Brain Tumor Classification by MRI Images Through Variants of Linear Discriminant Analysis

By

Afia Munawar

MS Statistics

School of Natural Sciences(SNS)

National University of Sciences and Technology. Islamabad

A dissertation that has been presented in substantial accomplishment of the graduation requirement **Master of Science**

in

Statistics

Supervised by: Dr. Tahir Mehmood

Department of Mathematics & Statistics

School of Natural Sciences
National University of Sciences and Technology
H-12, Islamabad, Pakistan

Year 2022

FORM TH-4

National University of Sciences & Technology

MS THESIS WORK

We hereby recommend that the dissertation prepared under our supervision by: **Afia Munawar**, Regn No. **00000330926** Titled: **"Brain Tumor Classification by MRI Images Through Variants of Linear Discriminant Analysis"** accepted in partial fulfillment of the requirements for the award of **MS** degree.

Examination Committee Members

1. Name: DR. FIRDOS KHAN Signature:

2. Name: DR. SHAKEEL AHMED Signature:

Supervisor's Name: DR. TAHIR MEHMOOD Signature:

Head of Department

19/08/2022
Date

COUNTERSINGED

Date: 22.08.2022

Dean/Principal

I dedicate this dissertation to my beloved late father. And my mother for her love, patience, and unfailing support to me over the years which I can not even imagine expressing in words. A passionate kindhearted personality who always instilled in me the belief in Allah Almighty, the value of hard effort, and the belief that much could be accomplished with little. I adore you both and am grateful for what you have done for me. This thesis is also dedicated to my brother. I am truly thankful for having you in my life.

Acknowledgments

First and foremost, praises and thanks to **Almighty Allah**, for his shower of blessings in opening new paths and showing me the right direction, and helping me to take steps outside of my comfort zone in order to complete my research work successfully. All praise be to the Holy Prophet Hazrat Muhammad (SAW), who created the opportunity for me to acknowledge my creator and his accomplishments to comprehend the meaning of life. According to the Holy Prophet, I am the light, and anybody who follows me will never be in the dark.

I want to thank my supervisor **Dr. Tahir Mahmood**, Department of Statistics, National University of Science & Technology Islamabad, for giving me the opportunity to do research in the field of my interest under his supervision. I have increased the breadth of my knowledge and skills which was only made possible due to his invaluable guidance and constant motivation.

I am highly obliged to my GEC members **Dr. Firdos Khan & Dr. Shakeel Ahmad** for providing research facilities.

Last but not least, I owe my respect and gratitude to my parents and my brother who facilitated me in every possible way so that I could focus on my studies and their sincere prayers lead to the successful completion of my thesis.

Afia Munawar.

June, 2022.

Abstract

This thesis is based on different methods of discriminant analysis applied to brain tumor data. Tumor detection is crucial to improving medical treatment. Magnetic Resonance Imaging (MRI) scans are crucial in several traits and therapeutic applications. For image-based classification problems, Linear Discriminant Analysis (LDA) is a potential candidate. In the current article, we have used the LDA variants including Flexible Discriminant Analysis(FDA), Mixture Discriminant Analysis(MDA), Sparse Discriminant Analysis(SDA), and Regularized Discriminant Analysis(RDA) for tumor classification based on MRI scans. For this MRI scans were first compressed with Principal Component Analysis (PCA), moreover PCA helps to remove the outlier samples. It appears the outlier removal slightly increases the brain tumor classification ability. Further, the above-mentioned methods have several parameters to tune, which was done by Cross-Validation. The meta-analysis based on 100 Monte-Carlo simulation runs reveals that MDA-PCA and SDA-PCA have significantly ($p - value \leq 0.05$) better able to classify the brain tumor on test data (82%), while RDA-PCA has worst ability to classify the brain tumor. The findings indicate the LDA variants can be used not only for brain tumor classification but also for image-based other classification problems.

Keywords: Discriminant Analysis, Classification, Variants, Brain Tumor, Mixture Discriminant Analysis (MDA), Principal Component Analysis(PCA)

Contents

List of Abbreviation

ML Machine Learning

AI Artificial Intelligence

RL Reinforcement learning

LDA Linear Discriminant Analysis

FDA Flexible Discriminant Analysis

RDA Regularized Discriminant Analysis

SDA Sparse Discriminant Analysis

MDA Mixture Discriminant Analysis

PCA Principal Component Analysis

ANOVA Analysis of Variance

SVM Support Vector Machine

MRI Magnetic Resonance Imaging

CTs Computer Tomography scan

PET Positron Emission Tomography

OASTS Osteochondral Autograft Transfer System

List of Figures

Chapter 1

Introduction

Undoubtedly a major component of the human body is the brain. Our total health, especially brain health, is essential. Unfortunately, a wide range of ailments could affect brain health. In brain surgery, a biopsy is performed in order to identify and classify any brain tumors. Technological advancements and machine learning methods may enable radiologists to diagnose malignancies non-invasively. To distinguish between tumorous and non-tumorous tumors, we classified the tumor using a deep learning-based method.

A crucial part of machine learning is Statistics. Analyzing raw data makes it easier to come to insightful conclusions. When we hear the term "Machine Learning," we think it might picture flying cars, robots, and Artificial Intelligence (AI), yet when we think of "Statistics," we might think of bell-shaped charts and the results of sports games. However, since they both involve data analysis, these two fields have a lot in common. Machine learning and statistical modeling can even be used in scenarios that are comparable to each other to solve a variety of problems.

1.1 What is Machine Learning

The field of machine learning focuses on developing algorithms that can learn from data so that programs and systems can carry out tasks without explicit sets of programmed instructions. For instance, image recognition technology frequently uses machine learning algorithms that parse enormous amounts of images and learn to recognize objects

and other features over time and after analyzing large volumes of image data. A sub-field of artificial intelligence research at first, machine learning has subsequently developed into a separate field of study within the subject of AI [1].
Machine learning is particularly useful in situations like this when the amount of data keeps increasing over time. As the study progresses, the increasing volume can help in the training of the algorithms, enabling the algorithm to get "smarter" and provide new, more accurate, or efficient output. Machine learning is comparable to computational statistics, which emphasizes predicting predictions via the use of computers and information technology. Not all machine learning, though, is regarded as statistical learning. Since it offers methodology, theory, and applications in several fields, the study of mathematical optimization aids the science of machine learning.

1.2 Background of Machine Learning

The concept "Artificial Intelligence" was created in the 1950s as a basic idea of human intellect being shown by machines [2]. Jerrold S. Maxmen stated in 1976 that Artificial Intelligence (AI) will usher in the twenty-first century. In the current era of technological advancement and the accessibility of gigantic data sets (also known as "big data"), AI has advanced beyond mere theory to real application on a massive scale.
Machine learning (ML), which is considered a subcategory of AI, demonstrates the empirical “acquisition” concerned with human intellect and has the potential to learn and enhance its assessment by using computer algorithms. The machine can take an input and estimate a result with repetitions and alterations to the algorithm. The algorithm's accuracy is therefore tested by trying to compare the outcomes with ground truth, which is repeatedly revised to perfect the ability to predict future events.

The challenge of developing a prediction based on data that incorporates algorithms and large data calculations is tackled by machine learning. It helps predict future predictions based on already known data. It is useful for interpreting the data according to our preferred information. Machine learning support interpreting data in many fields such as Science, Medicine, Economy, Policy-Making, etc. Equations in

mathematics, statistical analysis, and computer programming are used to implement machine learning. supervised, unsupervised, and reinforcement learning are the three basic categories of machine learning.

It is based on a proper programming language that deals with dependent and independent variables, error terms, and some estimation parameters. For getting accurate interpretations of data, researchers are still working to introduce new methods of prediction corresponding to the demand for data.
Machine learning algorithms are classified as supervised, unsupervised, and reinforcement learning.

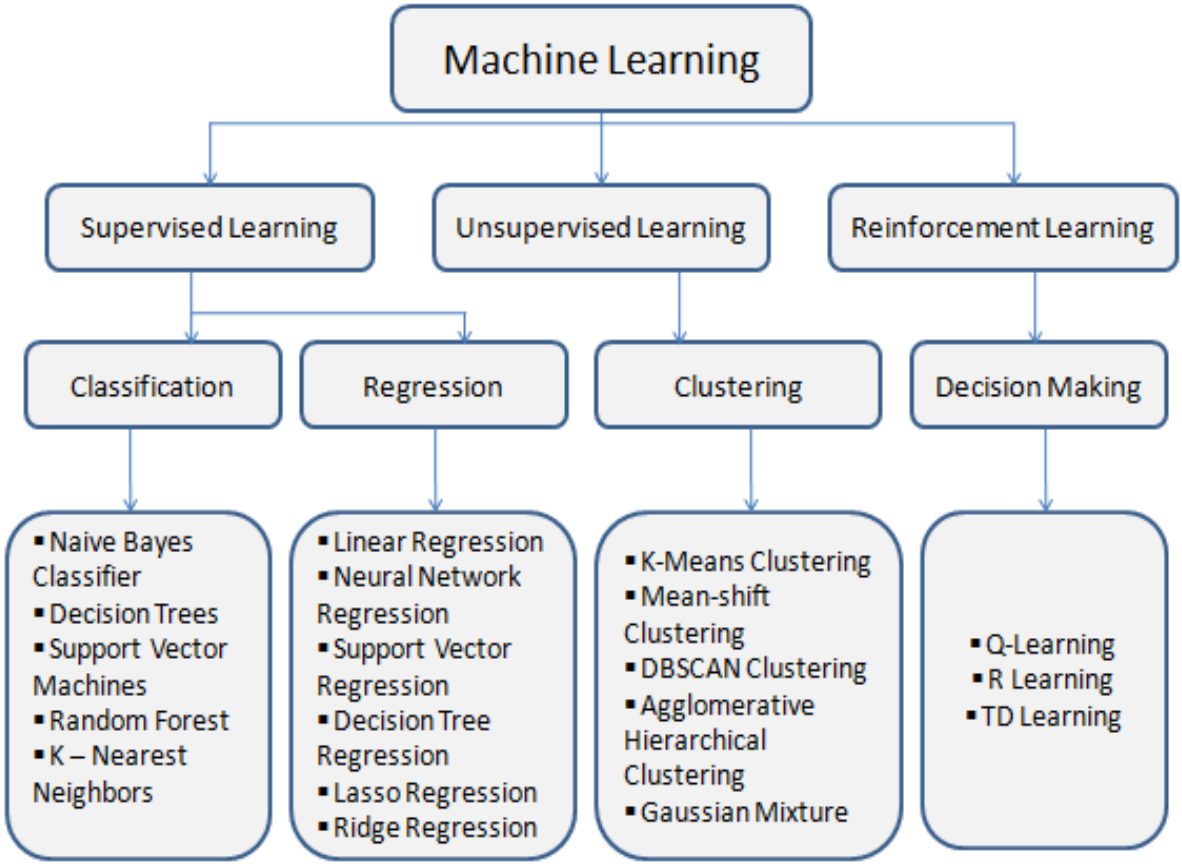


Figure 1.1: Flowchart of Machine Learning algorithms

1.2.1 Supervised Learning

There are two primary ways used in Machine Learning (ML) and Artificial Intelligence (AI): supervised learning and unsupervised learning. One employs labeled data to aid in result prediction, whilst the other does not. This is a major distinction.
In this branch of Machine Learning (ML), the term "supervised learning" refers to algorithms that learn from training datasets and are being trained to supervise the learning process [3]. A machine learning approach termed supervised learning uses input and output data. Input data corresponds to independent variables while output data corresponds to the predicted response. In supervised learning, there is a possibility to test the large data by working on small training data set. In machine learning, the algorithm provides the characteristics of final data sets similar to training data sets. Supervised learning helps know the relationship between the explanatory and response variable with accuracy. There are many practical applications of supervised learning algorithms such as text categorization, signature recognition, weather forecast, stock exchange predictions, face detection, etc [7].

1.2.2 Unsupervised Learning

Learning is a form of statistical learning in which only calculations are conducted on unlabeled input, leading to the construction of various structures. It consists of only independent variables. There is no response variable in unsupervised learning [3].
The algorithm sets the link between data sets in a random way. There is no need for a human to give any input. The formation of different structures from data makes this learning more useful. Unsupervised Learning algorithms are useful to handle complicated tasks. Examples of unsupervised learning techniques include clustering, anomaly detection, and in some cases neural networks.

1.2.3 Reinforcement Learning

A branch of machine learning, Reinforcement learning (RL), investigates how intelligent creatures should behave in a given environment in order to maximize the concept

of cumulative reward. Reinforcement learning, along with supervised and unsupervised learning, is one of the three major machine learning paradigms. By not requiring the display of labeled input/output pairs or the explicit correction of undesirable behaviors, reinforcement learning differs from supervised learning. The foundation of Reinforcement learning is sequential decision-making. Simply said, the current input's state influences the results of the previous input, which indicates the status of the next contribution[8].

Games of chess are an illustration of reinforcement learning.

1.3 Statistics vs Machine Learning

Statistics is a fundamental part of data analytics and machine learning. It enables us to examine and display data in order to uncover previously hidden patterns. The fundamental contrast between Machine Learning (ML) and Statistics is their intended application. Machine learning models are created to make the most accurate predictions feasible. Statistical models are intended to predict associations between variables. Various ML approaches are derived from statistics (for example, linear regression and logistic regression), as well as other fields such as calculus, linear algebra, and computer science. Statistics and Machine Learning are frequently grouped because they employ comparable methods to achieve a goal. However, the objectives they are attempting to achieve are significantly different. The goal of statistics is to conclude a population by studying a sample. Machine learning is used to identify patterns in data to produce reliable predictions.

Machine learning is used to predict future events while Statistics is used to find the relationship between data points. Applications of machine learning are forecasting, modeling, and predictions, and applications of Statistics are finding patterns and outliers in the data. To obtain a reliable model of machine learning or statistical analysis, one needs to understand the Statistics, ML algorithms, and fundamentals of mathematical concepts.

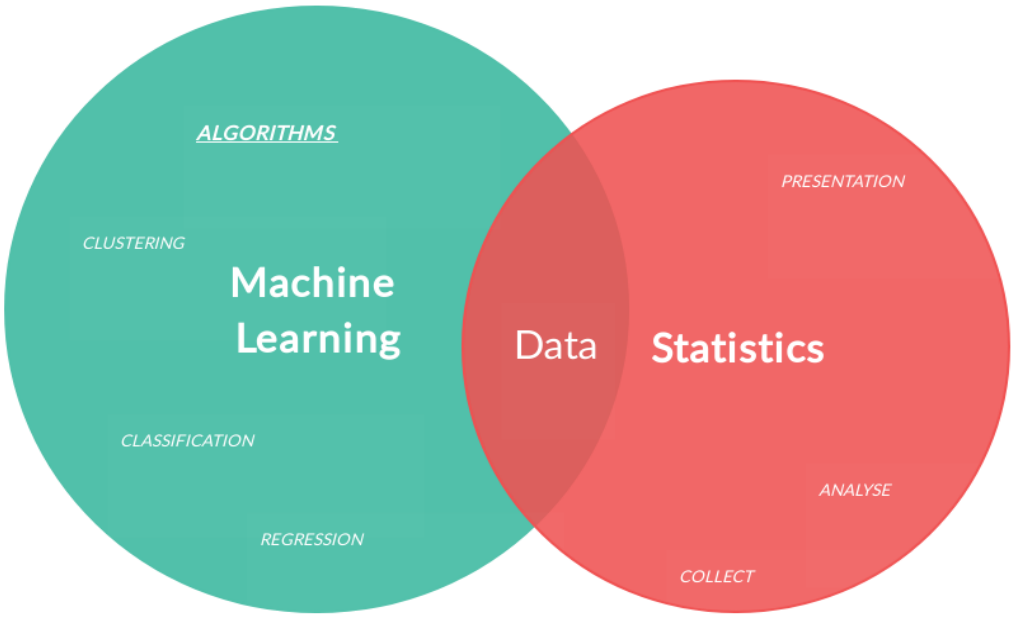


Figure 1.2: Relation between Statistics and Machine Learning

1.4 Multivariate Analysis

Research has been using machine learning (ML) and multivariate analytic techniques more and more recently. An important area of statistics called multivariate analysis evaluates several variables at once. Multivariate is a supervised Machine Learning algorithm that analyses multiple data variables [4]. In practice, multiple variables data sets appear commonly and we usually concern ourselves with several features of the observations.

Discriminant analysis is one of the multivariate statistical methods. This thesis concerns contributions to variants of Linear Discriminant Analysis with applications applying to a real-world data set.

1.5 Classification

Classification is the statistical problem of classifying an individual into one of several groups based on a collection of observations or measurements of the individual.

Classification is the process of classifying a set of data. It may be done on both structured and unstructured data. Predicting the class of provided data points is the first step in the procedure. The classes are also known as the goal, label, or categories.

Estimating the mapping function from discrete training data to discrete independent variables is the problem of classification predictive modeling. The basic purpose is to determine which class/category the new data belongs to.

For classification issues, a variety of methods and algorithms exist, including logistic regression, linear discriminant analysis, cluster analysis, and classification trees (Agresti 2002). These features are already included in most software programs. Each method has drawbacks and advantages of its own. In this dissertation, we employed the classification machine learning method, which would be supervised-based.

In statistics, data categorization refers to the process of grouping data according to their properties into homogenous and similar groupings. A supervised learning approach used to model and predict variables is classification. The explanatory variables could be quantitative or qualitative, and the response is a categorical variable. Classification techniques or methods are also called Classifiers.

1.5.1 Classification Algorithms

Classification is a supervised learning paradigm in machine learning that divides the number of observations into classes. Speech recognition, face identification, handwriting recognition, document categorization, and other recognition-based challenges are common in classification problems. It might be a binary classification problem or a multi-class classification problem.

In machine learning, there are several machine learning methods for classification. For all types of data sets, classification techniques are suitable. There is a substantial toolset that is used for classification purposes. The most often employed classification algorithms are listed here,

Fisher's linear discriminant

Quadratic classifiers

Support vector machines

Logistic regression

Random forests
Least squares support vector machines
Naive Bayes classifier
k-nearest neighbor
Decision trees

These classification algorithms are divided into mainly two categories, Linear Models and non-Linear Models. In this thesis, we'll discuss conventional discriminant analysis with its variants for classification.

1.5.2 Use Cases of Classification Algorithms

Different circumstances require the application of classification methods. Here are a few frequent applications for classification algorithms:
1) Email spam monitoring,
2) Speech Recognition
3) Identifications of Cancer tumor cells.
4) Drugs classification
5) Biometric identification, etc.

1.6 Problem Statement

Classify brain tumor images using different variants of Linear Discriminant Analysis algorithm based on Magnetic Resonance Imaging (MRI).

1.7 Aim & Objective

This dissertation aims to classify the tumorous and non-tumorous by MRI images correctly using LDA (Linear Discriminant Analysis) and its Variants.
Our objectives include:
Brain tumors classification by MRI X-rays.
Comparison of different variants of LDA

Find the most accurate classifier among all the classifiers.

1.8 Research Questions

How may brain tumors be categorized?
Can LDA accurately differentiate between cancers and non-tumors?
Can LDA and its Variants be used to categorize tumors and non-tumors with more accuracy?

1.9 Covered Topics

This thesis is organized as follows: We first briefly describe Linear Discriminant analysis (LDA) with its variants and Principal Component Analysis(PCA) classification. We provide an overview of the LDA algorithm as well as the description of different variants of LDA in chapter 2. The third chapter of this thesis is about the literature review. In chapter 4 we explained the mathematical form of the methods.
We explained the data application in chapter 5. Results and discussions are presented in chapter 6 and conclude with Chapter 7. The list of references is provided in the Bibliography section.

Chapter 2

Preliminaries

This chapter contains the technical setup for classification. There are two main approaches are used. Principal Component Analysis (PCA) is used for dimensionality reduction, the other one, Linear Discriminant Analysis (LDA), is used for classification propose.

In Section 2.1, we'll review the classical classification method, Linear Discriminant Analysis(LDA). In this section, we will first give an introduction to the conventional discriminant analysis and the limitation of this method. In section 2.3, we'll discuss variants one by one. A detailed discussion of Principal Component Analysis (PCA) is in section 2.4. Preprocessing and analysis of variance (ANOVA) are presented in sections 2.5 & 2.6, respectively.

2.1 Introduction

R.A. Fisher first introduced discriminant analysis in 1936. Fisher is a two-class classification method (Fisher 1936). 1948 [27], when C.R. Rao expanded it to integrate issues involving multiple classes. Regression analysis and discriminant analysis are similar, although the dependent variable in discriminant analysis is categorical as opposed to continuous (Draper and Smith 1981)[5]. LDA typically looks for linear combinations of predictor factors that best distinguish the acquisition of observational data. The goal of discriminant analysis is to predict class variables of individual observations based on a set of predictor variables. Discriminant functions are the term referring to these combinations (Mika et al. 1999).

2.2 Linear Discriminant Analysis

Ronald A. Fisher initially developed the linear discriminant in 1936 and has shown some useful applications as a classifier. He presented it for a two-class issue, and C.R. Rao later generalized it as "Multi-class Linear Discriminant Analysis" or "Multiple Discriminant Analysis" in 1948. In addition to being a powerful classification technique, Linear Discriminant Analysis provides a technique for dimension reduction. Prior to classification, it is used to whittle down the number of features to a more suitable or manageable quantity. One dependent variable is attempted to be expressed by LDA as a linear combination of other characteristics or measures. For ordinarily distributed predictor variables, LDA has been established [23].

Therefore, it makes sense to anticipate that LDA will produce better results when the normality requirements are satisfied. It can be applied in positioning and product management.[21, 23] Some studies use Linear Discriminant Analysis in bankrupt predictions, face recognition, credit scoring, and the like. LDA was formulated for a two-class problem, later it was generalized for multi-class problems. This approach projects the properties of a higher multidimensional space into a lower multidimensional space.

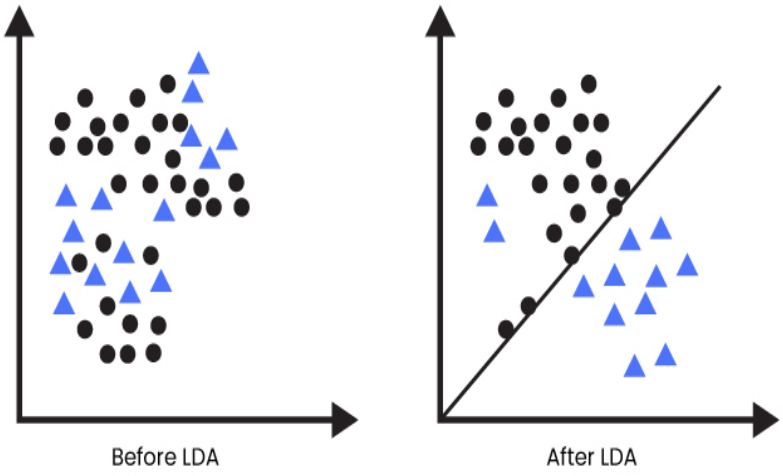


Figure 2.1: Classification of two classes by applying LDA in presented.

2.2.1 Assumptions

LDA assumes some criteria about the data[5]. Discriminant algorithms have the following assumptions.

1) There is no multicollinearity.

2) Variables should follow a gaussian distribution.

3) Requires homoscedasticity and full rank covariances.

Quadratic Discriminant Analysis (QDA)

The only significant difference between linear and quadratic discriminant analysis is the relaxation of the assumption that the covariances of all classes are different. Variables must follow a normal distribution but each class should have the same covariance is not necessary.

2.2.2 How does Linear Discriminant Analysis (LDA) work?

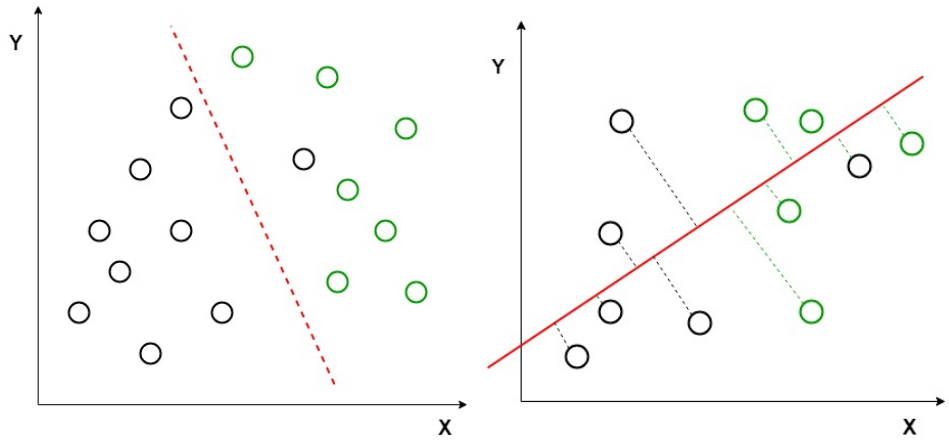
Example

Suppose we have 2 subjects, so we must effectively divide them. The classes may consist of several properties. As seen in the following image, if we categorize them using only one feature, overlapping may occur. As a result, we will constantly be adding features to assure accurate classification.

Figure 2.2: Two subjects are overlapping.

Let's say we wish to categorize two sets of data points that belong to two distinct classes. As shown in the above 2D graph, when the data points are lying on the 2D plane, there is not a perfect separation between the data points of the two classes. In this instance, in order to maximize the separability between the two classes, LDA

(Linear Discriminant Analysis) is used to transform the 2D graph into a 1D graph.



(a) (b)

Figure 2.3: LDA boundary between two subjects.

Through the use of both axes (X and Y), Linear Discriminant Analysis provides a new axis and projects data onto it to optimize the separation of the two classes and convert the 2D graph into a 1D graph.

As shown in the aforementioned graph, a new axis (shown in red) is produced and drawn in the 2D graph in a way that minimizes variation within each class and optimizes the distance between the means of the two classes. Simply said, the space between the data points for the two groups is widened by this newly constructed axis. As seen in the graphic below, this new axis is constructed using the aforementioned criteria and all of the data points for the classes are displayed on it.

However, Linear Discriminant Analysis (LDA) is unable to create a new axis that

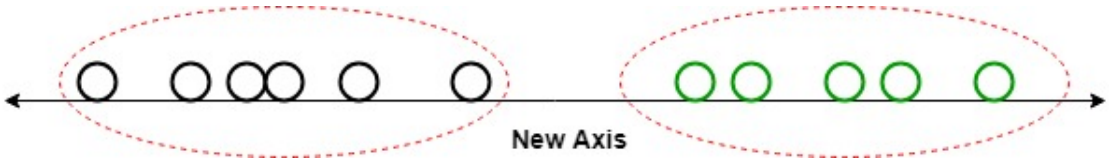


Figure 2.4: LDA divided data points of two groups separably.

renders both classes linearly separable when the distributions' means are shared. We

employ non-linear discriminant analysis in these circumstances.

2.2.3 Why We Use LDA?

• One of the most used classification methods, logistic regression, performs well for binary classification but struggles when dealing with multiple classification issues involving distinct classes. LDA manages these simultaneously fairly effectively.
• In order to reduce the number of features in data pre-processing, LDA may be employed, similar to PCA, which significantly reduces computing costs.
• LDA is also used in computer vision applications and for face detection. In Fisherfaces, LDA is used to extract pertinent data from a variety of faces. When used with eigenfaces, it produces potent results.

2.2.4 Applications

Applications of Linear Discriminant Analysis (LDA) are given below [6],

Face Recognition

In the science of computer vision, the identification of faces using a large number of pixel values as a representation of each face is a particularly typical application. Before the classification process, Linear Discriminant Analysis is used to condense the number of features to a more manageable amount (LDA). Each of the additional dimensions is produced by a linear transformation of pixel values that produces a new pattern. Fisher's faces are the linear representations generated by using Fisher's linear discriminant.

Biomedical Studies

Based on the patient's numerous characteristics and the medical care he is receiving, this discipline uses Linear Discriminant Analysis (LDA) to categorize the patient's illness status as mild, medium, or chronic. This enables the doctors to speed up or slow down their course of therapy.

Discriminant analysis is mostly used in medicine to assess a patient's degree of illness and predict how the sickness will progress. For instance, during retrospective analysis, patients are categorized into mild, moderate, and severe illness forms. The outcomes of clinical and laboratory studies are then examined to identify factors that are statistically different in the groups under investigation. Discriminant functions are created using these variables to assist categorize a patient's condition into mild, moderate, or severe forms.

Customer Identification

Let's say we want to figure out which people at a mall are more likely to purchase a specific item. We may compile all the characteristics of the clients by conducting a straightforward question and response poll. Here, a Linear Discriminant Analysis will assist in locating or choosing the variables that can characterize the traits of the clients most likely to purchase that specific product at the shopping center.

2.2.5 Limitations

1) Small sample size problem.
2) Discriminative information is not in the means of classes.
3) Sometimes not good for a few categories variables.
4) It requires a normal distribution assumption on features/predictors.

Due to Linear Discriminant Analysis's (LDA) limitations, we move toward the modified extensions of LDA. which work with the non- linear as well as non-normal datasets.

There are several modifications and extensions of Linear Discriminant Analysis. Each of them has been created to enhance the effectiveness and output of the Linear Discriminant Analysis.

2.3 Variants of LDA

Flexible Discriminant Analysis (FDA)

Mixture Discriminant Analysis (MDA)

Regularized Discriminant Analysis (RDA)

Sparse Discriminant Analysis (SDA)

2.3.1 Regularized Discriminant Analysis (RDA)

Regularized Discriminant Analysis (RDA) [29] was proposed by Friedman in [1989]. RDA (Regularized Discriminant Analysis) modifies the impact of various variables on LDA by introducing regularization. That can estimate the variance (actually covariance) of each class. Regularizing the estimation of the covariance reduces the impact of different factors on LDA. LDA and QDA are combined in RDA. By regularizing the group covariance matrices, RDA creates a classification rule and makes the model more resistant to multicollinearity in the data [9]. For a sizable multivariate data collection with highly associated predictors, it is quite helpful.

Similar to LDA, RDA turns the several covariances of QDA into a single covariance. As a consequence, the covariance matrices may be calculated more precisely when the number of observations exceeds the classification performance, possibly increasing the model's accuracy.

Regularized Discriminant Analysis (RDA) has two parameters α and δ. This operator uses LDA if the α argument is set to 1. Similar to how QDA is carried out when the α value is set to 0. similarly δ is used for the class covariance. Lower δ indicates both classes have a common variance and higher δ indicates both classes tend to have different variances.

2.3.2 Flexible Discriminant Analysis (FDA)

Only linear combinations of inputs are used in Flexible Discriminant Analysis (FDA) [28]. In 1994, (Robert Tibshirani, Trevor Hastie, and Andreas Buja) [28] introduced Flexible Discriminant Analysis (FDA). FDA accepts non-linear input combinations like splines. In the FDA classification model, the response variable is transformed using

optimum scoring to improve the data's suitability for linear separation. The blend of linear regression models serves as its foundation. It functions well with a lot of independent variables and detects interactions between variables automatically. Despite its complexity and robustness to outliers, it is a quick and effective method. FDA is useful for showing non-linear correlations between variables within each group and making classification more precise.
Using nonparametric regression and Linear Discriminant Analysis, FDA is a multigroup non-linear discrimination technique (LDA). LDA is extended in a non-parametric manner by FDA.

2.3.3 Mixture Discriminant Analysis (MDA)

By Hastie and Tibshirani (1996), Mixture Discriminant Analysis(MDA) [30] was proposed to model training data sets and to classify test data effectively when the structure of the training groups is heterogeneous. The well-known Linear Discriminant Analysis method's underlying conventional Gaussian assumptions are naturally extended by the Mixture Discriminant Analysis (MDA) model. This is an example of supervised classification using a mixture model. Both linear and non-linear boundaries can be used with MDA.
For each class, it is used to estimate the density. MDA has the ability to generate non-linear decision limits. In comparison to LDA and QDA, the MDA classifier successfully recognized the subclasses. Mixture Discriminant Analysis (MDA) is a non-parametric extension of LDA.

Although there are now fewer parameters to estimate, the LDA classifier is ill-posed if $(p > n)$ due to the unique characteristics of the sample covariance matrix, in which case either regularization methods, feature selection, or further model restrictions are employed, such as assuming Σ is diagonal (Ramey and Young, 2013; Dudoit et al., 2002).

Clemmensen et al. (2011)[31] argue that although LDA is often well-suited for simple, low-dimensional settings, linear decision boundaries are often insufficient to

separate classes in practice. Furthermore, a single Gaussian distribution may be insufficient in characterizing a single class. With the latter point in mind, Hastie, and Tibshirani (1996) introduced Mixture Discriminant Analysis (MDA), where $f_k(x)$ is the probability density function of the finite Gaussian mixture model.

2.3.4 Sparse Discriminant Analysis (SDA)

Sparse Discriminant Analysis (SDA) was proposed by (Line Clemmensen, Trevor Hastie, Bjarne Ersbøll, and Daniela Witten) in 2008[31]. The Linear Discriminant Analysis's optimal score interpretation serves as the foundation for the Sparse Discriminant Analysis. If the classes include subgroups or nonlinear boundaries are present, LDA can be expanded to do sparse discrimination using mixtures of Gaussian functions.
By trying to find the LDA score analysis with the best possible results, the SDA approach leverages over LDA to address the two issues that LDA indicated. This technique enhanced memory effectiveness, temporal complexity, and prediction accuracy. By applying a sparseness constraint to Linear Discriminant Analysis (LDA), Sparse Discriminant Analysis (SDA) enables the simultaneous performance of feature selection and classification.
LDA is not possible when the within-class covariance matrix lacks complete rank or when the number of features is excessive in comparison to the number of observations. In that case, Sparse Discriminant Analysis can be used instead of LDA.

2.4 Principal Component Analysis (PCA)

Karl Pearson created the Principal Component Analysis (PCA) [32] in 1901 as a mathematical equivalent of the principal axis theorem. Harold Hotelling then independently extended the concept and gave it his name in the 1930s.
To reduce the dimensionality of data, a non-parametric approach called Principal Component Analysis (PCA) is used. With its great flexibility, PCA can analyze datasets with various characteristics, such as multicollinearity, missing values, categorical data, and inaccurate measurements. The objective is to identify the key information in the data and represent it as a collection of summary indices known as primary components.

Using the fewest possible components and a mean squared error that is kept to a minimum, principal components offer a linear representation of the original data. The purpose of PCA is two-fold: dimension reduction and uncorrelated feature learning. The PCA algorithm is based on some mathematical concepts such as:
1) Variance and Covariance
2) Eigenvalues and Eigen factors

2.4.1 PCA Applications in Artificial Intelligence

Data from multiple dimensions are visualized using PCA.
It is applied to significantly reduce the dimensions in healthcare data..
PCA can play an important role in picture resizing.
It may be used in finance to anticipate returns and evaluate stock data.
PCA aids in pattern recognition in high-dimensional datasets.

2.4.2 Advantages of PCA

1) PCA helps to remove all the features that are correlated
2) With the number of features reduced with PCA, machine learning algorithm performance improves.
3) By removing the unnecessary features in the dataset, PCA helps to overcome overfitting.

2.4.3 Disadvantages of PCA

1) PCA reduces features into a smaller number of components. Each component becomes a linear combination of original features, which makes it less readable and interpretable.
2) Data information loss may occur if we don't choose the right number of components.
3) Because PCA is a variance maximizing technique, PCA requires features to be scaled before processing.

4) PCA is not robust against outliers.

2.4.4 Fast Principal Component Analysis (FAST-PCA)

The modern-day dataset's size and dimensions have made the centralized PCA [33] solutions expandable. To overcome the problems of PCA, a new approach has been introduced. this new approach for existing solutions for distributed PCA algorithms is called FAST-PCA. The new proposed approach FAST-PCA works more efficiently than PCA for dimension reduction with uncorrelated features.
Communication with the proposed algorithm is efficient and the algorithm is proven to converge linearly and exactly to the principal components, leading to a reduction in dimension as well as uncorrelated features.

2.5 Preprocessing

In machine learning, Preprocessing is the process of preparing raw data (cleaning and organizing it) for use in constructing and training machine learning models [22]. Essentially stated, data pre-processing in machine learning is a data mining approach that transforms raw data into a legible and understandable format. Preprocessing frequently includes picture enhancement and smoothing. Real-world data generally fail to capture specific attribute values or trends. It may contain mistakes or outliers having inconsistency and erroneous. Handling real-world data is a tedious and time taking process. So, pre-processing helps to clean, format, and organize the raw data and make it ready for use by machine learning models.

Feature Extraction

Feature extraction is used in image processing to reduce dimensionality. Feature extraction turned the input data into a piece of pertinent information when an algorithm's input data was too vast to analyze. The process of turning input data into a collection of features is called feature extraction. In order to create new, more significant features, feature extraction modifies the original features in some way. In this view,

feature extraction may be used to simplify data representation and minimize complexity by expressing a linear combination of each variable. Principal Component Analysis is the method of feature extraction that is most extensively utilized and quite well (PCA).

When we have a huge data collection and need to utilize fewer resources without sacrificing any crucial or pertinent information, the process of extracting the features is helpful. The data reduction achieved by feature extraction speeds up the learning and generalization phases of the machine learning process while also requiring less computer work to create the model.

2.6 Analysis of Variance (ANOVA)

ANOVA [10] is a statistical method for determining if there are any differences or correlations between the impact of independent factors on a dependent variable. When there are numerical input variables and a categorical target variable, as there would be in a classification problem, ANOVA is performed. The discriminant analysis employs continuous independent factors and a categorical dependent variable, whereas ANOVA uses categorical explanatory variables and a continuous response variable.

Discriminant Analysis and MANOVA, the multivariate form of ANOVA, are essentially equivalent. Both methods rely on identifying the linear combination of continuous variables (the dimension on the multivariate space) that distinguishes the categories specified by the categorical variable the most effectively.

2.6.1 One-way ANOVA

One-way (or unidirectional) and two-way ANOVA are the two primary variations. ANOVA also comes in several forms. For instance, MANOVA (multivariate ANOVA) is different from ANOVA in that the former evaluates several dependent variables simultaneously while the latter only does so for one. The number of distinct variables in any analysis of variance test—one or two—is referred to as one-way or two-way. A one-way ANOVA analyses how one factor affects just one response variable. It establishes if every sample is the same. To evaluate if there are any statistically significant differences

between the means of three or more predictors (unrelated) groups, the one-way ANOVA is performed.

2.6.2 Two-way ANOVA

The one-way ANOVA is expanded upon by the two-way ANOVA. One independent variable influences one or more dependent variables in a one-way relationship. A two-way ANOVA has two independent variables. A two-way ANOVA, for instance, enables a business to compare operational efficiency based on two independent factors, such skill level, and compensation. It is used to examine the simultaneous effects of two factors and see how the two elements interact.

Chapter 3

Literature Review

The methods that are relevant to our work are described in this chapter. There have been several studies on how to categorize brain tumors. Many methods, including fuzzy clustering means (FCM), support vector machines (SVM), convolution neural networks (CNN), artificial neural networks (ANN), and the expectation-maximization (EM) algorithm technique, have been proposed in recent years for the classification of brain tumors in MR images. The application of statistical techniques to the issues of high dimensionality, classification, face recognition, feature extraction, and data visualization has received a lot of attention lately [23, 27]. Discriminant analysis has been used for decades in many biomedical applications.

In 1994, (Trevor Hastie, Robert Tibshirani, and Andreas Buja) proposed FDA [28] One can identify a smaller number of discriminant coordinate functions that are "optimal" for dividing the groups when there are a lot of predictors. A classification map that divides the reduced space into areas that can be distinguished by group membership and have linear decision borders may be created using two of these functions. Richer nonlinear classification systems are the topic of this research. When utilizing ideal scorings to represent the groups, Linear Discriminant Analysis is identical to multi-response linear regression. By substituting any nonparametric regression approach for linear regression, we were able to create nonparametric versions of Flexible Discriminant Analysis (FDA).

Hastie and Tibshirani (1996) [30] a model for discriminant analysis based on a mix

of Gaussians, each of which has a shared covariance matrix, has been presented. The well-known linear and quadratic discriminant analysis methods are based on common Gaussian assumptions, which are naturally extended by the mixed discriminant analysis (MDA) model [30]. We show that the MDA classifier recognizes three classes, each of which has non-adjacent subclasses, whereas the typical Gaussian assumption used in linear and quadratic discriminant analysis is insufficient and results in subpar decision boundaries.

Line CLEMMENSEN et al proposed Sparse Discriminant Analysis (SDA) [31], a technique for carrying out Linear Discriminant Analysis while imposing sparseness criteria, allowing for simultaneous feature selection and classification. If class boundaries are nonlinear or if the subgroups within each class are present, Sparse Discriminant Analysis (SDA), based on the optimal scoring interpretation of LDA, can be expanded to perform sparse discrimination using mixtures of Gaussians.

Together with other researchers, Sankari et al. [35] developed a model for the most difficult task in cancer diagnosis: identifying a brain tumor. PCA, Route set theory, and the Wavelet approach have been used in the majority of this field's research studies. Convolutional neural networks were employed by the authors to resolve the issue. The authors suggested bias-field correction, intensity normalization, and image de-noising as methods for pre-processing images. The noise from the MRI is eliminated using Principal Component Analysis (PCA). For picture enhancement and feature extraction, histogram equalization is performed. Finally, CNN was applied to characterize the photos, which produced an improved outcome.

The author classified normal and abnormal brain MR images using the SVM classification approach [24]. For implementation purposes, Matlab 7.9 was used to extract the characteristics. Extracted data is utilized as a classification process input to determine if the findings are normal or abnormal. Normal photos are successfully categorized with an accuracy of 65%, whereas aberrant ones are not. Use of Radiant Basis Function (RBF) with categorization is the cause of the failure. This study claims that SVM

cannot produce reliable findings when dealing with big amounts of data.

As well using the SVM classification approach was another author[25]. 140 MR images of brain tumors were collected from an online brain tumor archive. For tumor detection, a large dataset is employed, which produced noticeably better results. Shape, intensity, and texture are used to extract features. PCA and LDA are two analytic methods that are used to minimize the characteristics after selection. The accuracy rate has increased to 98.77 percent.

In 2017[26], authors use biologically inspired BWT and SVM to perform image analysis for MRI-based brain tumor classification and feature extraction. They conclude that in comparison to manual detection carried out by radiologists or clinical professionals, the analysis for brain tumor identification is quick and accurate based on experimental findings done on various pictures.

With a similarity index of 96.20 percent, an overlap fraction of 95 percent, and an additional fraction of 0.025 percent, Kumar and Vijayakumar proposed the segmentation and classification of brain tumors based on PCA [33, 32] and RBF kernel-based SVM. This method's classification accuracy for determining the kind of tumor is 94 percent, with a total of 7.5 percent mistakes found.

In a different research, the thresholding segmentation method is used to segment MR images. Images are first converted to grayscale before being filtered to reduce noise and improve brightness or sharpness for better output. It is utilized as a classifier using the SVM classification approach to show whether the condition is benign or normal.

The authors [34] of this research suggested a four-step methodology for classifying brain tumors. They used pre-processing, segmentation, feature extraction, and classification in their strategy. A novel method called hybrid SVM is used to categorize tumors. On the OASTS dataset, this method was used. The OASTS dataset took significantly less time to classify than the other two. They used the hybrid methodology

in their method, which produced a high detection rate and quick classification speed.

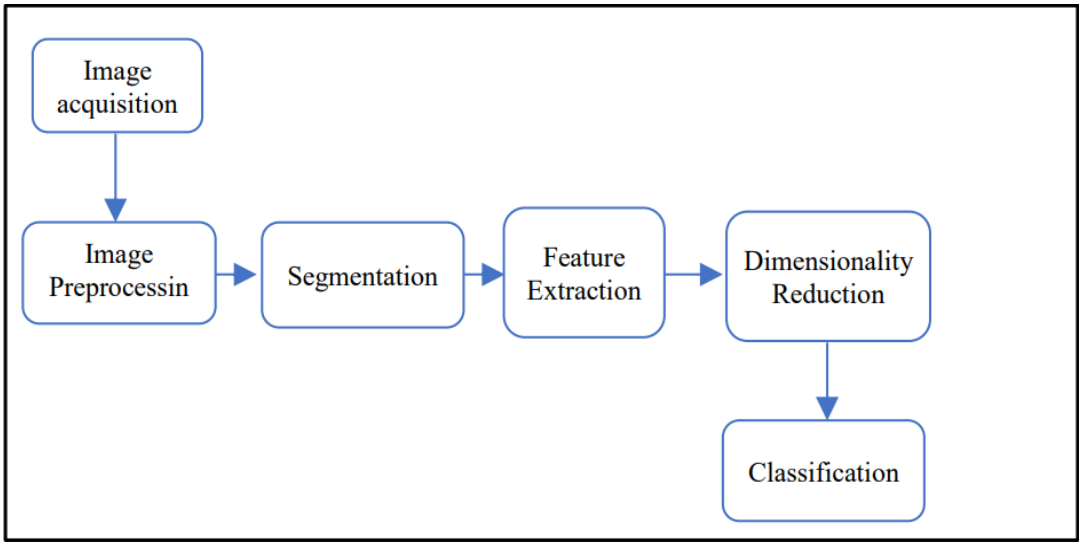


Figure 3.1: General system for brain tumor analysis

Chapter 4

Featured Methodology

4.1 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) [21] based on the Bayes theorem. The mathematical formation of Linear Discriminant Analysis (LDA) is
Given a data matrix X and a class variable Y, Bayes Theorem states that

$$P(Y = k|X = x) = \frac{P(X = x|Y = k)P(Y = k)}{P(X = x)} \quad for\ k = 1, 2, ..., K \tag{4.1}$$

Where,
$P(Y|X)$ the posterior probability
$P(X|Y)$ the likelihood
$P(Y)$ the prior probability of class
and $P(X)$ the prior probability of predictor.

To calculate the posterior probability from the likelihood and prior probabilities, using the Bayes rule assumption to obtain the estimation, which states that Model a $P(Y|X) = f_k(x)$ as a multivariate normal distribution so,
Suppose that $P(Y = k) = \pi_k$
$P(X = x|Y = k)$ is multivariate normal distribution with σ and μ_k

π_k = sample number in k class divided by a total number of samples. π_k is usually estimated by the empirical frequency of the training set.

$$f_k(x) = N_k(\mu_k, \sigma) \tag{4.2}$$

where

μ_k mean for all k

σ covariance matrix

By Bayes rule,

$$P(Y = k|X = x) = \frac{f_k(x)\pi_k}{P(X = x)} \tag{4.3}$$

$P(X = x)$ is a constant function. C represents those constant values that don't depend upon k so

$$P(Y = k|X = x) = C * f_k(x)\pi_k \tag{4.4}$$

Now $f_k(x)$ is,

$$P(Y = k|X = x) = \frac{C\pi_k}{2\pi_k^{p/2} * |\sigma|^{1/2}} \exp^{-1/2(x-\mu_k)^t\sigma^{-1}(x-\mu_k)} \tag{4.5}$$

Taking log both sides and get,

$$\zeta_k(x) = log\pi_k - 1/2\mu_k^t\sigma^{-1}\mu_k + X^t\sigma^{-1}\mu_k \tag{4.6}$$

So the linear boundary is,

$$log\pi_k - 1/2\mu_k^t\sigma^{-1}\mu_k + X^t\sigma^{-1}\mu_k \tag{4.7}$$

For 2-class classification

$$\zeta_k(x) = \zeta_l(x) \tag{4.8}$$

$$log\pi_k - 1/2\mu_k^t\sigma^{-1}\mu_k + X^t\sigma^{-1}\mu_k = log\pi_l - 1/2\mu_l^t\sigma^{-1}\mu_l + X^t\sigma^{-1}\mu_l \tag{4.9}$$

For multi-class classification,

$$log\pi_k - 1/2\mu_k^t\hat{\sigma}^{-1}\mu_k + X^t\hat{\sigma}^{-1}\mu_k = log\pi_l - 1/2\mu_l^t\hat{\sigma}^{-1}\mu_l + X^t\hat{\sigma}^{-1}\mu_l \tag{4.10}$$

Here ζ sign is used for posterior probability.

4.1.1 Regularized Discriminant Analysis (RDA)

Regularized Discriminant Analysis (RDA) [29] is a combination between LDA and QDA as it diminishes σ_j to a pooled variance σ by describing,

$$\hat{\sigma}_j(\beta) = \beta\hat{\sigma}_j + (1-\beta)\hat{\sigma} \tag{4.11}$$

and substituting σ_j with $\sigma_j(\beta)$ in the discriminant functions. Here, $\beta \in [0,1]$ is a tuning parameter defining whether the covariance should be estimated independently ($\beta = 1$) or should be pooled ($\beta = 0$). Furthermore, $\hat{\sigma}$ can also be minimized toward the scalar covariance by making

$$\hat{\sigma}(\lambda) = \lambda\hat{\sigma} + (1-\lambda)\hat{\sigma}^2 I \tag{4.12}$$

where $\lambda = 1$ indicates the pooled covariance and $\lambda = 0$ indicates the scalar covariance. Substituting $\hat{\sigma}_j$ by $\hat{\sigma}\beta, \lambda$ leads to a more general perception of covariance. Although RDA is a regularization approach, it is especially beneficial when numerous attributes may be associated.

RDA restricts QDA's distinct covariance to the common covariance of LDA. This increases covariance matrix estimates when the number of predictors is greater than the number of samples in the training data, resulting in improved model accuracy.

4.1.2 Flexible Discriminant Analysis (FDA)

Using optimal scoring, Flexible Discriminant Analysis [28] transforms the response variable to train the classifier for linear separation. It is a classification model based on a combination of linear regression models. Mathematical derivation of FDA by optimal scoring is,

Let X be the $N * K$ indicator matrix with mean zero and unit variance and Y be a class variable. Let $\Theta_{(J*K)}$ be a matrix of $J = 0, 1, 2....j$ score vectors for K classes. If we take Θ_0^* as a $N * K$ matrix of transformed values of the classes, then

$$\Theta_0^* = Y\Theta \tag{4.13}$$

Suppose an initial score matrix Θ_0 With J rows and K columns such that J≤K satisfies the constraints,

$$\Theta^T D_p \Theta = I \tag{4.14}$$

Where

Θ is the initial score matrix.

D_p be a diagonal matrix of class proportions.

$$D_p = \frac{1}{n} Y^T Y \tag{4.15}$$

If the scores were fixed, the average squared residual could be minimized by regressing Θ_0^* on data matrix X. Those responses are obtained by assigning scores to the classes such that the transformed class labels are optimally predicted by regression on X.

Make a transformation by fitting a multivariate response non-parametric regression of Θ_0^* on X. Which gives the estimated values $\hat{\Theta}_0^*$.

Obtain the eigenvector matrix Φ by multiplying the linear operator S($\hat{\lambda}$) with transformed value $\hat{\Theta}_0^*$.

$$\Phi = \Theta_0^{*T} \hat{\Theta}^* = \Theta_0^{*T} S(\hat{\lambda}) \Theta_0^* \tag{4.16}$$

Where $\hat{\lambda}$ is the hyper-parameter that is used to specify the amount and type of scoring. Hence the optimal scores,

$$\Theta = \Theta_0 \Phi \tag{4.17}$$

From equation (4.17), we can update the model by using optimal scores $n(x) \leftarrow \Phi^T n(x)$. $n(x)$ are the number of scores of fitted regression functions.

Flexible Discriminant Analysis (FDA) can be considered as the application of LDA on the matrix obtained with the non-parametric regression and on the transformed class matrix.

4.1.3 Mixture Discriminant Analysis (MDA)

By Hastie and Tibshirani (1996) [30] each subclass has a multivariate normal distribution with its own mean vector μ_{kr} and common covariance matrix σ. Assume that we have K classes in our classification system and the number of sub-classes in each class is $R_k, k = 1, 2, ..., K$. Then the mixture density for class k as

$$f_k(x) = P(X = x|G = k) = |2\pi\sigma_k|^{-1/2}\sum_{r=1}^{R_k}\pi_{kr}exp(-D(x, \mu_{kr})/2) \tag{4.18}$$

where X is a data matrix.
G refers to the class of a given object.

σ_k is a covariance matrix assumed common to the mixture sub-classes of class k.

π_{kr} and μ_{kr} are the mixing probabilities and mean of the r-th subclasses of the k class.
$D(x, \mu_{kr})$) refers to the Mahalanobis distance between x and μ_{kr}.

The conditional log-likelihood is

$$l^{mix}(\theta) = l^{mix}(\mu_{kr}, \sigma_k, \pi_{kr}) = \sum_{i=1}^{N} log(f_{gi}(x_i)) \tag{4.19}$$

The EM algorithm provides a convenient method for maximizing $l^{mix}(\theta)$ The EM steps are

E-step: For each class k, collect the samples for this class and compute the posterior probabilities of all the R_k components.

$$\frac{\pi_{kr}exp(-D(x, \mu_{kr})/2)}{\sum_{r=1}^{R_k}\pi_{kr}exp(-D(x, \mu_{kr})/2)} \tag{4.20}$$

M-step: Compute the weighted MLEs for all the parameters. Where $\hat{p}(C_{kr}|x, k) =$ Prob(x $\in$ r-th subclass of class $k|x, k$) is the function of $\hat{\mu}_{kr}$ and $\hat{\sigma}$.

$$\hat{\pi}_{kr} \propto \sum_{gi=k} p(c_{kr}|x_i, k), \quad \sum_{r=1}^{R_k}\hat{\pi}_{kr} = 1 \tag{4.21}$$

$$\hat{\mu}_{kr} = \frac{\sum_{gi=k} x_i p(c_{kr}|x_i, k)}{\sum_{gi=k} p(c_{kr}|x_i, k)} \tag{4.22}$$

$$\hat{\sigma} = \frac{1}{N} \sum_{k=1}^{K} \sum_{gi=k} \sum_{r=1}^{R_k} p(c_{kr}|x_i, k)(x_i - \mu_{kr})(x_i - \mu_{kr})^T \tag{4.23}$$

The notation $\sum_{gi=k}$ used for the sum of all the observations belonging to the k-th class, N refers to the total number of training data,

The equation (4.21) is the estimation step, whereas (4.22) to (4.24) are the maximization steps.

Using the posterior class probabilities via Bayes theorem, are of the form,

$$P(G = k|X = x) \sim \Pi_k Prob(x|k) \sim \Pi_k \sum_{r=1}^{R_k} \pi_{kr} exp(-D(x, \mu_{kr})/2) \tag{4.24}$$

where Π_k is the prior probability for class k estimated from the training data.

4.1.4 Sparse Discriminant Analysis (SDA)

In Sparse Discriminant Analysis (SDA) [31], let suppose,

X be an $n * k$ matrix of $X_{ij} = 1_{i \in (C_j)}$ and is orthogonal to all of the columns of X.

and suppose that each of the n observations fall into one of K classes. Suppose that each of the p features have been centered to have $\mu = 0$, and the features have been standardized to have equal variance if they are not measured on the same scale.

$$D_p = \frac{1}{n} X^T X \tag{4.25}$$

D_p be a diagonal matrix of class proportions.

For $k = 1$, Q_1 be a $k * 1$ matrix of 1's.

For $k = 1, 2, ..., m$ SDA for (λ_k, δ_k) as follows:

(a) $\lambda_k = (I - Q_k Q_k^T D_p)\lambda_*$ is a random vector of data matrix X.

Simplify $\lambda_k = \lambda_k^T D_p \lambda_k = 1$

(b) Iterating the function until a maximum number is obtained as convergence,

let δ_k be the solution to the generalized problem. And γ and φ are non-negative tuning

parameters.

$$minimized\delta_k[\frac{1}{n}||X\lambda_k - Y\delta_k||^2 + \gamma\delta_k^T\Omega\delta_k + \varphi||\delta_k||_1] \tag{4.26}$$

For fixed δ_k,

$$\hat{\lambda_k} = (I - Q_k Q_k^T D_\xi)D_\xi^{-1}X^T Y\delta_k, \qquad \lambda_k = \frac{\hat{\lambda_k}}{\sqrt{\hat{\lambda_k}^T D_\xi \hat{\lambda_k}}} \tag{4.27}$$

(c) If $k < m$, $Q_k + 1 = (Q_k, \lambda_k)$

The classification results by performing this technique with $n*q$ matrix $(Y\delta_1, Y\delta_2, ..., Y\delta_q)$

Once sparse discriminant boundaries have been obtained, we can plot the matrices $X\beta_1, X\beta_2$, and so on to perform data visualization in the reduced subspace. The classification rule is obtained by performing standard LDA on the $n * q$ reduced data matrix $(X\beta_1.....X\beta_q)$ with $q < k$.

Chapter 5

Data Explanation and Statistical Software

Datasets are essential to the results of classification and the learning process. Firstly, we discuss the application which we used to classify.

5.1 Brain Tumor

The brain is indeed one of the body's vital organs since it orchestrates all of the body's movements. Infections, strokes, and tumors are just a few of the many illnesses that may impact the human brain.

What is a brain tumor?

A brain tumor is a growth or expansion of abnormal cells formed in the parenchyma or surrounding brain tissue. Brain tumors are classified into two categories: primary tumors and secondary tumors. A primary brain tumor originates in the brain. It can either be benign (doesn't contain cancer cells) or malignant (does not contain cancer cells). A subsequent or metastatic brain tumor develops when cells from a primary malignant tumor that originated elsewhere in the body travel to the brain and begin to grow there[17].

X-rays, powerful magnets, or radioactive chemicals are used to make images of the brain for image testing of a brain tumor. Various scan kinds are often used to diagnose brain cancers. The most often utilized scan types to identify brain illnesses are

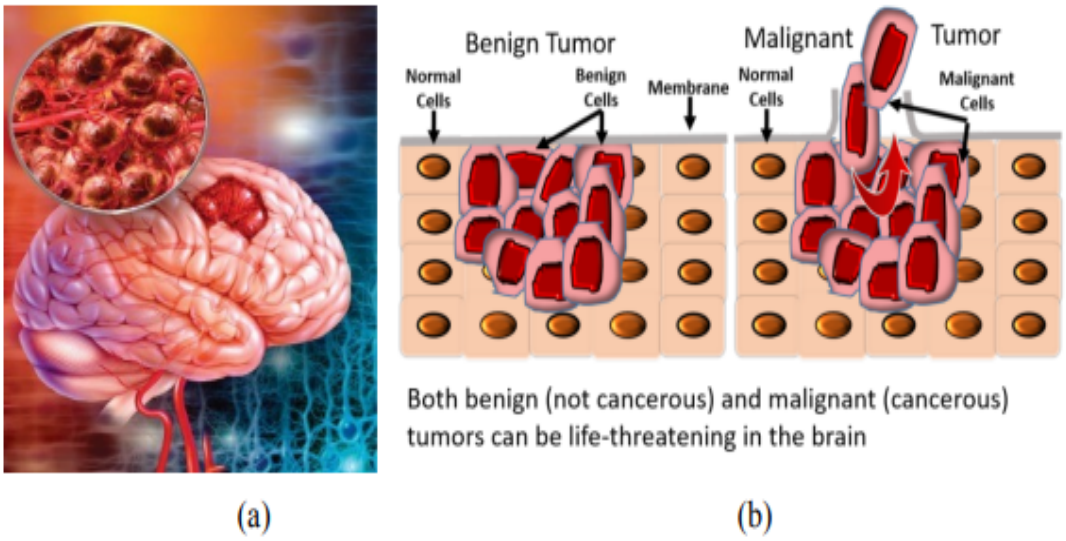


Figure 5.1: Example of brain image (a) and (b) with tumor

Positron Emission Tomography (PET), Computer Tomography (CT), and Magnetic Resonance Imaging (MRI). These images are so effective that they can provide crucial facts regarding a tumor's presence and location.

According to the World Health Organization (WHO), [11, 12]. , cancer ranks as the second most common cause of death worldwide. One of the main reasons why deaths among adults and children are rising globally is brain tumor, which places a significant strain on families and healthcare systems. Only brain cancer is to blame for about 13% of all fatalities worldwide. At the age of 15, 3,540 kids were diagnosed with brain tumors in the year 2020 [13].

The biopsy of a brain tumor is typically not performed before total brain surgery, in contrast to cancers located elsewhere in the body. First, benign tumors cannot metastasize outside of the brain. The majority of benign brain tumors are self-limiting and do not require treatment. Because of their location, they may create difficulties, however, surgery or radiation can be beneficial. The second is that malignant tumors are commonly referred to as brain cancer[19]. The brain may not be the only location of these malignancies. Malignant brain tumors will inevitably become a problem if left

untreated, and forceful treatment is nearly always necessary. Malignancies of the brain can be categorized into two groups. The brain is where primary brain cancer starts. The brain is affected by secondary or metastatic brain cancer that has spread from another part of the body [13, 14].

When bodily cells (in this example, brain cells) divide uncontrollably, cancer develops. Brain cancer patients may experience severe limitations that severely limit their everyday activities and lower their quality of life. It creates resistance while the brain is trying to operate. This anomaly or dysfunction is a sign of a brain tumor. The ability of radiologists to identify brain tumors early mostly depends on their experience. Radiologists frequently utilize magnetic resonance imaging (MRI) to assess brain tumors to see whether the proper classification is in place[15].

The first step in addressing the appalling number of deaths caused by brain tumors is the classification of normal and abnormal brain pictures obtained from MRI. However, due to the volume of data from MRI, manual classification is laborious, difficult to perform, takes a long time, and calls for experience. The spectator has a very tough time understanding and evaluating the photos to find the tumor. Consequently, it is essential to create and put into use an automated picture analysis system. It needs to be quicker, more precise, and simple to use when concluding MRI scans. Numerous autonomous and precise medical diagnosis approaches that employ sophisticated signal/image processing techniques and computationally clever machine learning algorithms have been the subject of research.

The application of statistical methods to MRI offers knowledge to medical professionals and aids in the diagnosis of a wide range of illnesses and ailments. Using supervised or unsupervised algorithms, the data from MRI pictures may be evaluated, processed, and divided into normal and pathological groups. However, how well the characteristics are extracted from the images and how useful they are in identifying the condition will decide how accurate the classification is. The preferred option to balance the disadvantages is to select the best technique that can determine the fewest,

most relevant features possible to obtain the complete characteristic anatomy of the tumor, thereby reducing the additional computational complications for unnecessary feature extraction. Extraction of meaningful features is important, but it also increases the computational burden of the classifier. The Fast Principal Component Analysis approach is one good method that considers the restrictions.

5.2 Medical Imaging and Brain Tumor Diagnostic Approaches

A prompt diagnosis aids in the therapeutic process. Different methods, such as brain imaging and brain biopsy, are used to diagnose tumors, their causes, and their consequences [16].

5.2.1 Computer Tomography (CT scan)

An essential imaging method in the medical profession, a CT scan provides results in a matter of seconds and often takes just a small fraction of the time. Compared to X-rays, it aids in delivering clearer information, but the danger of radioactive contamination is quite minimal.

5.2.2 Positron Emission Tomography (PET)

Positron emission tomography, or PET, is a procedure in which a radioactive substance is injected into the blood and a scanner picks it up to provide a picture. This method provides insight into the activity and operation of the brain. This technique uses toxic materials but is also cost-effective.

5.2.3 Magnetic Resonance Imaging (MRI)

Medical professionals can diagnose diseases and make decisions since MR pictures don't expose them to dangerous radiation [15]. Brain tumor identification and diagnosis preprocessing involve the use of MR images. Of all the scan kinds, an MRI is the most

commonly used for diagnosis. Instead of using X-rays, MRI [20] employs radio waves and powerful magnetic fields. The kind of bodily tissue and specific disorders determine how the radio wave radiation is absorbed and then released.

The pattern is converted by a computer into a very detailed representation of various bodily components. Gadolinium, a contrast agent and dye used in the scan, is administered into the patient's veins before the scan to provide more accurate results.

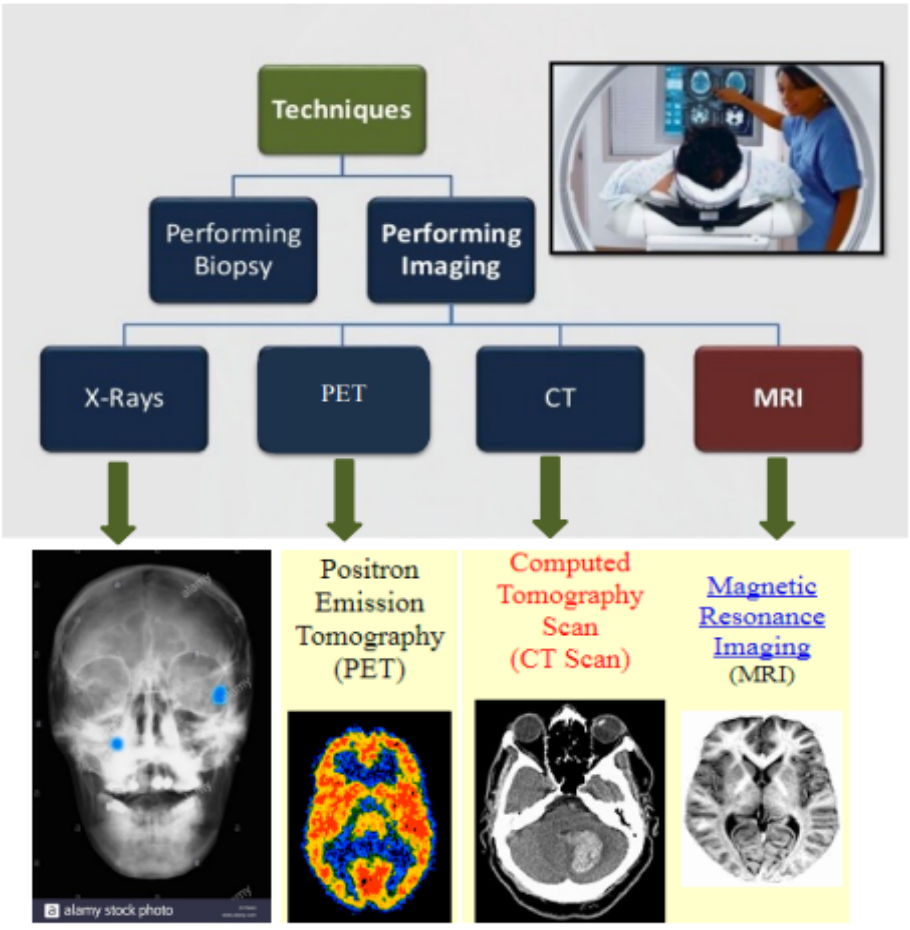


Figure 5.2: Different techniques of brain tumor imaging

5.2.4 Biopsy of Brain

A brain biopsy is a technique in which a hole is drilled in the skull and a portion of tissue and a tumor are extracted to be examined under a microscope to determine the kind, content, and source of the tumor. This method poses a serious threat to human life. To find the tumor and get a sample of tissue, a biopsy also uses imaging technology.

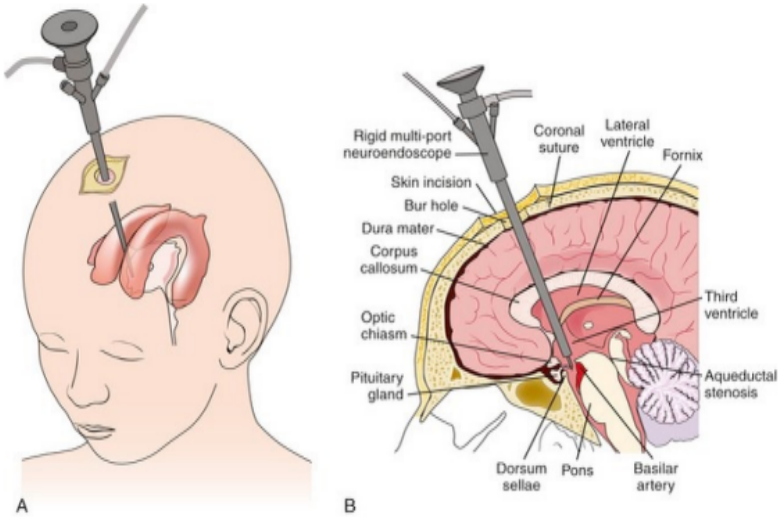


Figure 5.3: The procedure of brain biopsy is presented.

Why are MRI-scanning images preferred over other imaging methods?

1) MRI is non-invasive.

2) MRI is economical.

3) It provides good contrast of brain tumors.

4) MRI acquisition time (total body scan) is shorter than that of PET and X-ray.

5) MRI gives an excellent description of skeletal structure and organ systems behind them, such as the lungs behind the ribs and the brain underneath the skull.

5.3 Statistical Software

R is used for both computations statistical analysis and modeling. https://www.R-project.org/. Some R libraries, including MASS, rda, sda, and dplyr, are utilized in the proposed study for classification purposes.

5.4 Dataset

The data set was collected to classify brain tumors using Magnetic Resonance Imaging (MRI) obtained from the Kaggle link
https://www.kaggle.com/navoneel/brain-mri-images-for-brain-tumor-detection
253 MRI pictures were included in the data used to categorize the data for this thesis. There are 155 of them with the label "yes," indicating the presence of a tumor, and 98 with the label "no," indicating the absence of a tumor.
5.4 shows the images of tumor patient and 5.5 shows the image of normal patient.

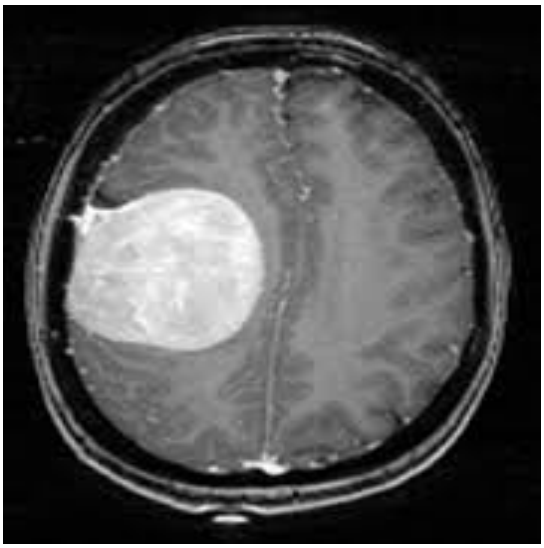


Figure 5.4: MRI x-ray of the patient when the tumor is present.

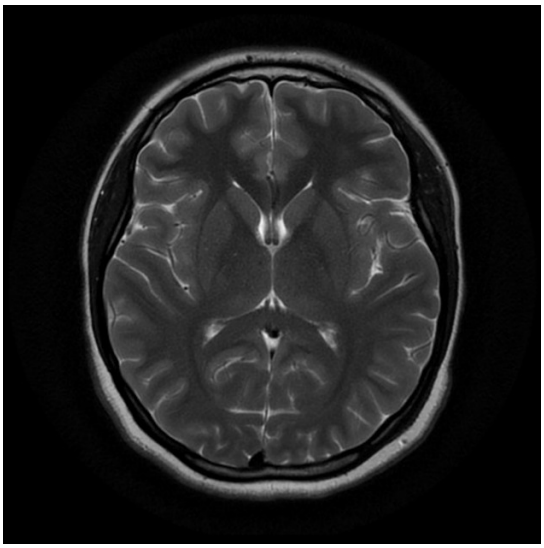


Figure 5.5: MRI x-ray of normal patient when the tumor is absent.

Chapter 6

Results & Discussions

In this thesis, Linear Discriminant Analysis with its variants and FAST Principal Component Analysis are the algorithms that are applied to brain tumor MRI. A detailed process has been done with MRI images given below.

The actual images of the database have different spatial and high resolutions. This determines challenges for computer hardware and have a high computational cost. Data is divided into two folders. 155 images in one folder of tumorous patients. The second folder represents the images of non-tumorous patients.

Table 6.1: The length and percentage of data folders are presented in the table.

Tumor Patients	155
Normal patients	98
% of tumor patient's folder	61.26482
% of normal patient's folder	38.73518

The input data of the proposed system is in different shapes (e.g., 225X225, 630X630, 180X280). In pre-processing, we changed the size of the images to 500X500. Then compared the original images with the resized one shown in the 6.1.

The 253 brains MRI resized images each of $500x500$ were used for the analysis. The re-sizing that was performed on the whole database without destroying much information is shown in 6.1. For two images, it is possible to observe both the original images and the resized versions. This procedure must be followed for every one of the patient's

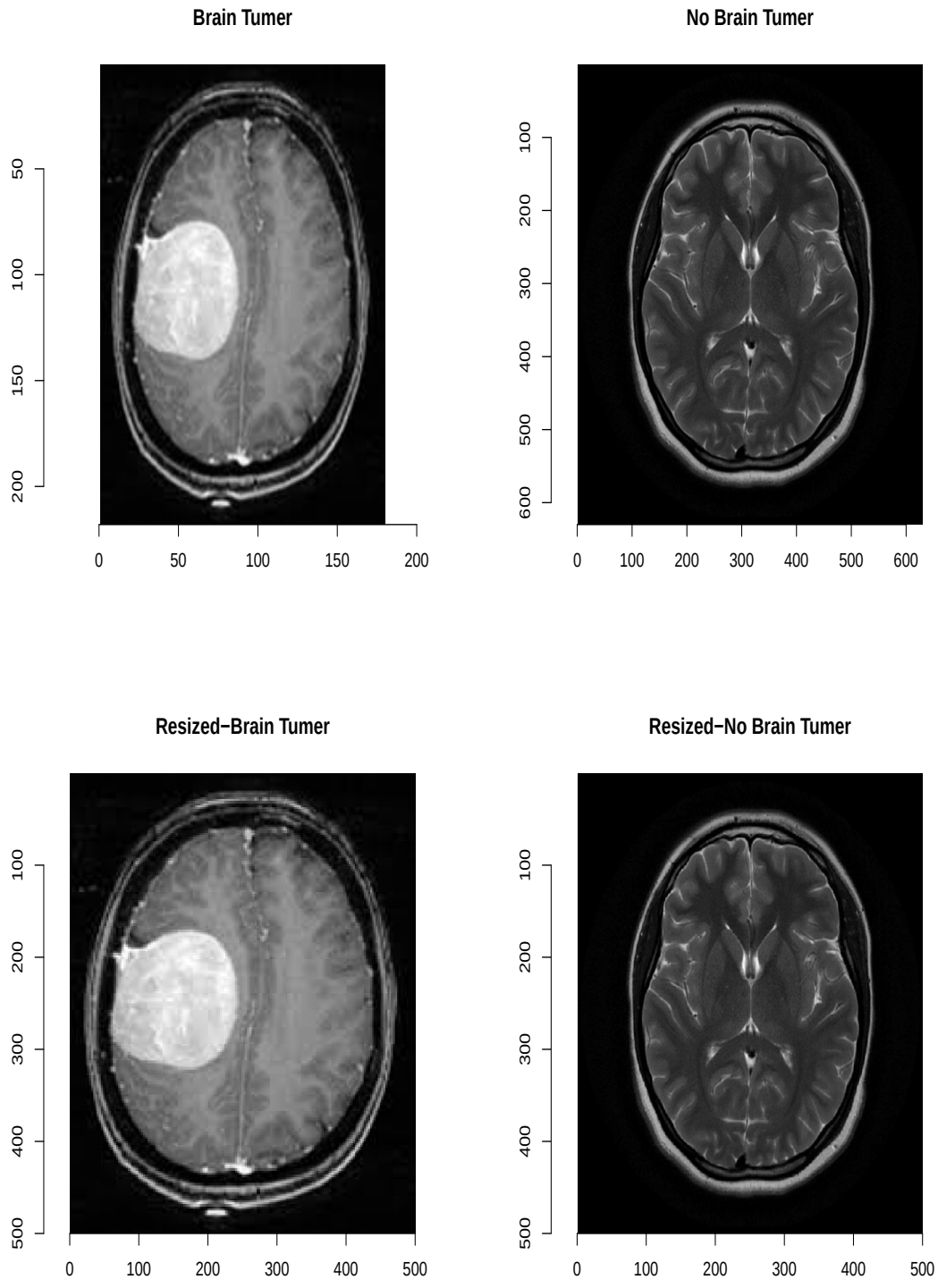


Figure 6.1: Pre-processing results are presented. The comparison between an original image and resized one.

images. R studio is used for all image processing as well as programming. Analyzing the images and converting them into a vector of pixel values rather than a matrix of pixel values is required since the classification technique relies on vectors.

Before building the discrimination model we used PCA for the removal of outliers. For this purpose PCA biplot is constructed for both classes separately and is presented in 6.2.

The biplot, which combines the principal component values and the loading vectors in a discrete biplot presentation, is a widely common method for visualizing the findings from PCA. The points lying outside of the outer circle can be marked as an outlier, similarly, points lying inside of the inner circle can be marked as collinear and can be removed from the analysis. It appears 227 MRI images, out of them 139 have a brain tumor, and the remaining 88 have no tumor.

The full data has 253 samples while the limit where outliers are removed has 227 samples.

In order to classify response class vector Y based on the MRI images based data matrix X, we have applied LDA and its variants over the full data and the data with removed outliers with PCA through Monte-Carlo simulation.

In the Monte-Carlo simulation part, we go through methods for assessing classifier performance. The training error rate is one of the performance metrics taken into account. Assume there are n training data points for Cross-Validation.

The results of this approach include high bias but low variance, whereas, in Monte Carlo cross-validation, each data point is evaluated arbitrarily many times. Repetitive random selection and statistical approaches are used to achieve the Monte Carlo Cross-Validation results.

This approach is analogous to arbitrary experiments, where the final result is not known in advance. In the natural sciences, social sciences, and technical domains, mathematical models are used to express system dynamics. These models frequently begin with a set of input parameters, which are then processed by the model's mathematical formulae to create one or maybe more outputs.

It is also known as repeated random sub-sampling Cross-Validation. It randomly di-

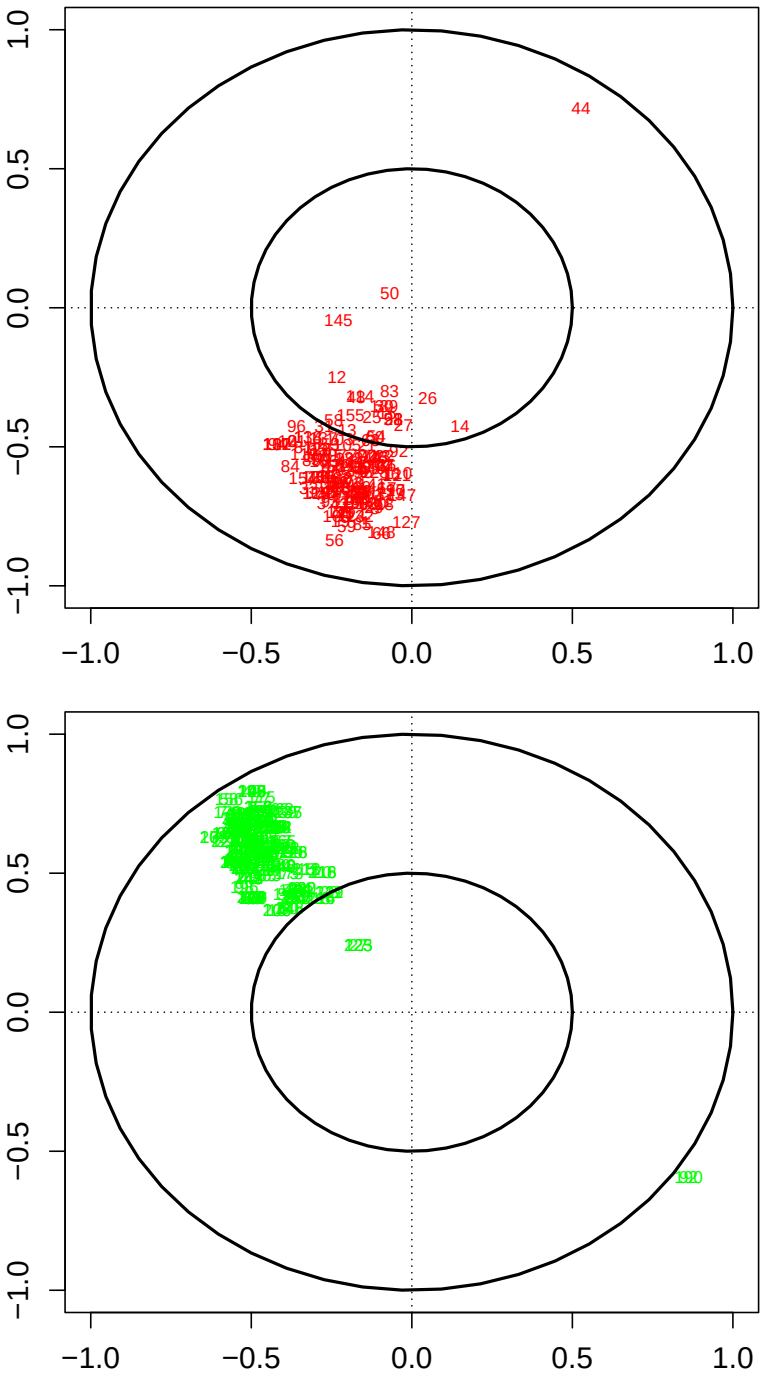
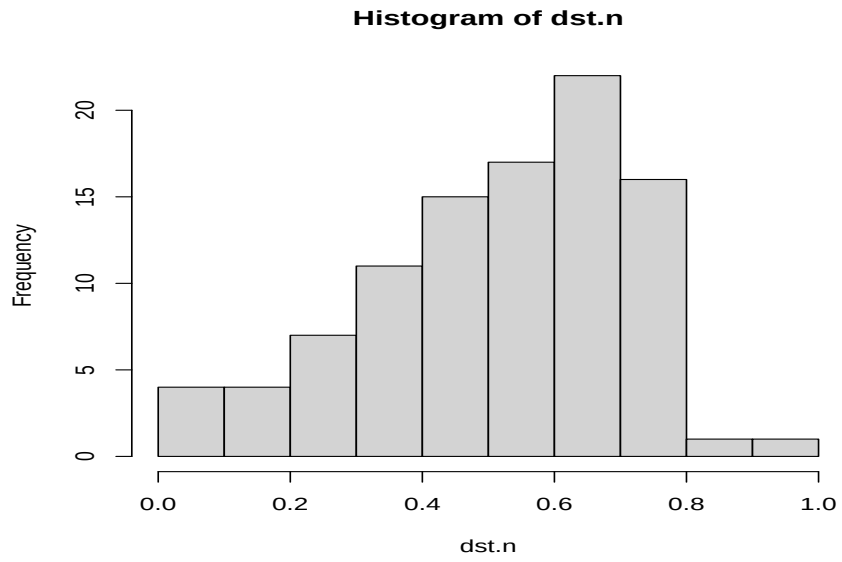


Figure 6.2: The right panel presents the PCA biplot of patients having a brain tumor, while the lower panel presents the PCA biplot of patients having no brain tumor.

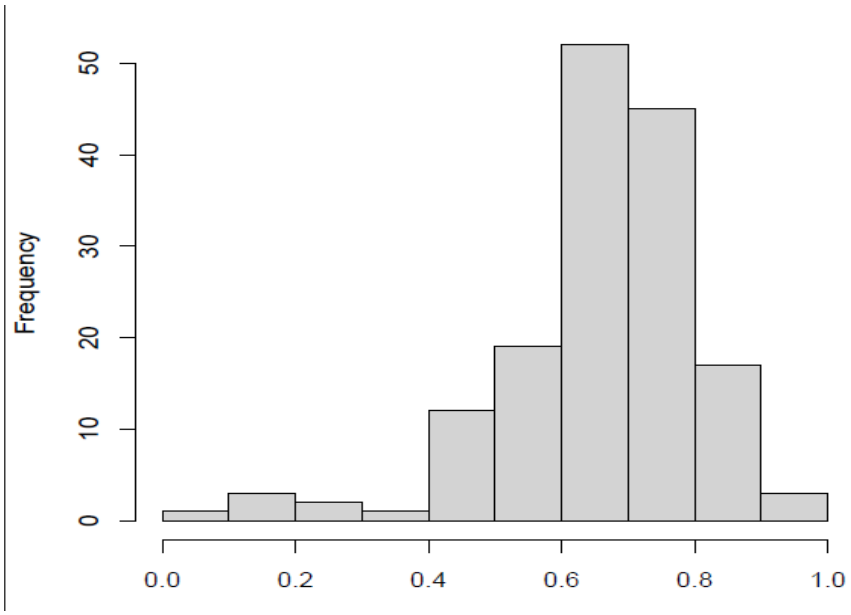
vides the training data (perhaps 70–30% or 60–40%). That iteration fitted the classifier to the train data set. The training data set and the validation set are separated from the data set. The validation set is used to calculate the test error, whereas the training set is utilized to train the model. Depending on the Cross-Validation approach, the procedure is repeated a few times by resampling the validation set and the training set. The model's test error is ultimately calculated as the mean of all test errors acquired by resampling.
We are going to take R package 'e1071' to help us to do the classification and prediction. In R, we can write a 'for loop' to build this Cross-Validation. For the comparison of different methods, we have used the Monte-Carlo simulation with 100 runs. When the runs were increased, the computer hanged from processing. It was assumed that the data was huge and required more computer hardware space. However, LDA, like other machine learning methods has a weakness, in that it is a time-consuming process. To deal with this problem, we re-scaled the data by PCA. 6.3 are shown the distribution of images after re-scaling by histogram.

The data is divided into 70% training and 30% testing. The accuracy of all fitted models is computed for both testing and training data. The calibration accuracy of training data is computed and presented through the graph.
For reliable comparison, we have used Monte-Carlo simulation with 100 runs, where in each run the MRI samples from each class were divided into 30% test and the rest in the training set. Training data was used to fit the MRI classifiers while the accuracy of these fitted models is computed from both test and training data, respectively called validation and calibration.

The validation and calibration accuracy from each run is computed and is presented in 6.4. The upper left panel compares the classifier's accuracy over training data when the full data set is used. In the left panel in the calibration graph, Methods are applied over full data when outliers are not removed and in the right panel, results are shown when methods are applied after removing outliers. Similarly, the result is presented in a validation graph.



(a)



(b)

Figure 6.3: (a) Re-scale the data of normal patients from 0 to 1. The distribution of normal patients is presented. (b) Re-scale the data of tumor patients from 0 to 1. The distribution of tumor patients is presented.

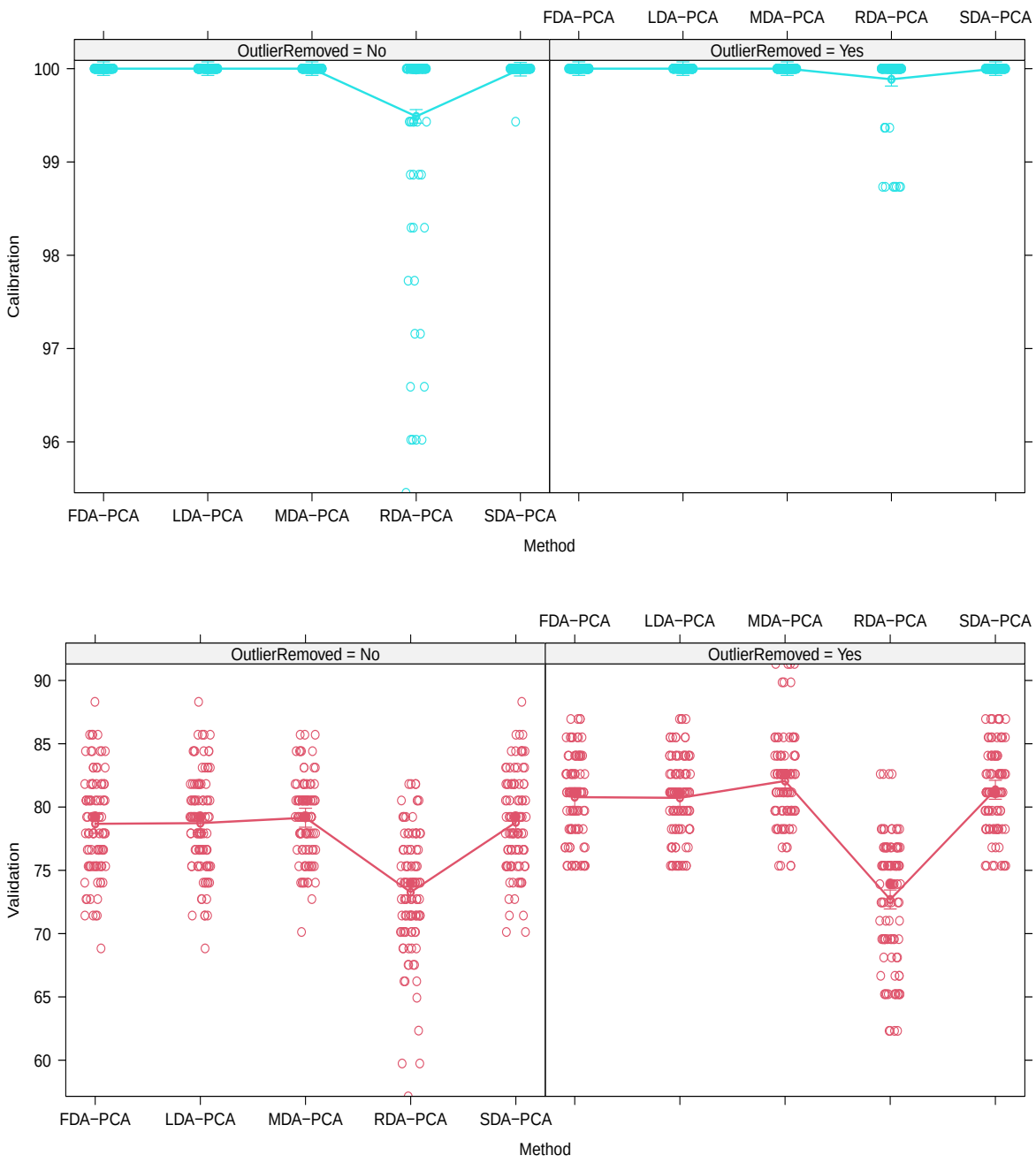


Figure 6.4: The calibrated and validated accuracy of all considered LDA variants with and without outlier removal is presented in the upper and lower panel respectively.

We removed outliers with PCA, and applied methods LDA-PCA, MDA-PCA, RDA-PCA, SDA-PCA, and FDA-PCA before removing outliers and after. All methods except RDA-PCA show 100 % accuracy while RDA-PCA has around 98 % accuracy in training data. The upper right panel compares the classifier's accuracy over training data when data with outlier removal is used. In 6.4, calibration graph, the results are presented. A similar pattern appears in validation. Notably, RDA-PCA has accuracy is improved by removing the outliers. The lower left panel compares the classifier's accuracy over test data when the full data set is used.
All methods except RDA-PCA show 80% accuracy while RDA-PCA has around 73 % accuracy. The lower right panel compares the classifier's accuracy over test data when the data with removed outlier is used. MDA-PCA outperforms all methods by showing 83.1 % accuracy on test data, while FDA-PCA, LDA-PCA, and SDA-PCA show 82.5 % accuracy, moreover RDA-PCA has around 71 % accuracy.

Over MDA-PCA classifier appears the winner with best for MRI image-based tumor classification over the test data and training data, whereas RDA-PCA appears the worst classifier. Well, each classifier being used here has some parameters to tune. For instance, the ratio 'R' presents the number of PCA scores being used with classifiers. We have used $R = (0.70, 0.73, 0.76, 0.79, 0.82, 0.85, 0.88, 0.91, 0.94, 0.97)$, lower R indicates lower number dimensions are used in final classifiers and higher R indicates higher number of dimensions are used.
In 6.5, the values of R are shown by a line graph. In the graph, different lines represent different methods. The Red line indicates RDA-PCA, which gives the worst results. On the other hand, the green line MDA-PCA shows better results. The upper panel presents the distribution of R for different methods. It appears PCA-RDA has the best accuracy with R=0.97, this means PCA-RDA is consuming a high amount of principal components and is not contributing to dimension reduction. FDA-PCA results from the best accuracy with R=0.79, while SDA-PCA demonstrates the maximum precision with R=0.91, moreover MDA-PCA illustrates the best accuracy with R=0.7 indicating is consuming the least level of principal components.

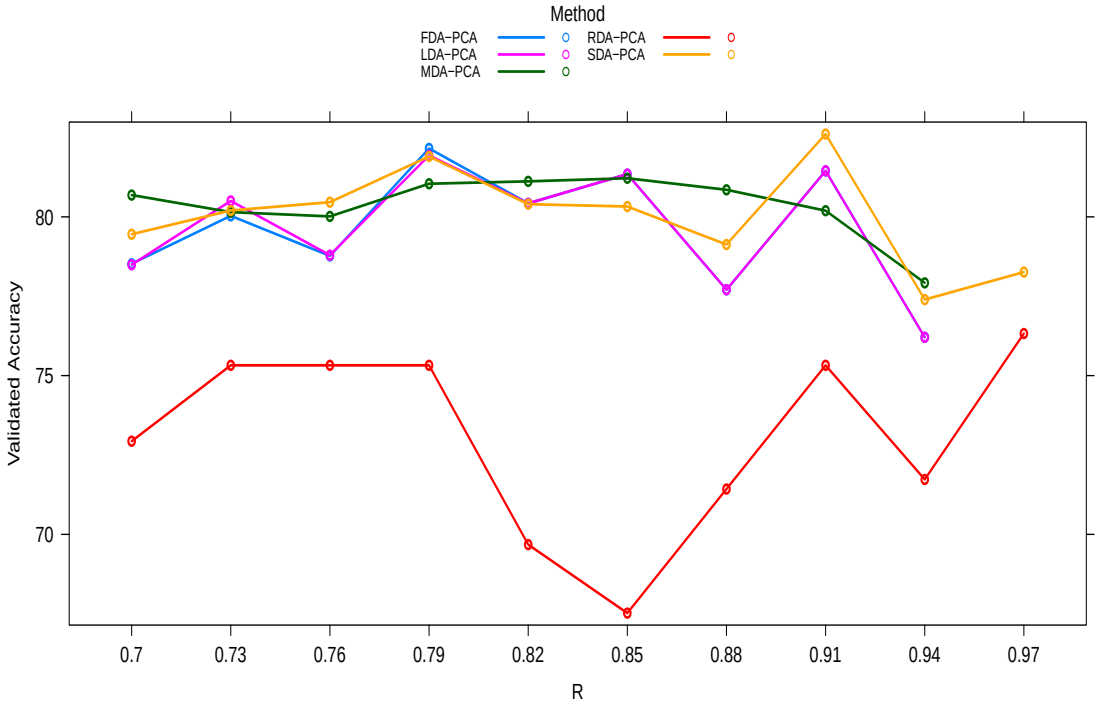


Figure 6.5: The distribution of R for respective methods is presented. The Red line indicates RD-PCA, which gives the worst results. On the other hand, the green line MDA-PCA shows a better result.

RDA requires tuning parameters δ and α which are the regularized parameter indicating the amount of balance between LDA and QDA 6.6. α value near zero indicates the operator performs Quadratic Discriminant Analysis (QDA), and α values toward 1 indicate the operator tends to perform Linear Discriminant Analysis (LDA).

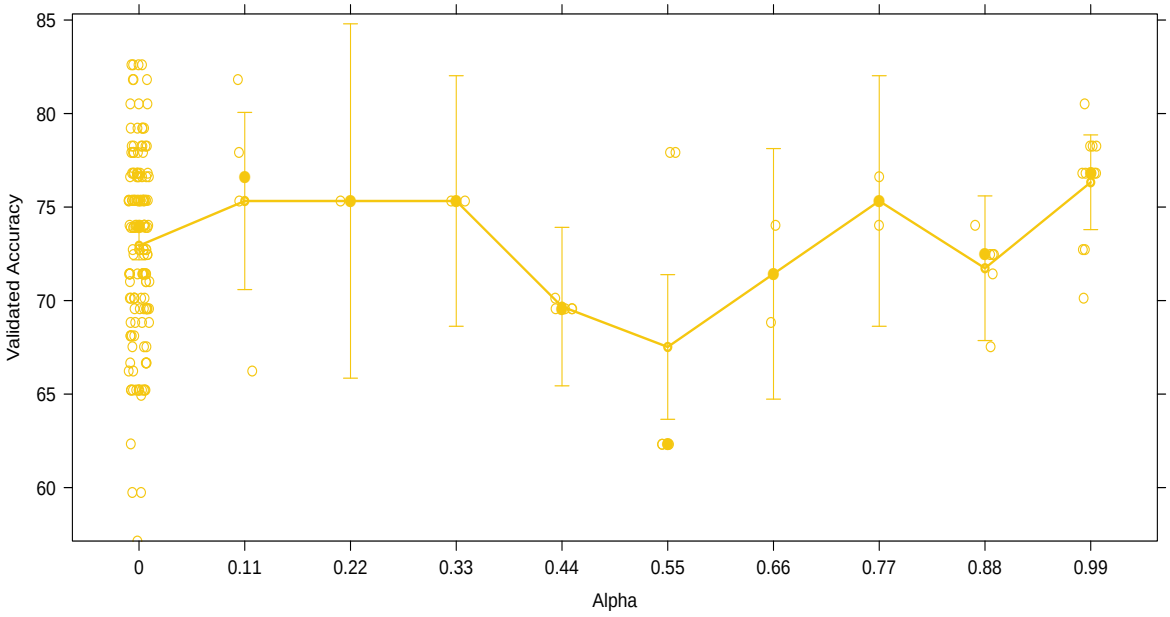


Figure 6.6: The distribution of *alpha* for respective methods is presented.

Higher δ indicates both classes have a common variance and lower δ indicates both classes tend to have different variances. To choose the optimal level, the accuracy of test data was computed at all levels the one which results in the best accuracy for a given Monte-Carlo simulation run is chosen. The optimal parameter being chosen from each run is presented in Figure 6.7. In RDA-PCA most of the average α and δ are close to 0 indicating RDA tends to be LDA for image-based classification.

Finally in order to statistically describe the impact of classification methods and outlier removal methods we have used the analysis of variance (ANOVA) approach. In ANOVA the response is taken as validated accuracy whereas the brain tumor classification methods and choice of outlier removal are taken as factors.

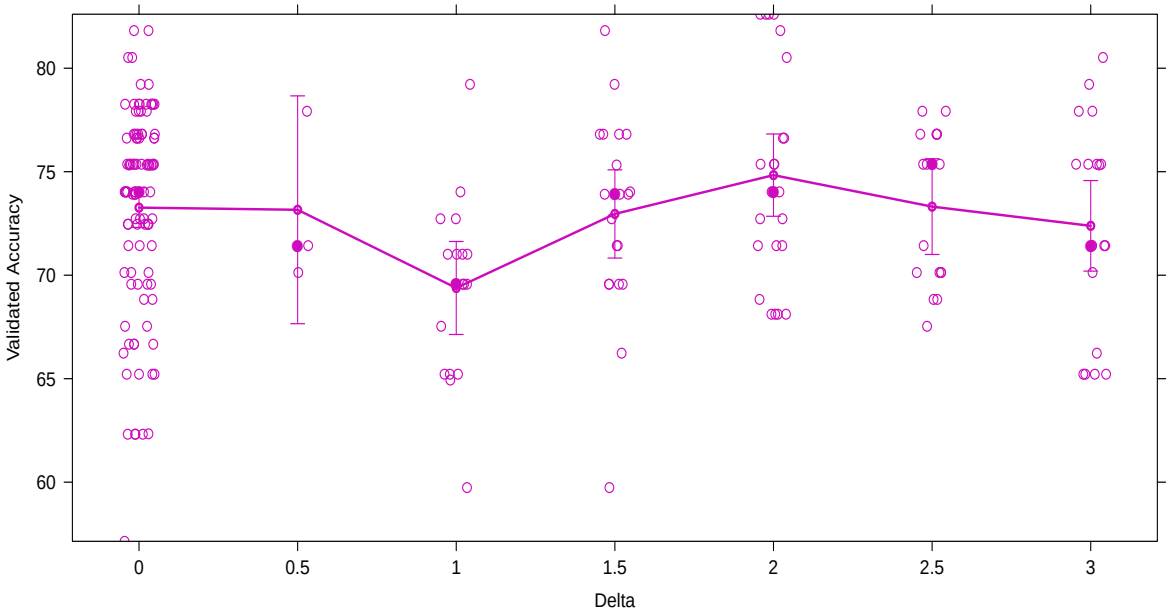


Figure 6.7: The distribution of δ for respective methods is presented.

To ascertain the statistically significant variations in some of the means, we use the p-value from the ANOVA results. And compare the p-value to determine the null hypothesis and decide on the statistically significant methods. The classification method has 5 levels LDA-PCA, FDA-PCA, RDA-PCA, MDA-PCA, and SDA-PCA, while the outlier removal has 2 levels yes or no. The ANOVA results are presented in Table 6.2. It appears that only RDA-PCA is significantly varying $p - value \leq 0.001$. The classification accuracy of SDA-PCA, MDA-PCA, and LDA-PCA is not statistically significant. Moreover, The classifiers have significantly $p - value \leq 0.001$ better accuracy when outliers are removed.

The coefficient's standard error serves as an estimation of its standard deviation. It effectively informs us of the degree of uncertainty surrounding our level of classification method. Simply dividing the coefficient by the standard error yields the t-statistic. Higher t-score numbers denote a significant difference between the two tested samples. All methods haven't a significant difference when outliers are presented. But the value of t-statistics is higher when the outliers are removed.

Table 6.2: The ANOVA results are presented, where the response is taken as validated accuracy whereas the brain tumor classification methods and choice of outlier removal are taken as factors. The classification method has 5 levels LDA-PCA, FDA-PCA, RDA-PCA, MDA-PCA, and SDA-PCA, while the outlier removal has 2 levels yes or no.

Factor	Level	Estimate	Std. Error	t-value	p-value
Intercept		99.96	0.03	3417.18	0.00
Method	LDA-PCA	-0.00	0.04	-0.00	1.00
	MDA-PCA	-0.00	0.04	-0.00	1.00
	RDA-PCA	-0.31	0.04	-8.28	0.00
	SDA-PCA	-0.00	0.04	-0.08	0.94
Outlier Removed	Yes	0.08	0.02	3.38	0.00

Chapter 7

Conclusion

MRI images of brain tumors numerous analyses have used feature selection and extraction for classification in the past. One approach for classification that is often utilized is Linear Discriminant Analysis. It is known to fail, though, if there are nonlinear decision boundaries between the classes. In this thesis, we utilized a couple of methods for generalizing LDA to a high-dimensional situation, resulting in a discriminant classifier that includes both robust and nonlinear solutions. Our suggestion is based on the straightforward optimal scoring framework, which recasts LDA as a classifier that takes less time and is easier to use.

The procedures of image segmentation, pre-processing, feature extraction, feature selection, and classification are part of the methodology provided in this thesis for the aforementioned study. For tumor diagnosis based on MRI images, we have employed LDA variations such as Flexible Discriminant Analysis (FDA), Mixture Discriminant Analysis (MDA), Sparse Discriminant Analysis (SDA), and Regularized Discriminant Analysis (RDA).

The brain MRI images can be classified by LDA-based methods coupled with PCA. The study compares 5 LDA-based methods levels LDA-PCA, FDA-PCA, RDA-PCA, MDA-PCA, and SDA-PCA. MDA-PCA shows the best classification accuracy while RDA-PCA shows the worst classification accuracy. Mixture Discriminant Analysis(MDA), a non-linear and robust extension of Linear Discriminant Analysis, with Principal Component Analysis (PCA) gives the best results as compared to other extensions. Moreover, it appears that PCA-based dimension reduction effectively classifies the MRI images related to a brain tumor. The proposed methods can further be used for other

image-based classification problems.

7.1 Limitations

Our thesis work has numerous limitations, including the following,
The brain tumor data set only contains 253 images. These LDA variants can also use for a large number of the dataset. In this thesis, we only worked with 2D images. 3D images can also be used for image-based classification. We might have used more conventional classifiers to enhance the accuracy. Tumor types could not be identified.

Bibliography

[1] J. M. Helm, A. M. Swiergosz, H. S. Haeberle, J. M. Karnuta, J. L. Schaffer, V. E. Krebs, A. I. Spitzer, and P. N. Ramkumar, "Machine learning and artificial intelligence: Definitions, applications, and future directions," *Current reviews in musculoskeletal medicine*, vol. 13, no. 1, pp. 69–76, 2020.

[2] title=Artificial intelligence, author=Winston, Patrick Henry, year=1992, publisher=Addison-Wesley Longman Publishing Co., Inc.

[3] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, vol. 112. Springer, 2013.

[4] title=Multivariate analysis, author=Bartlett, Maurice S, journal=Supplement to the journal of the royal statistical society, volume=9, number=2, pages=176–197, year=1947, publisher=JSTOR

[5] title=Fast and robust discriminant analysis, author=Hubert, Mia and Van Driessen, Katrien, journal=Computational Statistics & Data Analysis, volume=45, number=2, pages=301–320, year=2004, publisher=Elsevier

[6] title=Linear discriminant analysis, author=Xanthopoulos, Petros and Pardalos, Panos M and Trafalis, Theodore B, booktitle=Robust data mining, pages=27–33, year=2013, publisher=Springer

[7] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, "Interpretable machine learning: definitions, methods, and applications," *arXiv preprint arXiv:1901.04592*, 2019.

[8] J. S. Armstrong, "Illusions in regression analysis," *International Journal of Forecasting*, vol. 28, no. 3, p. 689–694, 2012.

[9] R. K. Paul, "Multicollinearity: Causes, effects and remedies," *IASRI, New Delhi*, vol. 1, no. 1, pp. 58–65, 2006.

[10] title=Analysis of variance (ANOVA), author=St, Lars and Wold, Svante and others, journal=Chemometrics and intelligent laboratory systems, volume=6, number=4, pages=259–272, year=1989, publisher=Elsevier

[11] Organization, W. H. *World health statistics 2008* (World Health Organization, 2008).

[12] Mattiuzzi, C. & Lippi, G. Current cancer epidemiology. *Journal of epidemiology and global health* **9**, 217 (2019).

[13] Johnson, K. J. *et al.* Childhood brain tumor epidemiology: a brain tumor epidemiology consortium review. *Cancer Epidemiology and Prevention Biomarkers* **23**, 2716–2736 (2014).

[14] Wadhwa, A., Bhardwaj, A. & Verma, V. S. A review on brain tumor segmentation of mri images. *Magnetic resonance imaging* **61**, 247–259 (2019).

[15] Sarhan, A. M. *et al.* Brain tumor classification in magnetic resonance images using deep learning and wavelet transform. *Journal of Biomedical Science and Engineering* **13**, 102 (2020).

[16] title=MRS of brain tumors: Diagrammatic representations and diagnostic approach, author=Shokry, Ahmed, journal=The Egyptian Journal of Radiology and Nuclear Medicine, volume=43, number=4, pages=603–612, year=2012, publisher=Elsevier

[17] Nazir, M., Shakil, S. & Khurshid, K. Role of deep learning in brain tumor detection and classification (2015 to 2020): A review. *Computerized Medical Imaging and Graphics* **91**, 101940 (2021).

[18] Bolón-Canedo, V., Sánchez-Maroño, N. & Alonso-Betanzos, A. Recent advances and emerging challenges of feature selection in the context of big data. *Knowledge-based systems* **86**, 33–45 (2015).

[19] Herholz, K., Langen, K.-J., Schiepers, C. & Mountz, J. M. Brain tumors. In *Seminars in nuclear medicine*, vol. 42, 356–370 (Elsevier, 2012).

[20] Mohan, G. & Subashini, M. M. Mri based medical image analysis: Survey on brain tumor grade classification. *Biomedical Signal Processing and Control* **39**, 139–161 (2018).

[21] Lin, G.-C., Wang, W.-J., Wang, C.-M. & Sun, S.-Y. Automated classification of multi-spectral mr images using linear discriminant analysis. *Computerized Medical Imaging and Graphics* **34**, 251–268 (2010).

[22] title=Data preprocessing in data mining, author=García, Salvador and Luengo, Julián and Herrera, Francisco, volume=72, year=2015, publisher=Springer

[23] Balakrishnama, S. & Ganapathiraju, A. Linear discriminant analysis-a brief tutorial. *Institute for Signal and information Processing* **18**, 1–8 (1998).

[24] title=An SVM based classification approach to speech separation, author=Han, Kun and Wang, DeLiang, booktitle=2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages=4632–4635, year=2011, organization=IEEE

[25] title=An SVM based classification method for EEG signals, author=Costantini, Giovanni and Casali, Daniele and Todisco, Massimiliano, booktitle=Proceedings of the 14th WSEAS international conference on Circuits, Corfu Island, Greece, volume=2224, year=2010

[26] title=SVM-based classification of segmented airborne LiDAR point clouds in urban areas, author=Zhang, Jixian and Lin, Xiangguo and Ning, Xiaogang, journal=Remote Sensing, volume=5, number=8, pages=3749–3775, year=2013, publisher=Molecular Diversity Preservation International (MDPI)

[27] Izenman, A. J. Linear discriminant analysis. In *Modern multivariate statistical techniques*, 237–280 (Springer, 2013).

[28] Hastie, T., Tibshirani, R. & Buja, A. Flexible discriminant analysis by optimal scoring. *Journal of the American statistical association* **89**, 1255–1270 (1994).

[29] Friedman, J. H. Regularized discriminant analysis. *Journal of the American statistical association* **84**, 165–175 (1989).

[30] Ju, J., Kolaczyk, E. D. & Gopal, S. Gaussian mixture discriminant analysis and sub-pixel land cover characterization in remote sensing. *Remote Sensing of Environment* **84**, 550–560 (2003).

[31] Clemmensen, L., Hastie, T., Witten, D. & Ersbøll, B. Sparse discriminant analysis. *Technometrics* **53**, 406–413 (2011).

[32] Maćkiewicz, A. & Ratajczak, W. Principal components analysis (pca). *Computers & Geosciences* **19**, 303–342 (1993).

[33] Geladi, P., Isaksson, H., Lindqvist, L., Wold, S. & Esbensen, K. Principal component analysis of multivariate images. *Chemometrics and Intelligent Laboratory Systems* **5**, 209–220 (1989).

[34] Gabriel, K. R. The biplot graphic display of matrices with application to principal component analysis. *Biometrika* **58**, 453–467 (1971).

[35] M. Tranmer and M. Elliot, "Multiple linear regression," *The Cathie Marsh Centre for Census and Survey Research (CCSR)*, vol. 5, no. 5, pp. 1–5, 2008.