NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

IE4476 Image Processing and Computer Vision
AY 2024/25 Semester 1

Dimensionality Reduction for Classification on Brain Tumor Dataset

Name: Aaryan Kandiah

Matriculation Number: U2223903E

Table of Contents

1. Introduction

Brain Tumors are a pressing concern for the international medical community with over 300,000 worldwide diagnoses every year (Cohen-Gadol, 2024). Accurate classification of brain tumors is vital for effective diagnosis, treatment planning and early detection. Individuals diagnosed with early-stage, localized brain tumors have a five-year survival rate of 35.5%. In contrast, those with cancers that have spread to regional lymph nodes have a lower five-year survival rate of only 21.3%. A reliable automated classification method for brain tumors is of critical importance to improve survival prospects for patients (Fabiano, 2020). This study aims to use three dimensionality reduction techniques, Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Autoencoder neural networks to improve the classification of accuracy of a brain tumor dataset. This study uses Preet Viradiya brain tumor dataset available on the Kaggle website.

Source: https://www.kaggle.com/datasets/preetviradiya/brian-tumor-dataset

2. Pre-processing

Pre-processing is an important step in the machine learning pipeline which cleans and transforms pixel values to standardize them on a consistent scale. Fig. 1 presents an image before and after preprocessing. The dataset was then split into training and test following a 80:20 ratio respectively.

- **Resizing**: Changing the dimensions of the images to a standard size 256 x 256 to ensure uniform input sizes for the Machine Learning model. It also helps to mitigate computational load as images of a smaller resolution require less memory usage.
- **Augmentation**: Randomly flipping and rotating training samples in the training dataset to simulate different perspectives of the same image. It exposes the machine learning model to a diverse variety of data, so that it becomes more robust and less prone to overfitting. It improves the overall generalization capability of the machine learning model.
- **Standard Scaling:** Standardizing the pixel intensity values by subtracting the mean from each pixel position and scaling to unit variance such that each feature has mean of 0 and standard deviation of 1. It is a form of normalization that reduces the impact of varying contrast and increases compatibility of pixel intensities across different images.
- **Grayscaling:** Converting RGB images to grayscale, reducing the dimensionality by two-thirds. This transformation lowers computational load while preserving essential information.

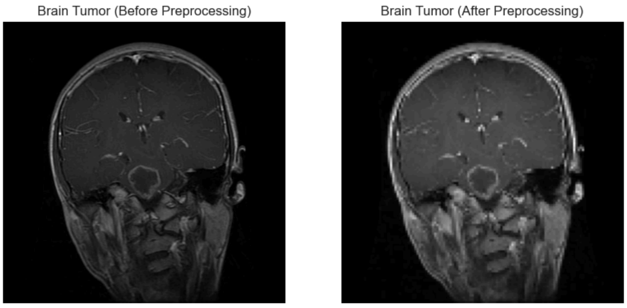

Brain Tumor (Before Preprocessing)
Brain Tumor (After Preprocessing)

Fig 1. Image of Brain Tumor Scan before and after preprocessing

3. Dimensionality Reduction Methods

In the brain tumor dataset, images are being handled where each image pixel represents a dimension, making the dataset a high-dimensional dataset. The curse of dimensionality states that as the number of dimensions increases, for the machine learning model to capture useful and relevant information, the quantity of data fed into the model must grow at a similar rate. When this is not possible, the accuracy of the prediction models suffer from a range of problems such as data sparsity, overfitting, performance degradation and more (Leygonie et al., 2023). Dimensionality reduction methods present a solution to the curse of dimensionality.

Prior to performing dimensionality reduction, additional preprocessing is required to meet the specific requirements of the technique being used. In this case, both the training and test sets were flattened from a multidimensional array of shape (128, 128, 1) to a 1D array with 16,384 elements. This was done by reshaping the data rather than multiplying the dimensions. Additionally, the data type was converted from float64 to float32 to reduce memory usage and improve computational efficiency.

4.1 Principal Component Analysis

Principal Component Analysis (PCA) is the fundamental unsupervised dimensionality reduction algorithm. PCA maps n-dimensional features to m-dimensional data where n > m. To determine the number of principal components for PCA, the number of components that preserve 95% of the variance in the dataset are chosen as the parameter. This choice is significant because it ensures that the dataset retains its variation for generalization purposes while effectively reducing its dimensionality. This number can be obtained by plotting the explained variance of the dataset against the number of dimensions in the images (Feng et al., 2021). From Fig.2 it can be seen that 1413 principal components will be needed to be preserved to retain 95% of variance in the dataset. Fig.3 shows the representation of an image before and after PCA has been applied upon it, the after image has been reconstructed using the original 128x128x1 dimensions of the image.

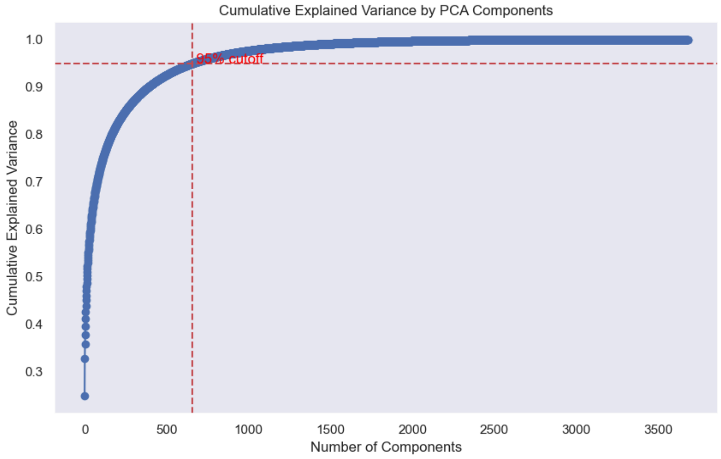


Fig 2. Cumulative Explained Variance by PCA Components

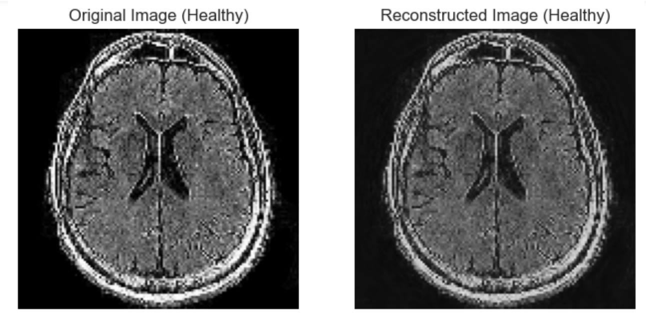


Fig 3. Picture of image before and after PCA

4.2 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a supervised dimensionality reduction algorithm designed to identify the linear combination of features that best separates two or more classes of objects. In the case of the brain tumor dataset, there are two classes: healthy and brain tumor. Since the maximum number of linear discriminants in LDA is equal to the number of classes minus one, n_components is set to 1 for this dataset.

4.3 PCA and LDA

This approach begins with applying Principal Component Analysis (PCA) to the data, followed by Linear Discriminant Analysis (LDA) for further dimensionality reduction. The rationale for this order is that PCA is particularly effective at reducing noise by transforming the original features into a set of linearly uncorrelated components. This preprocessing step can enhance the robustness of LDA, as it operates on a cleaner and lower-dimensional representation of the data. Parameter ‘n_components’ remained at 1 as per the aforementioned requirements.

4.4 Autoencoder

Autoencoders are a neural network-based dimensionality reduction technique that learns compressed representation of the dataset by reconstructing the input from lower-dimensional encoding. The model contains an encoder, which reduces the data’s dimensionality through a series of convolutional and pooling layers, and a decoder, which reconstructs the input from this reduced representation. As a result of multiple layers, this dimensionality reduction method could be more suitable to complex and non-linear data.

4. Performance Metrics

The classification_report tool from the sklearn library was used to generate key performance metrics, including precision, recall, and accuracy. Accuracy was selected as the primary metric for evaluating the effectiveness of different dimensionality reduction techniques because it reflects the model’s overall performance across all instances in the dataset. Given that the dataset is balanced, with a 45:55 ratio of normal to brain tumor images, there is minimal class bias, making accuracy a suitable choice. Additionally, as misclassifications in either class are equally costly for this dataset, the accuracy metric provides a straightforward and unbiased assessment of the model's effectiveness.

5. Classification Models

The dimensionality reduction techniques were evaluated using six different classification models, to allow for a comprehensive assessment of the different dimensionality reduction techniques. Table 1 presents the results and the rationale behind the opting to use the particular classifiers for analysis.

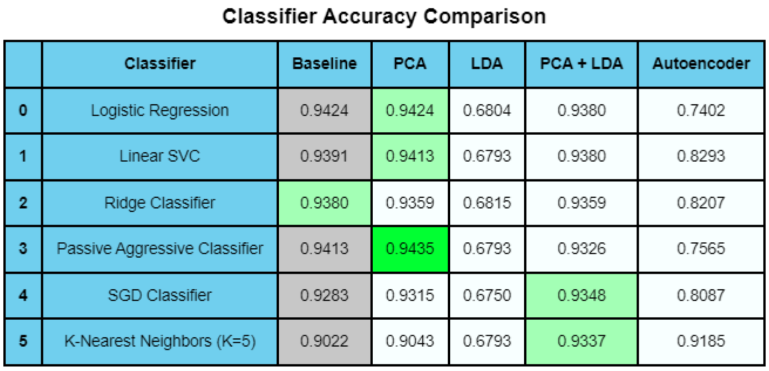
Classifier	Type	Rationale
Logistic Regression	Linear	Suited for binary classification tasks, effective on high-dimensional datasets like the brain tumor image dataset.
Linear SVC	Linear	Effectively separates classes with a hyperplane, advantageous for distinguishing normal and brain tumor
Ridge Classifier	Linear	Well-suited for binary classification, aligns with the dataset's two classes: normal and brain tumor.
Passive Aggressive Classifier	Linear	Efficient for large-scale datasets and can be ideal for high-dimensional image data such as brain tumor dataset where rapid adjustments to new data may be required.
SGD Classifier	Linear	Offers efficiency for large-scale datasets, particularly beneficial for high-dimensional image data like the brain tumor dataset.
KNN Algorithm	Non-linear	Non-linear model, enabling evaluation of the dimensionality reduction's impact across both linear and non-linear models.

Table 1. Classifier Rationale for Evaluating Dimensionality Reduction Techniques

6. Results and Analysis

6.1 Findings

Table 2 presents the results of the different classification models after the application of the different dimensionality reduction algorithms, alongside the baseline which is the classification results prior to any dimensionality reduction. A corresponding graph illustrating these results is included in Appendix 1.



Classifier Accuracy Comparison

	Classifier	Baseline	PCA	LDA	PCA + LDA	Autoencoder
0	Logistic Regression	0.9424	0.9424	0.6804	0.9380	0.7402
1	Linear SVC	0.9391	0.9413	0.6793	0.9380	0.8293
2	Ridge Classifier	0.9380	0.9359	0.6815	0.9359	0.8207
3	Passive Aggressive Classifier	0.9413	0.9435	0.6793	0.9326	0.7565
4	SGD Classifier	0.9283	0.9315	0.6750	0.9348	0.8087
5	K-Nearest Neighbors (K=5)	0.9022	0.9043	0.6793	0.9337	0.9185

Table 2. Classifier Accuracy Comparison for different dimensionality reduction algorithms

From the results, it is clear that the reduction of dimensions has not only significantly mitigated the computational load but also significantly enhanced the accuracy of the classification models. As shown in Table 2, both PCA and PCA + LDA were the most effective techniques in preserving and, in some cases, improving accuracy. The pastel green cells in the table highlight the dimensionality reduction technique that achieved the highest accuracy for each classification model, while the bright green cell indicates the overall highest accuracy across all models. With the exception of the ridge classifier, all other models benefited from the dimensionality reduction techniques, with PCA delivering the best results for three models and PCA + LDA being the most accurate for two.

PCA is an unsupervised learning algorithm with the objective of capturing and maximizing variance within the dataset. This means that it does not prioritize class boundaries in contrast to LDA; PCA focuses on general patterns within the data. A potential reason for PCA's excellent performance and LDA's poor performance is the assumption that the classes are linearly separable, which prevents it from fully capturing variations in the brain scan images. By effectively preserving the overall data variance, PCA enables better generalization by retaining subtle differences that LDA might overlook. This broader capture of detail contributes to PCA’s generally higher accuracy.

The PCA+LDA method is effective because PCA first reduces dimensionality by capturing the maximum variance, generating a more simplified and noise-reduced space. When the LDA is applied following PCA, it can then focus purely on class separation within this lower-dimensional space, which is less affected by noise. This sequential approach allows LDA to operate more effectively by working within a space that has already emphasized the most important features of the data.

The autoencoder performed well only with the k-Nearest Neighbors (kNN) model, as it captures non-linear relationships that align with kNN's ability to classify based on local patterns. In contrast, PCA likely outperforms the autoencoder due to its linear, interpretable approach that efficiently captures the main variance in the data.

6.2 Implications of Findings

This research shows that, given 4000 images, 16000 dimensions can be condensed into 1413 dimensions while preserving and even enhancing the accuracy of image classification. This suggests that with PCA, 40,000 images could be classified with a higher accuracy than 4000 images without PCA.

These findings have vast implications in the field of healthcare and automated diagnostics. Highly skilled surgeons have limited and precious time, and machines can eliminate the need for preliminary scanning and identify tumors when they are not visible to the human eye. The hope is that this research can be improved further, incorporate an enormous scope of different brain diseases and issues, and improve accuracy to a high enough level while keeping relatively low dimensions to be phased in real-world medical applications.

7. Citations

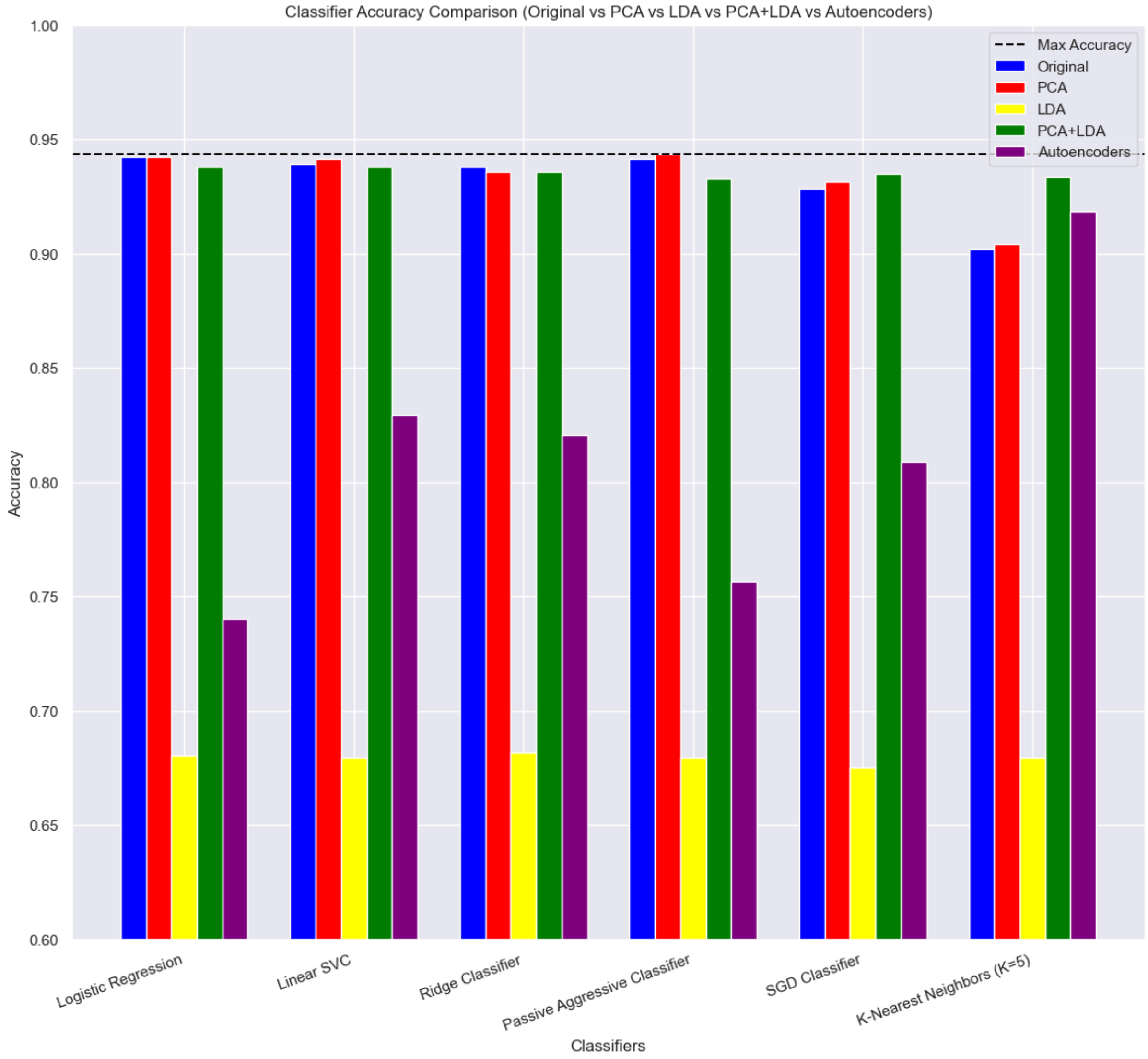
Cohen-Gadol, A., MD. (2024). 72 Must-Know Brain Tumor Statistics (2024). www.aaroncohen-gadol.com. https://www.aaroncohen-gadol.com/en/patients/brain-tumor/types/statistics

Fabiano, A. (2020, February). Survival for Brain Cancer. Roswell Park Comprehensive Cancer Center. Retrieved October 23, 2024, from https://www.roswellpark.org/cancer/brain/survival-rates#:~:text=In%20patients%20with%20early-stage,Distant%20disease

Leygonie, R., Lobry, S., Vimont, G., & Wendling, L. (2023). Transforming Multidimensional Data into Images to Overcome the Curse of Dimensionality. IEEE. https://doi.org/10.1109/icip49359.2023.10222192

Feng, S., & Wang, H. (2021). Comparison of PCA and LDA Dimensionality Reduction Algorithms based on Wine Dataset. In *2021 33rd Chinese Control and Decision Conference (CCDC)* (pp. 2791-2796).

Appendix 1: Classifier Accuracy Comparison for different dimensionality reduction algorithms



Appendix 2: Code