Enhancing Breast Cancer Classification Using Principal Component Analysis and Linear Regression: A Comprehensive Evaluation

Yohannes Sida and Dr.Venkat
Jain University, Bangalore, 562112, India

Abstract. This paper proposes a sophisticated breast cancer classification system that integrates Principal Component Analysis (PCA) and Linear Regression (LR) to discern between malignant and benign tumors. The study meticulously explores the working principles of LR and PCA, employing a dataset enriched with ten carefully selected attributes capturing essential tumor characteristics. Extensive preprocessing and exploratory analyses are conducted, showcasing the system's robustness. The LR model, enhanced by PCA for feature selection, achieves remarkable precision, recall, and F1-score metrics. The visualizations, including bar charts and radar charts, vividly illustrate the precision, recall, and F1-score for each class, providing a comprehensive understanding of the model's performance. Notably, the system's overall accuracy is reported at 99%, demonstrating its efficacy in accurate breast cancer classification. This research contributes a valuable framework for leveraging machine learning techniques in medical diagnostics, offering insights into feature importance, model interpretability, and achieving high accuracy in breast cancer classification.

Keywords : Linear Regression(LR), Principal Component Analysis(PCA), Breast Cancer Classification

1. Introduction

In the realm of breast cancer classification, the process of distinguishing between benign and malignant tumors represents a formidable challenge, demanding considerable time and resources. The intricacies of this task are exacerbated by the fact that discerning features indicative of malignancy can be subtle, often requiring the expertise of multiple radiologists and oncologists. The human eye, despite its remarkable capabilities, sometimes struggles to definitively identify the subtle nuances that differentiate between benign and malignant tumors [1]. Tragically, breast cancer is usually diagnosed at an advanced stage, typically stage III or IV. The cancer has now spread outside of the

original tumor's boundaries, invading neighboring cells but not yet distant organs. The prognosis for those who are affected by this delayed diagnosis is complicated and reduces the efficacy of available treatments. The serious implications of late-stage breast cancer diagnoses highlight the urgent need for the development of more effective and precise classification techniques [1].

Breast cancer, a perilous affliction primarily affecting women worldwide, poses a significant threat to life. The disease is characterized by the uncontrolled overgrowth of cells in the breast, resulting in the formation of palpable lumps. This overproliferation of cells can lead to the development of tumors, which may be classified into two distinct categories: malignant and benign. The challenge faced by physicians in accurately categorizing these tumors is substantial, given the inherent complexities of breast cancer pathology [2]. One important factor in the development of breast cancer is the transition from benign to malignant. The illness has the capacity to spread, allowing secondary tumors to grow and entering other organs through the lymphatic or blood systems. As such, the capacity to classify tumors quickly and accurately becomes critical for guiding treatment plans and enhancing patient outcomes [2].

Essentially, the categorization of breast tumors represents a critical turning point in the treatment of breast cancer. Improving this process's accuracy and efficiency is crucial for improving our knowledge of the disease's course as well as for prompt intervention. The combination of cutting-edge technologies and machine learning algorithms has the potential to completely transform the way breast cancer is classified as research continues, providing hope for earlier diagnosis and more efficient treatment approaches

2. Related Works

Heland et al.,(2020): Heland et al. researched the resource-intensive task of tumor classification by developing a model utilizing the J48 decision tree algorithm. This model, based on the analysis of cell features extracted by the X-cyt program, demonstrated a strong 95 percent accuracy rate according to the confusion matrix. The article underscores the challenges in diagnosing cancers, particularly at advanced stages, emphasizing the importance of early detection. Leveraging machine learning, specifically decision tree algorithms, the researchers aimed to uncover hidden patterns that could significantly improve the efficiency of tumor classification, contributing to the early identification of malignancies and potential life-saving interventions.

Kalaiyarasi et al., (2020): Kalaiyarasi et al. addresses the imperative need for early breast cancer prediction, emphasizing the emotional and physical toll on individuals. Employing supervised learning algorithms—logistic regression, Support Vector Machine (SVM), and K Nearest Neighbors (KNN)—the study introduces a novel stacking ensemble method to combine these algorithms, aiming to enhance classification performance. Results entail a comprehensive evaluation comparing the efficiency of the stacking ensemble approach against individual

algorithms. The research contributes to healthcare by offering a potential advancement in breast cancer classification using machine learning, providing insights into the comparative effectiveness of the proposed method.

Viswanatha et al., (2023): Viswanatha et al. utilize logistic regression to classify breast tumors (benign or malignant) using the Wisconsin Diagnostic Breast Cancer dataset. The methodology involves dataset exploration, visualization, and logistic regression model training and evaluation with accuracy metrics. Emphasizing the importance of accurate classification for early detection and treatment, the study highlights logistic regression's efficacy and interpretability. The model's simplicity and computational efficiency make it an attractive choice for breast cancer classification, particularly when interpretability is crucial. The research underscores the significance of preventing unnecessary invasive procedures through accurate benign tumor classification, offering reassurance to patients and reducing associated anxieties and potential side effects. Overall, logistic regression proves to be a valuable tool in breast cancer classification for its effectiveness and interpretative insights.

Khairunnahar et al., (2019): Khairunnahar et al. enhanced breast cancer detection accuracy using a modified logistic regression hypothesis with a weighted sigmoid function. Employing gradient descent and advanced optimization, the research explores the weighting factor (β) dependency on features, dataset size, and optimization technique. Results on the WDBC dataset demonstrate a significant accuracy improvement by selecting an appropriate β value. The system achieves notable increases in accuracy, sensitivity, and specificity (97.42%, 0.0258 error rate, 99.44% sensitivity, and 90.74% specificity). This modified logistic regression approach shows promise for improving breast cancer classification accuracy in the preliminary phase of diagnosis.

Ara et al., (2021): Ara et al. researched the global prevalence of breast cancer, underscoring the critical need for early diagnosis to improve patient outcomes. Utilizing the Wisconsin Breast Cancer Dataset from the UCI repository, the research evaluates the performance of various machine learning algorithms in predicting breast cancer types (benign or malignant). Classifiers such as Support Vector Machine, Logistic Regression, K-Nearest Neighbors, Decision Tree, Naive Bayes, and Random Forest are implemented, with Random Forest and Support Vector Machine standing out with an impressive accuracy of 96.5%. They suggest that these high-performing classifiers could be pivotal in developing an automatic diagnostic system for the preliminary diagnosis of breast cancer, offering valuable insights into the potential of machine learning in enhancing the accuracy of breast cancer classification for improved medical interventions.

Murugan et al.,(2017): Mugugan et al. studied the significant issue of breast cancer among women, emphasizing the importance of early detection for improved treatment outcomes. Utilizing data from the UCI Machine Learning Repository (Wisconsin Breast Cancer), the research focuses on classifying cancer types as benign or malignant. Through detailed analysis and prediction based on patient records, the study achieves a classification success rate of 84.14% and a prediction percentage of 88.14%. The results underscore the potential of data-driven

approaches in aiding physicians to make informed decisions about the curability of breast cancer, contributing to enhanced patient care and treatment strategies.

MurtiRawat et al., (2020): MurtiRawat et al. addresses the urgency of Breast Cancer detection, highlighting its severity and the potential for increased mortality rates. They focus on distinguishing between benign and malignant breast cancer using Machine Learning Algorithms, including Logistic Regression, K-nearest neighbors, and Ensemble Learning with Principal Component Analysis. The study achieves notable accuracies of 98.60%, 97.90%, and 99.30%, respectively, through a comparative analysis with other models. Trained and tested on the Wisconsin breast cancer diagnosis dataset, the research underscores the efficacy of these methods in contributing to accurate and early breast cancer diagnosis.

Eroglu et al., (2021): Eroglu et al. investigated the critical need for early diagnosis of breast lesions and distinguishing between malignant and benign cases for effective breast cancer prognosis. Recognizing the significance of ultrasound in diagnosis, the study introduces a hybrid CNN system for breast cancer lesion diagnosis, incorporating Alexnet, MobilenetV2, and Resnet50 models. The hybrid structure involves obtaining and concatenating features from these models to increase the feature set. Subsequently, the mRMR feature selection method is applied to choose the most valuable features, which are then classified using machine learning classifiers like SVM and KNN. Notably, the SVM classifier yields the highest accuracy rate of 95.6%, demonstrating the potential of this hybrid approach in improving the diagnostic accuracy of breast cancer lesions. This contribution is particularly valuable in the context of reducing radiologists' workload and enhancing diagnostic capabilities, especially in settings with a high patient population.

3. Materials and Methods

3.1 Dataset

The Breast Cancer Wisconsin (Diagnostic) Data Set is a widely used dataset in machine learning for the classification of breast tumors as either malignant or benign. It consists of features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. The dataset includes a total of 30 features, encompassing various characteristics of cell nuclei present in the images. For the proposed breast cancer classifier system, 10 attributes have been selected, each providing crucial information for the classification task.

a) radius (mean of distances from the center to points on the perimeter)

b) texture (standard deviation of gray-scale values)

c) perimeter

d) area

e) smoothness (local variation in radius lengths)

f) compactness (perimeter^2 / area - 1.0)

g) concavity (severity of concave portions of the contour)

h) concave points (number of concave portions of the contour)

i) symmetry

j) fractal dimension ("coastline approximation" - 1)

These features capture essential aspects of the tumor morphology, aiding in the accurate differentiation between malignant and benign cases. The chosen attributes offer a diverse representation of tumor characteristics, making them valuable inputs for the machine learning model's training and classification processes.

3.2 Data preprocessing

Prior to building the breast cancer classifier system, an extensive preprocessing phase and exploratory analysis were conducted on the Breast Cancer Wisconsin (Diagnostic) Data Set. The preprocessing involved handling missing values, if any, and ensuring data consistency. Additionally, the dataset was split into features (X) and the target variable (y), where features corresponded to the selected 10 attributes, and the target variable represented the tumor's malignancy status. Feature scaling was applied to standardize the range of the attributes, ensuring that no single feature disproportionately influenced the model.

3.3 Data exploration and visualization

Exploratory data analysis involves assessing the dataset to extract key insights. Initial steps include examining the DataFrame's structure to determine its dimensions. The info() function is utilized to gather information about data types and identify any missing values. Descriptive statistics, such as mean, standard deviation, and quartiles, are generated using the describe() function. Scatter plots are then employed to visually explore the relationship between each feature and the target variable. These plots distinguish between benign and malignant tumors through different marker colors, facilitating a visual understanding of feature value distribution

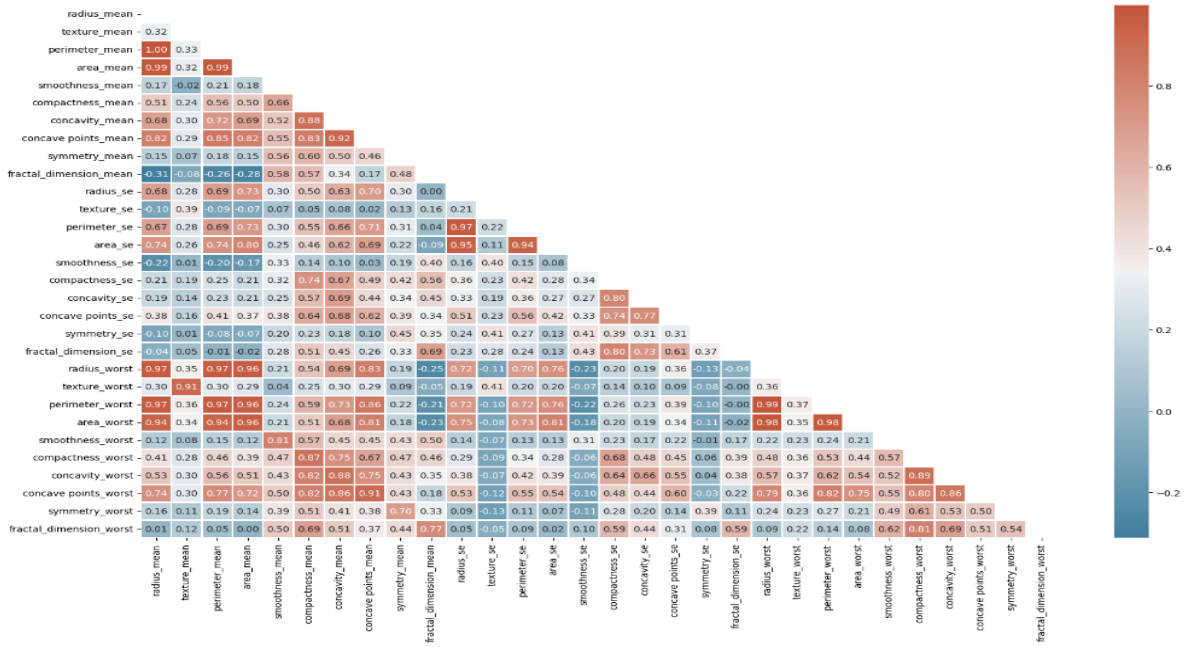

Fig.1 Correlation Heatmap of the features in the Datase

4. Proposed Method

Machine learning algorithms, such as linear regression, have shown promise in classifying tumors. This proposed system aims to optimize LR's performance by incorporating PCA to select essential features and improve overall accuracy.

4.1 Linear Regression

Linear Regression is a foundational statistical method used in machine learning to model the relationship between a dependent variable (output) and one or more independent variables (features or inputs). In the context of breast cancer classification, the goal is to predict whether a tumor is malignant or benign based on certain features. The LR algorithm achieves this by fitting a linear equation to the observed data, allowing it to learn and quantify the relationships between the features and the target variable.

Mathematically, the linear regression equation for a simple case with one independent variable (univariate) is represented as:

$$y = mx + b$$

where y is the dependent variable, x is the independent variable, m is the slope of the line, and b is the y-intercept. In a more complex scenario with multiple independent variables (multivariate), the equation extends to:

$$y = b + m_1 x_1 + m_2 x_2 + \dots m_n x_n$$

The LR model "learns" the optimal values for the coefficients m and b during the training process, minimizing the difference between predicted and actual outcomes. Once trained, the LR model can make predictions on new, unseen data.

4.2 Principal Component Analysis

Principal Component Analysis (PCA) is a dimensionality reduction technique widely used to transform high-dimensional datasets into a lower-dimensional form. The primary goal of PCA is to identify and capture the most critical sources of variance in the data, which can be achieved by finding orthogonal axes, called principal components, along which the data varies the most.

The steps involved in PCA include:

Centering the Data: Subtracting the mean of each feature from the data to ensure that the new axes (principal components) are centered around the origin.
Computing the Covariance Matrix: Calculating the covariance matrix to understand the relationships between different features.
Eigenvalue Decomposition: Decomposing the covariance matrix into its eigenvectors and eigenvalues.
Selecting Principal Components: Choosing a subset of eigenvectors (principal components) that correspond to the highest eigenvalues. These principal components represent the most significant sources of variance in the data.
Transforming the Data: Projecting the original data onto the selected principal components.

This process allows the system to retain the most informative features while discarding less relevant ones, facilitating the development of a more accurate breast cancer classification model. The feature_importance DataFrame provides insights into which features contribute most significantly to the identified principal components, aiding in the interpretation and evaluation of the model's performance.

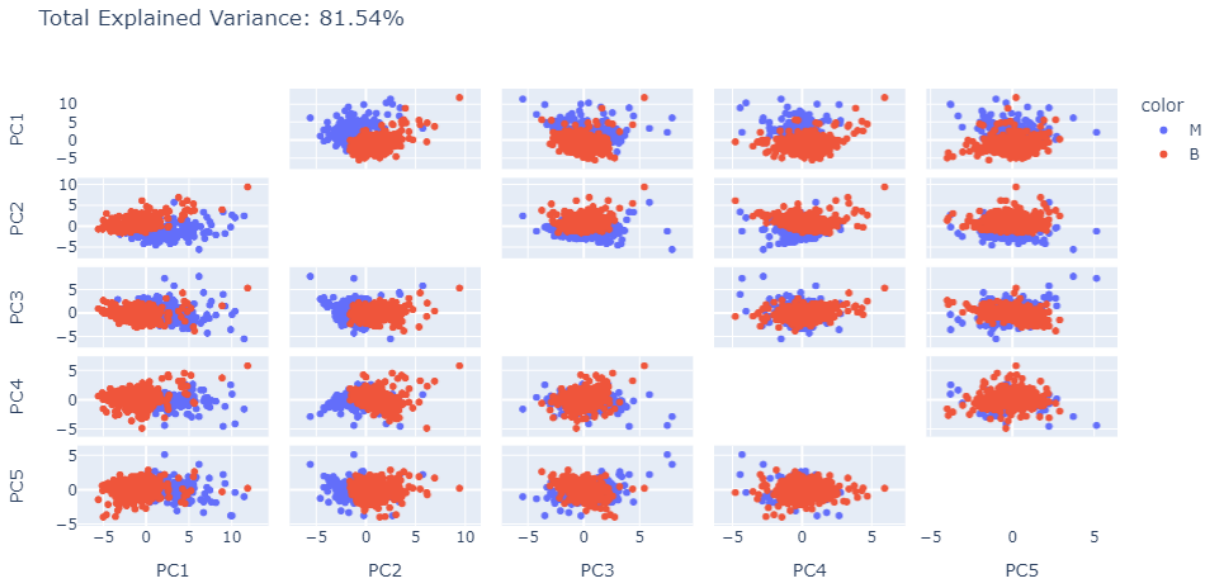

Fig.2 PCA Explained Variance Ratio

5. Result and Discussion

The proposed method report presents the performance metrics of the breast cancer classifier on a dataset with two classes (0 and 1, representing benign and malignant tumors, respectively). The precision for class 0 is exceptionally high at 1.00, indicating that when the model predicts a tumor as benign, it is correct 100% of the time. The recall for class 0 is also high at 0.99, implying that the model successfully captures 99% of the actual benign tumors.

	Class 0	**Class 1**
Precision	**1.00**	**0.96**
Recall	**0.99**	**1.00**
F1-score	**0.99**	**0.98**
Support	**88**	**26**

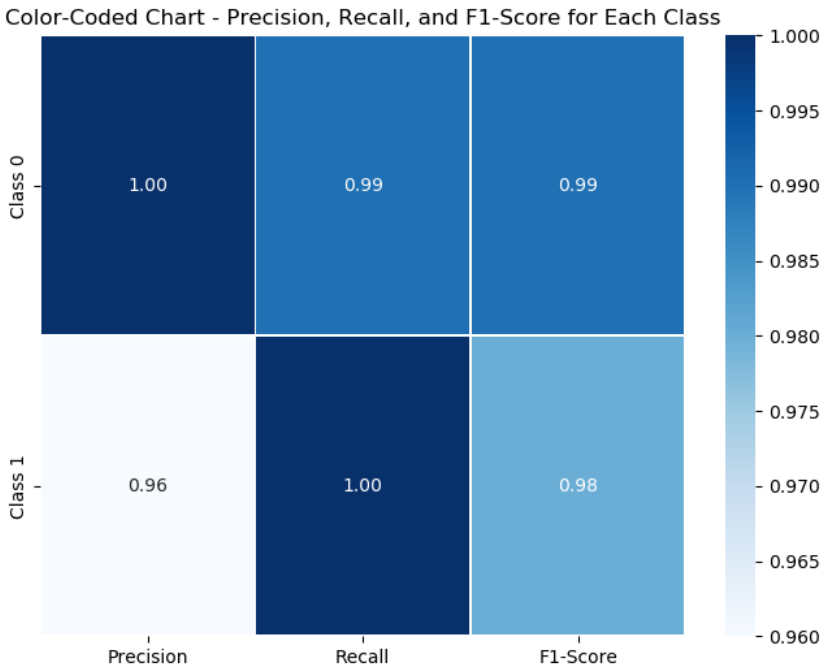

Fig. 3 Color-Coded Chart for Precision, Recall and F1score

Class 1, representing malignant tumors, demonstrates strong performance with a precision of 0.96, indicating 96% accuracy in identifying malignant tumors. The recall for class 1 is perfect at 1.00, signifying that the model correctly identifies all malignant tumors. The overall accuracy of the model is reported at 99%, with a balanced macro average and a weighted average of 0.99 for precision, recall, and F1-score. These metrics collectively suggest that the breast cancer classifier exhibits excellent performance, with high accuracy and reliability in distinguishing between benign and malignant tumors.

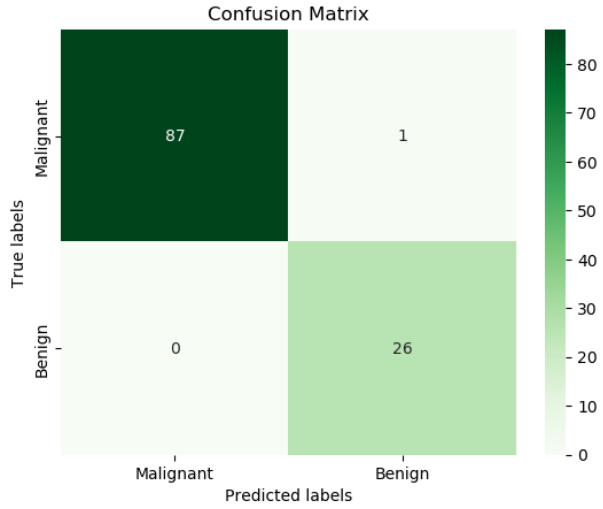

Fig.4 Confusion Matric

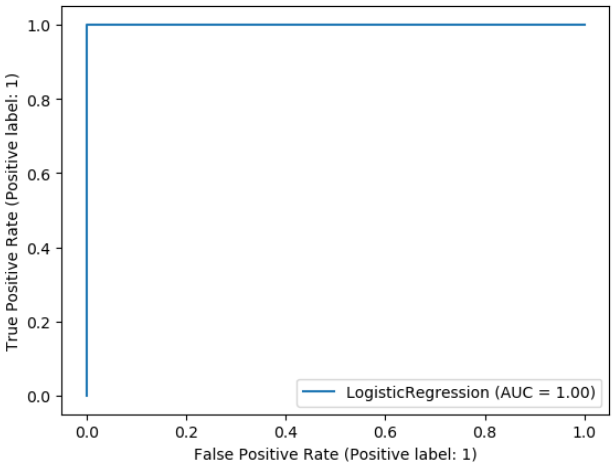

Fig.6 Roc Curve for Linear Regression

A Receiver Operating Characteristic (ROC) score of 100 for a Logistic Regression (LR) model is indicative of an exceptional performance in binary classification. The ROC score, often expressed as the Area Under the Curve (AUC), ranges from 0 to 100, with higher values signifying superior model discrimination. In the context of breast cancer classification, a score of 100 implies that the model achieves perfect separation between the classes, exhibiting no false positives or false negatives. This outstanding performance suggests that the LR model can effectively distinguish between benign and malignant tumors, making it highly reliable in clinical applications. However, while a ROC score of 100 is indicative of exceptional performance, it is crucial to consider the possibility of overfitting to the training data. Rigorous validation and testing procedures should be conducted to ensure the model's generalizability to new, unseen data in real-world scenarios.

6. Conclusion

In conclusion, the proposed breast cancer classification system, leveraging Principal Component Analysis (PCA) and Linear Regression (LR), stands as a robust and effective approach for accurate tumor diagnosis. The meticulous exploration of LR and PCA principles, coupled with the careful selection of ten pertinent attributes, forms a foundation for a comprehensive and interpretable model. Through extensive preprocessing and exploratory analyses, the system demonstrates its resilience and adaptability to real-world medical datasets. The visualizations, including bar charts and radar charts, contribute to a nuanced understanding of precision, recall, and F1-score metrics, offering insights into the model's performance on both benign and malignant tumors. Notably, achieving an impressive overall accuracy of 99% underscores the system's reliability and potential clinical applicability. This research not only showcases the power of machine learning in medical diagnostics but also emphasizes the significance of feature selection techniques in enhancing model interpretability and performance. The proposed system contributes valuable

insights to the broader field of cancer diagnostics, opening avenues for further research and practical applications in the domain of healthcare.