Arab American University
Faculty of Graduate Studies

A Feature Space Exploration for Improving Prediction of Demographics from Multimodal Neuroimaging Data

By

Obada Basem Abdalrahman Abdallah

Supervisor

Dr. Ahmad Hasasneh

Co- Supervisor

Dr. Jurgen Dammers

This thesis was submitted in partial fulfillment of the requirements for the Master's degree in Data Science and Business Analytics.

Feb /2025

Thesis Approval

A Feature Space Exploration for Improving Prediction of Demographics from Multimodal Neuroimaging Data

By

Obada Basem Abdalrahman Abdallah

This thesis was defended successfully on 20.2.2025 and approved by:

Committee members	Signature
1. Dr. Ahmad Mohammed Hasasneh: Supervisor	
2. Dr. Jurgen Dammers: Co- Supervisor	
3. Dr. Majdi Owda: Internal Examiner	Majdi Owda
4. Dr. Anas Samara: External Examiner	..AnasS....

Declaration

I hereby declare that this thesis, entitled " A Feature Space Exploration For Improving Prediction Of Demographics From Multimodal Neuroimaging Data " is the result of my original research work and has not been submitted for the award of any other degree or diploma in any university or institution. The work presented in this thesis is based on my findings and analysis under my supervisor's guidance.

Name: Obada Abdallah

Date: 11.8.2025

Signature: Obada Abdallah

Student ID: 202012860

Dedication

I dedicate this work too – firstly- to my family and especially to my mother, she always pushes me towards having a higher education degree. Her support always motivates me to keep going and overcome all challenges. For other family members, you were always by my side in good and bad times.

I would also like to dedicate this thesis to my supervisors, who have been a constant source of guidance and mentorship throughout this research work. Their knowledge, expertise, and insights have been invaluable in shaping my research and refining my ideas.

Finally, I dedicate this thesis to all who have supported and encouraged me, including my friends, colleagues, coworkers, and mentors. Your unwavering support and encouragement have been instrumental in my success, and I am grateful to have you in my life.

Thank you all for being a part of my journey, and I hope this thesis will serve as a testament to the hard work and dedication of all those who have contributed to its completion.

Acknowledgment

Would like to express my deepest gratitude to those who supported me in my thesis journey, they were always available to make it up.

First and foremost, I would like to thank Dr. Ahmad Hasasneh and Dr. Jurgen Dammers for their professional supervision, they added a lot to this project and it wouldn’t be here with them. I would like to thank the Arab American University especially my department through the program of Data Science and business analytics, all of you were a part of this.

Finally, my gratitude to my friends and the family, they were there once a help was needed, with their continuous support, I am able now to submit my thesis. All thesis efforts will be always appreciated.

Abstract

The gender and age classification based on multimodal brain imaging with LEMON and HCP datasets are being explored in this thesis. The main feature engineering pipelines include feature connectivity and network graph parameter analysis in feature selection. Feature connectivity provided the main contribution to be a dominant feature which resulted in achieving the majority of accuracy ratio. More to add, the network parameters feature had added an enhancement to the accuracy in a novel way, they have been tuned in a unique way by reflecting the original meaning of graph parameters on brain concept (considering the brain regions as network nodes). While LEMON dataset presented difficulties due to small sample size and gender class imbalance, balanced data augmentation combined with PCA for dimensionality reduction and the integration of network parameters improved classification accuracy. Gender classification on the LEMON dataset has been boosted to 78.5% accuracy via SVM by using MSDL atlas and balanced augmentation. PCA also served to enhance accuracy to 84%, and 82% for SVM and FFNN on LEMON dataset as well. Concerning age classification, the accuracy was high at 92% by using SVM and MSDL atlas with the consideration of merging the network graph metrics with connectivity features.

The HCP dataset outperforms with balanced classes and a large sample size by about 96% accuracy in gender classification using connectivity features and network parameters. Underlining the importance of dataset size, feature extraction, and balancing in improving predicted performance. The thesis emphasizes the effect of merging the feature connectivity with network graph parameters achieving excellent classification results. Future works may investigate possibilities of multi-modal integration, advanced methods of augmentation, and more sophisticated machine learning models for generalization and clinical application.

Table of Contents

Thesis Approval ……………………………………………………………………I

Declaration ………………………………………………………………………II

Dedication -- III

Acknowledgment ……………………………………………………………………IV

Abstract --- V

List of Tables-- VIII

List of Figures-- IX

List of Abbreviations--- XI

List of Equations-- XII

Chapter one: Introduction --1

1.1 Introduction --1

1.2 Problem Statement ---3

1.3 Research Objectives--3

1.4 Research Questions and Methodology ---3

1.5 Thesis Organization--4

Chapter Two: Literature Review --6

2.1 Introduction to fMRI---6

2.2 Introduction to the Machine Learning and Deep Learning Models -------------------8

2.2.1 Support Vector Machine ---8

2.2.2 Feed Forward Neural Network-- 11

2.3 Related Works -- 14

2.4 Summary and Discussion --- 20

Chapter Three: Model Design and Development --- 23

3.1 Introduction -- 23

3.2 Model Workflow --- 24

3.3 Dataset Description (LEMON and HCP) -- 25

3.3.1 LEMON Dataset --- 25

3.3.2 Human Connectome Project Dataset --- 26

3.4 Data Preprocessing and Preparation --- 27

3.4.1 For LEMON Dataset:--- 27

3.4.2 For HCP dataset: --- 29

3.5 Feature Selection and Engineering -- 30
3.5.1 Feature Connectivity-- 30
3.5.2 Network Graph Analysis-- 31
3.6 Proposed models-- 33
3.6.1 Support Vector Machine Model -- 33
3.6.2 Feed Forward Neural Network Model -- 34
3.7 Summary--- 37
Chapter Four: Results and Discussion--- 38
4.1 Results and Discussion -- 38
4.2 SVM with LEMON dataset --- 39
4.3 SVM with HCP dataset -- 52
4.4 Gender Classification using FFNN model with PCA – LEMON dataset ---------- 52
4.5 Key Findings and Results -- 53
4.6 Discussion --- 54
4.7 Summary--- 58
4.8 Recommendation --- 59
Chapter Five: Conclusion and Future Work -- 60
5.1 Conclusion--- 60
5.2 Future work --- 61
5.3 Summary--- 62
References --- 64
الملخص --- **Error! Bookmark not defined.**

List of Tables

TABLE 2: CLASSIFICATION RESULTS FOR AGE AND GENDER USING MSDL ATLAS AND SVM MODEL - LEMON DATASET 51
TABLE 3: CLASSIFICATION RESULTS FOR AGE AND GENDER USING SHEAFER ATLAS AND SVM MODEL - LEMON DATASET 51
TABLE 4: HCP - GENDER CLASSIFICATION RESULTS OVER DATA SUBSETS....52

List of Figures

FIGURE 1: FMRI SCAN WHERE THE RED AREAS SHOW ACTIVE BRAIN REGIONS (HUETTEL, SONG & MCCARTHY, 2004) 6
FIGURE 2: LINEAR (LEFT) AND NON-LINEAR SEPARATION (RIGHT) IN SVM CLASSIFICATION (GEEKSFORGEEKS , 2023) 9
FIGURE 3 FEED FORWARD NEURAL NETWORKS (FFNN) NEURON AND ACTIVATION FUNCTION WHICH CALCULATES THE OUTPUT BASED ON THE INPUT AND WEIGHTS (THORAT ET AL., 2016) 12
FIGURE 4 FFNN ARCHITECTURE CONSISTS OF INPUT, HIDDEN, AND OUTPUT LAYERS (THORAT ET AL., 2016) 13
FIGURE 5: REGIONS OF INTEREST (ROIS) USED FOR AGE AND GENDER CLASSIFICATION (MENDES ET AL.,2021) 15
FIGURE 6: DEEP LEARNING MODEL STRUCTURE (AL ZOUBI ET AL.,2022) 15
FIGURE 7: GENDER CLASSIFICATION USING SVM AND LSTM + CNN – COMPARISON ON HCP DATASET (FAN ET AL., 2020) 16
FIGURE 8: ON THE RIGHT IS A PERFORMANCE FOR ACCURACY ACROSS DIFFERENT TASKS, AND ON THE LEFT :TRAINING PERFORMANCE USING HDC AND RETRAINING (BILLMEYER AND PARHI, 2021) 19
FIGURE 9: MODEL WORKFLOW 25
FIGURE 10: LEMON DATASET - GENDER DISTRIBUTION 26
FIGURE 11: CORRELATION MATRIX BETWEEN BRAIN MSDL REGIONS BY (VAROQUAUX RT AL., 2011) 28
FIGURE 12: TIME-SERIES GENERATION PROCESS 29
FIGURE 13: MALES AND FEMALES DATA AUGMENTATION 29
FIGURE 14: CORRELATION CONNECTIVITY MATRIX 30
FIGURE 15: DEGREE CENTRALITY - METHOD OF CALCULATION 31
FIGURE 16: DIRECT CONNECTION BETWEEN THE ROIS 32
FIGURE 17: FFNN TOPOLOGY 36
FIGURE 18: GENDER CLASSIFICATION ACCURACY (LEFT) AND ROC CURVE (RIGHT) - MSDL ATLAS - SVM WITH CONNECTIVITY 39
FIGURE 19: GENDER CLASSIFICATION ACCURACY (LEFT) AND ROC CURVE (RIGHT) - SHEAFER ATLAS - SVM WITH CONNECTIVITY 40
FIGURE 20: AGE CLASSIFICATION ACCURACY (LEFT) AND ROC CURVE (RIGHT) MSDL ATLAS - SVM WITH CONNECTIVITY 40
FIGURE 21: AGE CLASSIFICATION ACCURACY (LEFT) AND ROC CURVE (RIGHT) - SHEAFER ATLAS - SVM WITH CONNECTIVITY 41
FIGURE 22: GENDER CLASSIFICATION ACCURACY (LEFT) AND ROC CURVE (RIGHT) MSDL ATLAS - SVM WITH CONNECTIVITY & NETWORK PARAMETERS 42

FIGURE 23: GENDER CLASSIFICATION ACCURACY (LEFT) AND ROC CURVE (RIGHT)- SCHEAFER ATLAS - SVM WITH CONNECTIVITY & NETWORK PARAMETERS..................42
FIGURE 24:AGE CLASSIFICATION ACCURACY (LEFT) AND ROC CURVE (RIGHT)- MSDL ATLAS - SVM WITH CONNECTIVITY & NETWORK PARAMETERS 43
FIGURE 25: AGE CLASSIFICATION ACCURACY (LEFT) AND ROC CURVE (RIGHT) - SCHEAFER ATLAS - SVM WITH CONNECTIVITY & NETWORK PARAMETERS 43
FIGURE 26: GENDER CLASSIFICATION ACCURACY (LEFT) AND ROC CURVE (RIGHT) - MSDL ATLAS - SVM WITH CONNECTIVITY & DATA AUGMENTATION 44
FIGURE 27: GENDER CLASSIFICATION ACCURACY (LEFT) AND ROC CURVE (RIGHT)- SCHEAFER ATLAS - SVM WITH CONNECTIVITY & DATA AUGMENTATION....................45
FIGURE 28: AGE CLASSIFICATION ACCURACY (LEFT) AND ROC CURVE (RIGHT) - MSDL ATLAS- SVM WITH CONNECTIVITY & DATA AUGMENTATION 45
FIGURE 29: AGE CLASSIFICATION ACCURACY (LEFT) AND ROC CURVE (RIGHT) - SCHEAFER ATLAS - SVM WITH CONNECTIVITY & DATA AUGMENTATION 46
FIGURE 30: GENDER CLASSIFICATION ACCURACY (LEFT) AND ROC CURVE (RIGHT) - MSDL ATLAS - SVM WITH CONNECTIVITY & BALANCED DATA AUGMENTATION 46
FIGURE 31: GENDER CLASSIFICATION ACCURACY (LEFT) AND ROC CURVE (RIGHT) - SCHEAFER ATLAS - SVM WITH CONNECTIVITY & BALANCED DATA AUGMENTATION..47
FIGURE 32: AGE CLASSIFICATION ACCURACY (LEFT) AND ROC CURVE (RIGHT) MSDL ATLAS - SVM WITH CONNECTIVITY & BALANCED DATA AUGMENTATION 47
FIGURE 33: AGE CLASSIFICATION ACCURACY (LEFT) AND ROC CURVE (RIGHT) - SCHEAFER ATLAS - SVM WITH CONNECTIVITY & BALANCED DATA AUGMENTATION 48
FIGURE 34: GENDER CLASSIFICATION ACCURACY (LEFT) AND ROC CURVE (RIGHT) - MSDL ATLAS - SVM WITH CONNECTIVITY, NETWORK PARAMETERS AND BALANCED DATA AUGMENTATION 49
FIGURE 35: GENDER CLASSIFICATION ACCURACY (LEFT) AND ROC CURVE (RIGHT) - SCHEAFER ATLAS - SVM WITH CONNECTIVITY, NETWORK PARAMETERS AND BALANCED DATA AUGMENTATION...49
FIGURE 36: AGE CLASSIFICATION ACCURACY (LEFT) AND ROC CURVE (RIGHT) - MSDL ATLAS - SVM WITH CONNECTIVITY, NETWORK PARAMETERS AND BALANCED DATA AUGMENTATION 50
FIGURE 37: AGE CLASSIFICATION ACCURACY (LEFT) AND ROC CURVE (RIGHT) - SCHEAFER ATLAS - SVM WITH CONNECTIVITY, NETWORK PARAMETERS AND BALANCED DATA AUGMENTATION 50
FIGURE 38: FFNN MODEL ACCURACY FOR LEMON PCA MODEL..53
FIGURE 39: CLASSIFICATION REPORT - SVM- LEMON PCA...53

List of Abbreviations

No.	Abbreviation	Description
1	MRI	Magnetic Resonance Imaging
2	fMRI	Functional Magnetic Resonance Imaging
3	ECG	Electrocardiogram
4	EEG	Electroencephalogram
5	SVM	Support Vector Machine
6	FFNN	Feedforward Neural Network
7	HCP	Human Connectome Project
8	MSDL	Multi-subject Dictionary learning
9	CNN	Convolutional Neural Network
10	ANN	Artificial Neural Network
11	ROI	Region of Interest
12	fALFF	Fractional Amplitude of Low-Frequency Fluctuations
13	ALFF	Amplitude of Low-Frequency Fluctuations
14	AUC	Area Under the Curve
15	PCA	Principal Component Analysis
16	ICA	Independent Component Analysis
17	DC	Degree Centrality
18	BOLD	Blood-Oxygen-Level Dependent

List of Equations

EQUATION 1: CLUSTERING COEFFICIENT CH3………………………………..…………. 41

Chapter 1

Introduction

1.1 Introduction

There is a long history of using neuroimaging in demographic prediction. Before the development of in vivo imaging techniques, our understanding of the brain was insufficient to perform what were considered complex prediction tasks. With the appearance of magnetic Resonance Imaging (MRI) in the early 1980s and Computed Tomography (CT) in the 1970s, this was a revolution in this domain enabling more advanced image analysis and altering neuropsychological research. These developments have improved our ability to understand and analyze brain structure and find out how it relates to demographic variables. Human brain activities can provide insights related to demographics (Bigler, 2017). The human brain holds cognitive functions, behavior, and emotions, which gives us an indication of the ability to extract information from such features that can be used as applications in human life (Giedd et al., 2015). This thesis aims to demonstrate the ability to predict demographics such as (gender and age) from multimodal neuroimaging data such as functional Magnetic Resonance Imaging (fMRI) and Electroencephalography (EEG).

The motivation of this thesis starts with **Neuroscientific Implications**, throughout the lifespan, the human brain experiences dynamic changes, and these changes differ between genders. These differences can be investigated to gain a deeper understanding of the anatomy and function of the brain (Gogtay et al., 2004). Moreover, the results can be used in **Clinical Applications** for gender and age prediction using non-invasive imaging techniques, which can help in the early diagnosis and intervention of neurodevelopmental disorders, psychiatric issues, and neurodegenerative diseases which can help in improving the healthcare sector (Dehaene et al., 2010). In addition, **Personalized Medicine treatment** can be achieved, as brain imaging can be used to predict Sex and brain age, opening the door to individualized treatment plans and therapies that are based on each person's particular brain profile (Finn et al., 2015). Furthermore, such results can provide **Educational Insights**, and understanding gender differences in learning and age-related cognitive development can change educational systems (Dehaene et al., 2010). In addition to that, this thesis is important to examine whether combining different modalities (such as fMRI and EEG) can enhance the accuracy of prediction for demographics such as gender and age from neuroimaging data, and trying to obtain the highest possible accuracy for such targets using an appropriate feature engineering process, taking into consideration the selection of the best model to be applied, and try to generalize the model to be used on different datasets providing accurate results as implemented in the datasets related to this thesis.

The prediction of demographics from multimodal neuroimaging data has different issues and limitations. Neuroimaging data (such as fMRI and EEG) has a sense of variability, it can be affected by physiological conditions in addition to other factors such as the differences in brain anatomy which introduce noise signals that can affect the accuracy of demographic predictions as indicated by (Van Horn & Toga, 2014). Moreover, there are complexities in analyzing brain activity, such as the need for computational resources to perform some analyses, especially when the data is large and includes many features. The nature of neuroimaging data isn't easy for humans to understand, it requires advanced computational methods such as deep neural networks (DNNs), which can help in predicting demographics. However, DNN models require a large dataset, which is another challenge in such a domain (Sejnowski, Churchland, & Movshon, 2014). In addition, the authors (O'Neil, 2016) have pointed out two important points, the first is the ethical concern that such prediction results can be misused in different domains and unethical ways, such as using the results of a model in the employment cycle to analyze some insights that ethically need permission from the person, and the second is data privacy, which is one of the main concerns in this domain, the researcher has to guarantee privacy and prevent unauthorized access to such sensitive data. Furthermore, there is a challenge in generating a reliable model for predicting demographics (such as age and gender) from multimodal neuroimaging data especially when it is used in a sensitive decision-making cycle such as criminal cases, the model – in such cases – must be capable to be generalized and produce accurate results (Jones et al., 2020). Finally, one of the main challenges in dealing with neuroimaging studies is the difficulty to the collect data. This process needs the participation of individuals, which isn't an easy task, and – in most cases- it requires a financial budget to collect such data (Finn et al., 2015).

In conclusion, this thesis aims to provide a computational model that can compete with other studies in terms of accuracy, it uses both machine learning and deep learning models. Moreover, it provides some insights into how to deal with small datasets and how we can develop models that are capable of dealing with limited data sizes. It also discusses the issue of generalizing such models so that they can be widely used in appropriate domains.

1.2 Problem Statement

The prediction of demographics from healthy subjects becomes an essential issue as it will help in understanding many phenomena related to the diseases and health issues that may be affected by gender or age classification as mentioned by (Singh et al, 2022). In addition, this thesis will address the challenges associated with this process due to the structure of neuroimaging data, which requires advanced methods for analyzing the patterns and extracting meaningful features. Furthermore, it will address the difficulty of dealing with limited datasets and how to ensure that the generated model is highly accurate and stable when applied to other similar datasets. Finally, this thesis will try to overcome the issues related to the noisy signals that come along with neuroimaging data. The research problem is about addressing a solution for predicting subject-specific demographics from healthy subjects with high accuracy compared to the published studies, the classification problem includes gender and age. In particular, this work will attempt to investigate the effect of combining different features from different modalities and how this can positively influence the accuracy of the classification.

1.3 Research Objectives

This research has three main objectives, the main focus is related to investigating the effectiveness of extracted features from EEG and fMRI resting state data to predict subject-specific demographics from healthy subjects such as gender and age, it also investigates the ability to mix different features from different modalities seeking for high accuracy for classification problems that can add to existing related work. Moreover, it identifies the most effective and important features that affect the accuracy. Also, it addresses methods to deal with small and unbalanced datasets without affecting the accuracy of the classification as the dataset size is one of the main concerns for researchers in such fields (Finn et al., 2015).

1.4 Research Questions and Methodology

This thesis has three main questions that need to be answered based on the results, they are:

- What are the most effective features to be used for predicting subject-specific demographics from the health subjects?

- Can different features from different pipelines be combined to improve the quality of the prediction?
- What are the possible methods to improve the accuracy of classification for gender and age from multimodal neuroimaging data when dealing with small and unbalanced datasets?

The methodology used in this research starts with having access to an online dataset called the LEMON dataset by (Babayan et al., 2019), which consists of fMRI data from 211 subjects, recorded during a task-free resting-state condition. In addition, another dataset called the HCP-1200 dataset published by Human connectome project, was used, which consists of 1200 subjects with the same measurement characteristics for fMRI (resting-sate recordings). This dataset was used to test the performance of the model on a larger dataset. Appropriate masks were then applied to these images to extract regions of interest (ROIs). We then went through the feature engineering process and data pre-processing phase to select the best features to be used as input for training Machine learning models, which were Support Vector Machine (SVM) and Deep Neural Network (DNN). An additional feature engineering process was added by extracting network graph features such as Degree Centrality (DC), Average Clustering Coefficients (CC), and Maximum Path Length (PL) and trying to combine them using connectivity features from ROI to get better accuracy for the gender and age classification processes. Finally, an augmentation process was applied to the LEMON dataset trying to increase the number of samples, the unbalanced classes were a real problem, and therefore a balanced augmentation method was used.

The use of two datasets provided an opportunity to test the model on different data, giving the model a real test of the generalization concept. The HCP data was tested on different sizes (200, 400, 800, and 100 subjects) to investigate the effect of increasing the number of samples in such a study, and to ensure that the comparison between the two datasets can be made at a similar size which is 200 subjects in common.

1.5 Thesis Organization

This thesis will start with an introduction chapter that includes an overview of fMRI images, the problem statement, the thesis objectives, the research questions, the used methodology, and finally the organization of the thesis. It will then go through a background and the current related work Chapter to gain insights about this topic and go through the existing work in such domain. In addition, it will go through the model design and development chapter providing more details about the used design and the development process that led to building a classification model for the research problem. The next chapter will focus on presenting the results and the related discussion, taking into account the details that led to such results, supported with a scientific explanation. Finally, at the end of the thesis, a conclusion chapter is provided. This chapter will summarize the thesis and offer insights into potential future work and potential related ideas stemming from this study. It also serves as an opportunity for future research.

Chapter Two

Literature Review

2.1 Introduction to fMRI

Functional Magnetic Resonance imaging (fMRI) is a neuroimaging modality capable of measuring ongoing brain activity. Rather than structural MRI, which focuses on the anatomical structure of the brain, even though structural MRI images are important in mapping the brain regions, fMRI provides information about the dynamics and changes in brain activity based on changes in the level of oxy- and deoxyhemoglobin in the blood vessels. When a brain region becomes more active, it consumes more oxygen, thus the blood flow will increase in that region to meet the increase in oxygen demand (Huettel, Song & McCarthy, 2004). fMRI records the so-called blood-oxygen-level-dependent (BOLD) signal, which is indirectly related to the underlying neural activity. This process is known as neurovascular coupling.

There are many benefits to using fMRI, for example, it provides non-invasive access to brain activity, with no exposure to low-energy radiation, such as X-ray scans (Van et al., 2020). More to add, it provides real-time brain activity scanning and it is suitable for pre-surgical planning (Taschereau, 2022). The image below shows an example of an fMRI scan highlighting active brain regions in terms of Oxygen consumption.

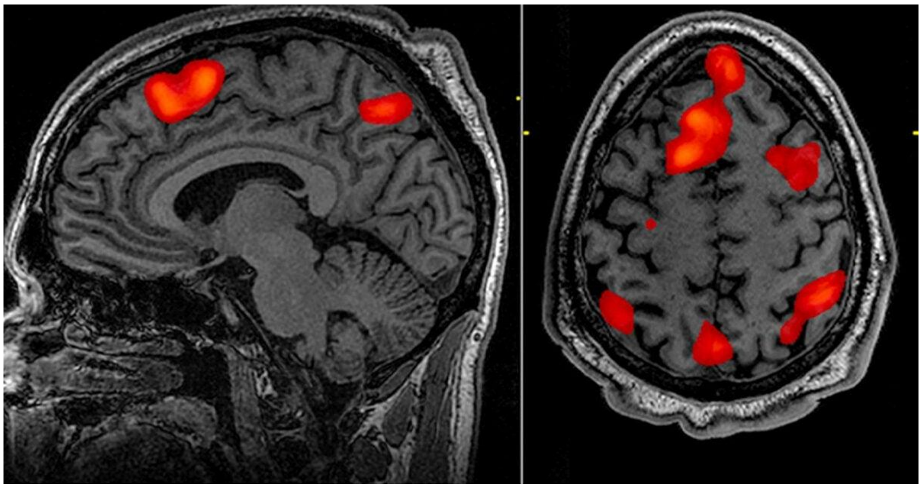


Figure 1: fMRI scan where the red areas show active brain regions (Huettel, Song & McCarthy, 2004)

There are several applications for fMRI scans, (Poldrack, 2011) indicates that fMRI scans are used to assess brain injury in addition to planning surgery for brain tumors. In research, fMRI scans are widely used to understand the relationship between brain regions, and different brain networks have been introduced to select regions of study. Understanding the connectivity of brain regions has opened the door for investigating the relationship between brain connectivity and biological differences such as age and gender (Al Zoubi et al., 2022).

The history of neuroimaging went into different development stages, with Wilhelm Conrad Röntgen's invention of X-ray imaging in 1895 being the first neuroimaging method. With the help of X-ray imaging, it is possible to image the skull, limbs, fractures, tumors, and other abnormalities. Paul Lauterbur and Peter Mansfield created magnetic resonance imaging (MRI) in the 1970s. In contrast to Mansfield's study, which focused on using magnetic gradient fields to encrypt spatial information in the MRI signal, Lauterbur's work on employing magnetic fields to produce spatial images was published in 1973 (Lauterbur, 1973). (Mansfield & Morris, 1982). The first MRI scans of the human body, including the brain, were made using a combination of these two techniques. In the 1950s, positron emission tomography (PET) was developed by Gordon Brownell and Michael Ter-Pogosian. They used radioactive isotopes to create images of blood flow in the brain, making it possible to study brain activity (Ter-Pogossian et al., 1977). Functional MRI (fMRI) was developed in the 1990s. To detect variations in cerebral activity, fMRI monitors changes in blood oxygen levels in the brain (Ogawa et al., 1990). More recently, the white matter tracts of the brain have been examined using diffusion tensor imaging (DTI). DTI detects the movement of water molecules within the brain and makes it possible to see the white matter pathways (Basser et al., 1994).

Network graph theory is increasingly used in fMRI analysis to facilitate the complex connectivity of brain regions. Using Graph theory analysis, functional connections between brain regions are represented by nodes and edges, where nodes represent the brain region itself and the edges represent the connections between these regions. Network graph theory has many metrics, such as the degree of centrality, betweenness, maximum path length, and clustering coefficient.

(Bullmore & Sporns, 2009) Indicates that the use of such metrics helped in understanding the complex connections between brain regions providing an understanding of neurological conditions such as autism, schizophrenia, and Alzheimer's disease. In addition, network graph analysis provides evidence of biological differences and helps in classifying the images based on gender and age group as indicated by (Lorenzini et al., 2023).

There are many research studies conducted on fMRI data, (Bowring, Maumet, C., & Nichols. ,2019) explored the effect of using different software tools on the results of fMRI tasks, the research provided a result showing that there is a significant effect the use of different software tools for analysis, this research provides insights that help in choosing the appropriate tool. Moreover, (Esteban et al.,2019) introduced a preprocessing pipeline called (fMRIPrep) for fMRI data, the model provided was robust and well-tested, the pipeline took into consideration the data consistency and reliability providing high and preprocessing quality. The authors (Liu et al., 2021) have published a comprehensive study on the use of machine learning in the classification of autism spectrum disorder (ASD), this study is a reference because of the valuable insights provided in the feature engineering process for fMRI analysis. Finally, (Gorgolewski et al., 2011) introduced a data preprocessing framework in Python called “Nipype”, this tool helps perform the different levels of fMRI data processing.

2.2 Introduction to the Machine Learning and Deep Learning Models

2.2.1 Support Vector Machine

Machine learning is a part of the Artificial intelligence (AI) concept, it involves many algorithms that were built based on mathematical models, and it helps in making decisions and predictions by learning the data behavior and pattern. The training process involves a tuning phase that helps in making sure the training process is running with the best parameters to provide optimal results. For supervised machine learning, the features are labeled to achieve two tasks mainly (classification and regression). In this thesis, support vector machine, and feed-forward neural networks will be used for sex and brain age classification for fMRI dataset,

SVMs comprise a set of supervised learning techniques utilized for classification and regression (Vapnik 1999). They are part of a broader family of generalized linear classification methods. An inherent feature of SVM is its ability to simultaneously minimize empirical classification errors and maximize geometric margins, earning it the designation of Maximum Margin Classifiers. SVM is based on the principles of Structural Risk Minimization (SRM). It involves mapping input vectors to a higher-dimensional space, where a maximal separating hyperplane is formulated. Two parallel hyperplanes are created on either side of the hyperplane that segregates the data. This segregating hyperplane maximizes the distance between the two parallel hyperplanes. The assumption is that a larger margin or distance between these parallel hyperplanes corresponds to an improved generalization error for the classifier (Vapnik 1999). The below image shows linear and non-linear separation by SVM.

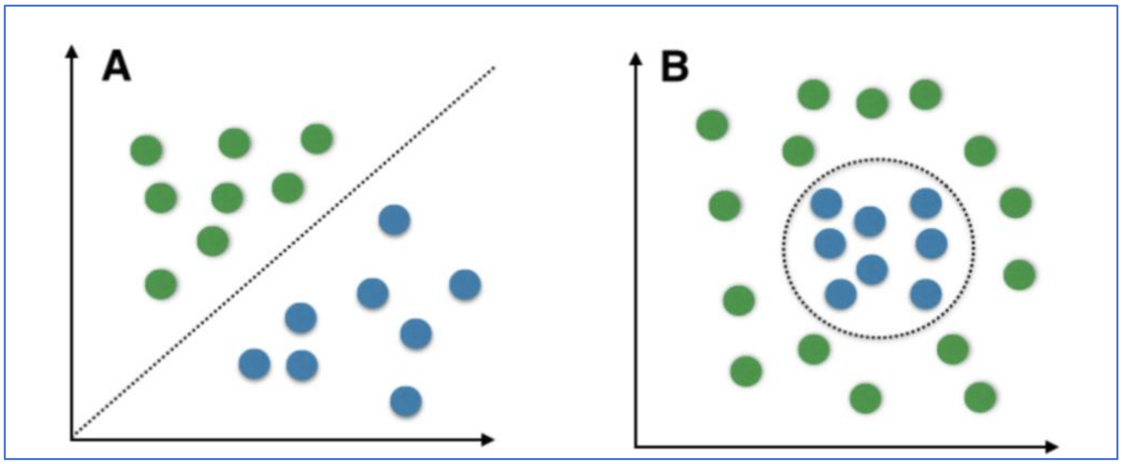


Figure 2: Linear (left) and non-linear separation (right) in SVM classification (GeeksforGeeks , 2023)

SVM stands as a contemporary machine learning approach rooted in statistical learning theory, falling under the category of computational methods introduced by (Suthaharan 2016). Based on the fundamental principle of structural risk minimization (SRM), SVM can derive decision-making rules that yield minimal errors for independent test sets, effectively addressing learning challenges (Deng, Wu, and Shao 2008). Its application extends to tackling issues like nonlinearity, local minima, and high dimensionality. In particular, in numerous practical scenarios, SVM consistently ensures superior accuracy for long-term predictions when compared to alternative computational methodologies such as artificial neural networks (ANN) and k-nearest neighbors (KNN).

SVM is a widely known method and has been applied with significance in conventional machining operations, specifically in the enhancement of performance and the anticipation of pivotal machining results. Efficiently predicting machining performance metrics like surface roughness, cutting force, and tool life requires the careful optimization of the machining process, which stands out as its most formidable aspect. Rather than focusing solely on the challenges inherent in machining, it has been acknowledged that optimizing the economics of the machining process involves the judicious selection of cutting conditions such as cutting speed, feed rate, and depth of cut (Aggarwal and Singh 2005). SVM has emerged as a widely adopted optimization and modeling technique in various machining processes, with common applications including the prediction of surface roughness, tool breakage, and tool wear. In (Hsueh and Yang 2009), researchers introduced a novel diagnostic approach for tool breakage detection in face milling, using SVM. The study incorporated process parameters such as depth of cut, feed per tooth, spindle speed, and cutting diameter. The kernel functions considered in this research encompassed linear, polynomial, and radial basis functions (RBF).

SVM finds diverse and valuable applications in modern machining operations, contributing to the optimization and efficiency of various processes. In contemporary machining processes involving chemical, thermal, or electrical methods for material removal, (L. Zhang et al. 2010) employed SVM with a multi-objective approach to create a hybrid model for optimizing processing parameters in micro-EDM (Electrical Discharge Machining). The researchers identified peak discharge current, pulse duration, pulse-off time, capacitance, electrode rotating speed, and servo reference speed as pivotal process parameters. These parameters exert diverse influences on processing time (PT) and electrode wear (EW). Both PT and EW are significant input targets in the optimization process.

The authors (D. Zhang and Sui 2011) presented a condition monitoring approach for rolling bearings utilizing an auto-regressive (AR) model in conjunction with SVM during Electrical Discharge Machining (EDM). The incorporation of SVM effectively addressed traditional classification challenges, encompassing local minimization, selection of neural network (NN) structure, and issues related to overfitting. SVM showed remarkable success in monitoring the bearing condition of mechanical components. In a separate study, researchers

(Sugumaran, Sabareesh, and Ramachandran 2008) established a fault diagnosis system for roller bearings in EDM machining, employing a neighborhood score multiclass SVM. Given the widespread usage of roller bearings as crucial rotary elements in machinery, the study employed the Radial Basis Function (RBF) as the kernel function. The methodology integrated a kernel-based neighborhood score multiclass SVM for classification, supplemented by a decision tree to address future selection processes. The investigation into the application of a multi-class SVM highlighted its efficacy in diagnosing fault conditions in rolling bearings, presenting a robust approach for fault diagnosis in EDM machining.

The researchers in (Akinnuwesi et al. 2023) proposed to use of a support vector machine model (SVM-PCa-EDD) to enable early differentiation between prostate cancer (PCa) and benign prostate hyperplasia (BPH). Trained on a PCa dataset, the SVM model addresses issues such as class imbalance and dimensionality reduction. In contrast to logistic regression models, SVM-PCa-EDD demonstrates notable performance with 90% accuracy, 80% sensitivity, and 80% specificity. This indicates its potential for facilitating early diagnosis, particularly in healthcare settings with limited resources. The study in (Çinar et al. 2009) also attempts to develop an expert system with a classifier for early organ diagnosis, avoiding the need for biopsy. It explores the relationship between BMI, smoking, and prostate cancer using data from 300 men. Classifiers like artificial neural networks (ANN) and SVM were used. The results showed that smoking increased the risk of prostate cancer, while BMI had no significant impact. The proposed system could assist family physicians in the Turkish Family Health System by generating risk maps and guiding patients toward appropriate treatments through the expert system.

2.2.2 Feed Forward Neural Network

Neural Networks (NNs) are a set of algorithms that mimic the way the human brain works. They are designed to recognize patterns depending on a large database of prior examples (Islam, Chen, and Jin 2019). Due to their high performance, adaptive capabilities, fault tolerance, and ability to be implemented on a large scale, NNs are used in various domains (Murat., 2006) (Ozanich, et al., 2020). These domains include pattern recognition, signal and image processing, and stock market prediction (Murat., 2006). Despite the great results on NN performance, the central critical aspect for network designers is to choose the appropriate network size for the desired application, since the size of the NN affects the learning time, complexity, and generalization capabilities (Enrico., 1994).

The basic building block of NN is the “neuron,” which can be defined as a processing unit. An NN is a group of neurons connected by synaptic weights, as shown in Figure 3 (Thorat et al., 2016). The neuron the weighted information from the neurons which are connected through these synaptic connections and produces an output using the weighted sum of the input signals through an activation function (Murat., 2006).

The basic building block of NN is the “neuron”, which can be defined as a processing unit. NN is a group of neurons connected through synaptic weights, as shown in Figure 4 (Thorat et al., 2016). The neuron receives the weighted information from the neurons it is connected to through these synaptic connections and produces an output using the weighted sum of the input signals through an activation function (Murat., 2006).

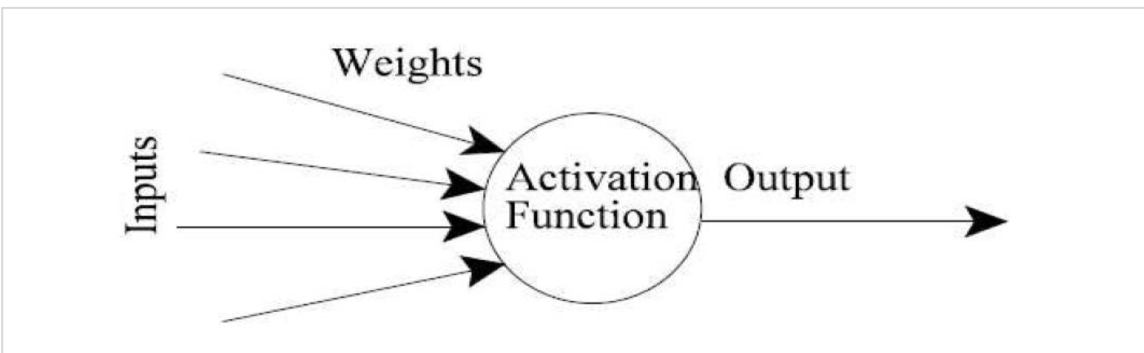


Figure 3 Feed Forward Neural Networks (FFNN) neuron and Activation Function which calculates the output based on the input and weights (Thorat et al., 2016)

Network architectures are separated into two categories depending on the connection type between the neurons, “feed-forward neural networks” (FFNN) and “recurrent neural networks” (RNN). The main difference is that if there is no feedback from the output neurons toward the input ones, then the network is referred to as FFNN (Murat., 2006) and (Islam et al., 2019).

FFNN is one of the simplest ANNs, as information travels through the network in only one direction, forward from the input to the hidden layers and output nodes, as shown in Figure 4. It is also known as a Multi-Layer Perceptron (MLP) (Islam et al., 2019). MLP consists of at least three layers: input, hidden, and output layers (Thorat et al., 2016).

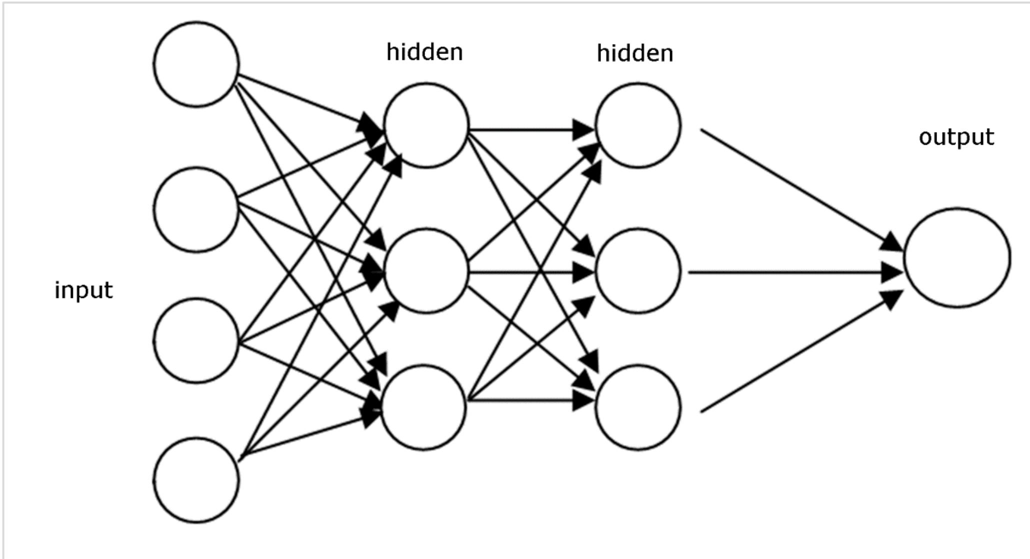


Figure 4 FFNN architecture consists of input, hidden, and output layers (Thorat et al., 2016)

In FFNN, the number of outputs and input nodes is fixed. The architecture (the number of hidden layers and neutrons in each) is determined according to the complexity of the training data (Aldakheel, Satari, and Wriggers 2021). The training process has two main types: supervised and unsupervised training. Unsupervised training is when the desired outcome is unknown, so the model is provided with a group of patterns and left to learn from them and separate the data within several specified iterations (Svozil, Kvasnička, and Pospíchal 1997).

Usually, the network training process for classification problems includes two main tasks: choosing the appropriate network architecture and adjusting the network weights; these weights are used to perform the classification process and separate the different classes within the Bayesian theorem framework (Hasasneh 2014). In the training phase, the neutron weights are initially initialized and updated until the predefined criteria are satisfied. These criteria can be reaching the maximum number of epochs, the maximum training time, and decreasing performance compared to the last time (Aldakheel, Satari, and Wriggers 2021), (Asadi and Kareem 2014).

The optimization problem for supervised learning can be expressed as the summation of the squared errors between the output activations and the target activations. At the same time, the time complexity is mainly related to the number of weighted functions in the hidden layers and the number of epochs (aka iterations) (Asadi and Kareem 2014).

FFNN is applied in many challenging areas, including computer vision, speech recognition, and natural language processing (NLP), where the classification of the target classes is complicated. This type of NN is easy to maintain and suitable for noisy data (Islam, Chen, and Jin 2019), (Li et al. 2022). There are many applications for FFNN in classification problems, (Cai et al., 2021) used the FFNN model to classify arrhythmia from electrocardiogram (ECG) data, the used dataset published by the Massachusetts Institute with 48 ECG records, the model provided an accuracy of 97%. Moreover, (Sen et al., 2014) achieved a high accuracy using FFNN model to classify the sleep stages using EEG data. (Fayaz et al., 2022) predicted the rainfall using FFNN model, the use of such a model shows a better error measure compared to other regression models such as SVM. (Shafiullah et al., 2018) Added research on the classification and detection of distribution fault grids using FFNN model, the model provides a significant accuracy of 99%.

2.3 Related Works

(Mendes et al.,2021) published a study, where the authors aimed to predict the chronological age of individuals in addition to gender classification, the participants' dataset contains biological measures including neuropsychiatric disorders, and the dataset contains structural magnetic resonance imaging (sMRI) data obtained from two different public datasets (ABIDE-II and ADHD-200). Individuals were categorized into three groups, healthy controls (HC) and groups with attention deficit hyperactivity disorder (ADHD) and autism spectrum disorder (ASD). In this study, gray and white matter were used for data processing along with voxel-based morphometry (VBM). Then, a 3D convolutional neural network was used to train the model for gender classification, age prediction, and mental health status estimation. For the ADHD-200 dataset, the results for these targets were as follows, Mean Absolute Error (MAE) = 1.43 years for age prediction and AUC=0.85 for gender classification. For ABIDE-II, the MAE) = 1.57 years for age prediction and AUC=0.89 for gender classification. The image below shows the extracted ROIs used for this study, (a) represents the top ROIs used for gender classification and (b) related to age prediction.

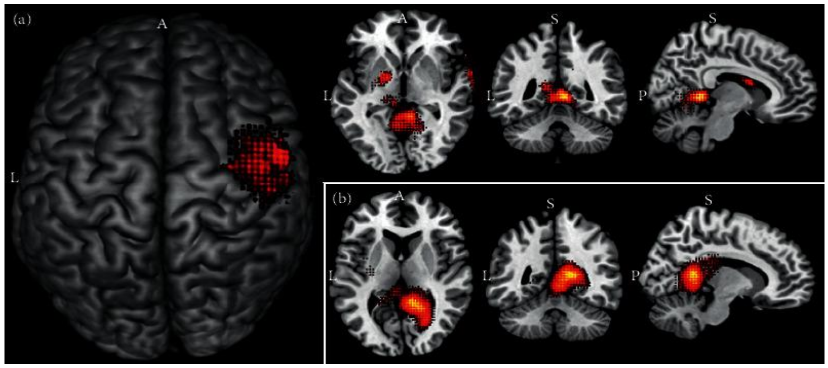


Figure 5: Regions of interest (ROIs) used for age and gender classification (Mendes et al.,2021)

Another study conducted by (Al Zoubi et al.,2022) performed gender classification on data consisting of 5500 resting-state fMRI data from five independent cohorts, focusing on the blood-oxygen-level-dependent (BOLD) signal and especially the fraction of Amplitude of Low-Frequency Fluctuations (fALFF) and the amplitude of low-frequency fluctuation (ALFF). The extracted features used as an input for logistic regression, Random Forest, XGBoost and Gaussian Naïve Bayes (GaussianNB), and Deep learning models to classify the gender for each sample, the ALFF and fALFF feature were extracted using a previously defined mask to extract the ROIs. The classification results were high with AUC > .89, and it was found that the linear models are best in this case rather than the nonlinear models. The diagram below shows the architecture of the deep learning model that is used for gender classification. It contains different layers that are responsible for reducing the number of features through the learning process by using a Max pooling block.

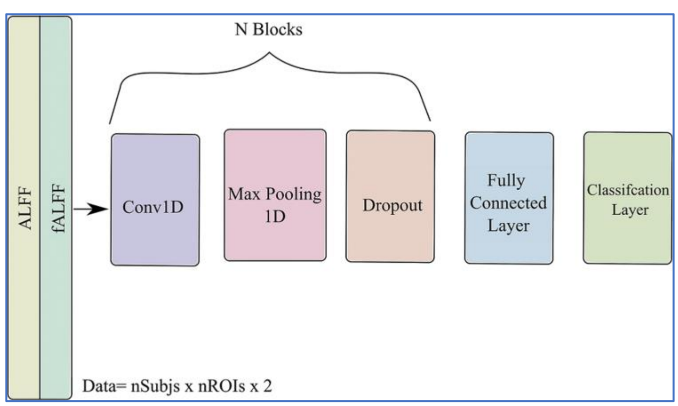


Figure 6: Deep learning model structure (Al Zoubi et al.,2022)

(Fan et al., 2020) published a paper that successfully increased the accuracy of gender classification using a dataset consisting of 1050 resting-state fMRI data gained from the Human Connectome Project (HCP). In addition to gender classification, this study also investigated the ability to predict the intelligence of the participants. An advanced deep learning model was used, combining long short-term memory (LSTM) and convolutional neural network (CNN) in addition to SVM (Figure 7). The model analyzes the dynamic feature connectivity (dFC) to analyze the temporal and spatial features that help in achieving the study aims. This study succeeded in achieving a significant accuracy for gender classification with about 93% and 0.31 Pearson's correlation coefficients for intelligence prediction, this study indicated higher accuracy metrics compared to other studies. The authors (Vergun et al., 2013) have worked on understanding the gender and age differences related to functional brain connectivity. The dataset consists of 65 individuals (three scans for each subject) which was obtained from the 1000 Connectome Project (Vergun et al., 2013). They used feature connectivity information extracted with ROIs, the SVM was used to classify the gender with an accuracy of 84%, and the SVM regressor was used to predict the age with an R-squared value of 0.419. Moreover, (Retico et al., 2016) used SVM to study the effect of gender on the neuroanatomy of children with autism, the data consisted of 152 scans for sMRI, and the SVM model showed a significant effect of gender for the target of the study. The authors (Rosa et al., 2015) also used the SVM as well to discriminate psychiatric patients from fMRI healthy subjects, the accuracy for major depressive disorder (MDD) reached out to 72%.

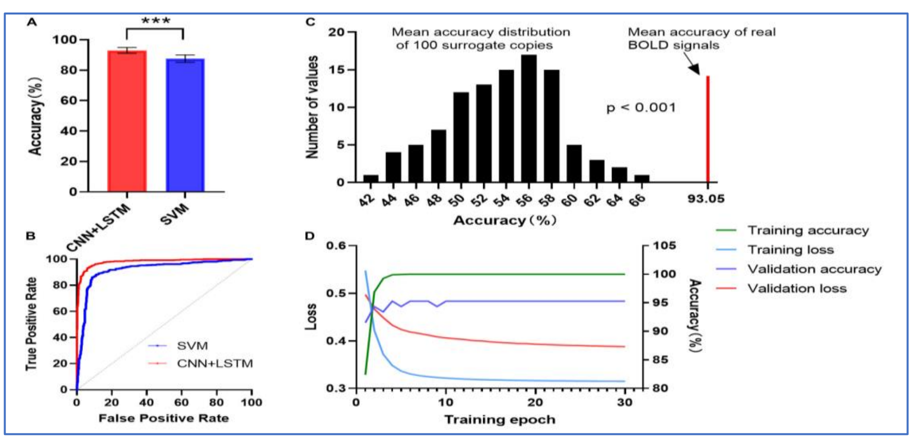


Figure 7: Gender Classification using SVM and LSTM + CNN – Comparison on HCP dataset (Fan et al., 2020)

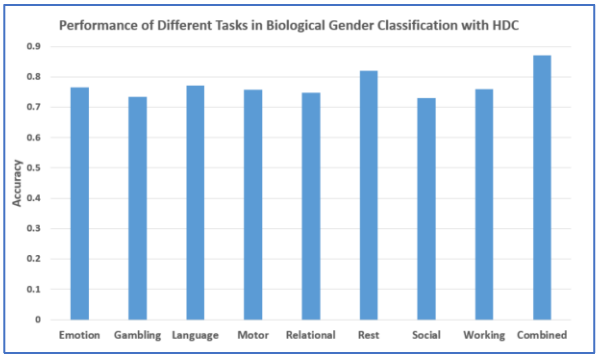
Another study by (Sen et al., 2020) described a biological gender classification for a dynamic functional connectivity model to classify the biological gender using the random forest classifier, the used dataset is HCP which is available online and the same as used in this thesis, it had been gathered as fMRI data in resting state, and the data was pre-processed by identifying the region of interest and applying the dynamic connectivity matrices and dynamic network analysis such as Independent Component Analysis (ICA) to be ready as an input for the regression model. The accuracy of the model was 97% for females and 87% for males. Another high accuracy was achieved by (Kaushik et al., 2018) on a model of predicting the age and the gender of 60 subjects (females and males) in relaxed positions with eyes closed by EEG analysis by using the deep LSTM, the accuracy for predicting the age and the gender was 93.7% and 97.5 respectively.

A study conducted by (Meszlényi et al., 2016) investigated the use of Dynamic Time Warping (DTW) distance to identify similarities between the different regions of the brain area based on BOLD signals, which reflect the connections between brain regions. The study found that using this approach enhanced the results compared to the traditional method of measuring similarities using correlation coefficients. The dataset was obtained from the Functional Connectomes Project with 1000 fMRI healthy subjects in resting state. In addition to that, the authors aimed at classifying both, gender and attention deficit hyperactivity disorder (ADHD). The classification results for gender and ADHD in both approaches (DTW and Traditional Correlation Coefficient) were found to be 0.74 for gender and 0.60 for ADHD when using DTW, and 0.42 for gender and 0.44 for ADHD when using correlation coefficient showing a significant enhancement by using DTW model, the classification was done using SVM model. More to add, (Zhang et al., 2018) published a paper on predicting gender for fMRI data using connectivity features, the dataset used was large with about 820 subjects obtained from the Human Connectome Project (HPC). The study aimed at investigating the effect of psychopathologies such as autism and depression on gender differences. A partial least squares (PLS) regression model was used to predict the gender with 0.881 ± 0.006 and $79.9\% \pm 0.9\%$ for AUC and accuracy, respectively. This study provides additional evidence of model robustness by selecting the training and testing data from different fMRI runs.

Within the same domain, (Weis et al.,2020) used SVM to classify gender for two datasets of fMRI data obtained from the Human Connectome Project, the first dataset consisted of 434 subjects with 217 males, and the second sample contained 310 samples with 155 males. The interesting point in this study is the analysis of the effect in classification results by using different ranges of voxel sizes (Voxels), the highest accuracy for the first sample was 75% and 72% for the second sample. This research demonstrated a clear relationship between brain connectivity data and gender. However, this research didn't provide clear evidence of model generalization and the ability to apply such models to different datasets. Another application of gender classification was performed by (Arslan et al.,2018). The researcher used Graph convolutional networks (GCNs) to classify gender from data obtained from UK biobank with 5000 fMRI healthy subjects in a resting state, this approach provided a good measure of accuracy with an average of 86% classification accuracy results across different folds. (Onal et al., 2020) used the human connectome fMRI data representing healthy subjects reordered through cognitive tasks, the study presents a multi-resolution analysis by converting fMRI signals into sub-bands using DWT, then representing brain regions by using a constructed mesh network, rider regression had been used to quantify such networks and connections. For each subject, multi-function mesh networks were generated taking into consideration the number of sub-bands and cognitive tasks reported during the data collection process. The machine learning model was successful in classifying gender. This integrated approach with mesh networks achieved a higher accuracy for classification tasks as compared to using a single sub-band task network. The average accuracy achieved for the integrated approach was found to be 92%.

(Billmeyer and Parhi, 2021) used another approach for biological gender classification, it was based on using Hyperdimensional (HD) computing which uses high dimensional vectors for the process of computation, this approach is popular for fast learning and robust data representation. This research utilizes a dataset for resting-state involved with task fMRI, the data is obtained from a human connectome project which is a database available online. The HD computational model can be used to perform classification tasks, it can find and predict strong features that can be used in achieving the classification target, this can be done using dFC. Such predicted features have been encoded in a high dimensional space using record encoding. The model was enhanced by using the concept of the adaptive retraining process.

The obtained accuracy of biological gender classification was 87%, which shows an enhancement in accuracy with a comparison with edge entropy measures which was 84% in the same case. The graphs below show detailed plots for the accuracy results.



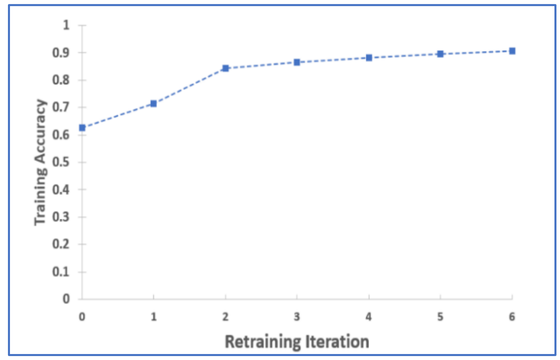


Figure 8: on the right is a performance for Accuracy across different tasks, and on the left :Training Performance using HDC and Retraining (Billmeyer and Parhi, 2021)

(Raison et al.,2021) utilizes the DNN model for classifying the gender, the research applied a resting-state fMRI data obtained from the human connectome project. The authors used a CNN model, which was trained over the given dataset. The gender was classified from the brain spatial maps. The CNN model provided a good result with about 0.1 binary-cross entropy (BCE) after 22 epochs. A similar approach in using DNN for gender classification was followed by (Zhao et al., 2020), the study aimed to find the gender differences using features obtained from the dynamic connectivity process, and the used datasets were obtained from the human connectome project (sp1200) with resting-state conditions, the full connectivity patterns had obtained, the study proposes feature ranking method using DNN to identify the most important features that can lead to the best optimal classification accuracy. In addition to that, Bayesian deep learning is used to provide uncertainty levels in addition to classification accuracy. The model provides a different range of accuracies from 83% up to 94% based on the number of ICA components used for functional connectivity input, the model robustness was tested through 50 random selections for training and testing data parts. In addition, (Supekar et al., 2022) achieved an average accuracy of 84% for gender classification, the DNN model was applied to resting-state data consisting of 773 healthy subjects which were collected through Autism Brain Imaging Data Exchange (ABIDE) in addition to a part of the data which had been collected by the researchers themselves.

For the use of FFNN model in classification problems with fMRI data, (Eslami et al., 2019) used such a model to classify autism spectrum disorder (ASD), the dataset was collected from Autism Brain Imaging Data Exchange containing 1035 subjects which were collected from 17 different brain centers, the machine learning results outperform the state of the art for the proposed dataset and task with a maximum accuracy of 82%. In addition, recent research conducted by (Pilli et al., 2023) worked on a dataset consisting of 115 sMRI scans obtained from the OpenNeuro database with three different age groups (young, older children, and adults), they sought to investigate the anatomical changes associated with age. The methodology was to apply segmentation for the brain scans into different tissues, Cerebrospinal Fluid (CSF), Gray Matter (GM), and White Matter (WM), the features were extracted from these segmented objects using a pre-trained model called ResNet-50 network and the classification was driven by ensemble deep random vector functional link (edRVFL) network with an accuracy of 93% for CSF, 98% for WM and 96% for GM.

Network graph theory plays an important role in classification problems for fMRI images as indicated by (Zuo et al.,2012). The authors (Saha et al., 2023) supported this with research showing that network graph parameters such as degree centrality and betweenness. Also, the authors (Lorenzini et al.,2023) have recently published a paper that investigates gender and age differences for fMRI data using the network graph analysis, the study shows that females show more contrast in the eigenvector centrality (EC) than males, and older participants show higher peaks compared to young ones. (Khazaee et al., 2014) Aimed to predict Alzheimer's disease (AD) by analyzing the connectivity feature for fMRI images supported by network graph theory analysis, 90 ROIs were used with SVM classifier which resulted in an accuracy of 97% for AD prediction, the number of participants 20 with multiple scans.

2.4 Summary and Discussion

Many of the aforementioned studies, such as (Mendes et al., 2021; Fan et al., 2020), have worked on different types of data (sMRI, resting-state fMRI) using different models such as 3D CNNs, SVM, and LSTM. These data show the importance of using different methods to obtain the best classification results for biological differences from fMRI data. In addition, (Al Zoubi et al., 2022; Vergun et al., 2013) showed the importance of using advanced learning models such as deep learning models to achieve more accurate predictive models. The comparison between using static (Meszlényi et al., 2016) and dynamic feature analysis on fMRI data provides insights into the different approaches to understanding brain connectivity.

Different gaps and limitations are obvious based on the literature review, (Mendes et al., 2021; Al Zoubi et al., 2022) studies show a limitation in model generalization across different datasets, the model performs well on the dataset it was trained on but fails to provide good accuracy once used for different datasets with different data acquisition and preprocessing. Another limitation that can be learned from (Fan et al., 2020) is the cost of interpretability when using advanced learning models such as the DNN model, such models usually provide high accuracy, and this raises the question of how to trade off this with model interpretability especially when these models serve critical sectors such as the medical one, the decision cycle is crucial and the models must be capable to be interpretable. Moreover, there is a real challenge in how the network graph theory analysis method can be integrated with other analysis methods to provide a combined and trustworthy model. (Meszlényi et al., 2016; Khazaee et al., 2014) Shows a promising methodology of using network graph theory and DWT, and there is much to be done in this area to provide better analysis and understanding of the connectivity of brain features.

This thesis will address the above limitations and provide recommendations to add to the existing knowledge. For the generalization issue, the model will be trained and tested on different datasets such as LEMON and HCP fMRI datasets. It will use the network graph theory in addition to feature connectivity analysis, and try to provide more insights and enhancements for combining different modalities to classify biological differences such as gender and age. (Zuo et al., 2012; Khazaee et al., 2014) Provided a good model that uses network graph theory analysis for classification problems related to brain data. The work will be extended and investigated in such a direction to understand the organization of brain networks related to gender and age. With the use of feature importance analysis, an enhancement can be provided for data interpretability along with model complexity which is an important point of balance between the concepts. Finally, the study will provide a standardization approach for data preprocessing which will help mitigate the variability issues and improve the comparability of the study in comparison with existing work. A feature importance analysis will help to identify the best features to be included to better meet the classification goals.

In summary, this thesis aims to apply the fundamental concepts of fMRI analysis to various datasets to enhance the existing results. The thesis will focus on the concept of combining the network graph theory with connectivity features to improve classification accuracy. Additionally, it will provide and compare analyses between different data sizes of data, which can be helpful when working with a small fMRI dataset and aiming to achieve the highest possible optimal accuracy. Finally, the model will be tested with a generalization concept to determine if it can adapt to different datasets.

Chapter Three

Model Design and Development

3.1 Introduction

This thesis utilized two datasets (LEMON and HCP) related to fMRI healthy subjects to perform classification tasks for sex and brain age groups. These datasets are available online and can be downloaded from the Human Connectome website. The group of healthy subjects is preprocessed and can be used directly for feature extraction with some additional preprocessing steps. First, the Multi-subject Dictionary learning (MSDL) atlas was applied to the LEMON dataset for each subject to extract the information related to 39 ROIs. Additionally, a correlation connectivity process was applied for each subject to generate a connectivity matrix with a dimension of 39x39. These data were vectorized to be used as an input for the machine learning models. Furthermore, network graph theory analysis was applied to each connectivity matrix providing different metrics such as degree centrality, clustering coefficient, path length, and betweenness, which were used in addition to the connectivity features to enhance the accuracy of classification. In addition to the MSDL atlas the well-known Schaefer atlas (with 100 ROIs) was used on the LEMON dataset to seek better feature extraction.

The LEMON dataset runs with unbalanced classes sexing terms of gender (with 33% of healthy subjects being female and 67% male), which poses a challenge for achieving balanced accuracy for both classes. In addition, the LEMON dataset is relatively small with about 200 subjects. To overcome these two issues, a balanced augmentation technique was used to increase the sample size and to achieve almost balanced classes (i.e., number of subjects per class). More importantly, it is expected that the accuracy result will increase due to the applied oversampling method. The risk of overfitting was mitigated using different techniques which are illustrated in this chapter. The HCP dataset is a relatively large dataset with almost 1000 subjects. Group independent component analysis (group-ICA) was applied to all data from which subject-specific sets with different numbers of independent components are provided in the HCP dataset. The dataset with 50 ICA time series was used for further analysis. The number of healthy subjects per class is almost equally balanced with about 45% to 55% for male-to-female classes, so there is no need for data augmentation.

The merged features from the connectivity matrix and network graph metrics were used as input for learning models. SVM and FFNN were used for brain age group and sex classification. Different tuning processes were applied to enhance the accuracy of classification, and a comparative study was performed on the results of both models. Moreover, the models were tested with different combinations of features, and different sizes of HCP data were used to understand the relationship between the size of data and accuracy performance, which will provide insights into how small datasets can be optimally utilized to provide robust models and high accuracy. Finally, a generalization test was performed to ensure that the models can be applied to different types of datasets.

3.2 Model Workflow

This thesis is about utilizing two datasets (LEMON and HCP) datasets which are related to healthy subjects (fMRI) images, the images go through a feature engineering process by extracting the feature connectivity matrices for the images. Moreover, network graph theory was used to extract network graph metrics such as (degree centrality, clustering coefficient, and shortest path length) to enhance the classification results. Furthermore, two classification models were used (SVM and FFNN) with a tuning process seeking the best accuracy.

LEMON dataset is unbalanced regarding the gender classes, this led to the introduction of an augmentation process by dividing the time-series array into smaller parts to increase the sample size and increase the number of low-class cases.

The below diagram illustrates the model workflow:

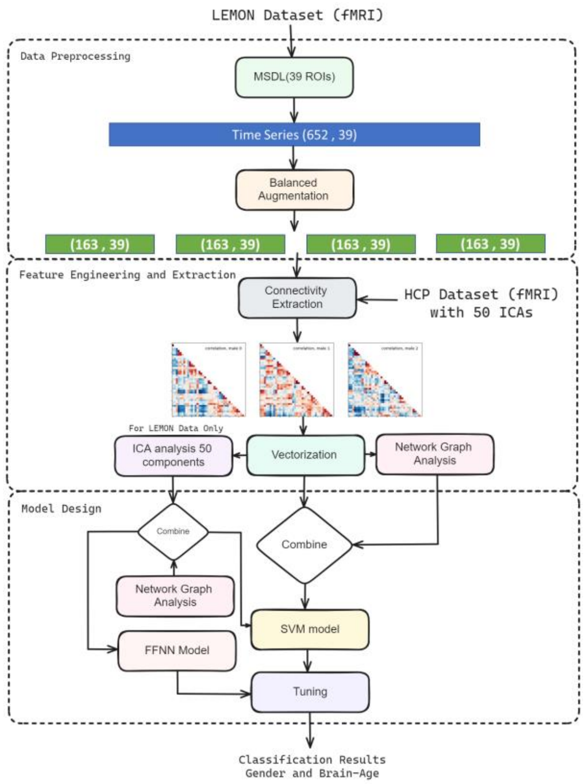


Figure 9: Model Workflow

The above model summarizes the thesis phases, it starts utilizing the fMRI data from two sources (HCP and LEMON), extracting the ROIs with time-series transformation, Feature engineering to extract important features, applying the ICA model to enhance the results, applying different algorithms for Gender and Age classification, and finally tunning the model to enhance the accuracy.

3.3 Dataset Description (LEMON and HCP)

3.3.1 LEMON Dataset

It is a publicly available dataset of 228 healthy subjects containing two age groups, young (with 145 subjects, age range 20-35 years, 45 females) and Elderly (with 74 subjects, age range 59-77 years, 37 females). The dataset was acquired cross-sectionally in Leipzig, Germany, between 2013 and 2015 to study the mind-body-motion interactions. The assessment took two days resulting in having a completed 3-Tesla fMRI resting-state. Additionally, a 62-channel Electroencephalogram (EEG) was acquired at rest. During a task-free resting fMRI, several measures were obtained, such as (heart rate, blood pressure, respiration, and pulse). Moreover, different tests were obtained like (blood samples, Anthropometrics, and drug tests). 21 questions are included which are related to personality traits, emotional behavior, addictive behavior, eating behavior, and tendencies. Finally, the dataset contains information about the gender and age group of each participant (which will be used as classification targets in this thesis) along with raw data and preprocessed versions of fMRI and EEG data (Babayan et al, 2019). Figure 1 shows the gender distribution, and Table 1 shows re-distributing the age groups into two groups only (Young and Elderly) to decrease the number of classes.

Table 1: LEMON dataset – Age Distribution

Age intervals	NO. of subjects	Groups
20-25	71	Young
25-30	59	
30-35	12	
35-40	1	
55-60	4	Elderly
60-65	18	
65-70	24	
70-75	19	
75-80	3	

Youth	143
Elderly	68
	211

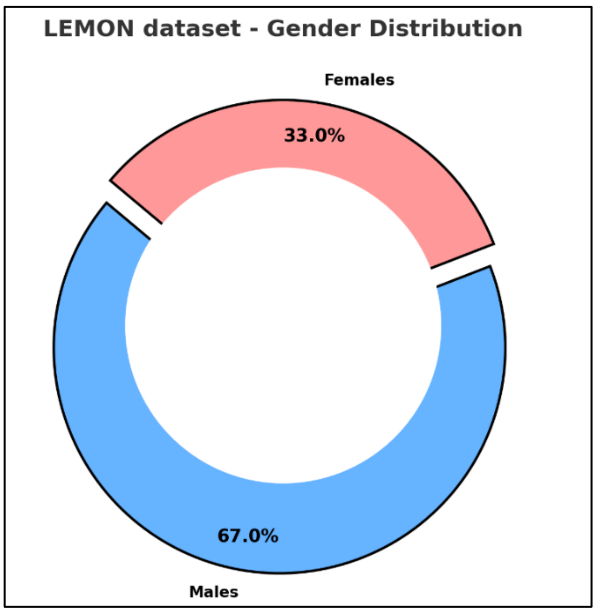


Figure 10: LEMON dataset - Gender Distribution

3.3.2 Human Connectome Project Dataset

This dataset is associated with the Human Connectome Project s-1200 (Human Connectome Project, 2017). It includes 1200 healthy subjects including 3-Tesla MR images obtained from 1206 young adult participants (1113 come with structural MR scans). Data were collected between 2012 and 2015. It also includes 7-Tesla MRI scans for 184 subjects and 95 subjects also have some resting-state MEG recordings. For the first time, behavioral retest data are available for 46 subjects. This data has widely been used by many researchers in Parkinson's disease studies, network graph theory studies, and gender classification (Fan et al., 2020), (Zhang et al., 2018), (Weis et al.,2020), (Onal et al., 2020), and (Raison et al.,2021). The dataset provides 7-Tesla scans, which contain more accurate data compared to 3-Tesla scans. Moreover, the dataset is large with biological information that can be used for several classification and regression tasks.

This data was preprocessed with different sets of independent components (ICA). In this thesis, 50 independent components were used, which represent the closest number to the 39 ROIs used in the LEMON dataset. The application of ICA is expected to enhance the accuracy of classification targets.

3.4 Data Preprocessing and Preparation

3.4.1 For LEMON Dataset:

LEMON fMRI files are available online as raw data files and a preprocessed one, the preprocessed versions were downloaded in neuroimaging information technology initiative (NIFTI) format.

1- **Deletion of corrupted files:** Some of the fMRI files were corrupted and had to be deleted due to incomplete processing of the data. The total number of recordings is 228, of which 17 are corrupted, resulting in 211 datasets for subsequent analysis.

2- **Application of an Atlas:** The Schaefer and MSDL atlases were used for analysis, where each atlas provides a different number of regions of interest (ROIs). ROIs are a subset of brain voxels that are chosen for a detailed analysis. They can be considered as an anatomical or functional representation depending on the study. The process of extracting the mean activity per ROI aims at extracting a representative time series per ROI instead of having the full data information which involves noisy data (Varoquaux et al., 2011), see figure that shows the correlation between the ROIs.

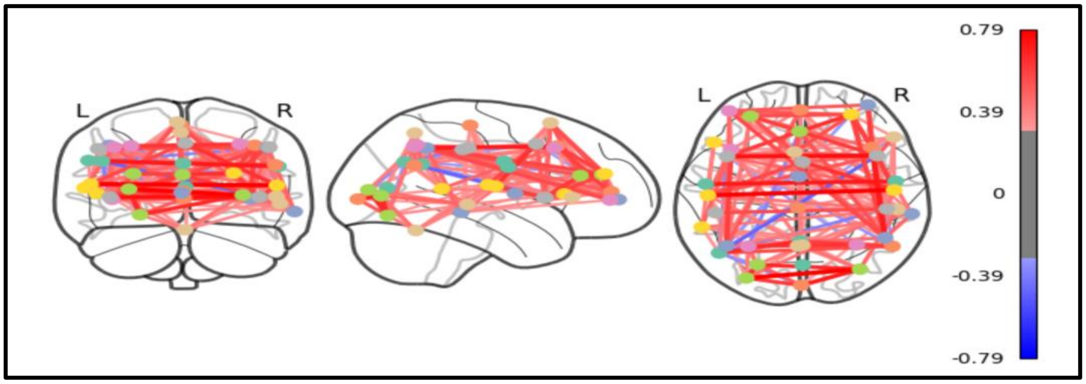


Figure 11: Correlation matrix between brain MSDL regions by (Varoquaux rt al., 2011)

3- MSDL atlas is a popular atlas that is available publicly and can be used on fMRI images. By using this atlas, 39 regions of interest will be tested providing a focused input of data.

This Atlas consists of 39 ROIs, which are part of the following networks: ['Aud', 'Aud', 'Striate', 'DMN', 'DMN', 'DMN', 'DMN', 'Occ post', 'Motor', 'R V Att', 'R V Att', 'R V Att', 'R V Att', 'Basal', 'L V Att', 'L V Att', 'L V Att', 'D Att', 'D Att', 'Vis Sec', 'Vis Sec', 'Vis Sec', 'Salience', 'Salience', 'Salience', 'Temporal', 'Temporal', 'Language', 'Language', 'Language', 'Language', 'Language', 'Cereb', 'Dors PCC', 'Cing-Ins', 'Cing-Ins', 'Cing-Ins', 'Ant IPS', 'Ant IPS'] by (ROBINSON, 2023).

4- **Time-Series Generation:** after applying the masks on the fMRI images, a time-series numpy array representing the time stamps with BOLD information for the ROIs. The generation of time series allows us to have a numerical input available that contains informative data that can help in feeding the learning models. The figure shows the process of generating time series per image. The dimension of the generated time-series is 652x39, which is time-stamps x ROIs, representing the BOLD signal information extracted for each ROI over the time of image recording.

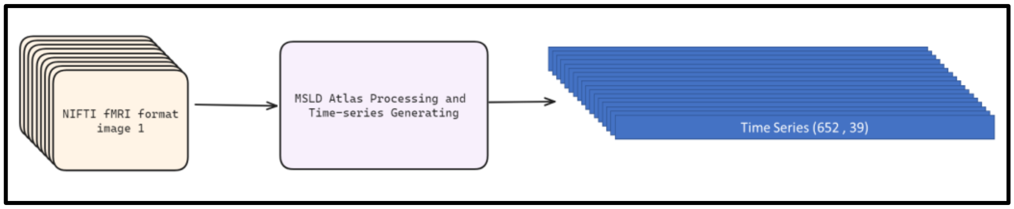


Figure 12: Time-Series Generation Process

5- **Data Augmentation:** due to the issue of unbalanced data described before, and because the LEMON dataset is relatively small, the balanced augmentation method was used to solve this issue. The augmentation concept is concerned with increasing the dataset size by generating new samples from the existing samples by different methods such as using mathematical calculations to generate more samples based on the existing sample information. Moreover, dividing the time-series data into different parts is a valid solution. In this thesis, the 652 time series from 39 ROIs (652x39) were divided into 4 samples resulting in sub-dimensions of (163x39). The splitting was done on the time stamp taking into consideration not losing the information within the time series as per measuring the effect of splitting on the accuracy. The LEMON dataset runs with unbalanced classes regarding sex (with 33% of healthy subjects being female and 67% male). Balanced augmentation was therefore used in a way that time-series slots were split 4 times for females and 2 times for males, resulting in having an almost equal number of time series per class (280 for each and 560 in total). Figure 5 illustrates the augmentation process.

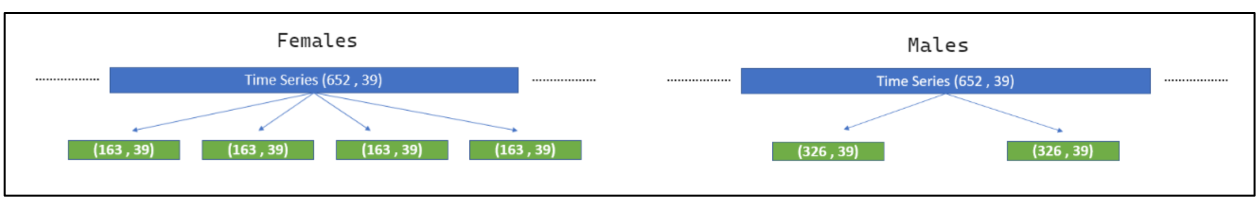


Figure 13: Males and Females Data Augmentation

3.4.2 For HCP dataset:

This dataset is preprocessed with 50 ICAs, , the sample size of the dataset is relatively high with 1200 subjects and the classes are balanced. So, the process of applying a mask (such as MSDL or Scheafer) and augmentation was suspended. Regarding the thesis workflow, the

HCP dataset is preprocessed and goes directly into the feature engineering process. The subject time-series dimensions are 4500x50, as HCP objects have a longer time record compared to the LEMON dataset.

3.5 Feature Selection and Engineering

3.5.1 Feature Connectivity

There is an interaction between the brain regions which can be used to extract features and use it to study some biological and mental cases which is the connectivity process (Iraji et al.,2016). In this thesis, the connectivity process was applied to the time-series input which produced a connectivity matrix with (39x39) for LEMON dataset and (50X50) for HCP, noting that only the lower triangle of the correlation matrix was used with a consideration of removing the diagonal values (N*(N-1)/2). It is about measuring the correlation between the ROIs for a specific time series, the high correlation figures indicate a high level of interaction between these regions. More to add, the correlation connectivity matrix was converted to a vector with a dimension of 1x741 for LEMON and 1x1225 for HCP, which represents an input for machine learning and deep learning models. Figure 6 illustrates the correlation connectivity matrix which is generated by calculating the correlation coefficient across all regions of interest. For example, the first row in the matrix shows the correlation between ROI-1 and all ROIs (from 1-39) and so on.

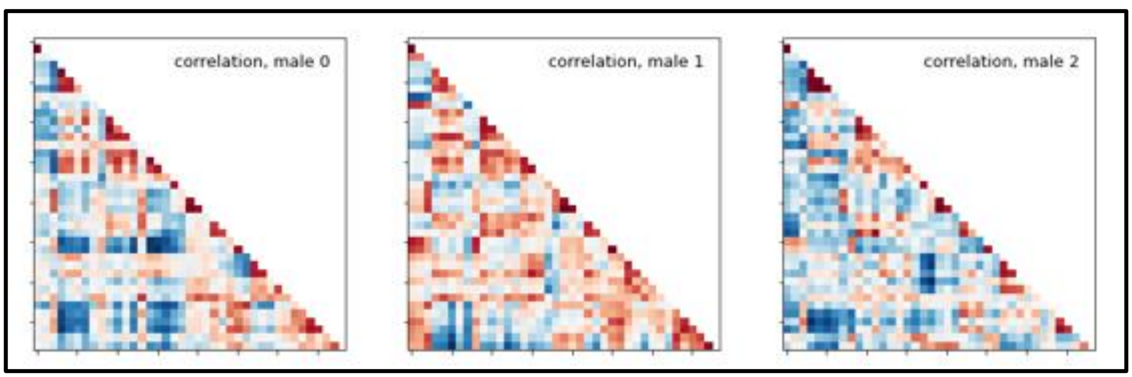


Figure 14: Correlation Connectivity Matrix

The number of features generated from the connectivity process is equal to the number of ROIx(ROI-1)/2, resulting in having a large number of features which makes it difficult to study the importance of such features or combining other features.

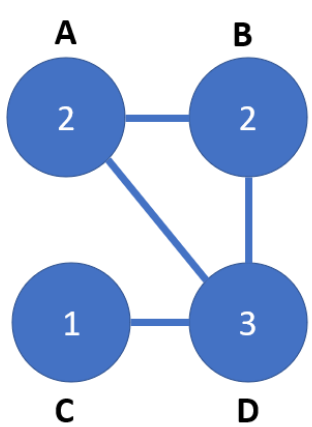
3.5.2 Network Graph Analysis

Additional features from network graph theory were added, such as degree centrality, clustering coefficient, and path length. These features are supposed to be combined with the connectivity set of features to enhance the accuracy of the classification targets. Below are some theoretical insights about the used network theory metrics in this thesis.

1- Degree Centrality

One of the network graph metrics can be defined as the number of edges connected to a node (which represents the ROI brain concept). It is a measure of global connectivity at the voxel level by measuring the number of instant functional connections a brain region has with the rest of the brain can be used to map brain hubs with high accuracy, repeatability, and sensitivity (Guo et al., 2020). Figure 6 shows the mechanism of calculating the degree of centrality. It is simply a matter of counting the number of direct connections with neighboring nodes.

In this thesis, the degree centrality represents the summation of correlation values for each ROI with other regions. Below figure 7 illustrates such a mechanism.



	A	B	C	D	Degree
A	0	1	0	1	**2**
B	1	0	0	1	**2**
C	**0**	**0**	**0**	**1**	**1**
D	**1**	**1**	**1**	**0**	**3**

Figure 15: Degree centrality - method of calculation

2- Clustering coefficient

It is one of the fundamental metrics of network graph theory, it represents the number of direct connections from the node itself within the network (Freeman and Linton., 1979). In this thesis, the correlation between the nodes was considered a direct connection if exceeding the average of the overall correlation vector. This metric provided additional features that helped in obtaining better performance of the model used. The figure below illustrates how the nodes are directly connected with neighbors.

$$CC = \frac{Nc}{Kc(Kc-1)}$$

Equation 1: Clustering coefficient

Kc= Degree of node "C" = 3

Nc= Number of links between neighbors

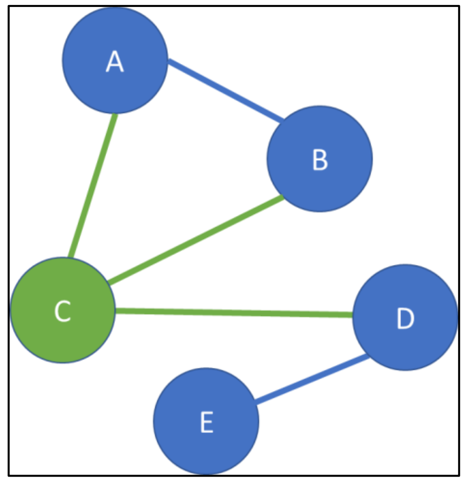


Figure 16: Direct Connection between the ROIs

In this thesis, the idea of finding a direct connection between ROIs is not applicable, we depended on the correlation values, and we introduced a correlation weight that represents the average value of the overall correlation values within the vector itself which represents the cross-correlation between all ROIs. This value was used as a threshold to consider a connection either direct or indirect connection. After that, the above equation was used to calculate the value of the clustering coefficient.

3- Path length

It is one of the metrics in network graph theory, it represents the number of edges between two nodes within the shortest path (Goldberg et al., 2005). In this thesis, the ROI is considered as a node, and the correlation with (a certain threshold) is considered as an edge to the selected node. In addition, the shortest path length has been calculated for all ROIs to be used as an additional feature to improve the performance of learning models.

3.6 Proposed models

3.6.1 Support Vector Machine Model

For Gender and Age-Group Classification

A support vector machine model was used to do the gender and age classification for the given dataset. SVM model has many advantages that serve the case study. First, it is effective in high-dimensional spaces (Cortes & Vapnik, 1995), which is very important within the given dataset as the number of features exceeds the number of records. More to add, regularization can help in avoiding overfitting which is vital especially when having features generated from other features. Moreover, SVM is very good in terms of tuning as it provides a variety of parameters that help in a wide range of tuning possibilities (Schölkopf & Smola., 2001). However, SVM model has many challenges, it is not efficient when considering noisy data (Ben-Hur & Weston., 2010), this effect was reduced by reducing the data dimension by using defined masks with ROIs. The following model design was used for both gender and age-group classification which considered the best parameters providing the best results:

Model Parameters:

Below parameters are the best parameters chosen after the fine-tuning process

1- **Kernel**: kernel='rbf'

Rbf kernel is widely used for non-linear data, it helps in mapping the input space into a higher dimension to handle the complex relationship between variables.

2- **Class Weight**: class_weight="balanced"

It adjusts the unbalancing rate between the classes, it gives the lower class a high weight to solve the issue of unbalanced classes.

3- **C (Regularization Parameter)**: C=10

It controls the trading off between low testing and low training errors, a higher ‘C’ value leads to better classification of training examples which may lead to overfitting.

4- Gamma: gamma=0.01

The gamma parameter affects the model’s performance directly, it measures how far the influence of a single training point reaches.

Model Design:

1- Data Augmentation

The code involves a data augmentation mechanism by splitting the time-series data into smaller parts which solved the issue of unbalanced classes, low class frequency was treated with a higher splitting rate.

2- Connectivity Measures:

The connectivity measure with Kind='correlation' was used to calculate the correlation matrix which was used as an input for the learning model after vectorization.

3- Graph metrics

Different graph metrics were used as additional features to enhance the machine learning results such as degree centrality, shortest path length, clustering coefficient, betweenness, and closeness centrality. These parameters merged with the connectivity feature which enhanced the classification results.

4- Cross-validation and Grid search

'StratifiedShuffleSplit' was used in cross-validation to ensure that each fold contains the same proportion of classes. Different numbers of folds were used which generated different values of accuracy, the best was considered based on the results.

3.6.2 Feed Forward Neural Network Model For Gender and Age-Group Classification

FFNN model was used because of its advantage and suitability for non-linear problems (LeCun et al., 2015), it is a scalable model that allows to addition of more layers if needed. However, FFNN model can easily be overfitting the results (Srivastava et al., 2014), so a need for a robust regularization method is a must to avoid such a problem.

In this thesis, FFNN was used because it is suitable for non-linear problems, it generates da good results in solving the issue of low accuracy obtained from SVM model.

Model Parameters and Design:

1- **Input data**

Principle Component Analysis (PCA) dimensionally reduction was used to reduce the input data to 32 principal components which helped in simplifying the input structure allowing the training process to cove more variances. Moreover, network graph metrics merged with PCA input to enhance the results of the Feed Forward Neural Network model.

PCA is a dimensional reduction technique that is widely used in data analysis, it helps in reducing the number of features taking into consideration holding as much information within the data space (Jolliffe & Cadima., 2016). In this thesis, the PCA analysis was applied to achieve the goal of reducing the large number of features which makes the learning more focused and enhances the results. To avoid the overfitting issue, PCA analysis was applied to training and testing data parts separately.

2- **Data Normalization**

'StandardScaler' was used in the normalization process with (mean=0, variance=1) which is vital to enhance the performance of the neural network model, this normalization technique is provided by the "Scikit-learn" library (Pedregosa et al., 2011), the idea of Standard scaler is to standardize the features into the properties of the normal distribution with a mean of 0 and a standard deviation of 1 (Sola & Sevilla., 1997).

3- **Model Architecture**

- **Input layer:** the first input layer has 46 neurons and the ReLU function is used as an activation function.
- **Hidden layer:** the second layer "hidden" has 32 neurons and the ReLU function was used as well in this layer.
- **Output layer:** A single neuron was used in the output layer with a sigmoid activation function which is suitable for binary classification problems.

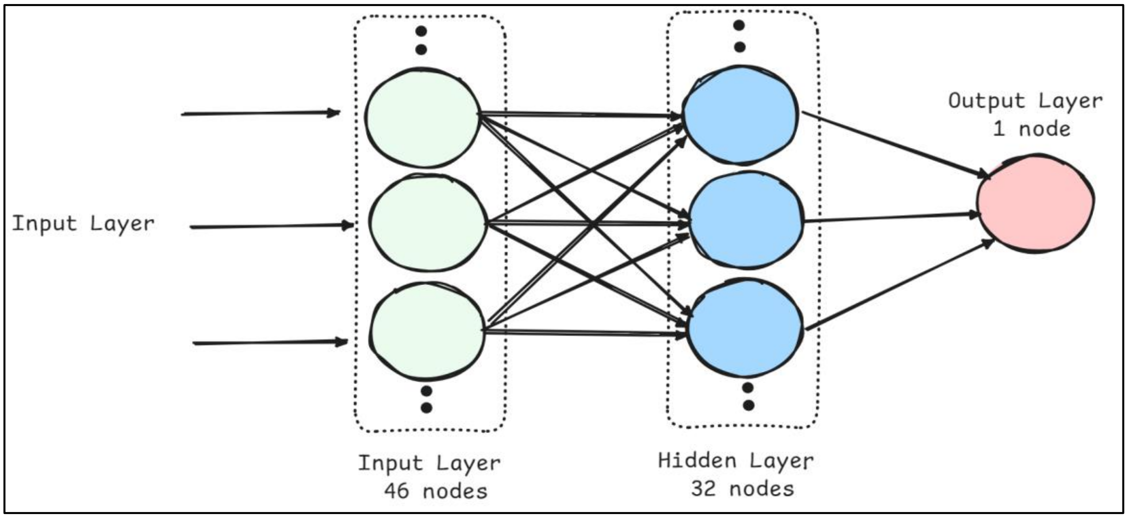


Figure 17: FFNN Topology

4- Optimizer

Adam Optimizer was used with a tunable learning rate as a part of the hyper-parameter tuning process, this optimizer is efficient in training the deep learning models, this optimizer has an adaptive learning rate that helps in achieving the best training performance. More to add, it has a feature of bias correction, especially during the initialization phase which prevents the model from being biased to the zero in initial phases (Kingma et al., 2014).

5- Loss Function

The "Binary Cross-entropy" function was used as a loss function which is efficient for binary classification cases, it calculates the difference between the actual and predicted values. Moreover, the Binary Cross-entropy function is efficient when dealing with the unbalanced dataset, this was a motivation to use it for LEMON dataset (Ruby et al., 2020).

It has several advantages. It is computationally efficient, works with probabilistic models, and is differentiable which makes it compatible with ADAM optimizer (Hurtiket al., 2022).

6- Hyperparameter tuning

Grid search and Cross-validation were used to tune the deep learning model seeking better results. The used cross-validation came with 5 folds.

The grid search parameters were as per the following:

- Learning Rate (lr): [0.001, 0.01, 0.1].
- Batch Size: [16, 32, 64].
- Epochs: [10, 20, 30, 50, 100].

3.7 Summary

This chapter outlines the model design and development process. The thesis utilizes two datasets (LEMON and HCP) which are related to fMRI recordings of healthy subjects to classify sex and brain age. The LEMON dataset has a problem represented by unbalanced classes between males and females. This problem led us to use an augmentation method to balance the classes. The HCP data is larger and doesn't have the problem of unbalanced classes.

MSDL and Schefer atlases were used and applied to LEMON dataset as a part of the feature engineering phase, the ROIs were extracted and used to generate connectivity matrices. Network graph parameters were extracted and merged with connectivity features to improve the classification results.

SVM and FFNN were used to classify the sex and brain age groups. The models were tuned with different hyper-parameters to ensure the optimal classification results from each model.

Chapter Four

Results and Discussion

4.1Results and Discussion

LEMON dataset is one of the important datasets that contain different modalities and provides a row data format for the images allowing the users to do their preprocessing and preparation. The dataset is available online with free access, it comes with different modalities such as fMRI, EEG, and Electrocardiogram (ECG) with gender and age-group information for each participant (Babayan et al., 2019). The participants completed a 3 Tesla fMRI providing 211 noncorrupted images, these images were used as a data source for this thesis. Moreover, the issues with this dataset are the small sample size and the unbalanced gender classes between males and females which imposes a challenge for classification tasks. For HCP dataset, it provides 1200 fMRI images with 3-T measurement for all subjects and 7-Tesla MRI scans for 184 subjects, this dataset includes some other modalities scans as 95 subjects have some resting-state MEG recordings. This dataset provides a bigger number of subjects in comparison with LEMON, which results in enhancing the classification results for gender. The dataset is preprocessed with 50 PCA (Human Connectome Project, 2017).

The main feature engineering pipeline used in this thesis is connectivity feature analysis, which implies a correlation between the different ROI generated based on the used mask (MSDL/Scheafer). (Iraji et al.,2016) indicates that there is a connection between the brain region of interest that can be used in studying some biological and mental cases related to the brain. So, feature connectivity analysis is the main pipeline for the feature engineering process in this thesis. Another feature of engineering is the network graph theory. By using such theory, some features were added to the connectivity features seeking enhancement in classification accuracy for gender and age.

The LEMON dataset comes with unbalanced classes regarding gender, 67% of the classes are males and 33% are females. For age classes, the data provided 9 classes starting from 20-25 till 75-80, the classes were aggregated to have two classes only (Youth and Elderly) which helps in finding a separation line resulting in a good separation result by the proposed models.

Below are the results of the study as per feature engineering process and modeling:

4.2 SVM with LEMON dataset

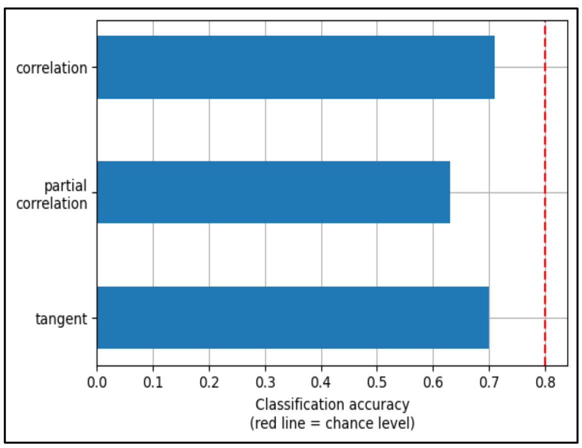
1- Feature connectivity only

Two Atlases were used in this process (MSDL and Sheafer), and no augmentation process was involved in this part, the results are as per the following:

Gender Classification:

- **For MSDL Atlas (39 ROIs)**

The best accuracy achieved for gender classification using SVM model (with MSDL Atlas applied to extract ROIs) was 71% and AUC= 0.81 (average recall= 60% and the best was with 5 folds), Augmentation wasn't used in this part and the only feature extraction method was connectivity, below figures illustrate such results:



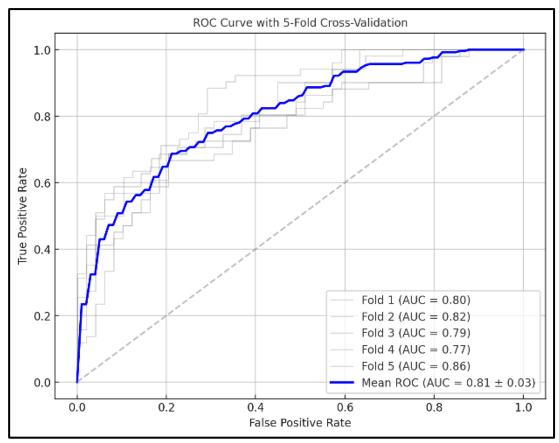


Figure 18: Gender Classification Accuracy (left) and ROC curve (right) - MSDL ATLAS - SVM with connectivity

- **For Sheafer Atlas (100 ROIs)**

The best accuracy achieved for gender classification using Scheafer Atlas and SVM model (without applying an augmentation process) was 72% and AUC=0.75 (average recall= 61%).

Scheafer Atlas has a wider range of ROIs than MSDL Atlas, this was promising in extracting more futures resulting in better classification accuracy. However, by using Scheafer atlas, no significant improvement was achieved. The below figures illustrate the classification accuracy and ROC curve:

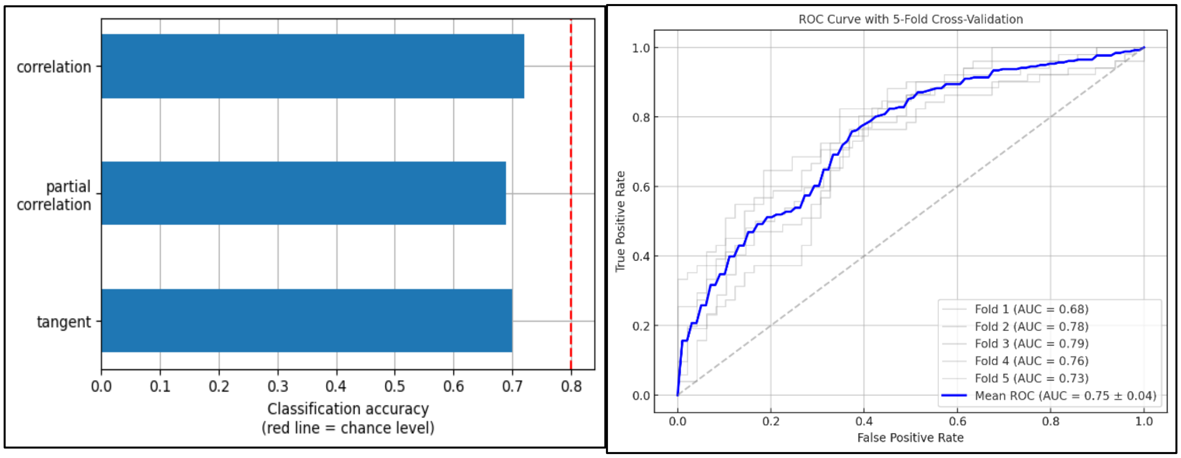


Figure 19: Gender Classification Accuracy (left) and ROC curve (right) - Sheafer ATLAS - SVM with Connectivity

Age Classification

- For MSDL Atlas (39 ROIs)

By using MSDL atlas, SVM model achieved an accuracy of 91% with AUC=0.90 (average recall= 88%), Below figures illustrate the classification accuracy and ROC curve:

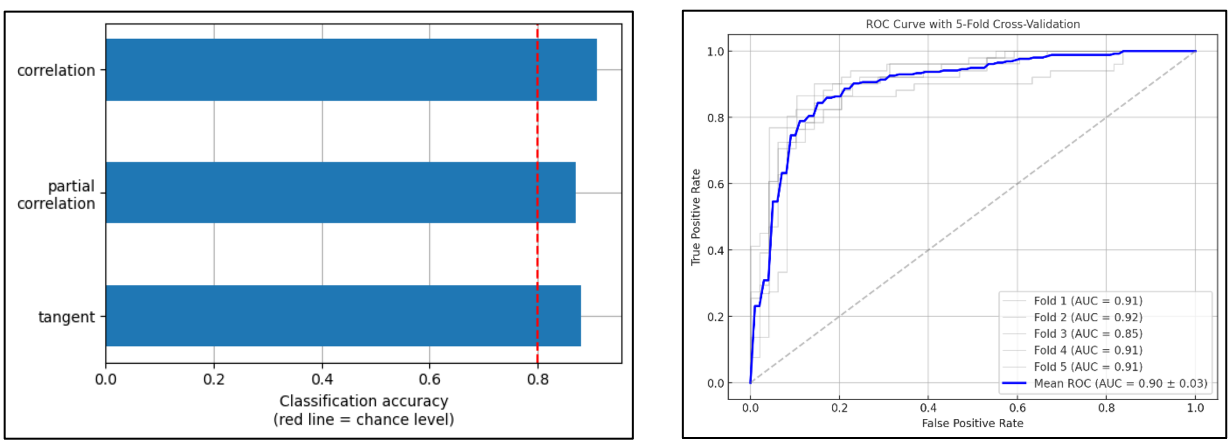
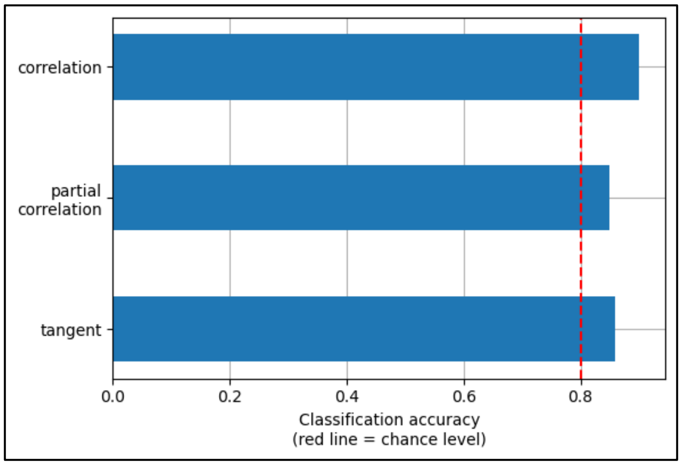


Figure 20: Age Classification Accuracy (left) and ROC curve (right) MSDL ATLAS - SVM with Connectivity

- **For Sheafer Atlas (100 ROIs)**

By using Scheafer atlas, SVM mode achieved an accuracy of 90% and AUC=0.95 (average recall=89 % and the best was with 5 folds), The Below figures illustrate the classification accuracy and ROC curve:



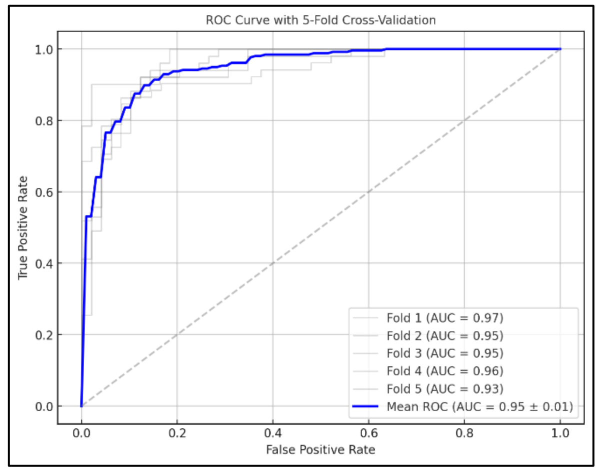


Figure 21: Age Classification Accuracy (left) and ROC curve (right) - Sheafer ATLAS - SVM with Connectivity

2- Feature connectivity with network parameters

In this section, network graph parameters such as degree centrality, average clustering, path length, the gender and age-group classification accuracy have been recalculated to study the effect of such network parameters, below are the results:

Gender Classification

- **For MSDL Atlas (39 ROIs)**

By using the feature connectivity in addition to the network parameters, and by using MSDL atlas, the gender classification accuracy enhanced to 74% with AUC = 0.79 (average recall= 64%), below figures illustrate the gender classification accuracy and ROC curve:

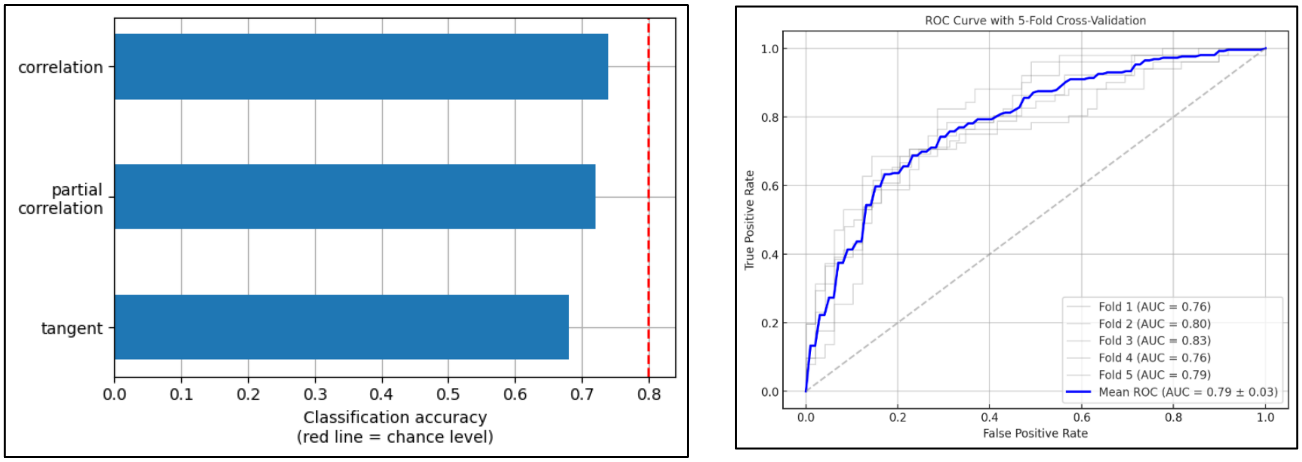


Figure 22: Gender Classification Accuracy (left) and ROC curve (right) MSDL ATLAS - SVM with connectivity & network parameters

- **For Sheafer Atlas (100 ROIs)**

By using a sheafer atlas with 100 ROIs, the accuracy was 70% and AUC= 0.76 (average recall= 59%), below figures illustrate the gender classification accuracy and ROC curve:

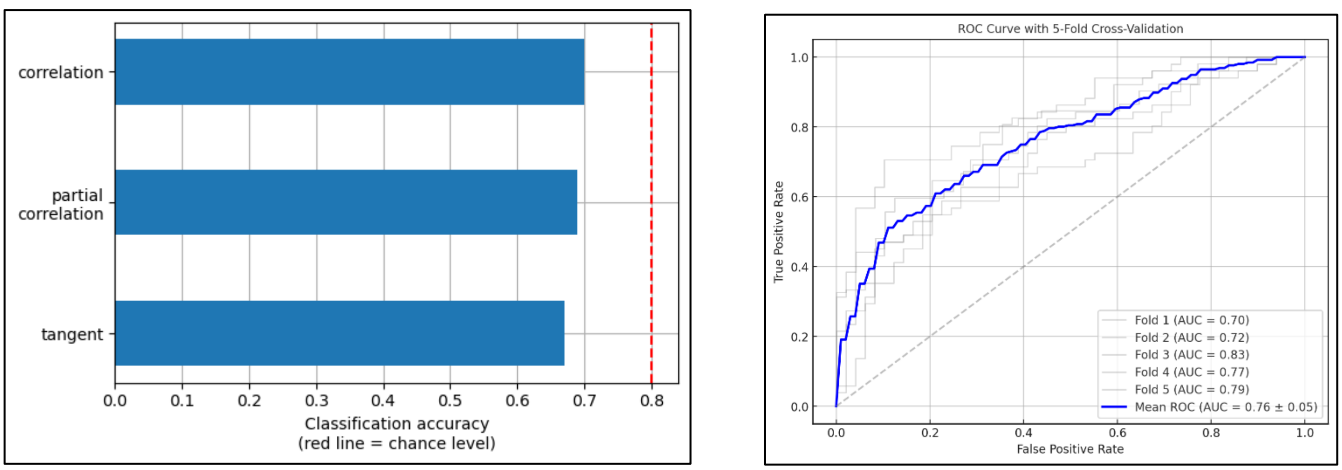
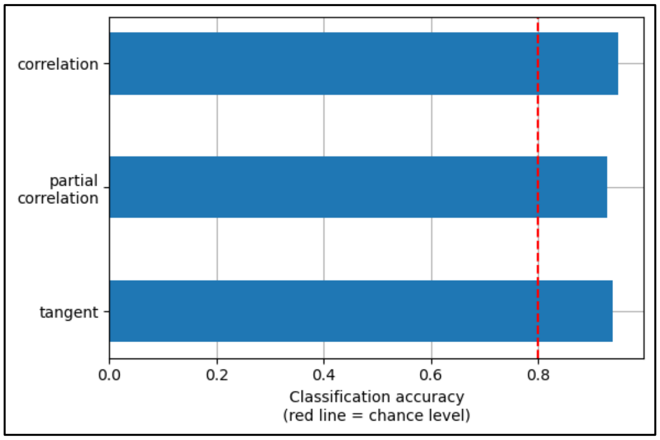


Figure 23: Gender Classification Accuracy (left) and ROC curve (right)- Scheafer ATLAS - SVM with connectivity & network parameters

Age Classification

- **For MSDL Atlas (39 ROIs)**

The age group classification using MSDL atlas and with network parameters provided the highest accuracy with 92% and AUC=0.95 (average recall= 90%), below figures illustrate the age classification accuracy and ROC curve:



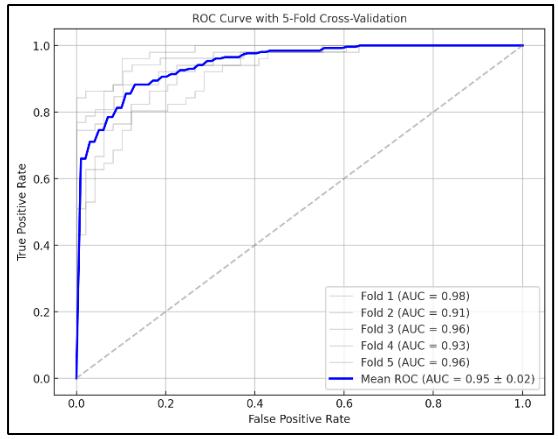
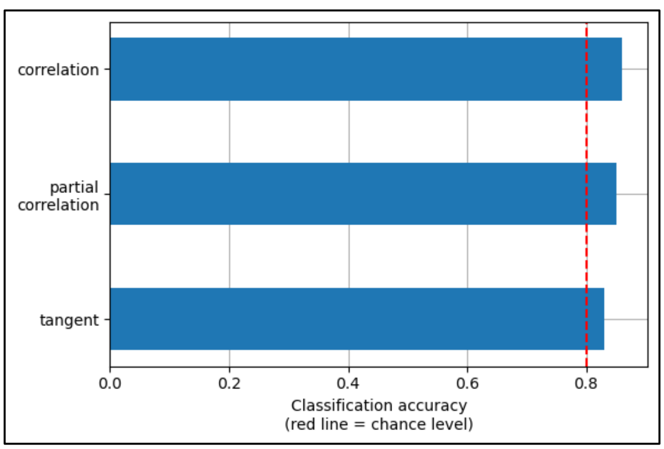


Figure 24:Age Classification Accuracy (left) and ROC curve (right)- MSDL ATLAS - SVM with connectivity & network parameters

- **For Sheafer Atlas (100 ROIs)**

By using the Schaefer atlas, and with network parameters existence, the age-group classification was good as well with an accuracy of 86% and AUC= 0.86 (average recall= 84%), below figures illustrate the results:



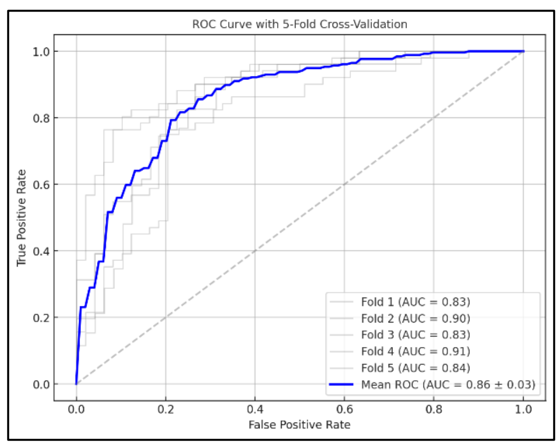


Figure 25: Age Classification Accuracy (left) and ROC curve (right) - Scheafer ATLAS - SVM with connectivity & network parameter

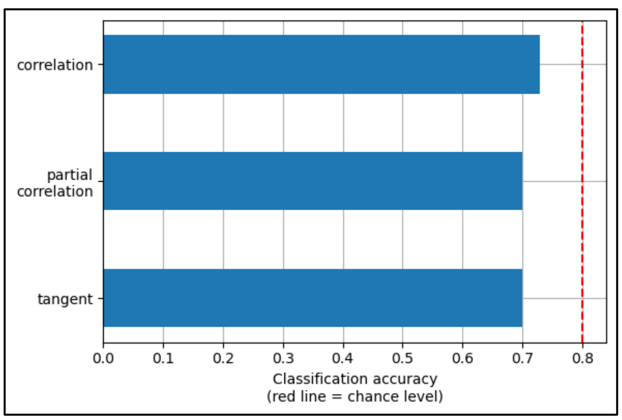
3- Feature connectivity with data augmentation

In this section, feature connectivity had been used as the only feature engineering pipeline but with data augmentation, the augmentation process had already been described in Chapter 3, the results were as per the following:

Gender Classification

- **For MSDL Atlas (39 ROIs)**

By using the MSDL atlas along with data augmentation, there was almost no change in accuracy with 73% with AUC=0.78 (average recall= 60%), below figures illustrate the results:



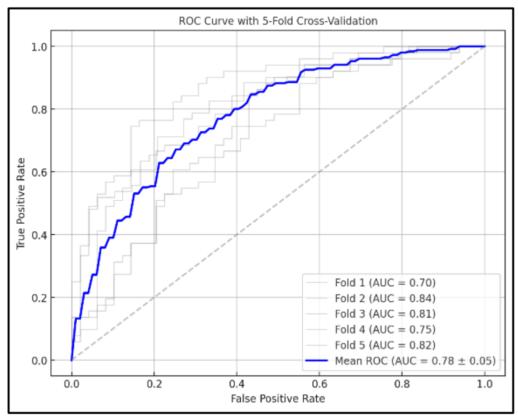


Figure 26: Gender Classification Accuracy (left) and ROC curve (right) - MSDL ATLAS - SVM with connectivity & data augmentation

For Sheafer Atlas (100 ROIs)

By using a sheafer atlas with data augmentation, the accuracy increased to 76% with AUC=0.83 (average recall= 62%), below figures illustrate the results:

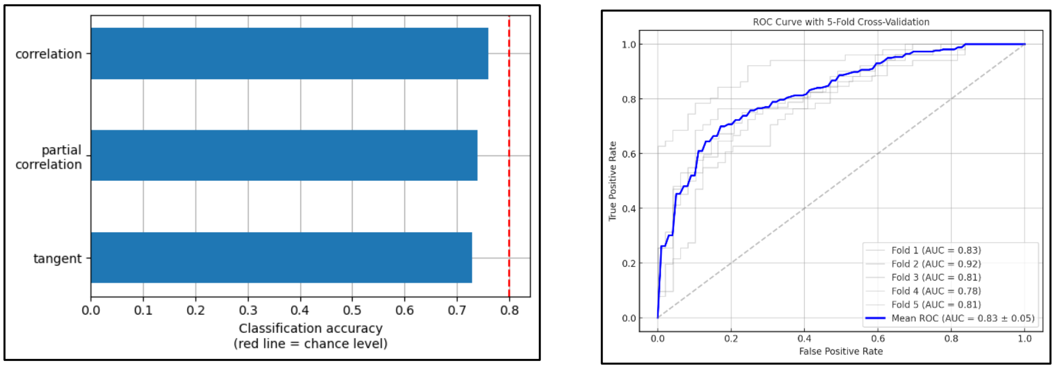


Figure 27: Gender Classification Accuracy (left) and ROC curve (right)- Scheafer ATLAS - SVM with connectivity & data augmentation

Age Classification

- **For MSDL Atlas (39 ROIs)**

By using MSDL atlas, it was clear that data augmentation wasn't helpful with age classification, the age classification accuracy decreased to 89% with AUC= 0.97 (average recall= 88%)**,** below figures illustrate the results:

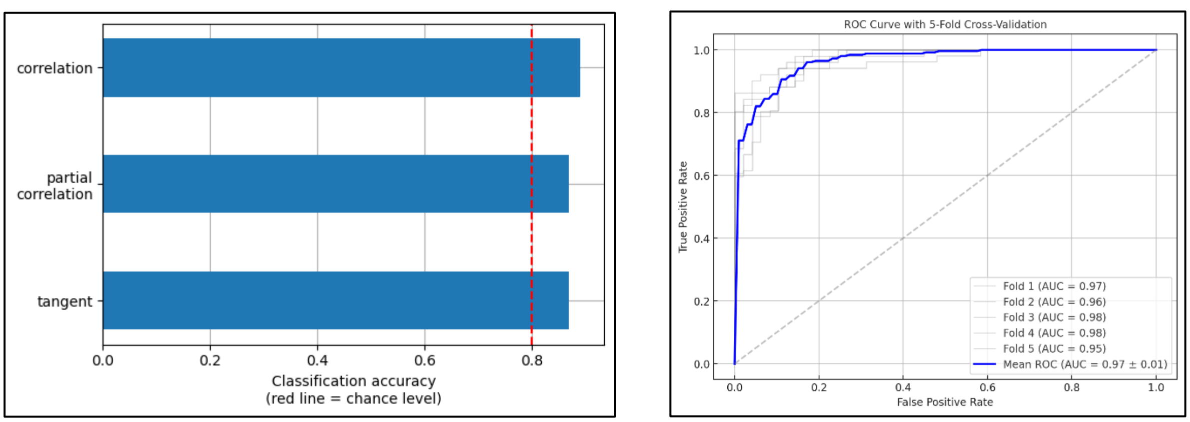


Figure 28: Age Classification Accuracy (left) and ROC curve (right) - MSDL ATLAS- SVM with connectivity & data augmentation

- **For Sheafer Atlas (100 ROIs)**

By using the sheafer atlas, we had no progress as well, the accuracy was less than the MSDL atlas, which was 83% with AUC=0.98 (average recall= 80%), below figures illustrate the results:

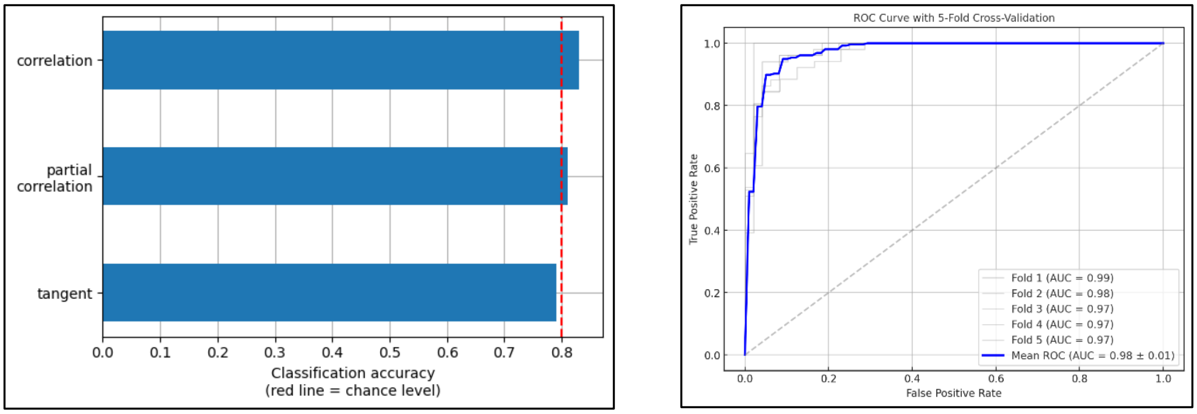


Figure 29: Age Classification Accuracy (left) and ROC curve (right) - Scheafer ATLAS - SVM with connectivity & data augmentation

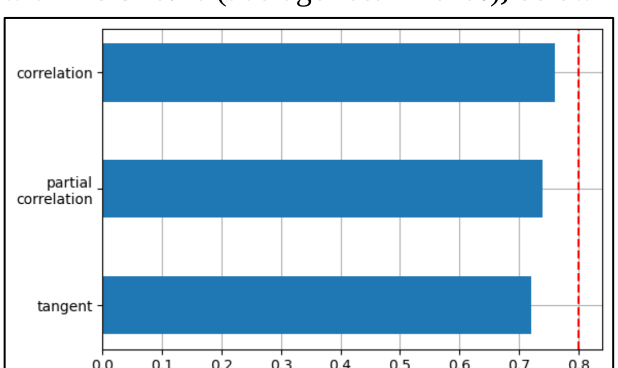
4- Feature connectivity with balanced data augmentation

As long as the female class is the minor one, a balanced augmentation was used to increase the female samples in a higher ratio than males making sure they are balanced at the end – described in Chapter 3, the results were as per the following:

Gender Classification

- **For MSDL Atlas (39 ROIs)**

By using balanced augmentation with MSDL atlas, the accuracy was increased hitting 76% with AUC=0.76 (average recall= 64%)**,** below figures illustrate the results:



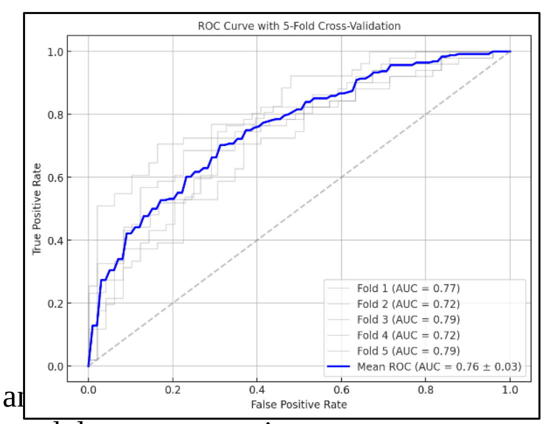


Figure 30: Gender Classification Accuracy (left) ar[…] SVM with connectivity & balanced data augmentation

- **For Sheafer Atlas (100 ROIs)**

By using balanced augmentation with the Sheafer atlas, the accuracy almost didn't change, it was 76% with AUC=0.81(average recall= 65%), below figures illustrate the results:

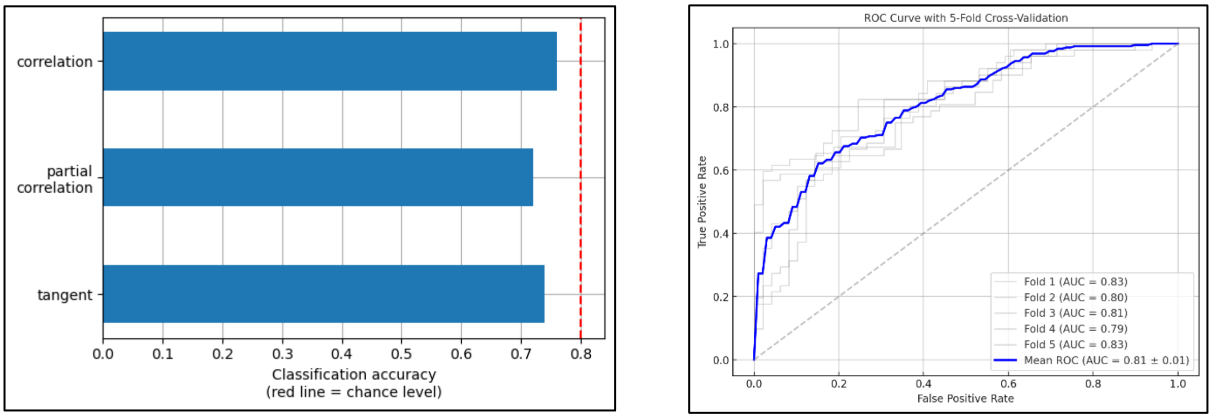


Figure 31: Gender Classification Accuracy (left) and ROC curve (right) - Scheafer ATLAS - SVM with connectivity & balanced data augmentation

Age Classification

- **For MSDL Atlas (39 ROIs)**

By using balanced augmentation with MSDL atlas, the accuracy of age classification decreased to 85% with AUC=0.97 (average recall= 83%), below figures illustrate the results:

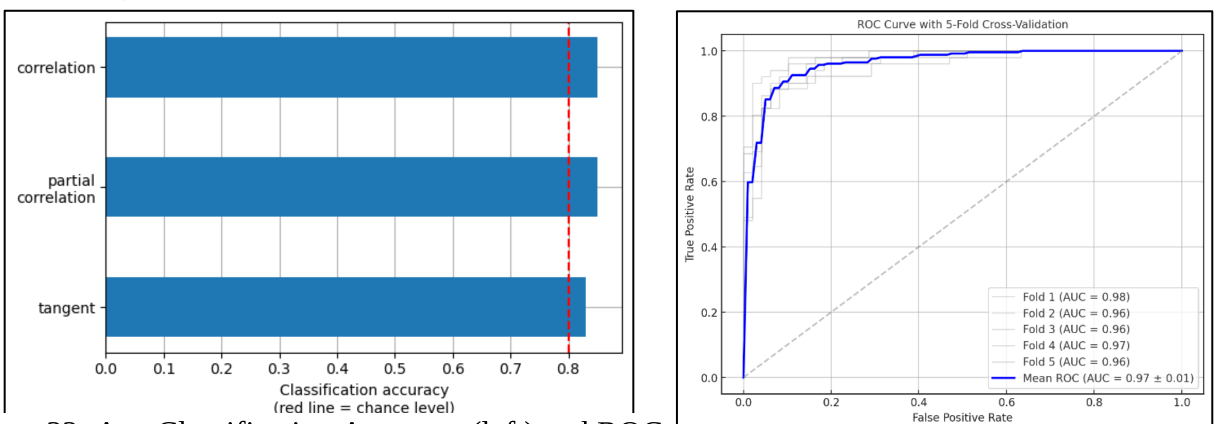


Figure 32: Age Classification Accuracy (left) and ROC curve (right) MSDL ATLAS - SVM with connectivity & balanced data augmentation

- **For Sheafer Atlas (100 ROIs)**

By using balanced augmentation with the Sheafer atlas, the accuracy almost didn't change, it was 84% with AUC=0.9 (average recall= 81%), below figures illustrate the results:

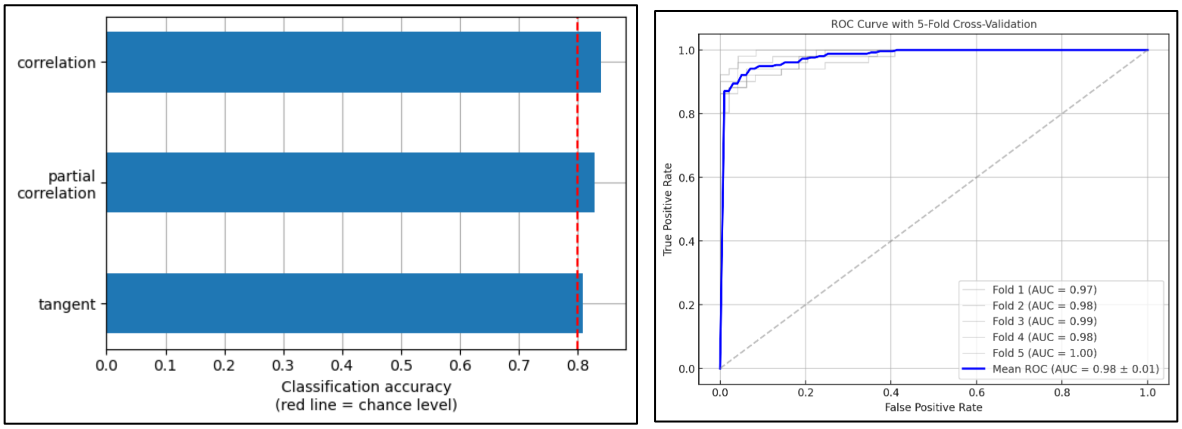


Figure 33: Age Classification Accuracy (left) and ROC curve (right) - Scheafer ATLAS - SVM with connectivity & balanced data augmentation

5- **Feature connectivity with network parameters and balanced data augmentation:**

This is the last case with LEMON dataset using SVM model, the accuracy had been recalculated for age and gender classification, and the results were as per the following:

Gender Classification

- **For MSDL Atlas (39 ROIs)**

By using MSDL atlas, the gender classification was the highest by 78.5% with AUC=0.78 (average recall= 67%), below figures illustrate the results:

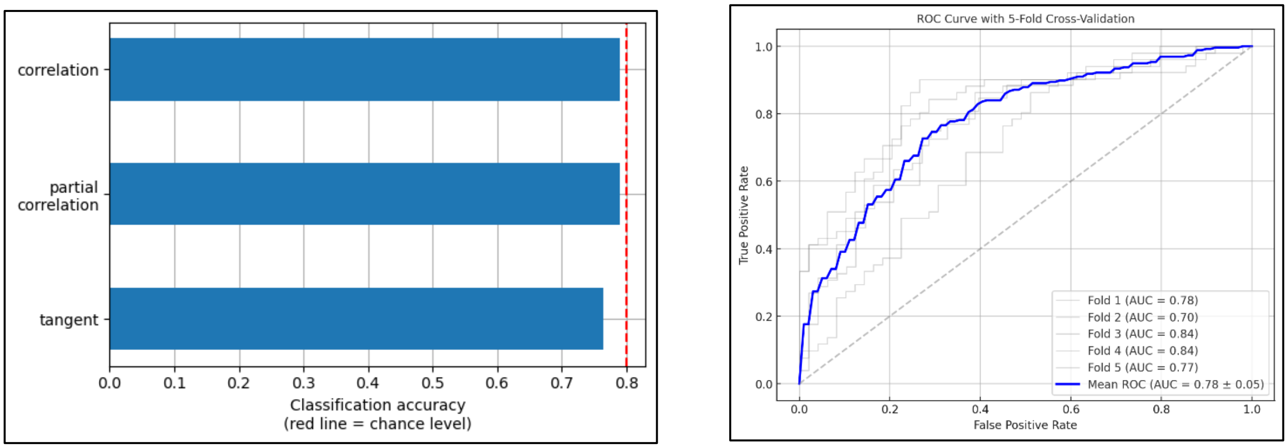


Figure 34: Gender Classification Accuracy (left) and ROC curve (right) - MSDL ATLAS - SVM with connectivity, network parameters and balanced data augmentation

- **For Sheafer Atlas (100 ROIs)**

With sheafer atlas, the accuracy decreased up to 74% with AUC=0.74 (average recall= 61%)**,** below figures illustrate the results:

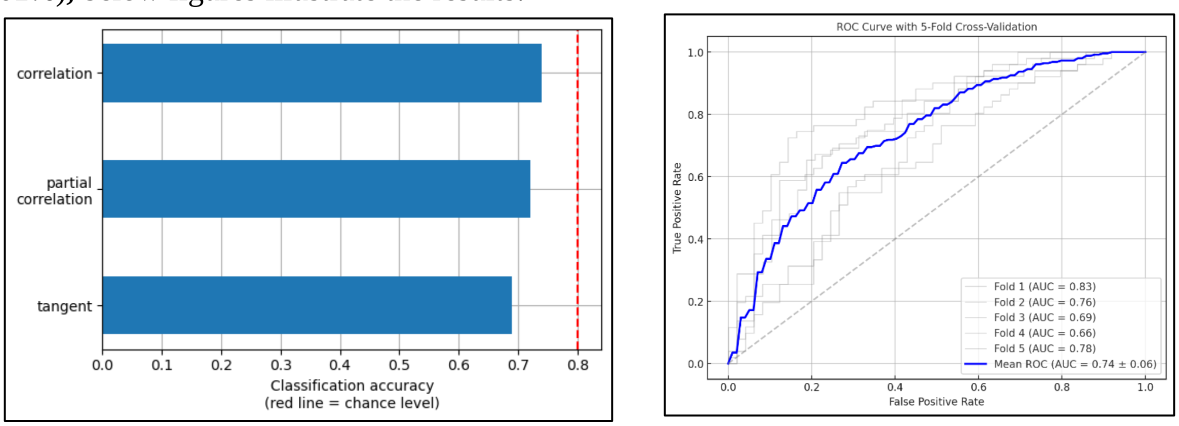
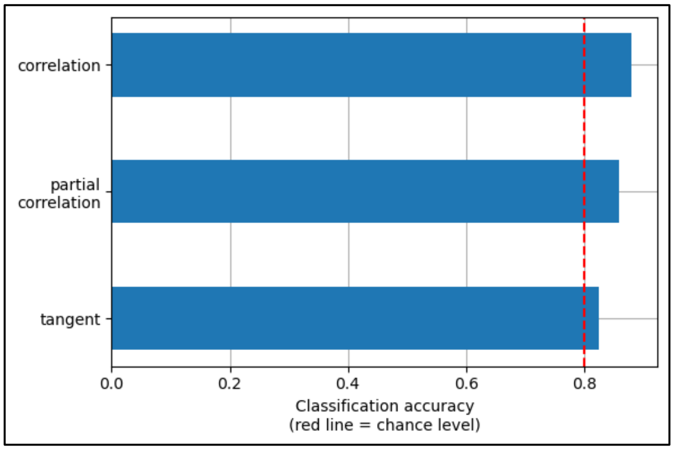


Figure 35: Gender Classification Accuracy (left) and ROC curve (right) - Scheafer ATLAS - SVM with connectivity, network parameters and balanced data augmentation

Age Classification

- **For MSDL Atlas (39 ROIs)**

There is no significant change in age classification accuracy using MSDL atlas, the accuracy was 88% with AUC=0.91(average recall= 89%)**,** below figures illustrate the results:



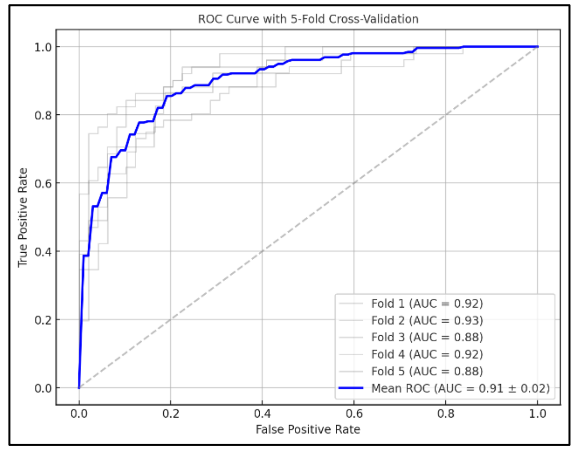
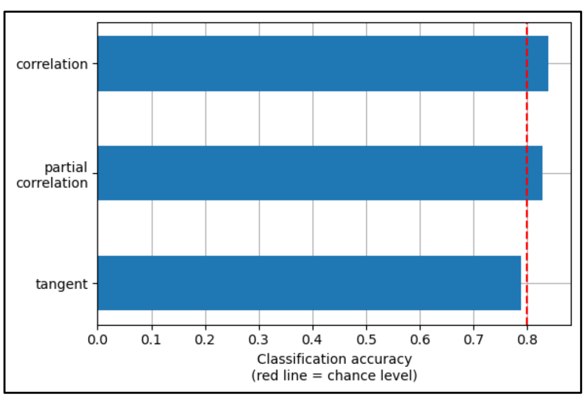


Figure 36: Age Classification Accuracy (left) and ROC curve (right) - MSDL ATLAS - SVM with connectivity, network parameters and balanced data augmentation

- **For Sheafer Atlas (100 ROIs)**

The accuracy decreased to 84% with AUC= 0.93 (average recall= 85%)**,** below figures illustrate the results:



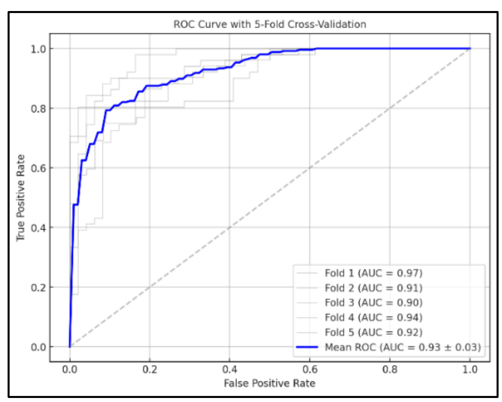


Figure 37: Age Classification Accuracy (left) and ROC curve (right) - Scheafer ATLAS - SVM with connectivity, network parameters and balanced data augmentation

Below two tables below summarize all accuracy results related to the gender and Age classification results for LEMON dataset using SVM model:

Table 2: Classification results for Age and Gender using MSDL atlas and SVM model - LEMON dataset

Atlas	Feature Engineering	Gender		Age	
		Accuracy	AUC	Accuracy	AUC
MSDL	Feature Connectivity	0.71	0.78± 0.06	0.91	0.94± 0.03
	Feature Connectivity + Network parameters	0.74	0.79± 0.07	0.92	0.95± 0.03
	Feature Connectivity +Data Augmentation	0.73	0.78± 0.12	0.891	0.97± 0.01
	Feature Connectivity + Balanced Data Augmentation	0.76	0.76± 0.06	0.85	0.97± 0.01
	Connectivity + Balanced Data Augmentation + Network parameters	0.785	0.78± 0.08	0.88	0.91± 0.02

Table 3: Classification results for Age and Gender using Sheafer atlas and SVM model - LEMON dataset

Atlas	Feature Engineering	Gender		Age	
		Accuracy	AUC	Accuracy	AUC
Sheafer	Feature Connectivity	0.72	0.74± 0.06	0.9	0.95± 0.03
	Feature Connectivity + Network parameters	0.701	0.76± 0.07	0.86	0.86± 0.01
	Feature Connectivity +Data Augmentation	0.76	0.83± 0.04	0.837	0.98± 0.00
	Feature Connectivity + Balanced Data Augmentation	0.76	0.81± 0.07	0.84	0.98± 0.00
	Connectivity + Balanced Data Augmentation + Network parameters	0.74	0.74± 0.06	0.841	0.93± 0.03

4.3 SVM with HCP dataset

Because LEMON dataset is relatively small, HCP data was a choice to extend the study and apply the analysis on a larger dataset that may provide better results. With 1200 objects, and 50 PCA components applied, the SVM model was trained on different subsets **(200,400,600,800,1000)** seeking better gender classification results.

Gender Classification

The below table shows the gender classification results for different subsets of HCP dataset:

Table 4: HCP - Gender classification results over data subsets

Feature's Combination	200		400		600		800		1000	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
Connectivity Feature only	0.92	0.90	0.91	0.91	0.93	0.95	0.96	0.95	0.95	0.96
Connectivity + degree centrality	0.91	0.95	0.90	0.99	0.93	0.97	0.95	0.97	0.94	0.96
Connectivity + average clustering	0.92	0.97	0.91	0.99	0.93	0.97	0.96	0.99	0.95	0.98
Connectivity + path length	0.92	0.96	0.90	0.98	0.93	0.96	0.96	0.97	0.95	0.97
Connectivity + path length + average clustering	0.92	0.95	0.90	0.91	0.93	0.95	0.96	0.95	0.95	0.96
Connectivity + path length + degree centrality	0.92	0.97	0.91	0.97	0.93	0.96	0.96	0.96	0.95	0.97
Connectivity + degree centrality + average clustering	0.91	0.98	0.89	0.98	0.93	0.97	0.95	0.98	0.94	0.98
Degree centrality only	0.43	0.42	0.47	0.54	0.43	0.47	0.41	0.45	0.43	0.45
Average clustering only	0.61	0.55	0.50	0.45	0.66	0.60	0.68	0.62	0.69	0.65
Path length only	0.50	0.45	0.50	0.56	0.51	0.46	0.54	0.48	0.53	0.48
Degree centrality + path length + average clustering	0.45	0.49	0.49	0.54	0.44	0.48	0.48	0.49	0.49	0.47
Path length + degree centrality	0.49	0.42	0.49	0.47	0.51	0.49	0.52	0.48	0.53	0.49
Average clustering + degree centrality	0.46	0.51	0.47	0.56	0.49	0.55	0.50	0.54	0.52	0.51
Average clustering + path length	0.49	0.49	0.50	0.46	0.45	0.48	0.47	0.46	0.48	0.48
All features	0.92	0.97	0.89	0.98	0.93	0.95	0.95	0.98	0.94	0.98

4.4 Gender Classification using FFNN model with PCA – LEMON dataset

Due to the poor results related to gender classification over LEMON dataset, another model (FFNN) was used along with PCA, this test was applied after considering the balanced data augmentation, the idea of using the principal component analysis obtained from HCP dataset experience. So, the same concept was tested seeking better results.

The feature engineering for the dataset was about using the feature connectivity matrix along with network parameters (degree centrality, average clustering, and path length). More to add, 50 components were generated to make it an apple-to-apple comparison with HCP data that was processed with the same number of components. The results enhanced achieving 82% testing accuracy.

Below figure-41 shows the model accuracy over the changing epochs:

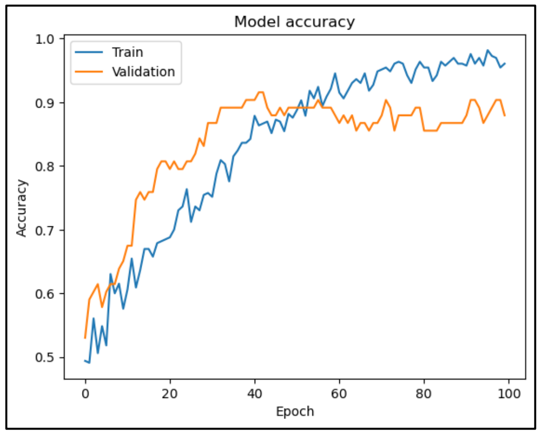


Figure 38: FFNN Model accuracy for LEMON PCA model

More to add, SVM was used again on LEMON dataset after Appling the PCA to check if the accuracy will be enhanced as well by using SVM the results demonstrated a clear positi ve change achieving a testing Accuracy of 84% with the Best Parameters: {'C': 0.1, 'gamma ': 0.1}. The below figure shows the classification report for the SVM results:

Classification Report:

	precision	recall	f1-score	support
0	0.97	0.71	0.82	55
1	0.75	0.98	0.85	49
accuracy			0.84	104
macro avg	0.86	0.84	0.84	104
weighted avg	0.87	0.84	0.83	104

Figure 39: Classification Report - SVM- LEMON PCA

4.5 Key Findings and Results

LEMON Dataset Limitations and Performance:

Gender classification using SVM did not start very strong with the connectivity feature alone: 71% and 72%, respectively, for MSDL and Schaefer atlases.

However, the age classification produced a substantially more significant boost to an average 91% accuracy in MSDL and 90% in Schaefer.

Impact of Network Parameters:

The addition of network graph parameters (degree centrality, average clustering, and path length) enhanced the gender classification accuracy by 74% and age classification accuracy by 92% on the MSDL atlas. For the Schaefer atlas, the accuracy of gender classification slightly decreased.

Balanced Data Augmentation (LEMON Dataset):

Significant improvements were seen when balanced data augmentation was employed with a 78.5% gender classification accuracy with the MSDL atlas, the highest obtained for this dataset. For the Schaefer atlas, accuracy was 76%, which is quite lower compared to the previous.

PCA and Dimension Reduction:

Performing PCA at 50 components for the features of LEMON dataset generally improved the performances of classification:

Gender classification for SVM stood at 84% and Gender classification for FFNN peaked at 82%. Legitimate reduction of PCA increased the quality of classification along with the original features.

HCP Dataset Performance:

On the larger HCP dataset (1200 samples), the gender classification would perform much better than on LEMON, with 96% accuracy and AUC=99% from SVM. This was consistent in all cases from data increment and then combining connectivity features with network parameters such as average clustering.

4.6 Discussion

In this section, the results will be deeply discussed showing the limitations, achievements, and possible improvements, hereunder are the points of discussion:

SVM and FFNN models were chosen as the main models for this case study due to the fact that they provide the best results. fMRI images are high-dimensional containing many voxels and regions, this makes SVM fit with such cases because of its ability to find the boundaries making benefits from kernel capabilities of separation (Vapnik et al., 2013). More to add, SVM is efficient with selected feature problems, and this is the case in the project. Also, SVM is doing well with data that is treated with reduction processes such as ICA and PCA (Formisano et al., 2008). FFNN, fits with a small dataset because of its simplicity in design, it doesn't involve complex computational processes such as convolutional neural network (CNN) (LeCun et al., 2015). Additionally, FFNN model comes with a simple design which allows the user to control the overfitting problem more easily than other models (Srivastava et al., 2014). Moreover, FFNN has a proven success in neuroimaging problems, this was one of the reasons to try this model (Vieira et al., 2017). Finally, other machine learning and deep learning methods were used but with worse results (Random Forest with less than 70% accuracy and DNN with almost 65%).

For **LEMON** dataset masked with MSDL atlas, and with feature connectivity only as a feature engineering pipeline, SVM gender classification accuracy wasn't high (71%), this is mainly because of the small size of the dataset in addition to the unbalancing issue between males and females. When another atlas (Schafer) was used with a higher ROI, the results almost didn't change with a notable drop in AUC value due to the increase in data complexity resulting from increasing the data dimensions. For gender classification and under the same feature engineering boundaries, the accuracy was much better achieving 91% and 90% for MSDL and Scheafer atlases. The justification for the high accuracy results in age-group classification rather than gender is the fact of re-grouping the age classes in LEMON dataset. Moreover, (Raz et al., 2006) indicate that the brain undergoes significant changes (functional and structural) with aging, these differences can be obvious to be distinguished from the machine learning model between young and older people.

To enhance the results, we came out with the idea of using the network graph theory, which allows doing mathematical relation between the region of interests that will generate different metrics such as degree centrality, average clustering, and path length.

These metrics are considered as additional features along with the main connectivity features mentioned in the point above, SVM was used and the results for the gender classification were enhanced using MSDL Atlas achieving an accuracy of 74%. The results of gender classification with Scheafer atlas came with no enhancement, this is because of the complexity in dimensions resulting from features increasing. For Age classification, the SVM accuracy slightly increased by 1% achieving an accuracy of 91% for MSDL in addition to an enhancement for AUC, while the accuracy and AUC decreased for scheafer atlas hitting 86% for the same suggested reasons mentioned in previous points.

As one of the main issues in the classification is the small sample size, a data augmentation technique was used to split the time-series data (before applying the connectivity transformation and without adding any network parameters) into two parts for each image which doubled the dataset size, the results for SVM gender classification using MSDL atlas didn't enhance (73%) while it recorded an enhancement for scheafer atlas achieving 74% accuracy. The enhancement in the accuracy of the latter is because removing the network parameters and increasing the row dimensions will decrease the effect of feature's dimensions. For age classification, the SVM accuracy was decreased in both MSDL and Scheafer atlases while the AUC recorded a slight increase, the decrease in accuracy can be justified due to the increase in noise that may introduced by data augmentation, which makes it a little bit harder for the SVM to distinguish between the classes. For the AUC, the increase can be justified with the fact augmentation effect in generalization, data augmentation helps in enhancing the ranking prediction which is the key factor of AUC (Raz et al., 2006).

Data augmentation was proposed seeking to enhance the results, but the results didn't enhance with MSDL atlas due to the effect of unbalanced classes mainly for gender. So, a balanced augmentation technique was proposed to overcome this issue (without adding network graph parameters in this case), male classes were augmented by splitting each time series into two objects, and female classes were divided into 4 subjects, this approach made the two classes are balanced having about 500 subjects in total. The SVM classification for gender using MSDL atlas was enhanced reaching 76%, with almost no change for Scheafer case. This enhancement was clearly due to balancing the classes. For age classification, SVM classification using MSDL and Scheafer atlas decreased to the level of an accuracy of 85% because the balanced augmentation was done based on gender class distribution which affected the age classes. As a result of that, age classification results were better without this approach.

As a final step here, the network parameters were added to the case above having connectivity features with network parameters under a balanced augmentation umbrella. The SVM gender classification recorded the highest accuracy (using MSDL) hitting nearly 79% while Schaefer atlas recorded an accuracy of 74%. There is a fact here that Scheafer atlas doesn't respond costively to the augmentation process – in comparison with MSDL – and this can be justified due to the high number of dimensions used there (100 ROIs). For age classification, MSDL and Scheafer with SVM classification didn't provide an enhancement in the results with an average of 86% accuracy.

For HCP dataset, the gender classification was targeted here in order to get better results from a bigger dataset. The classification was done on different sizes of patches from the whole dataset, 200 (similar to LEMON), 400, 600, 800, and 1000 were processed to monitor how the accuracy can be enhanced when changing the sample size. For each cycle, the network parameter was merged with connectivity features then we can be able to compare the results with LEMON ones. No atlases were used here as the data had already been processed with PCA as the atlas process needs a specific format. The best gender classification accuracy achieved here was by using SVM with connectivity feature in addition to the average clustering as one of the network parameters, the accuracy reached 96% with AUC=99%. It turned out that when increasing the size of the data, the accuracy consistently increased.

Seeking to enhance the accuracy of the gender classification with LEMON dataset, the PCA was applied (as a lesson learned from HCP) with 50 components on the dataset after extracting the feature connectivity and merging them with network graph parameters, the results enhanced by FFNN to be 82% and 84% for SVM accuracy.

For the LEMON dataset, the gender classification wasn't too high due to the fact of the small size. The steps of enhancement by merging network parameters, data augmentation, dimension reduction, and balancing mechanism helped to enhance the results up to 84%. In both datasets, and both targets (gender and age) the connectivity features are the very important ones. When removing them and depending on the network parameters as a standalone feature, the accuracy dramatically drops to low levels as shown in Table. The dimension reduction technique (PCA) helped in enhancing the results by reducing the feature space allowing the model to focus on the most relevant features (Jolliffe et al., 2016).

Data augmentation is not preferable for medical problems (Kobayashi et al., 2016). The technique used in this study was good because no new sample was generated by applying sampling operation or any mathematical processes, it was about splitting the time series in a way to make sure it won't lose the information that is necessary for the classification.

4.7 Summary

This thesis provided different insights. First, for LEMON dataset, SVM and FFNN models provided the best gender classification accuracy over other tested models. More to add, and for fMRI images, a bigger sample size leads to better accuracy. Moreover, the data augmentation technique helped in enhancing the results, especially with balancing the classes. Additionally, adding the network parameters graph helped in increasing the accuracy up to 79% for MSDL gender classification using SVM. Also, PCA pushed the accuracy higher achieving 84% with SVM as well. Schafer atlas wasn't helpful because of the complexity of the provided dimensions, the atlas was processed with 100 ROI providing a large number of features after the connectivity matrix generation. For age classification, the accuracy was high (about 91%) with feature connectivity only. Adding the network parameters to HCP data provided a slight increase in accuracy to 92%. The age classes had better distribution, especially after re-grouping them into two classes which helped the SVM model to distinguish between them. For HCP data, it was processed with PCA, and only gender classification was applied, the results were higher than LEMON because of the high sample size, balanced classes, and dimension reduction. The accuracy reached 96% for SVM gender classification using connectivity features with an average clustering coefficient as a supporting variable.

The experience of HCP data pointed towards using PCA technique in LEMON dataset after applying the connectivity feature and merging them with network parameters, the accuracy reached 84% for SVM and 82% for FFNN. Finally, the feature connectivity process is the dominant feature for this case study, network parameters alone couldn’t provide a good result but they helped in enhancing the process.

4.8 Recommendation

SVM and FFNN can be utilized in neuroimaging problems due to their nature of processing which is suitable for small and complex spaces. More to add, Data augmentation is very important in such a domain, fMRI data collection is very complex, takes time, and comes with noisy parameters. Finding a way to make the dataset bigger will help in enhancing the results and the generalization purpose.

MSDL atlas can be tested more because of the small number of regions provided. It will be helpful in case of studying the feature's importance. Moreover, Dimension reduction is very important in such case studies, it helps in reducing the dimensions especially when using the connectivity feature which provides a high number of variables. Additionally, Network graph parameters are very helpful as supporting features, this topic can be utilized more by trying other parameters on different datasets.

It is worth to fitch ways and solutions to overcome the issue of unbalanced classes; it has a main effect on the classification results especially in gender classification as observed in this thesis. Another point is to focus on the grouping of age classes is very effective, other techniques should be tried rather than considering binary classes only. It is worth mentioning that there is a need to collect other information rather than biological ones when handling the data collection process. This can help in studying other mental disorders or issues. Also, it is important to focus on building models that can be generalized on other datasets as this is the ultimate goal of having a robust model that can be helpful with different data sources.

Finally, trying to merge different datasets to have a bigger sample size will help in providing better accuracy and focusing on merging different modalities to seeking better results, LEMON dataset has EEG and ECG data, there is a need to check if merging such modalities along with fMRI data will be effective.

Chapter Five

Conclusion and Future Work

5.1 Conclusion

In this thesis, gender and age classification using the LEMON and HCP datasets are treated and approached from the standpoint of feature connectivity and network metrics as possible means of feature engineering. During this project, several modalities have been put to test and evaluated, ranging from SVM and feedforward neural network models to balanced data augmentation and dimensionality reduction using PCA. The results have shown that there is importance of data size, feature extraction method, and class-balancing techniques in furthering classification accuracy, especially for these neuroimaging datasets, preferably obtained through fMRI.

The LEMON dataset had some issues such as small sample size, unbalanced gender classes (67% male, 33% female), and gender ratio impact, which greatly affects the performance of classification tasks, especially in gender classification. However, balanced data enrichment combined with network parameter enhancement could improve gender classification (78.5% for MSDL atlas). It is suggested that these two methods effectively – somewhat- eliminated the dataset's limitations. Given that feature connectivity turned out to be the most important feature engineering pipeline across both datasets, its capability of capturing correlations between ROIs provided resilient input features that could have fed machine learning models well. Network parameters provided evidence of the ability to enhance the classification results, especially for the gender.

The illustration has shown that PCA was key to the solution when handling high-dimensional fMRI data. The decision to use 50 components was worthy of separate treatment and value; the classification accuracy increased up to 84% on SVM and 82% on FFNN for the LEMON dataset, proving that dimensionality reduction is indeed important in dealing with high-dimensional datasets and making models converge on relevant features.

Achieving gender classification accuracy of 96% with SVM and an almost equally successful FFNN model underscored the HCP dataset's advantages: a more balanced class distribution of gender (1200 subjects); implementing PCA in preprocessing indicating that PCA would work for the LEMON models and still push classification performance. Two non-overlapping atlases, MSDL and Schaefer, were considered. It has been noticed that the MSDL atlas; with fewer ROIs (39), outperformed the Schaefer atlas (100 ROIs) every time in terms of gender classification. This could be attributed to the excellence of simpler models with fewer features regarding smaller datasets such as LEMON. On the other hand, due to the higher dimensionalities of the Schaefer atlas, any formal improvement was hardly received by augmentation. Age classification outperforms gender classification without any doubt, with an average accuracy of 91% for feature connectivity with the MSDL atlas and 90% for Schaefer atlas. The classification boosted a bit with age group re-scaling to "Youth" and "Elderly," both of which are biologically reasonable since there are distinctive structural and functional brain changes that come with advancing age. The addition of network parameters improved the performance of the MSDL atlas to 92%; however, Schaefer's accuracy decreased by 86%.

One of the challenges was the small sample size for LEMON and class imbalance posed limitations to classification within the dataset. This insignificant limitation pointed to the need for large datasets in neuroimaging. More to add, Schaefer's atlas wasn’t effective in the study. Moreover, intending to the class imbalance issue, data augmentation indeed introduced noise, in particular, for the age classification. This implies that augmentation techniques need to be cautiously designed and evaluated so as not to make the learned model perform the other way.

5.2 Future work

Acquiring or combining additional fMRI datasets is critical to enhance and increase the robustness and generalizability of models. Larger dataset sizes will also allow the use of more sophisticated models such as deep neural networks that require a significant amount of training data. Moreover, LEMON has EEG and ECG data in addition to fMRI, Investigating the integration of these modalities is recommended for richer features. By combining data from different modalities, one could potentially exploit additional information, hence furthering classification performance. Additionally, It is recommended to investigate additional network graph parameters such as betweenness centrality, modularity, and eigenvector centrality to better augment the feature space. Machine learning techniques like feature selection and feature importance analysis could assist in pinpointing the most relevant features for the purposes of classification.

More to add, SVM and FFNN can be considered good for the classification of the LEMON data. Future work could possibly involve the application of more advanced models, such as CNNs and transformers, which match large-dimensional problems. Those advanced models might enhance the accuracy of the research results. Beyond data augmentation, methodological alternatives such as cost-sensitive learning and oversampling (via SMOTE) for handling class imbalance might be considered useful. They may give an added advantage where the class is highly imbalanced.

The binary age grouping permits some wonderful validation; finer classification of age classes could provide a more extensive insight into age-associated changes that affect the brain. Additionally, exploring the classification of subparts of age classes within the gender dimension may enable providing more patterns, which may play an important role in improving classification. Also, In the forthcoming work, the validation of the established models on independent neuroimaging datasets is mandatory for robust model performance. This will be the first step toward correctness. If successful, this will help in putting down a framework for gradually moving these methods to a consistent application across various datasets.

5.3 Summary

The study provided proof of concept for feature-wise analysis, network graph metrics, and machine learning models for gender classifications by fMRI data and age analysis. The lack of a sufficient sample size with imbalanced class representation was problematic, although maintaining symmetry in the data. It always remains a consideration in the concluding discussion on the possibility of data size, feature selection, and architecture of models in neuroimaging design. Further future work should be directed towards data engineering, consideration of multimodal data integration, and go through extreme performances through the use of advanced machine learning models. Finally, it is possible to check the ability to merge different datasets to increase the dataset size involving robust models which can lead to better classification performance which will finally result in generalizing the model on other datasets. This can be very helpful in having many applications that can be used in the neuroimaging domain.

The research questions and the thesis answered all of them. Firstly, regarding the question of the most important features used, the thesis had shown the connectivity feature is the most important feature, adding the network graph metrics helps in enhancing the results but cannot be effective if used alone with feature connectivity. Secondly, the question of merging different pipelines was answered by merging the feature connectivity with network graph theory metrics and enhancing the results. The last question was about how can the results of classification be enhanced, this question was answered by the use of an augmentation process and merging the network graph theory. Moreover, the study had shown the ability to have better results with a larger sample size. More to add, merging different modalities such as ECG and EEG may help in enhancing the results. Finally, the possibility of merging different datasets in order to have a bigger dataset can be helpful in providing a high accuracy rate for gender and age problems.

All of these points were mentioned as future work for the upcoming projects that can implemented which will lead to having a robust model that can be generalized on other datasets.

References

Aggarwal, Aman, and Hari Singh. 2005. “Optimization of Machining Techniques—a Retrospective and Literature Review.” Sadhana 30: 699–711.

Al Zoubi, O., Misaki, M., Tsuchiyagaito, A., Zotev, V., White, E., Paulus, M., & Bodurka, J. (2022). Machine Learning Evidence for Sex Differences Consistently Influences Resting-State Functional Magnetic Resonance Imaging Fluctuations Across Multiple Independently Acquired Data Sets. Brain connectivity, 12(4), 348–361. https://doi.org/10.1089/brain.2020.0878

Aldakheel, Fadi, Ramish Satari, and Peter Wriggers. 2021. “Feed-Forward Neural Networks for Failure Mechanics Problems.” Applied Sciences (Switzerland) 11(14).

Arslan, S., Ktena, S. I., Glocker, B., & Rueckert, D. (2018). Graph saliency maps through spectral convolutional networks: Application to sex classification with brain connectivity. In Graphs in Biomedical Image Analysis and Integrating Medical Imaging and Non-Imaging Modalities: Second International Workshop, GRAIL 2018 and First International Workshop, Beyond MIC 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 2 (pp. 3-13). Springer International Publishing.

Asadi, Roya, and Sameem Abdul Kareem. 2014. “Review of Feed Forward Neural Network Classification Preprocessing Techniques.” AIP Conference Proceedings 1602: 567–73.

Babayan, A., Erbey, M., Kumral, D., Reinelt, R. D., Reiter, A. M. F., Röbbig, J., Schaare, L. H., Uhlig, M., … Gaebler, M., Villringer, A. (2019). A Mind-Brain-Body dataset of MRI, EEG, cognition, emotion, and peripheral physiology in young and old adults. Scientific Data, 6, 180308.

Basser, P. J., Mattiello, J., & LeBihan, D. (1994). MR diffusion tensor spectroscopy and imaging. Biophysical Journal, 66(1), 259-267.

Ben-Hur, A., & Weston, J. (2010). A user’s guide to support vector machines. In Data mining techniques for the life sciences (pp. 223-239). Humana Press.

Bigler, E. D. (2017). Structural neuroimaging in neuropsychology: History and contemporary applications. Neuropsychology, 31(8), 934-953.

Billmeyer, R., & Parhi, K. K. (2021, October). Biological gender classification from fmri via hyperdimensional computing. In 2021 55th Asilomar Conference on Signals, Systems, and Computers (pp. 578-582). IEEE.

Bullmore, E., & Sporns, O. (2009). Complex brain networks: graph theoretical analysis of structural and functional systems. Nature Reviews Neuroscience, 10(3), 186-198.

Cai, J., Zhou, G., Dong, M., Hu, X., Liu, G., & Ni, W. (2021). Real-time arrhythmia classification algorithm using time-domain ECG feature based on FFNN and CNN. Mathematical problems in engineering, 2021, 1-17.

Çinar, Murat, Mehmet Engin, Erkan Zeki Engin, and Y Ziya Ate\csçi. 2009. "Early Prostate Cancer Diagnosis by Using Artificial Neural Networks and Support Vector Machines." Expert Systems with Applications 36(3): 6357–61.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297.

Dehaene, S., Cohen, L., Morais, J., & Kolinsky, R. (2015). Illiterate to literate: behavioral and cerebral changes induced by reading acquisition. *Nature Reviews Neuroscience*, 16(4), 234-244.

Deng, Chao, Jun Wu, and Xinyu Shao. 2008. "Reliability Assessment of Machining Accuracy on Support Vector Machine." In Intelligent Robotics and Applications: First International Conference, ICIRA 2008 Wuhan, China, October 15-17, 2008 Proceedings, Part II 1, , 669–78.

Enrico, Guido, Guido Enrico, and Guido Enrico. 1994. "Why Network Size Is so Important." Ieee Potentials: 27–31.

Eslami, T., Mirjalili, V., Fong, A., Laird, A. R., & Saeed, F. (2019). ASD-DiagNet: a hybrid learning approach for detection of autism spectrum disorder using fMRI data. Frontiers in neuroinformatics, 13, 70.

Fan L, Su J, Qin J, Hu D and Shen H (2020) A Deep Network Model on Dynamic Functional Connectivity With Applications to Gender Classification and Intelligence Prediction. Front. Neurosci. 14:881. doi: 10.3389/fnins.2020.00881.

Fayaz, S. A., Zaman, M., & Butt, M. A. (2022). Knowledge discovery in geographical sciences—A systematic survey of various machine learning algorithms for rainfall prediction. In International Conference on Innovative Computing and Communications: Proceedings of ICICC 2021, Volume 2 (pp. 593-608). Springer Singapore.

Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M. M., Papademetris, X., & Constable, R. T. (2015). Functional connectome fingerprinting: Identifying individuals using patterns of brain connectivity. Nature Neuroscience, 18(11), 1664–1671.

Formisano, E., De Martino, F., & Valente, G. (2008). Multivariate analysis of fMRI time series: classification and regression of brain responses using machine learning. Magnetic resonance imaging, 26(7), 921-934.

Freeman, L.C. (1979) 'Centrality in social networks conceptual clarification', Social Networks, 1(3), pp. 215-239.

GeeksforGeeks. (2023, February 2). Introduction to Support Vector Machines (SVM). https://www.geeksforgeeks.org/introduction-to-support-vector-machines-svm.

Giedd, J. N., Raznahan, A., Alexander-Bloch, A., Schmitt, E., Gogtay, N., & Rapoport, J. L. (2015). Child psychiatry branch of the National Institute of Mental Health longitudinal structural magnetic resonance imaging study of human brain development. *Neuropsychopharmacology*, 36(1), 252-254.

Gogtay, N., Giedd, J. N., Lusk, L., Hayashi, K. M., Greenstein, D., Vaituzis, A. C., ... & Thompson, P. M. (2004). Dynamic mapping of human cortical development during childhood through early adulthood. Proceedings of the National Academy of Sciences, 101(21), 8174-8179.

Goldberg, A. V., & Harrelson, C. (2005, January). Computing the shortest path: A search meets graph theory. In SODA (Vol. 5, pp. 156-165).

Guo, M., Ren, Y., Yu, H., Yang, H., Cao, C., Li, Y., & Fan, G. (2020). Alterations in Degree Centrality and Functional Connectivity in Parkinson's Disease Patients With Freezing of Gait: A Resting-State Functional Magnetic Resonance Imaging Study. Frontiers in neuroscience, 14, 582079. https://doi.org/10.3389/fnins.2020.582079.

Hasasneh, Ahmad. 2014. "Robot Semantic Place Recognition Based on Deep Belief Networks and a Direct Use of Tiny Images Par Robot Semantic Place Recognition Based on Deep Belief Networks and a Direct Use of Tiny Images." (May).

Hsueh, Yao-Wen, and Chan-Yun Yang. 2009. "Tool Breakage Diagnosis in Face Milling by Support Vector Machine." Journal of materials processing technology 209(1): 145–52.

Huettel, S. A., Song, A. W., & McCarthy, G. (2004). Functional Magnetic Resonance Imaging. Sinauer Associates.

Human Connectome Project. (2017, March 1). 1200 Subjects Data Release. Human Connectome. Retrieved January 22, 2024, from https://humanconnectome.org/study/hcp-young-adult/document/1200-subjects-data-release.

Hurtik, P., Tomasiello, S., Hula, J., & Hynar, D. (2022). Binary cross-entropy with dynamical clipping. Neural Computing and Applications, 34(14), 12029-12041.

Iraji, A., Calhoun, V. D., Wiseman, N. M., Davoodi-Bojd, E., Avanaki, M. R., Haacke, E. M., & Kou, Z. (2016). The connectivity domain: Analyzing resting state fMRI data using feature-based data-driven and model-based methods. Neuroimage, 134, 494-507.

Islam, Mohaiminul, Guorong Chen, and Shangzhu Jin. 2019. "An Overview of Neural Network." American Journal of Neural Networks and Applications 5(1): 7.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202.

Jones, O. D., Wagner, A. D., Faigman, D. L., & Raichle, M. E. (2013). Neuroscientists in court. Nature Reviews Neuroscience, 14(10), 730-736.

Kaushik, P., Gupta, A., Roy, P.P. and Dogra, D.P., 2018. EEG-based age and gender prediction using deep BLSTM-LSTM network model. IEEE Sensors Journal, 19(7), pp.2634-2641.

Khazaee, A., Ebrahimzadeh, A., & Babajani-Feremi, A. (2014, November). Automatic classification of Alzheimer's disease with resting-state fMRI and graph theory. In 2014 21th Iranian Conference on Biomedical Engineering (ICBME) (pp. 252-257). IEEE.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980. Retrieved from https://arxiv.org/abs/1412.6980.

Kobayashi, Y. (2016). A corrector for the sample Mahalanobis distance free from estimating the population eigenvalues of covariance matrix. In Neural Information Processing: 23rd International Conference, ICONIP 2016, Kyoto, Japan, October 16–21, 2016, Proceedings, Part II 23 (pp. 224-232). Springer International Publishing.

Lauterbur, P. C. (1973). Image formation by induced local interactions: examples employing nuclear magnetic resonance. Nature, 242(5394), 190-191.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. nature, 521(7553), 436-444.

Li, Zewen et al. 2022. “A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects.” IEEE Transactions on Neural Networks and Learning Systems 33(12): 6999–7019.

Lorenzini, L., Ingala, S., Collij, L. E., Wottschel, V., Haller, S., Blennow, K., ... & Wink, A. M. (2023). Eigenvector Centrality Dynamics From Resting-State fMRI: Gender and Age Differences in Healthy Subjects.

Mansfield, P., & Morris, P. (1982). NMR imaging in biomedicine: introduction. Journal of Magnetic Resonance, 48(2), 285-293.

Mendes, S. L., Pinaya, W. H. L., Pan, P., & Sato, J. R. (2021). Estimating Gender and Age from Brain Structural MRI of Children and Adolescents: A 3D Convolutional Neural Network Multitask Learning Model. Computational intelligence and neuroscience, 2021, 5550914. https://doi.org/10.1155/2021/5550914.

Meszlényi, R., Peska, L., Gál, V., Vidnyánszky, Z., & Buza, K. (2016, August). Classification of fMRI data using dynamic time warping based functional connectivity analysis. In 2016 24th European signal processing conference (EUSIPCO) (pp. 245-249). IEEE.

MURAT H. SAZLI. 2006. "A Brief Review of Feed-Forward Neural Networks." : 11–17. https://doi.org/10.1501/commua1-2_0000000026.

O'Neil, C. (2016). Weapons of math destruction: How big data increases inequality and threatens democracy. Crown.

Ogawa, S., Lee, T. M., Kay, A. R., & Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. Proceedings of the National Academy of Sciences, 87(24), 9868-9872.

Onal Ertugrul, I., Ozay, M., & Yarman Vural, F. T. (2020). Gender classification using mesh networks on multiresolution multitask fMRI data. Brain imaging and behavior, 14, 460-476.

Ozanich, Emma, Peter Gerstoft, and Haiqiang Niu. 2020. "A Feedforward Neural Network for Direction-of-Arrival Estimation." The Journal of the Acoustical Society of America 147(3): 2035–48.

Pan, D., Zheng, H., Xu, F., Ouyang, Y., Jia, Z., Wang, C., & Zeng, H. (2023). MSFR-GCN: A Multi-scale Feature Reconstruction Graph Convolutional Network for EEG Emotion and Cognition Recognition. IEEE Transactions on Neural Systems and Rehabilitation Engineering.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Pilli, R., Goel, T., Murugan, R., & Tanveer, M. (2023). Association of white matter volume with brain age classification using deep learning network and region wise analysis. Engineering Applications of Artificial Intelligence, 125, 106596.

Plompen, A. J., Cabellos, O., De Saint Jean, C., Fleming, M., Algora, A., Angelone, M., ... & Žerovnik, G. (2020). The joint evaluated fission and fusion nuclear data library, JEFF-3.3. The European Physical Journal A, 56, 1-108.

Poldrack, R. A. (2011). Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. Neuron, 72(5), 692-697.

Raison, A., Bourdon, P., Habas, C., & Helbert, D. (2021, October). Explicability in resting-state fMRI for gender classification. In 2021 Sixth International Conference on Advances in Biomedical Engineering (ICABME) (pp. 5-8). IEEE.

Raz, N., & Rodrigue, K. M. (2006). Differential aging of the brain: patterns, cognitive correlates and modifiers. Neuroscience & Biobehavioral Reviews, 30(6), 730-748.

Retico, A., Giuliano, A., Tancredi, R., Cosenza, A., Apicella, F., Narzisi, A., ... & Calderoni, S. (2016). The effect of gender on the neuroanatomy of children with autism spectrum disorders: a support vector machine case-control study. Molecular autism, 7, 1-20.

ROBINSON, B. (2023). Classification of chronic pain using fMRI data: Unveiling brain activity patterns for diagnosis. Turkish Journal of Electrical Engineering and Computer Sciences, 31(6), 1061-1078.

Rosa, M. J., Portugal, L., Hahn, T., Fallgatter, A. J., Garrido, M. I., Shawe-Taylor, J., & Mourao-Miranda, J. (2015). Sparse network-based models for patient classification using fMRI. Neuroimage, 105, 493-506.

Ruby, U., & Yendapalli, V. (2020). Binary cross entropy with deep learning technique for image classification. Int. J. Adv. Trends Comput. Sci. Eng, 9(10).

Saha, P. (2023). Eigenvector Centrality Characterization on fMRI Data: Gender and Node Differences in Normal and ASD Subjects. Journal of Autism and Developmental Disorders, 1-12.

Scholkopf, B., & Smola, A. (2001). Learning with Kernels: support vector machines, regularization, optimization, and beyond. Adaptive computation and machine learning series.

Sejnowski, T. J., Churchland, P. S., & Movshon, J. A. (2014). Putting big data to good use in neuroscience. Nature Neuroscience, 17(11), 1440-1441.

Sen, B. and Parhi, K.K., 2020. Predicting biological gender and intelligence from fMRI via dynamic functional connectivity. IEEE Transactions on Biomedical Engineering, 68(3), pp.815-825.

Şen, B., Peker, M., Çavuşoğlu, A., & Çelebi, F. V. (2014). A comparative study on classification of sleep stage based on EEG signals using feature selection and classification algorithms. Journal of medical systems, 38, 1-21.

Shafiullah, M., & Abido, M. A. (2018). S-transform based FFNN approach for distribution grids fault detection and classification. IEEE Access, 6, 8080-8088.

Singh, N. M., Harrod, J. B., Subramanian, S., Robinson, M., Chang, K., Cetin-Karayumak, S., ... & Gollub, R. L. (2022). How Machine Learning is Powering Neuroimaging to Improve Brain Health. Neuroinformatics, 1-22.

Sola, J., & Sevilla, J. (1997). Importance of input data normalization for the application of neural networks to complex industrial problems. IEEE Transactions on nuclear science, 44(3), 1464-1468.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958.

Sugumaran, V, G R Sabareesh, and K I Ramachandran. 2008. “Fault Diagnostics of Roller Bearing Using Kernel Based Neighborhood Score Multi-Class Support Vector Machine.” Expert Systems with Applications 34(4): 3090–98.

Supekar, K., de Los Angeles, C., Ryali, S., Cao, K., Ma, T., & Menon, V. (2022). Deep learning identifies robust gender differences in functional brain organization and their dissociable links to clinical symptoms in autism. The British Journal of Psychiatry, 220(4), 202-209.

Suthaharan, Shan, and Shan Suthaharan. 2016. “Support Vector Machine.” Machine learning models and algorithms for big data classification: thinking with examples for effective learning: 207–35.

Svozil, Daniel, Vladimír Kvasnička, and Jiří Pospíchal. 1997. “Introduction to Multi-Layer Feed-Forward Neural Networks.” Chemometrics and Intelligent Laboratory Systems 39(1): 43–62.

Taschereau-Dumouchel, V., Cushing, C. A., & Lau, H. (2022). Real-time functional MRI in the treatment of mental health disorders. Annual review of clinical psychology, 18, 125-154.

Ter-Pogossian, M M, Phelps, M E, Brownell, G L, Cox, Jr, J R, Davis, D O, & Evens, R G (1977) Reconstruction tomography in diagnostic radiology and nuclear medicine. United States.

van Beek, E. J. R., Kuhl, C., Anzai, Y., Desmond, P., Ehman, R. L., Gong, Q., Gold, G., Gulani, V., Hall-Craggs, M., Leiner, T., Lim, C. C. T., Pipe, J. G., Reeder, S., Reinhold, C., Smits, M., Sodickson, D. K., Tempany, C., Vargas, H. A., & Wang, M. (2019). Value of MRI in medicine: More than just another test?. Journal of magnetic resonance imaging : JMRI, 49(7), e14–e25.

Van Horn, J. D., & Toga, A. W. (2014). Human neuroimaging as a “Big Data” science. Brain Imaging and Behavior, 8(2), 323-331.

Vapnik, V. (2013). The nature of statistical learning theory. Springer science & business media.Babayan, A., Erbey, M., Kumral, D., Reinelt, R. D., Reiter, A. M. F., Röbbig, J., Schaare, L. H., Uhlig, M., … Gaebler, M., Villringer, A. (2019). A Mind-Brain-Body dataset of MRI, EEG, cognition, emotion, and peripheral physiology in young and old adults. Scientific Data, 6, 180308.

Vapnik, Vladimir. 1999. The Nature of Statistical Learning Theory. Springer science & business media.

Varoquaux, G., Gramfort, A., Pedregosa, F., Michel, V., & Thirion, B. (2011). Multi-subject dictionary learning to segment an atlas of brain spontaneous activity. In Information Processing in Medical Imaging (pp. 562–573). Springer Berlin Heidelberg.

Vergun, S., Deshpande, A. S., Meier, T. B., Song, J., Tudorascu, D. L., Nair, V. A., ... & Prabhakaran, V. (2013). Characterizing functional connectivity differences in aging adults using machine learning on resting state fMRI data. Frontiers in computational neuroscience, 7, 38.

Vieira, S., Pinaya, W. H., & Mechelli, A. (2017). Using deep learning to investigate the neuroimaging correlates of psychiatric and neurological disorders: Methods and applications. Neuroscience & Biobehavioral Reviews, 74, 58-75.

Weis, S., Patil, K. R., Hoffstaedter, F., Nostro, A., Yeo, B. T., & Eickhoff, S. B. (2020). Sex classification by resting state brain connectivity. Cerebral cortex, 30(2), 824-835.

Zhang, C., Dougherty, C. C., Baum, S. A., White, T., & Michael, A. M. (2018). Functional connectivity predicts gender: Evidence for gender differences in resting brain connectivity. Human brain mapping, 39(4), 1765-1776.

Zhang, Dan, and Wentao Sui. 2011. “The Application of AR Model and SVM in Rolling Bearings Condition Monitoring.” In International Conference on Computer Science and Information Engineering, , 326–31.

Zhang, J., & Luo, Y. (2017, March). Degree centrality, betweenness centrality, and closeness centrality in social network. In 2017 2nd international conference on modelling, simulation and applied mathematics (MSAM2017) (pp. 300-303). Atlantis press.

Zhang, Lingxuan, Zhenyuan Jia, Fuji Wang, and Wei Liu. 2010. “A Hybrid Model Using Supporting Vector Machine and Multi-Objective Genetic Algorithm for Processing Parameters Optimization in Micro-EDM.” The International Journal of Advanced Manufacturing Technology 51: 575–86.

Zhao, G., Hwang, G., Cook, C. J., Liu, F., Meyerand, M. E., & Birn, R. M. (2020). Deep Learning and Bayesian Deep Learning Based Gender Prediction in Multi-Scale Brain Functional Connectivity. arXiv preprint arXiv:2005.08431.

Zuo, X. N., Ehmke, R., Mennes, M., Imperati, D., Castellanos, F. X., Sporns, O., & Milham, M. P. (2012). Network centrality in the human functional connectome. Cerebral cortex, 22(8), 1862-1875.

الملخص

تتناول هذه الدراسة استكشافًا شاملاً لتحليل البيانات متعددة الوسائط من صور الرنين المغناطيسي الوظيفي (fMRI) بهدف تحسين التنبؤ بالسمات الديموغرافية مثل الجنس والعمر. تم الاعتماد على مجموعتي بيانات وهما LEMON وHCP، اللتين تحتويان على بيانات تصوير دماغي مختلفة لتوفير معطيات دقيقة وشاملة .تم تنفيذ عدة تقنيات لاستخراج الميزات، شملت تحليل الاتصال الوظيفي بين مناطق الدماغ(ROI) ، بالإضافة إلى استخدام مقاييس نظرية الرسم البياني مثل المركزية والتكتل. ومن أجل تحسين الأداء، تم إدخال تقنيات تعزيز البيانات لتوسيع مجموعة العينات، مع استخدام تقنيات تقليل الأبعاد مثل PCA لتحسين كفاءة النماذج.

أظهرت النتائج أن ميزات الاتصال الوظيفي كانت الأساس الأكثر أهمية لتحقيق دقة عالية في التصنيف. حيث حقق نموذج SVM نسبة دقة تصل إلى 78.5% لتصنيف الجنس باستخدام بيانات LEMON بعد تحسين التوازن بين العينات، بينما ارتفعت النسبة إلى 96% باستخدام بيانات .HCP أما بالنسبة لتصنيف العمر، فقد سجلت النماذج دقة أعلى بسبب الفروقات البيولوجية بين الفئات العمرية.

توصي الدراسة بدمج مجموعات بيانات مختلفة لزيادة حجم العينات، واستخدام تقنيات أخرى لتحسين جودة البيانات وتوازنها. كما تقترح استكشاف دمج المزيد من المعطيات مثل EEG و ECGمع صور الرنين المغناطيسي لتحقيق أداء أفضل في التطبيقات المستقبل