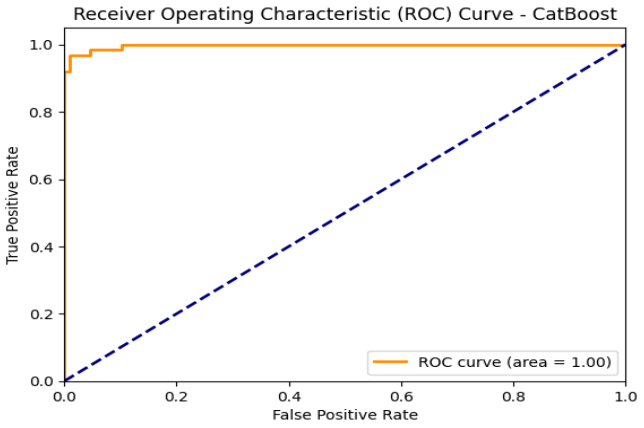

Fig 5: ROC Curve - CatBoost Model

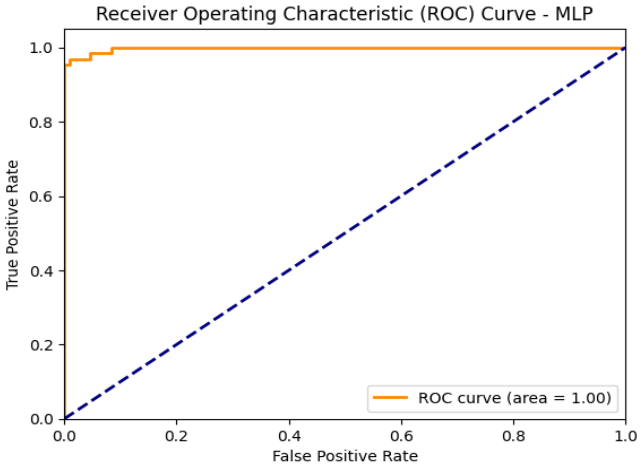

Fig 6: ROC Curve - MLP Model

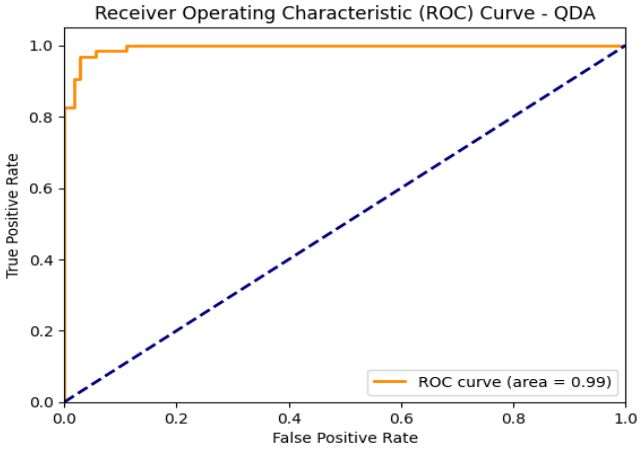

Fig 7: ROC Curve - QDA Model

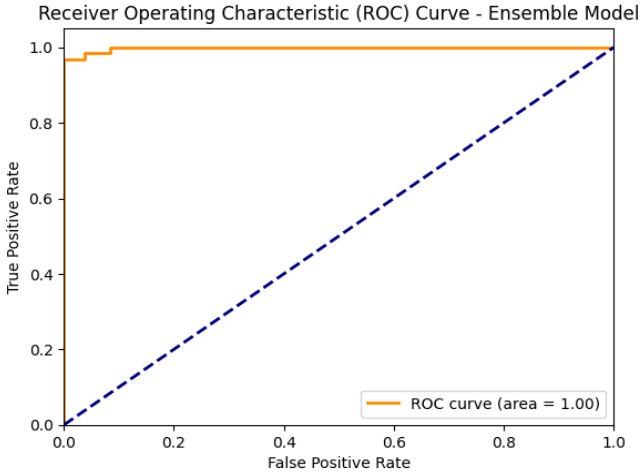

Fig 8: Model ROC Curve - Ensemble Model

V. Conclusion

This paper has discussed the use of several machine learning algorithms (LightGBM, CatBoost, Multi-Layer Perceptron (MLP), Quadratic Discriminant Analysis (QDA), and an Ensemble model) on Breast Cancer Wisconsin (Diagnostic) dataset in binary classification of benign or malignant tumors. Systematic preprocessing, feature selection, and model evaluation allowed the research to prove that all models performed well, with the accuracy and AUC-ROC values larger than 94 per cent and 0.99, respectively, for the best models. It was observed that the MLP had the best accuracy (0.9825) whereas the Ensemble model had the best AUC-ROC value (0.9981) which indicated that the Ensemble model had the capability of leveraging model strengths to have a robust classification.

The findings can shed light on the possibilities of applying machine learning methods to early detection of breast cancer that can lead to successful treatment planning. Even though the dataset was largely balanced, the class specific evaluation indicated false negatives (the relevant measure in the context of medical diagnosis) were reduced to a minimum by the best models.

Potential future research directions include a larger collection of patient data that covers a wider range of populations, and/or more complete clinical and genetic information, as well as the use of explainable AI in making the tool easier to read by medical experts. In general, this study contributes to the efficacy of machine learning as an effective decision-support tool in the field of oncology with a robust indication of increasing the accuracy of diagnoses and patient outcomes.

Future research will seek to confirm the results across more, multi-institutional cohorts and add more clinical and genomic characteristics to enhance the external validity. We intend to use explainability mechanisms (e.g. SHAP in case of tree models, and Grad-CAM in case of any image-based pipeline) to promote clinician interpretability as well. Finally, adaptive ensemble weighting and future clinical assessment will be studied to minimize false negatives on real-world deployment.

REFERENCES

[1] Mangasarian, O.L., and Bennett, K.P. (1992). Discrimination of Two Linearly Inseparable Sets Using Robust Linear Programming. Software and Optimization Techniques, 1, 23–34.

[2] Wolberg, W.H., Mangasarian, O.L., & Street, W.N. (1995). Using fine-needle aspirates and machine learning algorithms, breast cancer can be diagnosed. 163–171 in Cancer Letters, 77(2–3).

[3] Wolberg, W.H., Mangasarian, O.L., and Street, W.N. (1993). extraction of nuclear features for the diagnosis of breast tumors. 861–870 in IS&T/SPIE 1905.

[4] Ting, I.H., Das, R., Zomorodi-Moghadam, M., and Abdar, M. (2015). Evaluation of Classification Algorithms' Performance on the Breast Cancer Dataset. Unlocking Informatics in Medicine, 2, 8–17.

[5] Güneş, S., and K. Polat (2007). Using a least squares support vector machine, breast cancer can be diagnosed. 694– 701 in Digital Signal Processing, 17(4).

[6] Pal, S., and Chaurasia, V. (2017). A new method for detecting breast cancer that makes use of data mining tools. Computer and Communication Engineering: An International Journal of Innovative Research, 5(1), 1–7.

[7] Bahaj, M., and Khourdifi, Y. (2018). utilizing machine learning methods to forecast the recurrence of breast cancer. 104–113 in Procedia Computer Science, 127.

[8] Sahoo, G., Sarma, K.K., & Kumar, A. (2019). Machine learning algorithms are used to diagnose breast cancer. 8(4), 128–132, International Journal of Engineering and Advanced Technology.

[9] G. Murtaza and colleagues (2019). Research problems and state of the art in deep learning-based breast cancer categorization using medical imaging modalities. Medical Artificial Intelligence, 102, 101758.

[10] Al Moatassime, H., Asri, H., Mousannif, H., & Noel, T. (2016). predicting and diagnosing breast cancer risk using machine learning algorithms. 1064–1069 in Procedia Computer Science, 83.

[11] AI Dermatologist: Skin Cancer Detection with Deep Learning Technique," in Proceedings of the 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), June 2024, by A. U. H. Rupak, R. Ghosh, S. S. Khan, M. F. Ishrak, A. P. Alo, and M. S. H. Sabuj. ResearchGate doi: 10.1109/ICCCNT61001.2024.10723914.

[11] M.Z. Alom and associates (2019). A cutting-edge survey on deep learning for the processing of medical images. Access, IEEE, 7, 148010–148027.

[12] I.H. Sarker (2021). Algorithms, practical uses, and future research areas in machine learning. 2, 160; SN Computer Science.

[13] Li and colleagues (2020). Gene expression profiles are used in machine learning methods for the diagnosis of breast cancer. BMC Bioinformatics, 21, 1–13.

[14] Wishart, D.S., and Cruz, J.A. (2006). Machine learning applications for cancer prognosis and prediction. Informatics for Cancer, 2, 59–77.

[16]. R.C. Deo (2015). medical machine learning. Circulation, 1920–1930, 132(20).

[17] Husain, W., and Jothi, G. (2015). A review of data mining in healthcare. Computer Science Procedia, 72, 306–313.

[18] Deep Learning-Based Alzheimer's Disease Detection Using 2D MRI Images," by A. Hanzala, M. S. Rahman, S. S. Khan, M. M. Rahman, M. Sajjad, and A. K. Gupta, in Proceedings of the 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), June 2024, doi: 10.1109/ICCCNT61001.2024.10725482.

[18] Zhang, Z. (2016) (20). K-nearest neighbors: an introduction to machine learning. Translational Medicine Annals, 4(11), 218

[19] **Dataset Link:** https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data?utm_source