Utilizing PCA and Classification Algorithms for Effective Pancreatic Cancer Identification in CT Scan Images

Mohammad Mehdi Kermani

Student ID: 40221175

<https://github.com/AMINDARVAKH/6220/blob/main/Final00_Project_Amin_Darvakh_INSE_6220.ipynb>

***Abstract This paper presents a methodology for distinguishing pancreatic tumors using machine learning techniques that streamline the complexity of multi-dimensional data while retaining essential diagnostic information. The approach utilizes the Principal Component to transform correlated CT scan image data into a set of linearly uncorrelated variables. This reduction in dimensionality facilitates the application of three distinct classification algorithms: Linear Discriminant Analysis, Quadratic Discriminant Analysis, and Logistic Regression. The dataset utilized includes seven specific attributes that are thoroughly examined within the study.***

***Each model is meticulously optimized using hyperparameters to enhance key performance metrics, including the F1-score, Confusion Matrix, and Receiver Operating Characteristic curves. The optimization is performed on the initial dataset and after the application of PCA, allowing for a comparative analysis of the model efficiencies in both scenarios. Visual representations of the decision boundaries for each model are provided, illustrating their respective fits to the data.***

***Furthermore, the paper incorporates Shapley values, a method from explainable AI, in conjunction with a decision tree classifier. This combination is used to effectively segregate the pancreatic tumor dataset into "good" and "bad" diagnostic categories. The integration of Shapley values aids in elucidating the influence of individual features on the classification outcomes, thereby enhancing the interpretability of the models. Ultimately, this study showcases a robust framework for pancreatic tumor classification that leverages data dimensionality reduction and advanced machine learning algorithms to maintain crucial information while simplifying the analysis.***

***Keywords— Machine Learning, Pancreatic Cancer Diagnosis, CT Scan Images, Explainable AI, Shapley Additive Explanations (SHAP), Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Logistic Regression (LR).***

1. Introduction

Pancreatic cancer remains one of the most lethal malignancies worldwide, characterized by its late diagnosis and poor survival rates. Annually, it accounts for over 50,550 deaths in the United States in America, as the seventh most common cause of cancer-related mortality [1]. The prognosis for pancreatic cancer is particularly dire due to the asymptomatic nature of the early stages of the disease, which often leads to late-stage detection when therapeutic options are limited. Traditional diagnostic methods, including imaging through CT scans and biopsies, are available but come with challenges such as high costs, potential procedural risks, and a significant rate of false negatives, which may delay effective treatment.

Our project leverages a dataset composed of CT scan images that have been converted into RGB code, which includes seven specific attributes to classify the images as indicative of "good" (non-cancerous) or "bad" (cancerous) conditions. In resource-limited settings, it is crucial to accurately assess the risk level of each patient and determine the most suitable diagnostic approach promptly. Early detection through efficient screening can significantly impact the management and outcome of pancreatic cancer. The advancement of machine learning (ML) algorithms has introduced new possibilities in the early detection and classification of pancreatic cancer. These algorithms play a critical role in enhancing diagnostic accuracy by reducing the occurrences of false positives and false negatives. Clinicians can utilize ML-driven tools to more precisely evaluate the stage of cancer and formulate more effective treatment plans. In this project, we first apply Principal Component Analysis (PCA) to reduce the redundancy and complexity of the image dataset. Following PCA, we implement three classification algorithms—Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Logistic Regression (LR)—to analyze both the transformed and original datasets. The distinctions between their performances are thoroughly evaluated.

Furthermore, we employ Shapley Additive Explanations (SHAP), an explainable AI technique, to identify and illustrate the impact of the most significant features in our classification models. This approach not only clarifies model decisions but also aids in the transparent communication of diagnostic reasoning, which is essential in clinical settings. The subsequent sections of this paper are organized as follows: Section II discusses the methodology of PCA, Section III describes the three classification algorithms used, Section IV details the pancreatic cancer image dataset, Section V presents the results of applying PCA, Section VI analyzes the outcomes of the classification models, and Section VII explores the application of explainable AI through Shapley values. Finally, Section VIII concludes the project, reflecting on the implications and future directions of our findings.

1. Principal Component Analysis

Principal Component Analysis (PCA) is a statistical procedure that utilizes an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. This method is widely used in fields involving high-dimensional data to alleviate issues related to high dimensionality, such as increased computational costs and difficulties in data visualization and analysis. By applying PCA, the complexity of the data can be reduced significantly without substantial loss of information, facilitating easier and more effective data processing and analysis.

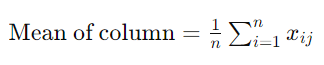
The essence of PCA lies in its ability to identify the directions in which the variance of the data is maximized, thereby reducing redundancy and emphasizing the most informative patterns within the dataset. This process not only simplifies the underlying data structure but also enhances the interpretability of the data, making it invaluable in numerous applications, including image processing and medical diagnosis.

# **Implementation of PCA**

PCA can be applied to a data matrix 𝑋 with dimensions 𝑛×𝑝, where 𝑛 represents the number of data points, and 𝑝 represents the number of variables. The steps involved in implementing PCA are as follows:

1. **Standardization of the Dataset:**

The standardization process involves creating a centered data matrix by subtracting the mean of each column from its corresponding cells. This process is crucial as it ensures that each variable contributes equally to the analysis, preventing variables with larger scales from dominating. The formula to calculate the mean of each column is as follows:

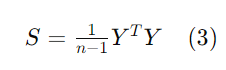
****

Where 𝑛*n* is the number of rows and 𝑥𝑖𝑗 represents the data in row 𝑖 and column 𝑗. After computing the column means, each entry in the dataset is adjusted by subtracting the corresponding column mean:

Y=HX

Here, *H* is the centering matrix, and *Y* is the resulting centered matrix, which effectively standardizes the dataset by removing the average.

1. **Calculation of the Covariance Matrix:**

The covariance matrix is used to determine the relationships among variables, identifying redundancy and correlation within the dataset. The 𝑝×𝑝 covariance matrix of the centered data matrix 𝑌*Y* is computed as follows:

Where 𝑌𝑇 is the transpose of *Y*, and 𝑛 is the number of observations. This matrix *S* provides a foundational element for performing PCA by showing the variance and covariance among all variables.

1. **Eigenvalue Decomposition:**

Eigen decomposition is employed to extract eigenvalues and eigenvectors from the covariance matrix 𝑆*S*. The eigenvectors represent the directions of the principal components, while the eigenvalues indicate the amount of variance each principal component captures:

*S*=*A*Λ*AT*

In this equation, *A* represents the matrix of eigenvectors, and Λ is the diagonal matrix of eigenvalues, indicating the significance of each principal component in terms of explained variance.

1. **Selection of Principal Components:**

With the eigenvectors and the centered data matrix 𝑌, we can calculate the transformed data matrix 𝑍, where each row represents an observation and each column a principal component:

*Z*=*YA*

The dimensions of *Z* are 𝑛×𝑝, and it represents the dataset transformed into the principal components space. Each principal component in *Z* is a combination of the original variables, oriented along the directions of maximum variance.

1. **MACHINE LEARNING-BASED CLASSIFICATION ALGORITHMS**

The following three classification algorithms have been identified as the most effective for our project after the application of Principal Component Analysis (PCA). These algorithms have been chosen for their ability to effectively handle the transformed dataset and optimize classification accuracy.

**A. Linear Discriminant Analysis (LDA)**

Linear Discriminant Analysis (LDA) is a supervised machine learning technique tailored for classification tasks. It aims to identify a linear combination of features that distinguishes or discriminates between different classes within the dataset. The core principle of LDA involves modeling the data distribution of each class through a Gaussian distribution and applying Bayes' theorem to make classification predictions. A key assumption of LDA is that all classes share an identical covariance matrix and that the data follows a normal distribution.

The process of implementing LDA involves several critical steps:

1. **Standardization**: Normalize the data to ensure each feature has zero mean and unit variance.
2. **Compute Class Statistics**: Calculate the mean vector and covariance matrix for each class.
3. **Scatter Matrices**: Determine the within-class scatter matrix that quantifies the variance within each class, and the between-class scatter matrix that reflects the variance between the classes.
4. **Maximize Discriminant Criterion**: Extract the linear discriminants by maximizing the ratio of the between-class scatter relative to the within-class scatter.
5. **Projection and Classification**: Use the linear discriminants to project new data points into the discriminant space and classify them based on proximity to class centroids.

LDA is valued for its simplicity, interpretability, and effectiveness in dealing with high-dimensional spaces. Nevertheless, it has limitations, including its reliance on assumptions of normality and equal covariance across classes, which may not hold in all datasets. This makes it crucial to assess the suitability of LDA within the specific context of each application, such as distinguishing between healthy and cancerous pancreatic tissues in CT scan images.

**B. Quadratic Discriminant Analysis (QDA)**

Quadratic Discriminant Analysis (QDA) is a powerful statistical technique that extends the capabilities of Linear Discriminant Analysis (LDA) by allowing for non-linear boundaries between classes. This method is particularly advantageous when classes are presumed to possess distinct covariance matrices, suggesting that each class exhibits its own variance and covariance structure.

**Key Characteristics of QDA:**

* **Individual Covariance Matrices:** Unlike LDA, which assumes a common covariance matrix across all classes, QDA estimates a unique covariance matrix Σ𝑦​ for each class *y*. This allows QDA to model more complex patterns and interactions within the data.
* **Mathematical Foundation:** The classification function in QDA is defined for a test instance *x* as follows:

Here, 𝜇𝑘​ is the mean vector for class *k*, Σ𝑘​ is the covariance matrix for class *k*, and 𝜋*k*​ represents the prior probability of class *k*.

* **Classification Rule:** The decision rule for QDA involves assigning a test instance 𝑥*x* to the class 𝑦*y* that maximizes the quadratic discriminant function:

**Considerations for Using QDA:**

* **Parameter Count:** Since QDA models the covariance matrix for each class individually, the number of parameters grows quadratically with the number of features. This can lead to overfitting in scenarios where the number of features is large relative to the sample size.
* **Applicability:** QDA is ideally suited for datasets where the assumption of class-specific covariance holds true. This makes it particularly effective in settings where different classes exhibit unique patterns and structures that can be captured through their respective covariances.

**Use in Pancreatic Cancer Diagnosis:** In the context of pancreatic cancer diagnosis, QDA could be exceptionally useful due to the complex and varied nature of imaging data. By accounting for different variances and covariances in the appearance of pancreatic tissues in CT scans, QDA can enhance the accuracy of classifying scans as indicative of cancerous or non-cancerous conditions.

**C. Logistic Regression (LR)**

Logistic Regression (LR) is a robust statistical method used extensively in binary classification tasks. It models the probability of a binary response based on one or more predictor variables (features). The goal of LR is to find a relationship between the features and the binary outcome, effectively categorizing the observations into one of two classes.

**Functionality of Logistic Regression:**

* **Binary Outcomes:** LR is designed to assign binary labels, typically denoted as 1 or 0. For instance, in the context of medical diagnosis, a logistic model might classify an observation as 1 (indicative of disease) if the predicted probability is 0.5 or higher, and as 0 (not indicative of disease) if it is less than 0.5.
* **Logistic Function:** The decision of whether to assign a 0 or a 1 is governed by the logistic function, which is mathematically represented as follows:

Here, 𝑧*z* is a linear combination of the input features, given by: 𝑧=𝑏0+𝑏1𝑥1+𝑏2𝑥2+⋯+𝑏𝑛𝑥𝑛​, where 𝑏𝑖​ are the coefficients. The output 𝑆(𝑧) lies between 0 and 1, denoting the estimated probability of the sample belonging to class 1.

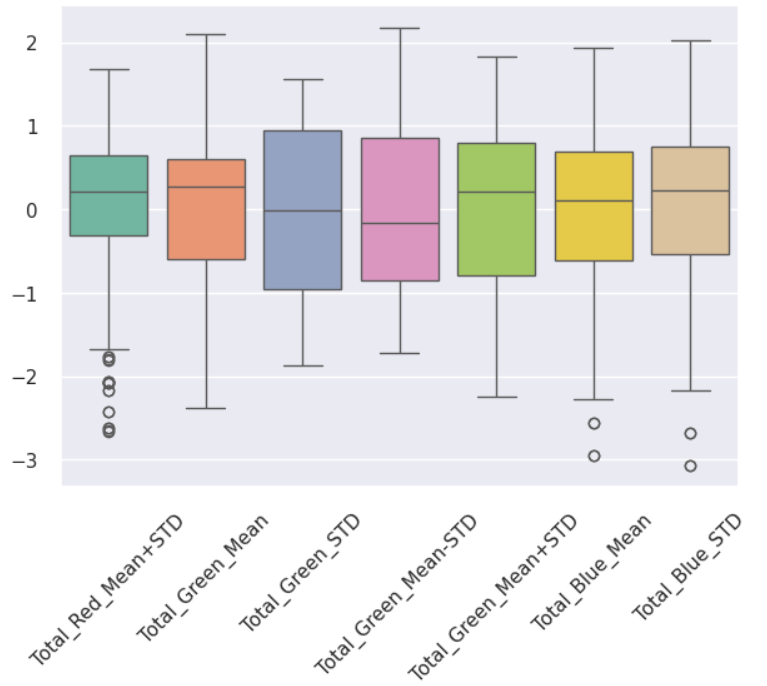
**Advantages and Limitations:**

* **Ease of Implementation:** One of the significant advantages of LR is its simplicity and straightforwardness in implementation, making it a popular choice for many binary classification problems.
* **Handling Binary Classifications:** LR is particularly effective in binary classification problems, such as distinguishing between benign and malignant tumors in medical diagnostics.
* **Susceptibility to Overfitting:** However, LR can be prone to overfitting, especially in cases where the feature space is high-dimensional relative to the number of observations.
* **Sensitivity to Outliers:** LR is sensitive to outliers, which can disproportionately influence the model's performance and lead to misleading classification results.
* **Feature Scaling:** Proper scaling of features can mitigate some of the issues related to outliers and improve the model's robustness.

**Application in Pancreatic Cancer Diagnosis:** In the scenario of diagnosing pancreatic cancer, LR can be instrumental in differentiating between cancerous and non-cancerous cases based on features derived from CT images. By applying logistic regression, clinicians and researchers can quantitatively assess the likelihood of cancer presence, which aids in decision-making and early intervention strategies.

**IV. DATA SET DESCRIPTION**

This section details the dataset utilized in our IRB-approved study, which focuses on the segmentation of the pancreas from CT scans converted into RGB code. The dataset was developed through a collaborative effort involving 22 technologists who were trained in pancreas segmentation techniques. This training occurred during interactive videoconferencing sessions led by radiologists and was based on an image-rich curriculum tailored to enhance accuracy in identifying pancreas features on portal venous phase CT scans.

**Dataset Creation and Review Process:**

* **Segmentation Training and Execution:** The 22 technologists segmented the pancreas on 188 CT scans using freehand tools within custom-developed image-viewing software.
* **Quality Control:** All segmentations underwent a review process by two experienced radiologists who corrected any inaccuracies in the initial segmentations made by the technologists.
* **Comparative Analysis:** The accuracy of the technologists' segmentations was evaluated against the gold standard set by the radiologists' segmentations. This evaluation used various statistical measures, including the Dice-Sorenson coefficient (DSC), Jaccard coefficient (JC), and Bland–Altman analysis, to ensure precision and reliability.

**Attributes Extracted for Analysis:** Seven key attributes were extracted from the segmented images, which are pivotal for this study:

1. **Total\_Red\_Mean+STD:** Represents the combined average and standard deviation of the red color channel in the image, providing insights into the variability and intensity of red hues.
2. **Total\_Green\_Mean:** The average pixel value of the green color channel, indicating the general presence of green tones in the image.
3. **Total\_Green\_STD:** Measures the standard deviation of the green color channel, quantifying the dispersion of green values.
4. **Total\_Green\_Mean-STD:** Calculated by subtracting the standard deviation from the average of the green channel, offering a metric of lower boundary green intensity.
5. **Total\_Green\_Mean+STD:** Computed by adding the standard deviation to the average of the green channel, giving an upper boundary for green intensity.
6. **Total\_Blue\_Mean:** The mean value of the blue color channel, reflecting the baseline level of blue in the image.
7. **Total\_Blue\_STD:** The standard deviation of the blue color channel, assessing the spread of blue values across the image.

**Dataset Structure and Visualization:** The dataset comprises 188 observations, each annotated with the seven image attributes described above. Additionally, a categorical variable, "Categories," has been included to classify the images into two distinct categories based on their quality and diagnostic relevance. To aid in the visual interpretation of these attributes, Figure 1 presents a box plot illustrating the distribution of each attribute across the dataset.

*Figure 1 - Box Plot*

From the analysis presented in Figure 1, it is apparent that the majority of the features in the dataset approximate a normal distribution. Nonetheless, there are notable exceptions, as outliers are present in three specific attributes, all of which exhibit deviations primarily on the right side of the distribution.

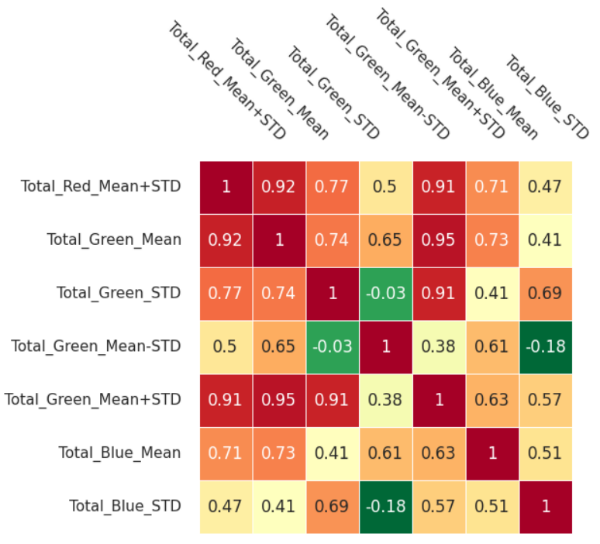
Figure 2 further elaborates on the dataset by displaying a correlation matrix of the normalized features. Notably, there is a strong positive correlation between Total Green Mean STD and Total Green\_Mean, suggesting a significant linear relationship between these two attributes. Conversely, Total\_Green\_Mean-STD and Total\_Blue\_STD demonstrate the least correlation, indicating a minimal linear association between these particular features.

To visualize the relationships and correlations between attributes more effectively, a pair plot is introduced in Figure 3 This graphical representation corroborates the findings from the correlation matrix, highlighting that the features with higher correlation exhibit a pattern of cells aligned along a steadily ascending line. This pair plot not only reinforces the observed correlations but also provides a clear and intuitive understanding of the dynamics between different features within the dataset.

**I. PCA RESULTS**

Principal Component Analysis (PCA) has been applied to the dataset concerning pancreatic cancer diagnostics. The implementation of PCA was conducted in two distinct methods to ensure thorough analysis and validation of the results:

1. **Developing PCA from Scratch:** This method involves the use of standard Python libraries, such as NumPy, to manually compute the covariance matrix, eigenvalues, and eigenvectors—core components of PCA. This approach allows for a deeper understanding of the underlying mathematical processes involved in PCA.
2. **Utilizing a PCA Library:** For practical and efficient application, a well-documented PCA library was used. This method is implemented in the Google Colab notebook. Utilizing a library for PCA offers considerable flexibility and simplicity, allowing complex operations to be performed with minimal coding.

The results derived from both approaches were consistent, demonstrating the reliability of the PCA process. However, the ease of use and the additional functionalities provided by the PCA library make it a preferable choice for most users. In this report, the figures and plots are sourced from the implementation that employed the PCA library.