http://dx.doi.org/10.12785/src/1570967447

Enhancing Breast Cancer Detection through Machine Learning: A Comparative Analysis of Logistic Regression and Linear Discriminant Analysis

Carlo V. Belvis
Department of Electronics Engineering
Batangas State University
Batangas City, Philippines
21-06118@g.batstate-u.edu.ph

Claire Ann R. Javierto
Department of Electronics Engineering
Batangas State University
Batangas City, Philippines
21-07562@g.batstate-u.edu.ph

Ma. Bernadette O. Santos
Department of Electronics Engineering
Batangas State University
Batangas City, Philippines
21-01242@g.batstate-u.edu.ph

Crizelle R. Datinggaling
Department of Electronics Engineering
Batangas State University
Batangas City, Philippines
21-05999@g.batstate-u.edu.ph

Christian Van R. Rivera
Department of Electronics Engineering
Batangas State University
Batangas City, Philippines
21-07960@g.batstate-u.edu.ph

***Abstract*— This paper intends to analyze classification models that can predict whether breast cancer is cancerous or noncancerous through the attributes of the selected dataset with the application of machine learning algorithms. The high-dimensional Breast Cancer Wisconsin (Diagnostic) Dataset is reduced through principal component analysis (PCA). Six major components were extracted with a cumulative variance of 91.063%. The selected classifier models that were evaluated were logistic regression and linear discriminant analysis (LDA). After training and testing, the model's performance was evaluated using a confusion matrix where metrics such as accuracy, precision, recall, and F1-score were calculated. Experimental results obtained a commendable prediction with 99.7% accuracy for logistic regression and 95.4% accuracy for LDA. The LDA model was observed to have greater misclassified diagnoses than logistic regression. Thus, it is inferred that in binary classification problems like breast cancer detection, logistic regression produces more promising predictions.**

Keywords—breast cancer detection, machine learning, principal component analysis, logistic regression, linear discriminant analysis

I. Introduction

As a formidable adversary to human health, cancer has claimed millions of lives worldwide. Within this complex landscape of malignancies, breast cancer has emerged as one of the most prevalent and concerning health issues, particularly among women. The development of breast cancer is linked to specific cell changes in the breast. Breast cancer is a malignant tumor arising in the breast tissue, wherein the breast cancer cells originate from the milk ducts. It also has the potential to spread through blood vessels and lymph vessels, in such instances breast cancer disperse throughout different parts of the body. The classification of breast cancer lies in the characteristics of the cell. Malignant often show irregular shapes, while benign tend to maintain an appearance similar to a normal cell. Having a deeper understanding of its characteristics greatly contributes to detection and diagnosis. It is considered to be a significant global health concern wherein in 2020, breast cancer took approximately 685,000 lives from women worldwide. Alarming cases of breast cancer have been recorded globally, as incidence rates continuously arise especially from less-developed regions with almost two-thirds of recorded deaths. [1] reveals that breast cancer constitutes 25% of all cancer cases detected in women worldwide. The Philippines exhibits a similar pattern wherein, in 2015, a report by the Philippine Cancer Society states that 20,267 new cases of breast cancer, representing 33% of all cancer cases [2]. Even more concerning, an estimated 7,384 fatalities were attributed to breast cancer in the same year, making it the third most common cause of cancer-related deaths. According to [3], traditional approaches to monitoring and diagnosing diseases rely on identifying specific signal characteristics by human observers. As the number of patients needing intensive care and monitoring increases, various processes have emerged to compensate for faster analysis towards condition recognition.

Furthermore, medical advances lessen the number of incidences and deaths from breast cancers as research gained traction from years of data from the medical and scientific fields. Obtaining the necessary data and information is crucial for patients dealing with breast cancer. However, the obtaining process of data is a great challenge. Advanced tests are available to help identify the best clinical trials for patients, but still complicated as tissue samples need to be sent to different laboratories for testing. Advanced and highly developed healthcare systems are currently the best practice for early detection and diagnosis of breast cancer. More developed countries tend to exhibit higher survival rates compared to less-developed countries with their early detection strategies and early diagnosis. The Philippines poses the same challenge as the country has limited access to healthcare. However, it does not mean that it cannot be resolved as further research has opted for machine automation for breast cancer detection using data analysis and machine learning. Decades of data from cancer patients have made a possibility for machine learning to create breast cancer detections possible. Such significant processes have paved a new way to define, diagnose, and treat breast cancer patients with less variation from human error. Progression in breast cancer treatment opens the application in developing countries that are lacking in treatment facilities, such as the Philippines.

The utilization of machine learning algorithms for healthcare and dimensionality reduction is an emerging field of inquiry in the Philippines and growing countries. Different medical fields, universities, and research organizations in the Philippines are studying and researching the applications of machine learning algorithms for healthcare. An analysis from [4] shows that their model still exhibits promising results even when Principal Component Analysis is incorporated into their classification models, providing efficiency for less data reading. This is supported thoroughly in the context of breast cancer detection, wherein the results from data mining for classification purposes show significant promise in breast cancer detection [5]. The study of [6] researched the most effective approach for predicting breast cancer. This involved a thorough investigation into methods for simplifying the feature space and exploring the application of Principal Component Analysis (PCA) to lessen the number of dimensions. This is supported by a study conducted by [7], wherein together with Principal Component Analysis, they suggested a hybrid method for diagnosing breast cancer, which involves first reducing the high-dimensional feature space using Linear Discriminant Analysis (LDA) and then applying the resulting reduced set of features to a Support Vector Machine (SVM). Their approach demonstrated an accuracy of 98.82%.

On the other hand, utilizing logistic regression is the best method to categorize how breast cancer will be categorized under benign or malignant as the latest study conducted by [8] claimed that it is effective in identifying malignancy as it can assess the probability of a tumor being malignant based on multiple predictor variables, including tumor size, shape, and various characteristics. It is used to categorize tumors, distinguishing between those that are benign and those that are malignant, based on the tumor's characteristics. If the predicted probability exceeds a common threshold, typically set at 0.5, the tumor is classified as malignant; otherwise, it is classified as benign.

With the growing cases of breast cancer the Philippines experiences, there is a call for advancing approaches to breast cancer diagnosis, specifically by integrating machine learning algorithms. This calls forth the present study to understand the issues related to feature selection, dimensionality reduction, and optimizing machine learning algorithms with its advancement in medical imaging and artificial intelligence to be a great help, particularly in healthcare. The present work intends to compare the classification models, specifically the logistic regression and LDA, through testing it to the reduced dataset to distinguish which provides a more precise prediction.

II. Methodology

A. Breast Cancer Wisconsin Dataset

The dataset used in the present study, often referred to as the WDBC dataset [9], is a well-known collection of medical data from the UCI Machine Learning Repository that is used in the world of machine learning and data analysis to help identify breast cancer tumors as either malignant (cancerous) or benign (non-cancerous) based on various features extracted from images of breast mass lesions. The class distribution contains a total of 569 instances, 212 of which have been classified as malignant (1) and 357 as benign (0), providing important data for training machine learning models to categorize tumors using the following features. This dataset contains two categorical data types which are the identification and diagnosis of the patients. Following this, it also includes continuous data which are the nuclear features such as radius, texture, perimeter, area, smoothness, compactness, concavity, symmetry, and fractal dimension, that are numerically modeled and are commonly employed for training machine learning models to classify tumors. These shape features within the dataset are extracted from digitized images of breast cell samples, which are taken using a fine needle aspirate (FNA) and then examined under a microscope. The mean value, extreme value, and standard error were computed for each image resulting in thirty (30) features.

These approaches help the researchers tell the difference between malignant and benign tumors, which is crucial for diagnosis. However, dealing with so many features can be complicated, thus, the use of techniques to reduce the complexity and ease data analysis is performed systematically. This requires the utilization of different methods or techniques to help identify the patterns that can distinguish between cancerous and non-cancerous cases.

B. Machine Learning Algorithms

Machine learning algorithms are essential in exploiting their potential to improve lives by accurately and precisely predicting breast cancer. These methods can analyze an extensive quantity of data to identify correlations or patterns and insights.

1. Dimensionality Reduction

   By reducing the number of features (or dimensions) while maintaining the most significant and original data or information, dimensionality reduction is utilized in data analysis and machine learning to simplify complex datasets. This method is frequently used to enhance the accuracy and performance of the predictive models by making the data more manageable. In this light, principal component analysis (PCA) emerged as a key technique for attaining this dimensionality reduction. The reduction involves extracting principal components which are the directions where the data varies, creating a new set of uncorrelated variables to identify the relationships between the data features.

   In the study, it is applied to the high dimensional WDBC dataset where the original data is transformed using a covariance matrix. To find the major components, the application of numerical analysis i.e. eigendecomposition, wherein the eigenvectors and eigenvalues are computed corresponding to the amount and direction of the variance of the data, is employed. In most cases, only selected factors are retained, reducing the complexity of data. The final step is to make the initial data into lower dimensional data based on the given principal components that make a collection of variables that fully characterize the data.

Dimensionality reduction is being utilized in this study to enable the transformation of high-dimensional data into lower-dimensional data. With the data present, this approach is utilized in this study to determine whether breast cancer is benign or malignant. To make the data easier to understand and visualize, and enhance the accuracy of generalization of the performance of the predictive models, the researcher effectively reduced the dimensionality of a huge amount of information regarding breast cancer patients.

2. Logistic Regression

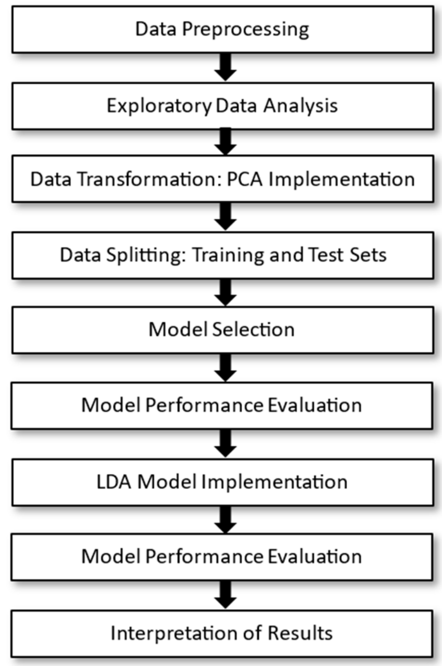
Logistic regression is a linear classification model and a type of regression analysis that is particularly well-suited for binary classification problems, where it aims to determine whether an outcome falls into one of two categories, such as malignant (cancerous) or benign (non-cancerous). The model quantifies the likelihood of a data point being assigned to a given class and is notably useful for its simplicity and comprehensibility. Utilizing Logistic regression in this study may calculate the outcome of breast cancer based on its given features. Tumor size, form, location, and other various characteristics and features are considered for tumor categorization. Once the model has been fitted, it is possible to create predictions for new data using the predictor values.

3. Linear Discriminant Analysis (LDA)

One of the supervised learning algorithms is the Linear Discriminant Analysis (LDA), which is often used for dimensionality reduction and classification. Even so, dimensionality reduction does have caveats; there will be a minor margin of error in the accuracy. In this study, LDA is used to discern malignant and benign tumors by distinction of unique features that make up the tumors. Its main objective is to find and identify the linear combinations of the features to differentiate each class and make it easier to classify the data. By modeling the distribution of features and computing for the means and variances in each separate class, it can create discriminant functions to classify new data samples. In this case, LDA helps determine if a patient is at risk by analyzing traits such as tumor size, shape, and texture. It gives a definitive and fail-safe method for early diagnosis and treatment planning, improving the patient's survival rate and overall healthcare management. Its ability to leverage statistical properties of data makes it an important component of breast cancer classification models, ultimately contributing to more accurate diagnoses and better patient care.

C. Workflow

The structural procedure performed in this study to build and test a classification model to make predictions about the type of cancer is demonstrated in Figure 1. The WDBC dataset is imported in the program in Python language derived from an automated system created to classify whether a diagnosis is malignant or benign [10]. First, the target values were defined as the diagnosis of whether the breast cancer is benign (0) or malignant (1). The intricacies of the data were explored through visualizing the distribution and trends of the features or predictors and their relationship with the target variables.



In order to simplify the complexity of the dataset, data transformation is performed where PCA is employed. The original data were converted into a covariance matrix to correlate the features while lowering its dimension with capturing most of the important information. The extraction of the principal components is achieved through eigendecomposition. Here, the eigenvectors and eigenvalues of the matrix were calculated by the program to represent the directions and amounts of the maximum variance of the transformed data. The present study used a cumulative explained variance of 92 percent. After transforming, the train test splitting of data into 75 percent training subset and 25 percent test subset is applied. These data were then used to fit the training models i.e. Logistic Regression, Gradient Boosting, Decision Tree, Random Forest, Linear SVM, and K-Nearest Neighbors. The selected model is evaluated using a confusion matrix which will then be applied to calculate the evaluation metrics such as accuracy, precision, recall, F1-score, and receiver operating characteristic curve.

Consequently, the data is used to train the LDA model to compare results with the predictive accuracy of the Logistic regression classifier. In assessing the performance of the linear discriminant analysis (LDA) model, several key metrics are considered, such as predictive accuracy, type I and II errors, false negatives, false positives, and an error rate calculation. The model's predictive accuracy was determined to be 95.4 percent which indicates the proportion of correct predictions out of the total dataset. However, within this accuracy, 4.6 percent of type

I and II errors were found, of which 4.45 percent were false positives and 0.15 percent were false negatives. The percentage of an error rate is calculated as the proportion of misclassified cases (16 out of 455) in relation to the total data set which has an error rate of 3.516 percent.

III. Results

The results were obtained using Google Colab, a free cloud-based Jupyter notebook environment apt for machine learning research. After cleansing the data by removing unnecessary columns i.e. patient ID, and checking for missing values, the target values were encoded. Table I shows the cancer classifications or the target variables used for the algorithm in the study.

TABLE I. Target Variables for the Algorithm

Diagnosis	Data Count	Encoded Value
Malignant	212	0
Benign	357	1

Figures 2 to 4 show the spread of the selected features i.e. texture, compactness, and concave points, which are great indicators of malignancy. The mean texture values displayed were based on the cancer classification (malignant=0 and benign=1). On the other hand, the compactness and concave points are plotted based on their mean values ranging from 0.2 and 0.1 or above for the malignant class, respectively. These presentations suggest that the malignant classification has a relatively higher mean for the mentioned attributes.

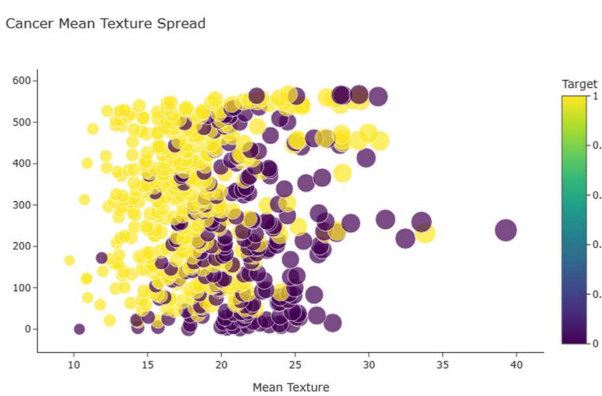


Fig. 2. *Spread of the Mean Texture values. The intensity of the color is ba on the encoded target variables.*

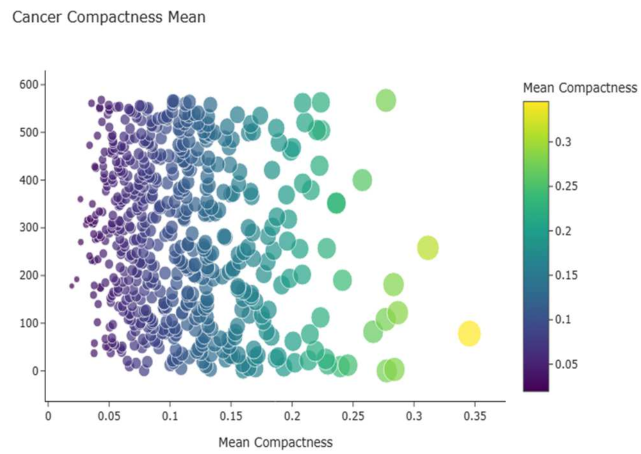


Fig. 3. *Spread of the Mean Compactness values.*

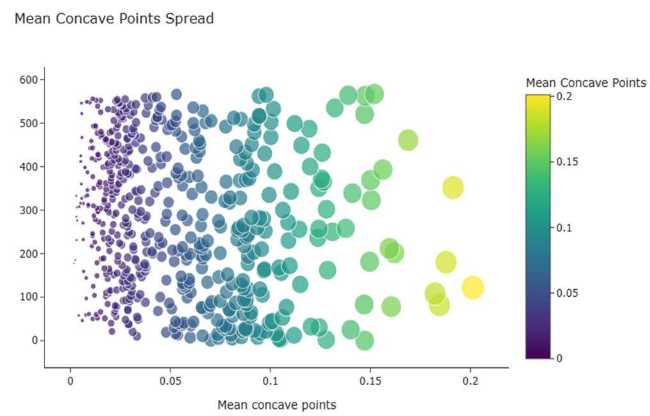
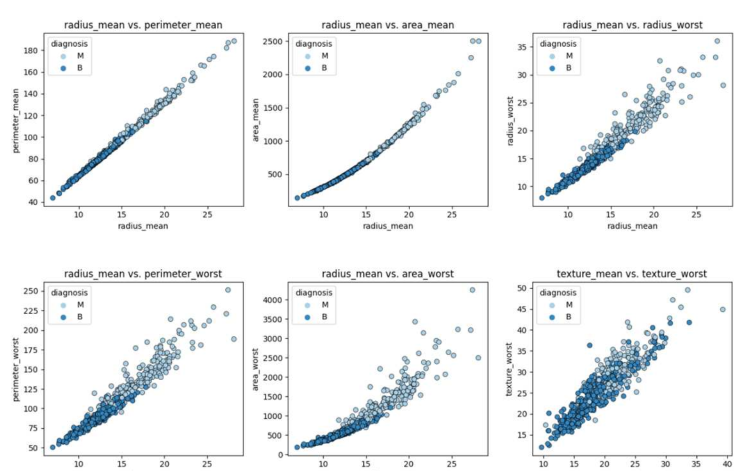


Fig. 4. *Spread of the Mean Concave Points values.*

The predictors exhibiting a high positive correlation (⩾ 0.90) were extracted from the presented heatmap. Figure 5 depicts the consistent trend of linear relationships between the predictors, with 0.90 or higher correlation, which are observed to correspond to an increased likelihood of a malign diagnosis.



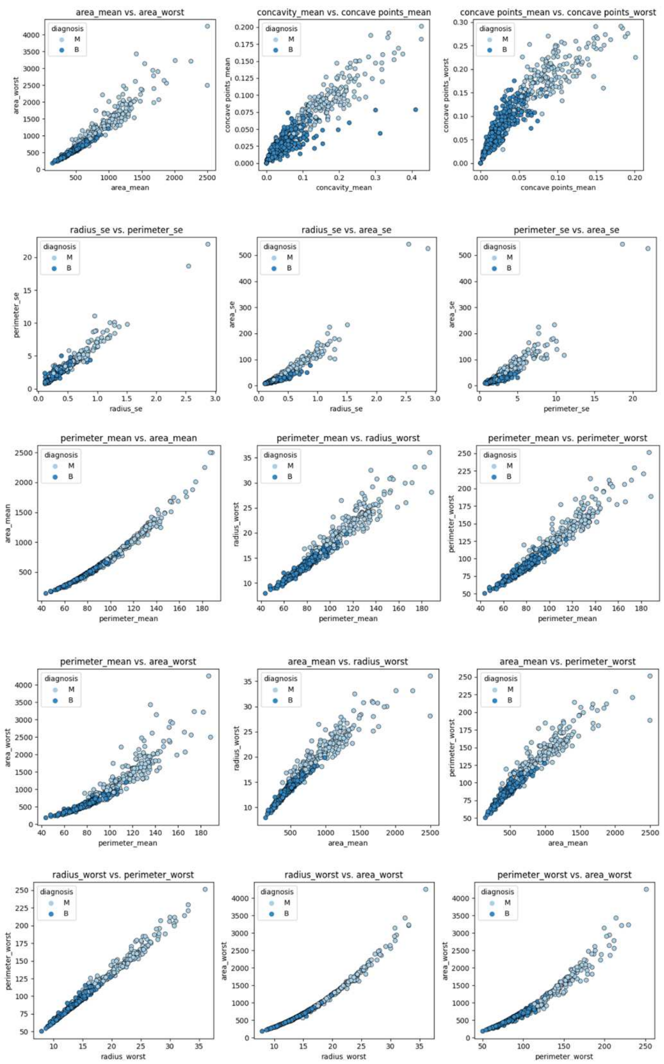


Fig. 5. *Scatter Plot of the Highly Correlated Features (Correlation Coefficient ⩾0.90).*

Transforming the data followed the distribution visualization of the different features. The high-dimensional dataset is reduced by utilizing principal component analysis (PCA). The cumulative explained variance used for this study is 92% to capture a significant portion of the variability of the data. The eigenvectors and eigenvalues were computed and subsequently, a cumulative percentage of 91.063 was achieved considering six (6) principal components as displayed in Figure 6.

The spread of the transformed lower-dimensional data defined by the retained components is projected in Figure 6. It is observed that the reduction reaffirmed the previous trend observed from the distribution visualizations of the features or predictors where higher values are highly associated with a greater likelihood of malign diagnosis.

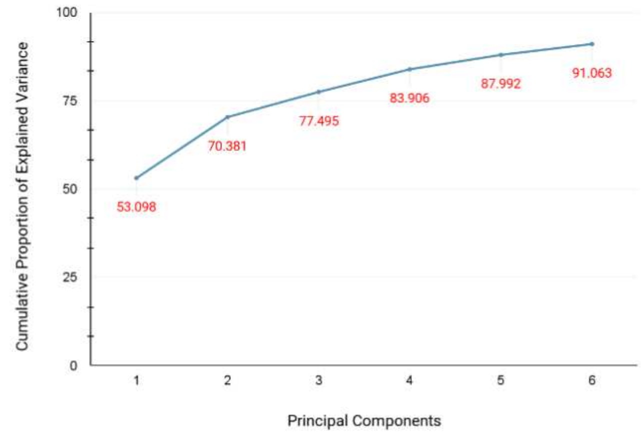


Fig. 6. *Cumulative Percentage of Explained Variance of each Principal Component.*

The spread of the transformed lower-dimensional data defined by the retained components is projected in Figure 7. It is observed that the reduction reaffirmed the previous trend observed from the distribution visualizations of the features or predictors where higher values are highly associated with a greater likelihood of malign diagnosis.

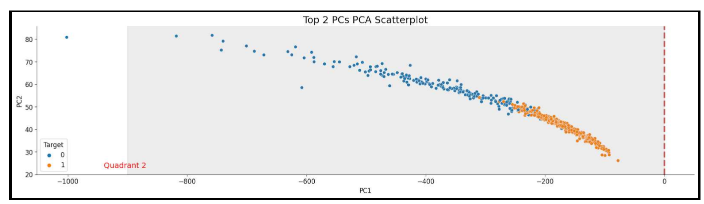


Fig. 7. *Scatter plot of PCA transformed data using First and Second Principal Components.*

After transforming, the data is split into a training and testing set of 75 and 25 percent, respectively to fit and evaluate the training models. Table II shows the accuracy scores of the classifiers that served as a basis for the selection of the model for the transformed training data. The first four classifiers scored 1.000 which indicates a promising classification. These baseline scores, however, are only the initial point of evaluation without hyperparameter tuning.

TABLE II. BASELINE SCORES OF THE SELECTED CLASSIFICATION MODELS

Classification Model	Accuracy Score
Logistic Regression	1.000
Gradient Boosting	1.000
Decision Tree	1.000
Random Forest	1.000
Linear SVM	0.993
K-Nearest Neighbors	0.958

Apparently, the study used the logistic regression model to perform predictions on the transformed data. Figure 8 presents the model accuracy of logistic regression through a confusion matrix. It suggests that the model exhibited a commendable prediction of 99.7 percent. The remaining portion (0.3 percent) of the prediction are false positives where the model made diagnoses as malignant instead of a non-cancerous tumor. The classification report of this model is further depicted in Table III which implies that 1 case out of the training set of 426 cases is misclassified with an error rate of 0.235.

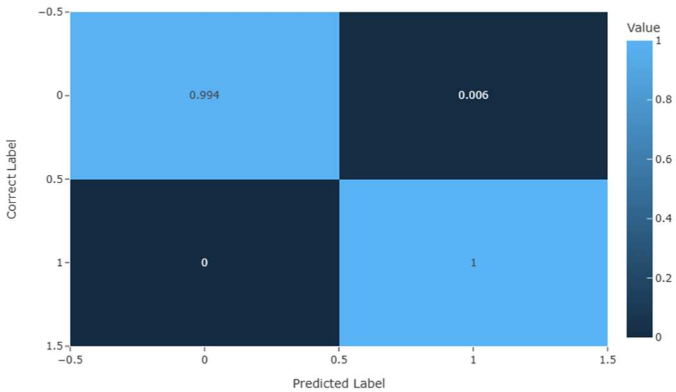


Fig. 8. *Model Performance Evaluation of Logistic Regression Model.*

TABLE III. LOGISTIC REGRESSION CLASSIFICATION REPORT

	Precision	Recall	F1-score	Support
Malignant	1.00	0.99	1.00	158
Benign	1.00	1.00	1.00	268
Accuracy			1.00	426
Macro ave	1.00	1.00	1.00	426
Weighted Ave	1.00	1.00	1.00	426

In line with this, Linear Discriminant Analysis (LDA), a supervised learning algorithm, was implemented on the dataset to compare results with the selected model. Figure 9 showcases the predictions performed by LDA. The model provided a predictive accuracy of 95.4 percent. However, within this accuracy, 4.6 percent of type I and II errors were also observed. It is observed that 0.15 percent of the data were false negatives and 4.45 percent were false positives. The misclassified diagnoses are 16 cases out of 455 which has an error rate of 3.516.

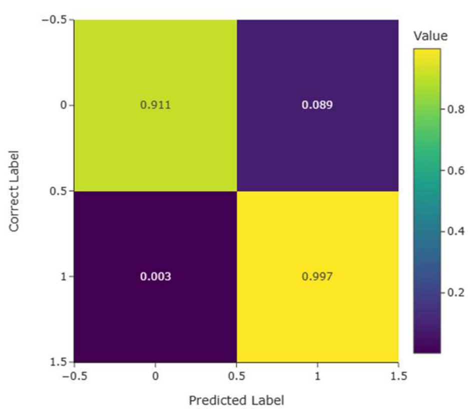


Fig. 9. *LDA Prediction Performance Evaluation.*

From these results, it is observed that although it can be employed in classification problems, Linear Discriminant Analysis (LDA) is unable to produce a new axis that allows the classifications to be linearly separable as it generates a high percentage of error, unlike the logistic regression model.

IV. DISCUSSION

The usage of machine learning algorithms for healthcare has ethical considerations. According to [11], since there is a limitation in the prediction of ML in terms of future machine actions it can't be held accountable; it may have errors if it doesn't have sufficient information needed which means that it should only be a basis or a guide. As a way to avoid bias, fairness must also be put into consideration. Because algorithms solely depend on past data, there might be a case where a specific nation has fewer healthcare medical facilities thus creating an impact on the data. In terms of privacy, data is accumulated through EHR (Electronic Health Record) which is the main cause of privacy concerns because of the way the data is being collected without proper consent. Another consideration is the access to information of the stakeholders such as that it should be traceable, monitored, and limitation wise of the system. Lastly is the effect on the jobs, as the demand for work shifts to other areas due to machine learning.

For the data splitting, there are generated scores that help to identify the grounds that will be set in terms of model selection. Despite getting an ideal classification score, the downside is that it lacks a hyperparameter tuning which reduces the number of layers and parameters that support its credibility or higher accuracy rate. In the transformed data using Linear Regression, a confusion matrix is utilized and based on the results, the error rate is considerably low which is 0.235 thus this shows that the model is reliable.

The logistic regression model is a more efficient choice in breast cancer detection in creating a diagnosis of whether the tumor is malignant or benign because based on the results, linear discriminant analysis (LDA) shows a higher value of error. According to [12], logistic regression is best used between analyzing two variables which are the dependent and independent variables which it gives a predictor or explanatory variable thus helping it to generate conclusions regarding the data. On the other hand, no premise could be produced in LDA in the case of the explanatory variables. Both LDA and Logistic regression have their strengths and weaknesses. In the case of LDA, it produces a higher efficiency rate when utilized for 5 or more variables and a lower efficiency rate when used for binary cases. With that, in LDA the efficiency and the number of categories are proportional. In contrast to LDA, utilizing logistic regression with 2-3 variables will result in higher accuracy which means that it is inversely proportional since a lower variable indicates higher efficiency. In the case of this study, with binary variables that are malignant and benign, it is preferable to use logistic regression.

The study used a data set thus; it must be put into consideration the possibility of bias in it. In that case, it must be examined to maintain the accuracy as well as the study's fairness. In addition to that, since the findings that were used in the study are from a single data set which is Wisconsin, there is a possibility that the implication of the result may not be applicable in other contexts.

Based on the results, it is recommended to provide additional research for the improvement of the study such as providing hyperparameter tuning. As stated by [13], there are significant large differences between the model parameter and hyperparameter that greatly affect the efficiency. In the model parameters, its modification solely relies on the nature of the data which means it is uncontrollable. However, in hyperparameters, few parameters are employed to manage the behavior of the model to give its best performance.

V. Conclusion

In conclusion, using (PCA) to reduce the dataset's complexity in the study proved effective and beneficial. It preserved the essential patterns and relationships between the features from the original data, making it a valuable approach for simplifying and working with the dataset. Its flaws include the potential blurring of diagnostic boundaries as it condenses high and low-value ranges, reduced interpretability of the transformed principal components, and the threat of losing crucial information due to strong relationships between variables, primarily through feature discretization [14]. These limitations emphasize that while PCA is effective for dimensionality reduction, there may be more suitable approaches, particularly in cases where maintaining clear distinctions and interpretability of features is crucial.

It is observed that, although Linear Discriminant Analysis (LDA) can be employed in classification problems, it results in a high percentage of errors, unlike the logistic regression model. In many cases, having five categories is sufficient, however, the advantages of using Logistic Regression (LR) become particularly pronounced when dealing with only two or three categories. It offers flexibility to handle non-linear relationships and does not assume specific data distributions, making it well-suited to the present study or work. The impact of categorizing the covariates depends on their relationship with the outcome variable. This becomes especially relevant in binary classification tasks, where cancer is identified as benign or malignant. However, logistic regression shines when the assumptions of LDA do not hold, offering reliable results no matter how the data is distributed [12]. While logistic regression aims to find a linear decision boundary that best separates classes, LDA focuses on maximizing class separability by projecting data onto new axes.

The results and findings of this study have several implications and practical applications in breast cancer detection. This review has the potential to serve as a vital reference for researchers of all experience levels interested in learning-based breast cancer classification across multiple imaging modalities, including those just starting in the field. These findings can guide future researchers and some clinical practices, improving the accuracy and reliability of classification models in determining whether a tumor is malignant or non-cancerous.

Acknowledgment

The authors wish to extend their gratitude to the University of Bahrain for the invaluable support received throughout this research endeavor. The university's dedication and provision of resources significantly contributed to the successful completion of this study.

References

[1] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent, and A. Jemal, "Global cancer statistics, 2012," *CA: A Cancer Journal for Clinicians*, vol. 65, no. 2, pp. 87–108, Feb. 2015, doi: https://doi.org/10.3322/caac.21262.

[2] A.V. Laudico, M.R. Mirasol-Lumague, V. Medina, C.A. Mapua, F.G. Valenzuela, E. Pukkala, "Philippine cancer facts and estimates," in *Philippine Cancer Society; Manila,* 2015. [Online] Available: http://www.philcancer.org.ph/wp-content/uploads/2017/07/2015-PCS-Ca-Facts-Estimates_CAN090516.pdf.

[3] E. D. Übeyli, "Implementing automated diagnostic systems for breast cancer detection," Expert Systems with Applications, vol. 33, no. 4, pp. 1054–1062, Nov. 2007, doi: https://doi.org/10.1016/j.eswa.2006.08.005.

[4] A. M. M. . Baes, A. J. M. . Adoptante, J. C. A. . Catilo, P. K. L. . Lucero, J. F. P. Peralta, and A. L. P. de Ocampo, "A Novel Screening Tool System for Depressive Disorders using Social Media and Artificial Neural Network", Int J Intell Syst Appl Eng, vol. 10, no. 1, pp. 116–121, Mar. 2022.

[5] S. A. Mohammed, S. Darrab, S. A. Noaman, and G. Saake, "Analysis of Breast Cancer Detection Using Different Machine Learning Techniques," Data Mining and Big Data, pp. 108–117, 2020, doi: https://doi.org/10.1007/978-981-15-7205-0_10.

[6] S. J. S. Gardezi, A. Elazab, B. Lei, and T. Wang, "Breast Cancer Detection and Diagnosis Using Mammographic Data: Systematic Review," Journal of Medical Internet Research, vol. 21, no. 7, p. e14464, Jul. 2019, doi: https://doi.org/10.2196/14464.

[7] D. A. Omondiagbe, S. Veeramani, and A. S. Sidhu, "Machine Learning Classification Techniques for Breast Cancer Diagnosis," IOP Conference Series: Materials Science and Engineering, vol. 495, p. 012033, Jun. 2019, doi: https://doi.org/10.1088/1757-899x/495/1/012033.

[8] C. Srividya, M. Madhubala, U. Neelaveni, A. R. Kommareddy, and D. Davuluri, "BREAST CANCER DETECTION USING LOGISTIC REGRESSION," European Chemical Bulletin, vol. 12, no. 5, pp. 3259–3267, doi: https://doi.org/10.48047/ecb/2023.12.si5a.0214.

[9] "Breast Cancer Wisconsin (Diagnostic) Data Set," *www.kaggle.com.* https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

[10] "PCA and LDA Implementation," kaggle.com. https://www.kaggle.com/code/bhatnagardaksh/pca-and-lda-implementation/notebook

[11] K. M. S. Staff, "5 Ethical Considerations for Machine Learning in Healthcare," *KMS Healthcare*, Mar. 23, 2023. https://kms-healthcare.com/5-ethical-considerations-for-implementing-machine-learning-in-healthcare/#:~:text=Ethical%20Considerations%20for%20Machine%20Learning%20Applications%20in%20Healthcare (accessed Dec. 15, 2023).

[12] M. Pohar, M. Blas, and S. Turk, "Comparison of logistic regression and linear discriminant analysis," Advances in Methodology and Statistics, vol. 1, no. 1, pp. 143–161, Jan. 2004, doi: https://doi.org/10.51936/ayrt6204.

[13] S. Pandian, "A Comprehensive Guide on Hyperparameter Tuning and its Techniques," Analytics Vidhya, Feb. 22, 2022.

[14] W.-M. Lee, "Using Principal Component Analysis (PCA) for Machine Learning," Medium, Feb. 10, 2023. https://towardsdatascience.com/using-sprincipal-component-analysis-pca-for-machine-learning-b6e803f5bf1e.