

UNIVERSITY OF BAHRAIN

College of Engineering



**Department of Electrical &
Electronics Engineering**

Enhanced Performance of the Early Detection of Attention Deficit Hyperactivity Disorder Using Artificial Intelligence

A Thesis Submitted in Partial Fulfilment of the Requirements for the
M.Sc. Degree in Artificial Intelligence Systems

Submitted by
Mohamad Zafar Iqbal Badar
University ID: 20180910

Supervised by
Dr. Zuhair Bahri
Associate Professor
University of Bahrain

Kingdom of Bahrain
November, 2025

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

Thesis Discussion Committee Decision

UNIVERSITY OF BAHRAIN
Deanship of Graduate Studies & Scientific Research



جامعة البحرين
عمادة الدراسات العليا والبحث العلمي

Thesis Discussion Committee Decision

Form Code: UDNGS013

Student Name:	Mohamad Zafar Iqbal Badar
Student ID. No.	20180910
Department:	Electrical & Electronics
College:	College of Engineering
Program:	M.Sc. in Artificial Intelligence Systems
Student GPA:	3.95 / 4.00
Thesis Title:	Enhanced Performance of the Early Detection of Attention Deficit Hyperactivity Disorder Using Artificial Intelligence

The Discussion Committee met on Nov 9, 2025 in S40-093 upon discussing the student whose details are stated above in public. The discussion committee recommends the following:

The student is awarded the grade Pass in the thesis course EEM699.

☐ Without making any changes to the thesis.

☒ After making the modifications attached with this decision.

Discussion Committee Members Approval:

1.	Supervisor Name:	Zouhir Bahri	Sign.:	
2.	Internal Examiner:	Dr. Ayman Al-Khazraji	Sign.:	
3.	External Examiner:	Prof. Gheith Abandah	Sign.:	

Abstract

Attention Deficit Hyperactivity Disorder (ADHD) is a highly prevalent neurodevelopmental condition characterized by persistent patterns of inattention, hyperactivity, and impulsivity. Despite its prevalence, ADHD remains largely underestimated in society and is often confused for naughtiness in children's attitude. This is aggravated by the fact that ADHD diagnosis is still largely dependent on subjective clinical assessments, which often result in delayed diagnosis, misdiagnosis, and untreated symptoms. This study investigates the potential use of various Artificial Intelligence (AI) techniques including traditional machine learning (ML), deep learning (DL), and ensemble methods to enhance the early detection of ADHD, especially in children. This is accomplished through a comparative evaluation of different data modalities such as handwriting, drawing trajectories, social media posts, electroencephalography (EEG) signals, and neuroimaging.

The experimental framework systematically evaluated each modality with rigorous preprocessing, hyperparameter optimization, and cross-validation. On handwriting data, support vector machine (SVM), XGBoost, and a stacking ensemble achieved equivalent top performance (accuracy $\approx 90\%$, F1-score = 0.9, recall = 1.0), with SVM offering the optimal balance of accuracy and computational efficiency. On drawing-based classification, K-Nearest Neighbors (KNN) delivered highest discriminative performance, with optimal results on all metrics (accuracy = 100%, F1-score = 1.0), followed by SVM at 95.45%, suggesting high potential for behavioral-task-based screening. For textual data, transformer-based language models significantly outperformed statistical and sequential baselines. MentalRoBERTa achieved 87.22% accuracy (F1-score = 0.873), with ensemble strategies providing slight but consistent improvements. EEG-based classification favored ensemble methods integrating KNN, SVM, and Random Forest, achieving 93.26% accuracy and near-perfect recall (0.97), making them suitable for confirmatory screening in clinical settings. Neuroimaging classification using transfer learning demonstrated that ResNet50V2 achieved 83% accuracy, highlighting the limited effectiveness of MRI-based ADHD detection given the challenges in MRI data collection.

The results collectively support a staged diagnostic pathway that begins with handwriting and drawing analysis for low-cost, non-invasive screening, incorporates digital text analysis for longitudinal monitoring, utilizes EEG as an objective confirmatory modality, and lastly reserves neuroimaging for complex mental disorder cases. Emphasis is placed on models with high recall to minimize false negatives and support early intervention. The study also highlights interpretability, computational efficiency, ethics, and future research in multi-modal fusion and privacy-preserving learning. Hopefully, this work will contribute to the growing body of research on AI-assisted mental health diagnostics and brings more awareness to society about this widespread yet overlooked problem.

Publications and Participations

A research paper has been submitted to the 9th Institution of Engineering and Technology (IET) Smart Cities Symposium and is currently under review. Additionally, several other manuscripts are in preparation and will be submitted to reputable conferences and journals in the near future.

- *“Mental Health Disorder Detection Using Deep Learning and Natural Language Processing” – Accepted for presentation and publication at the 9th IET Smart Cities Symposium.*

Dedication

To my parents and my wife,

Your unwavering support, countless sacrifices, and heartfelt prayers have been the foundation of my journey. Your values, love, and constant encouragement have shaped who I am today. Thank you for always believing in me and standing by my side.

Acknowledgments

I would like to express my deepest gratitude to my supervisor, Dr. Zuhair Bahri, for his continuous support, insightful guidance, and continuous encouragement throughout this research. His dedication and mentorship have been instrumental in the successful completion of this thesis, and I am truly honored to have been one of his top priorities.

I extend my heartfelt thanks to my parents, whose love, prayers, and constant support have been the backbone of my journey from the very beginning. I am also profoundly grateful to my wife, whose patience, understanding, and steadfast support gave me strength and motivation during the most demanding phases of this work.

My sincere appreciation also goes to my family members, relatives, and friends for their encouragement and care throughout this endeavor.

I am thankful to my colleagues and academic peers for their collaboration, valuable insights, and thoughtful discussions, which enriched my research experience. Special thanks to my friends, who stood by me with both intellectual and emotional support during the most challenging moments.

Lastly, I would like to thank Electrical & Electronics Engineering department, college of Engineering, and University of Bahrain for providing a supportive academic environment and the resources that made this research possible.

This work is only the beginning of my journey.

Table of Contents

Thesis Discussion Committee Decision	A
Abstract	B
Publications and Participations.....	C
Dedication.....	D
Acknowledgments	E
Table of Contents.....	F
List of Tables	I
List of Figures	J
Abbreviations	M
Chapter 1. Introduction	1
1.1 Motivation	3
1.2 Objectives and Contributions.....	5
1.3 Traditional ADHD Diagnostic Methods.....	7
1.4 Review of Available Literature	11
1.4.1 AI and Machine Learning in Mental Health Detection.....	11
1.4.2 AI Models for ADHD Detection Using Behavioral and Cognitive Data.....	14
1.4.3 Neuroimaging-Based ADHD Detection.....	15
Chapter 2. Background Material	17
2.1 Brain Structure.....	17
2.2 How ADHD Affects Productivity and Organization	18
2.3 Functional Brain Imaging Modalities.....	19
2.4 Machine Learning Models	21
2.4.1 Traditional Machine Learning Models.....	22
2.4.1.1 Logistic Regression	23
2.4.1.2 Support Vector Machine.....	24
2.4.1.3 Decision Tree.....	26
2.4.1.4 Random Forest.....	27
2.4.1.5 K-Nearest Neighbors (KNN).....	27
2.4.1.6 XGBoost.....	28
2.4.1.7 Multinomial Naïve Bayes.....	29
2.4.1.8 Linear Discriminant Analysis (LDA).....	30
2.4.1.9 Adaptive Boosting (AdaBoost)	30

2.4.1.10 LogitBoost	30
2.4.2 Parameters vs Hyperparameters	31
2.4.3 Machine Learning Pipeline	31
2.4.4 Generalization	32
2.4.5 Overfitting vs Underfitting	33
2.4.6 Performance Measures	34
2.5 Deep Learning Models	35
2.5.1 Convolutional Neural Networks (CNNs) for Image-Based Analysis	36
2.5.2 Recurrent Neural Networks (RNNs) for NLP-Based Classification	40
2.5.3 Hybrid Deep Learning Models	41
2.5.4 Transfer Learning	41
2.6 Ensemble Learning	43
2.7 Natural Language Processing	45
2.7.1 Data Cleaning	46
2.7.2 Tokenization	46
2.7.3 Word Embeddings	47
2.8 Generative Adversarial Network	47
2.9 Optical Character Recognition	49
Chapter 3. Methodology	50
3.1 Data Collection	50
3.1.1 Handwriting Dataset	50
3.1.2 Drawing Dataset	51
3.1.3 Reddit based Mental Health Dataset	52
3.1.4 EEG Dataset	54
3.1.5 Neuroimaging Dataset	55
3.2 Exploratory Data Analysis	56
3.2.1 Handwriting Dataset	56
3.2.2 Drawing Dataset	61
3.2.3 Reddit based Mental Health Dataset	62
3.2.4 EEG Dataset	70
3.2.5 Neuroimaging Dataset	72
3.3 Pre-Processing	76
3.3.1 Handwriting Dataset	76
3.3.2 Drawing Dataset	79
3.3.3 Reddit based Mental Health Dataset	80
3.3.4 EEG Dataset	82
3.3.5 Neuroimaging Dataset	87
3.4 Model Development and Explanation	88
3.4.1 Handwriting Dataset	88
3.4.2 Drawing Dataset	90
3.4.3 Reddit based Mental Health Dataset	92
3.4.4 EEG Dataset	94

3.4.5 Neuroimaging Dataset.....	96
3.5 Experimental Setup	98
Chapter 4. Experimental Results and Performance Evaluation	100
4.1 Handwriting Dataset Results	100
4.1.1 Traditional Machine Learning Models.....	100
4.1.2 Gradient Boosting (XGBoost).....	101
4.1.3 Stacking Ensemble	102
4.1.4 Comparative Summary.....	103
4.1.5 Optical Character Recognition and Graphology Analysis	105
4.2 Drawing Dataset Results	106
4.2.1 Traditional Machine Learning Models.....	107
4.3 Reddit-Based Mental Health Dataset Results	109
4.3.1 Baseline Statistical Model.....	109
4.3.2 Transformer Models.....	111
4.3.3 Sequence Models.....	114
4.3.4 Convolutional Models	115
4.3.5 Ensemble Models	116
4.3.6 Comparative Summary.....	117
4.4 EEG Dataset Results	118
4.4.1 Traditional Machine Learning Models.....	119
4.4.2 Ensemble Models	121
4.4.3 Deep Learning Architectures	122
4.4.4 Comparative Summary.....	123
4.5 Neuroimaging Dataset Results	124
4.5.1 Deep Learning Architectures	125
4.5.2 Comparative Summary.....	128
4.6 Recommendations.....	129
Chapter 5. Summary and Conclusion.....	133
5.1 Summary	133
5.2 Future Points of Expansion	135
Reference	138

List of Tables

Table 2.1 Difference Between Traditional and Transfer learning.....	42
Table 2.2 Comparison of GAN and WGAN Network	48
Table 4.1 Performance of all ML Models on Handwriting Dataset	103
Table 4.2 Optimized Hyperparameters for ML Models on Drawing Dataset	107
Table 4.3 Performance of all ML Models on Drawing Dataset	108
Table 4.4 Performance of all Models on Mental Health Dataset	118
Table 4.5 Optimized Hyperparameters for ML Models on EEG Dataset	119
Table 4.6 Performance of all Models on EEG Dataset.....	124
Table 4.7 Performance of DL Models on Neuroimaging Dataset.....	129

List of Figures

Figure 2.1 Difference between a Healthy Brain and a Brain with ADHD	18
Figure 2.2 Comparison of Model Fitting Behavior in Machine Learning.....	34
Figure 2.3 Performance Metrics in Machine Learning.....	35
Figure 2.4 Functions curves of ReLU	38
Figure 2.5 Functions curves of Leaky ReLU	39
Figure 2.6 Traditional Learning Vs. Transfer Learning	41
Figure 2.7 Wasserstein GAN Network Architecture with a Generator (G) and a Critic (C).....	48
Figure 3.1 Handwriting samples of a girl without ADHD and with ADHD. Adopted from [71]	51
Figure 3.2 Handwritten drawing data collection procedure. Adopted from [72]	52
Figure 3.3 Electrode placement for 10-20 system.....	55
Figure 3.4 Label Distribution in Handwriting Dataset	57
Figure 3.5 Feature Correlation Heatmap of Handwriting Dataset.....	58
Figure 3.6 Feature Importance Analysis of Handwriting Dataset.....	59
Figure 3.7 Sub-space using the first two Principal Components in the Handwriting Dataset	60
Figure 3.8 Label Distribution in Drawing Dataset	62
Figure 3.9 Class Distribution of Reddit Based Mental Health Dataset	63
Figure 3.10 Samples of Reddit Based Mental Health Dataset	64
Figure 3.11 Word Cloud of Reddit Based Mental Health Dataset	65
Figure 3.12 Sentiment Analysis of Reddit Based Mental Health Dataset.....	67
Figure 3.13 Pronoun Usage and Emotional Language Analysis of Reddit Based Mental Health Dataset	69
Figure 3.14 Raw EEG Signals for Person ID 000001 (ADHD group).....	71
Figure 3.15 MRI Samples of participants from Typically Developing Group	73
Figure 3.16 MRI Samples of participants from ADHD Group	73
Figure 3.17 Five evenly spaced Axial, sagittal, and coronal mid-plane slices from Participant of Typically Developing group in ADHD200 Preprocessed Dataset	74

Figure 3.18 Five evenly spaced Axial, sagittal, and coronal mid-plane slices from Participant of ADHD group in ADHD200 Preprocessed Dataset	74
Figure 3.19 MRI Samples from Curated Dataset of Sagittal View Slices	75
Figure 3.20 Top Features by Chi-Square Score in Handwriting Dataset	78
Figure 3.21 Cumulative Explained Variance of Principal Components in Handwriting Dataset	79
Figure 3.22 Comparison of EEG Signals for ADHD Sample before and after preprocessing	83
Figure 3.23 Comparison of PSD Signals for ADHD Sample before and after preprocessing	84
Figure 3.24 Top Features by ANOVA in EEG Dataset	86
Figure 3.25 End-to-End ML Workflow for ADHD Detection using a Handwriting-Based Dataset	90
Figure 3.26 End-to-End ML Workflow for ADHD Detection using a Drawing-Based Dataset	91
Figure 3.27 End-to-End ML Workflow for ADHD Detection using a Reddit based Mental Health Dataset	93
Figure 3.28 End-to-End ML Workflow for ADHD Detection using an EEG Dataset	96
Figure 3.29 End-to-End ML Workflow for ADHD Detection using Neuroimaging Dataset	98
Figure 4.1 Confusion Matrices of top-performing models (SVM and Random Forest) on Handwriting Dataset.....	101
Figure 4.2 Confusion Matrix of XGBoost on Handwriting Dataset.....	102
Figure 4.3 Model Performance of all Models on Handwriting Dataset	104
Figure 4.4 Comparison of ROC curves of all Models on Handwriting Dataset	105
Figure 4.5 OCR based Graphology Analysis Visualization	106
Figure 4.6 Confusion Matrix and ROC Curves of the KNN Models	108
Figure 4.7 Confusion Matrix and ROC Curves of the SVM Models	109
Figure 4.8 Confusion Matrix of MNB Classifier on Mental Health Dataset	110
Figure 4.9 ROC Curve of MNB Classifier on Mental Health Dataset	110
Figure 4.10 Confusion Matrix of BERT on Mental Health Dataset.....	111
Figure 4.11 Confusion Matrix of Mental BERT on Mental Health Dataset .	112
Figure 4.12 Confusion Matrix of RoBERTa on Mental Health Dataset	113

Figure 4.13 Confusion Matrix of Mental RoBERTa on Mental Health Dataset	114
Figure 4.14 Confusion Matrix of BiLSTM-Attention Model on Reddit Dataset	115
Figure 4.15 Confusion Matrix of CNN Model on Reddit Dataset	116
Figure 4.16 ROC-AUC curves of all traditional ML models on EEG Dataset	120
Figure 4.17 Confusion matrices of top-performing models (SVM, KNN, and Random Forest) on EEG Dataset.....	121
Figure 4.18 confusion matrices and ROC-AUC curves of soft-voting ensemble	122
Figure 4.19 Confusion matrices of CNN and CNN-LSTM models on EEG Dataset	123
Figure 4.20 Confusion Matrix of DenseNet121 Model on Neuroimaging Dataset	126
Figure 4.21 Predictions made by DenseNet121 Model on Neuroimaging Dataset	126
Figure 4.22 Confusion Matrix of ResNet50 Model on Neuroimaging Dataset	127
Figure 4.23 Sample Predictions made by ResNet50 Model on Neuroimaging Dataset	128

Abbreviations

AdaBoost	Adaptive Boosting
ADHD	Attention-Deficit/Hyperactivity Disorder
AI	Artificial Intelligence
ANN	Artificial Neural Network
ANOVA	Analysis of Variance
ASD	Autism Spectrum Disorder
ASR	Artifact Subspace Reconstruction
ASRS-V1.1	Adult ADHD Self-Report Scale Version 1.1
AUC	Area Under the Curve
BernoulliNB	Bernoulli Naive Bayes
BERT	Bidirectional Encoder Representations from Transformers
Bi-GRU	Bidirectional Gate Recurrent Units
Bi-LSTM	Bidirectional Long Short-Term Memory Network
CatBoost	Categorical Boosting
CE	Conformité Européenne
CNN	Convolutional Neural Network
CPTs	Continuous Performance Tests
DSM	Diagnostic and Statistical Manual of Mental Illnesses
DSM-5	Diagnostic and Statistical Manual of Mental Illnesses (Fifth Edition)
DSM-IV	Diagnostic and Statistical Manual of Mental Illnesses (Forth Edition)
DT	Decision Tree
EEG	Electroencephalography
FDA	Food and Drug Administration
fMRI	Functional Magnetic Resonance Imaging
FMRIB	Functional Magnetic Resonance Imaging of the Brain
FN	False Negative
FP	False Positive

FSL	Functional Magnetic Resonance Imaging of the Brain (FMRIB) Software Library
GAN	Generative Adversarial Network
GBM	Gradient Boosting Machines
GM	Gray Matter
GPT	Generative Pre-training Transformer
Gray Matter	fractional Amplitude of Low-Frequency Fluctuations
GRU	Gated Recurrent Unit
Hz	Hertz
ICA	Independent Component Analysis
KKI	Kennedy Krieger Institute
KNN	K-Nearest Neighbors
LDA	Linear Discriminant Analysis
Leaky ReLU	Leaky Rectified Linear Unit
LightGBM	Light Gradient-Boosting Machine
LIME	Local Interpretable Model-agnostic Explanations
LLM	large language models
LR	logistic regression
LSTM	Long Short-Term Memory
MEG	Magnetoencephalography
ML	Machine Learning
MLP	Multi-Layer Perceptron
MNB	Multinomial Naive Bayes
MNE	Magnetoencephalography and Electroencephalography
MNI	Montreal Neurological Institute
NB	Naive Bayes
NeuroIMAGE	NeuroIMAGE Sample
NIfTI open file format	Neuroimaging Informatics Technology Initiative – open file format
NLP	Natural language Processing
NLTK	Natural Language Toolkit
NYU	New York University Child Study Center

OCR	Optical Character Recognition
OHSU	Oregon Health & Science University
PCA	Principal Component Analysis
PCI	Personal Characteristic Information
PET	Positron Emission Tomography
PSD	power spectral density
PTSD	Post-Traumatic Stress Disorder
ReLU	Rectified Linear Unit
ResNet	Residual Neural Network
RF	Random Forest
RoBERTa	Robustly Optimized BERT Pretraining Approach
ROC	Receiver Operating Characteristic
ROI	Region of Interest
RS-fMRI	Resting state – Functional Magnetic Resonance Imaging
SASI	Self-Assessment Symptom Inventory
SFFS	Sequential Forward Floating Search
SGD	Stochastic Gradient Descent
SHAP	Shapley Additive exPlanations
SMOTE	Synthetic Minority Over-sampling Technique
sMRI	Structural Magnetic Resonance Imaging
SVM	Support Vector Machine
TBR	Theta / Beta Ratio
TD	Typically Developing
TL	Transfer Learning
TN	True Negative
TP	True Positive
VGG-Net	Visual Geometry Group Network
WGAN	Wasserstein Generative Adversarial Network
WHO	World Health Organization
XGBoost	Extreme Gradient Boosting

Chapter 1. Introduction

Attention-Deficit/Hyperactivity Disorder (ADHD) is a neurodevelopmental condition that affects both children and adults, with symptoms typically emerging in early childhood. Its onset and progression are shaped by a combination of genetic, environmental, and developmental influences. These factors influence how the brain develops and functions, often leading to difficulties with attention, impulse control, and behavior regulation. The impact of ADHD can vary widely from person to person and may change over time as individuals grow and adapt. Some individuals develop coping mechanisms that help them manage daily challenges, while others continue to struggle with attention, impulsivity, and organization [1]. For some, symptoms become less disruptive, leading to improved social interactions and adaptation. However, for others, the demands of adulthood can intensify these difficulties, impacting relationships, careers, and overall well-being [2]. This variability highlights the importance of recognizing ADHD as a highly individualized condition, with each person exhibiting a unique combination of strengths and challenges, namely being mistaken for bad manners by their parents.

Considering these complexities, it is crucial to understand how ADHD is categorized within recognized mental health frameworks. It is recognized as a mental health condition according to the Diagnostic and Statistical Manual of Mental Illnesses (Fifth Edition) (DSM-5), which defines mental illness broadly as health issues that lead to significant alterations in emotions, thoughts, or behaviors, causing considerable distress in social, work, or family settings [3]. While ADHD fits this definition of mental illness, it is more commonly referred to as mental disorder, usually aggravating to depressive state at a later age. The terms “mental illness” and “disorder” are frequently used interchangeably.

The core symptoms of ADHD manifest differently in everyone. However, certain patterns frequently appear across cases [3]. The variations in how ADHD symptoms are expressed throughout development indicate that our

current understanding of ADHD in childhood and adolescence may not apply to young adulthood [4]. The symptoms are commonly categorized into three primary types based on predominant symptoms [5]:

- **Predominantly Inattentive Presentation:** It is characterized by difficulty sustaining focus, following through on tasks, and organizing activities.
- **Predominantly Hyperactive-Impulsive Presentation:** It is marked by excessive movement, restlessness, difficulty remaining seated or quiet when appropriate, acting without considering consequences, interrupting others, and making hasty decisions.
- **Combined Presentation:** It involves a mix of both inattentive and hyperactive-impulsive symptoms.

Diagnosing ADHD remains complex due to its recent understanding and overlapping characteristics with other developmental disorders, such as autism spectrum disorder (ASD) [6]. Traditional diagnostic methods rely heavily on behavioral assessments, teacher and parent reports, and clinical evaluations, making them susceptible to subjectivity and inconsistency. These limitations, combined with the variable presentation of ADHD across an individual's lifespan, highlight the need for more accurate and objective diagnostic tools. Some of the symptoms as highlighted in Diagnostic and Statistical Manual of Mental Disorders are as follows [3]:

- **Symptoms in children:** A child with ADHD might,
 - Daydream a lot
 - Forget or misplace items
 - Squirm or fidget excessively
 - Talk excessively
 - Make careless mistakes or engage in unnecessary risks
 - Struggle to resist impulses or temptation
 - Have difficulty taking turns
 - Experience challenges in forming and maintaining positive relationships with others
 - Exhibit impulsivity and impatience

- **Symptoms in adults:** An adult with ADHD might experience
 - Difficulties in managing daily tasks
 - Consistently losing or misplacing things
 - Struggling with organization and neatness
 - Problems staying on top of responsibilities
 - Feeling overwhelmed or frustrated with everyday tasks
 - Frequent job changes
 - Procrastination
 - Struggling with time management
 - Relationship problems due to forgetfulness, hyperfocus, or disorganization
 - Strong emotional reactions to minor problems
 - Extreme agitation and restlessness when instructed to wait or stay still
 - Participating in highly stimulating sports or fast-paced activities such as video games
 - Internal restlessness, difficulty with time management, problems maintaining focus during meetings, disorganization, and impulsive decision-making

The advancements in artificial intelligence (AI) present an opportunity to refine early ADHD detection by introducing data-driven methods that enhance accuracy and reduce bias. These models analyze behavioral patterns, neurophysiological data, as well as cognitive tests that assist with diagnosis in a more standardized manner. The following sections explore the motivations for integrating AI into ADHD detection, the objectives of this study, the challenges associated with AI-based diagnosis, and alternative approaches. Additionally, a review of existing literature will provide insights into how machine learning (ML) algorithms can improve early detection and address limitations in traditional diagnostic practices.

1.1 Motivation

Every year, thousands of kids struggle with undiagnosed ADHD, which is often mistaken for laziness, poor behavior, or poor discipline. As a result of this, parents are frustrated, educators are overwhelmed, and kids sense a

feeling of failure when really, they are going through an unseen neurodevelopmental disorder. Early intervention and identification are very critical, as research shows that children diagnosed and supported at early ages have significantly better long-term outcomes [7]. The current average age of diagnosis remains between five and nine, and many children, and those from minority or lower-income backgrounds, are frequently overlooked [8, 9]. By the time they receive a diagnosis, they may have already experienced academic struggles, low self-esteem, and secondary mental health issues such as anxiety or depression.

A major factor contributing to delayed or missed diagnoses is the subjectivity of current ADHD assessments. Traditionally, diagnosis has been confirmed on behavioral observations, reports by parents and teachers, and clinical evaluation, all of which could introduce some bias. The biases can be introduced as one teacher interprets that a child is inattentive, whereas another would see him/her as normal childhood behavior. Studies indicate that ADHD diagnosis rates can vary by more than 30% between different clinicians, evaluating the same case [10, 11]. Moreover, without comparing ADHD with ASD, many symptoms of ADHD overlap with other conditions such as with anxiety disorders or learning disabilities, further complicating the diagnostic process [12].

Apart from the late or incorrect diagnosis, getting appointments from healthcare centers to access the ADHD assessment from a professional takes a lot of time and remains a significant barrier worldwide. In the meantime, a child's symptoms may worsen, affecting his/her academic performance, emotional well-being, and social interactions. In lower-income and rural areas, these delays can be even longer, leaving many children undiagnosed and without support. The lack of timely assessment not only impacts individuals but also places strain on educational and healthcare systems, as untreated ADHD is associated with higher dropout rates, increased mental health issues, and long-term socioeconomic consequences.

AI offers a powerful means of tackling these challenges. In contrast to conventional diagnostic practices that depend heavily on self-reported information, AI-based assessment systems can evaluate measurable

behavioral indicators, including cognitive functioning, speech characteristics, physical movement, and neurophysiological markers. ML models can detect subtle indicators of ADHD that may go unnoticed in clinical settings, leading to earlier and more accurate identification. Additionally, AI can provide consistent diagnostic criteria across different populations, reducing biases related to gender, race, and socioeconomic status. By integrating AI-based assessments into schools, primary care clinics, and community health programs, early detection can become more accessible, scalable, and cost-effective, ensuring that fewer children fall through the cracks.

Beyond improving diagnostic accuracy, AI-driven ADHD detection has the potential to reduce the economic burden associated with misdiagnosis and delayed intervention. According to analyses by Russell *et al.* and Holden *et al.* [13, 14], the additional annual cost to the healthcare system for each individual with ADHD is estimated at USD 1,430. Similar findings have been reported in recent studies from Germany, which indicate that the total healthcare costs for children and adolescents with ADHD exceed those of their non-ADHD peers by approximately USD 1,742 in the first year following diagnosis [15].

These findings emphasize not only the negative impact of ADHD on individuals, but also the significant cost impact on healthcare systems and society. By switching from a reactive methodology to a proactive AI-enhanced approach, we hope to lower these expenses and improve outcomes for individuals, families, and society at large. So, instead of waiting for symptoms to escalate, AI-powered tools can assist parents, educators, and clinicians to intervene sooner, more consistently, and with greater precision, ultimately transforming ADHD diagnosis from a slow, uncertain process into an efficient, data-driven system that changes lives.

1.2 Objectives and Contributions

This research aims at exploring and developing AI-based solutions that should hopefully enhance the early detection and diagnosis of ADHD through the following objectives and contributions:

1. To highlight remarkable characteristics of ADHD that may be integrated into ML models.

2. To design and implement diverse ML models (including supervised, unsupervised, and deep learning approaches) for ADHD detection, comparing their performance metrics (e.g., accuracy, precision, recall, F1-score) in order to identify optimal algorithmic approaches.
3. To identify the most informative data sources for AI-based ADHD detection, considering both traditional clinical measures and novel digital biomarkers.
4. To generate and utilize synthetic data, where applicable, for robust model testing, addressing data scarcity challenges, and evaluating model generalizability beyond training datasets.
5. To analyze feature importance across different models to identify the most informative markers for ADHD detection, potentially revealing novel biomarkers or behavioral indicators.
6. To investigate the potential of ensemble methods that integrate multiple AI approaches with the goal of achieving better performance than individual models.
7. To assess the feasibility of implementing AI-based screening tools in diverse settings, including schools, primary care clinics, and specialty mental health services.

During the development of the research methodology, several key questions were formulated to shape the direction of this study that aims at providing direct or indirect answers to these research questions, each of which contributes meaningfully to the field of ADHD classification. The study is guided by the following key questions:

Question 1. Can deep learning models outperform traditional machine learning approaches in ADHD detection when applied to multimodal datasets (neuroimaging, behavioral, text-based, and motion data)?

Question 2. What is the role of ensemble learning techniques in enhancing the reliability and accuracy of AI-based ADHD classification?

Question 3. How does the inclusion of synthetic data augmentation impact the performance, generalizability, and fairness of ADHD detection models?

Question 4. Can Natural Language Processing (NLP) models train on ADHD-related textual datasets (e.g., Reddit and Twitter Mental Health

Datasets) towards the effective identification of linguistic markers associated with ADHD symptoms?

1.3 Traditional ADHD Diagnostic Methods

Traditional ADHD diagnosis relies primarily on behavioral assessments, clinical interviews, and standardized questionnaires completed by parents, teachers, and medical professionals. These methods of diagnosis are based on specific criteria that may change many times per year [16]. While these methods are well established, they have significant limitations, including subjectivity, variability in reporting, and cultural biases. Some of these assessment tools that are either used in self-assessment or as part of a clinical diagnosis process include:

1. ASRS-V1.1 (Adult ADHD Self-Report Scale Version 1.1)

The Adult ADHD Self-Report Scale (ASRS-V1.1) was created by the World Health Organization (WHO) in collaboration with experts in ADHD research [17]. It is specifically intended for adults to self-evaluate ADHD symptoms, as they may develop masking or compensatory behaviors over time.

The ASRS is composed of 18 items that are directly aligned with the diagnostic criteria for ADHD as outlined in the DSM-IV (later updated for DSM-5). The first 6 items are considered the most predictive of ADHD and are referred to as the ASRS Screener. Each item asks respondents to rate the frequency of certain behaviors or difficulties they have experienced over the past 6 months.

The assessment evaluates both inattention symptoms (difficulty concentrating, forgetfulness, misplacing items) and hyperactivity/impulsivity symptoms (fidgeting, restlessness, interrupting) on a frequency scale from “Never” to “Very Often.” The scores above established thresholds suggest potential ADHD and indicate the need for a comprehensive clinical evaluation rather than providing a formal diagnosis.

2. DSM-5 ADHD Symptom Checklist

The DSM-5 checklist is based on the DSM-5, which is the standard reference used by mental health professionals in diagnosing psychological conditions [3]. This checklist provides specific criteria that must be met for an effective and hopefully successful ADHD diagnosis. There are separate criteria for children (symptoms must be present before age 12) and adults (where the presentation might be less obvious but still affects daily functioning).

The DSM-5 criteria divide ADHD symptoms into two categories: Inattention (poor attention to detail, organizational difficulties, distractibility) and Hyperactivity-Impulsivity (fidgeting, excessive movement, interrupting). Each category contains nine symptoms, with a diagnosis requiring at least 6 symptoms (children) or 5 symptoms (adults) from either category, which are present in multiple settings or causing significant functional impairment.

The checklist assesses symptom presence and severity through specific behavioral questions, requiring symptoms to persist for at least 6 months and exceed developmentally appropriate levels. Meeting these criteria helps mental health professionals to determine if an individual reaches the diagnostic threshold for ADHD, while also considering potential comorbid conditions like anxiety or depression.

3. Copeland Symptom Checklist (ADHD/ADD)

The Copeland Symptom Checklist was developed by Dr. Edward Copeland, a psychologist specializing in ADHD [18]. It is designed to identify ADHD symptoms in both children and adults. Unlike the ASRS, which is primarily for adult self-report, the Copeland Checklist can be filled out by parents, teachers, or adults themselves [18]. It provides a more behavioral observation-based approach, often used in schools or clinical settings.

The checklist evaluates symptoms across various environments (home, school, social interactions), focusing on attention sustainability and task completion, distractibility, restlessness and fidgeting, and conversational interruption and social intrusion. Observers rate behavior

frequency/severity on a scale from “Never” to “Very Often”. It captures symptom manifestation across different contexts, as ADHD tends to impact multiple areas of a person’s life, not just one (e.g., both school and home for children, or both work and relationships for adults). Higher observed behavior frequencies suggest a potential ADHD/ADD diagnosis, with the checklist serving as a preliminary assessment tool guiding further clinical evaluation.

4. Self-Assessment Symptom Inventory (SASI)

The Self-Assessment Symptom Inventory (SASI) is a broader self-report tool used to assess a variety of psychological symptoms, not just ADHD. It can help screen for a range of mental health issues, including mood disorders, anxiety disorders, and ADHD. SASI can be administered to individuals who may be experiencing symptoms but are not sure if they are related to ADHD, depression, or anxiety, offering an initial point of exploration for several conditions [19].

This comprehensive inventory includes questions about mood regulation (sadness, irritability, hopelessness), cognitive symptoms (focus difficulties, forgetfulness, organizational challenges), and physical symptoms (restlessness, fatigue, sleep disturbances). Individuals rate symptom severity over a specific timeframe (typically two weeks to six months) to assess chronicity and functional impact. SASI results help determine whether symptoms are more likely attributable to ADHD or other conditions like generalized anxiety disorder or depression, supporting clinicians in differential diagnosis.

All of the above tools are screening instruments that have been designed to help identify the presence of ADHD symptoms in children, adolescents, or adults. However, none of these tools alone are sufficient for a conclusive and accurate diagnosis. An accurate ADHD diagnosis generally requires a comprehensive evaluation that includes:

- Clinical interviews.
- Behavioral observations.
- Psychological testing

- Input from multiple sources (e.g., parents, teachers, coworkers)

The ASRS-V1.1 and the Copeland Symptom Checklist focus specifically on ADHD symptoms, while the DSM-5 Checklist provides the official criteria used for diagnosing the disorder. The SASI, on the other hand, takes a broader approach, evaluating a wide range of symptoms that may overlap with other mental health conditions.

One alternative approach is neurophysiological testing, including electroencephalography (EEG) and functional magnetic resonance imaging (fMRI). EEG is a methodology that assists with obtaining signals from the brain related to different states, such as from the scalp surface region, and are ordered as delta, theta, alpha, beta and gamma dependent on signal frequencies ranging between 0.1 Hz to 100 Hz [20]. The children with ADHD exhibit distinct brainwave activity patterns, particularly in theta and beta wave ratios, which could serve as biomarkers for diagnosis [21]. Similarly, fMRI scans have identified structural and functional differences in brain regions associated with attention, impulse control, and executive function. However, despite their potential, these methods are expensive, time-consuming, and not practical for large-scale clinical use. fMRI requires specialized equipment and infrastructure, making it costly and impractical for widespread screening. While EEG is less expensive, it still demands proper equipment and expertise for interpretation, limiting its feasibility for routine ADHD screening.

Another emerging alternative is the use of digital cognitive assessments and gamified behavioral tasks. These involve computerized tests that measure response times, attention spans, and impulsivity levels in real-time, offering a more objective evaluation of ADHD-related behaviors [22]. Digital tools, such as continuous performance tests (CPTs), track an individual's ability to sustain attention and inhibit responses [23]. These tools have shown promise in supplementing traditional diagnostic methods. However, these assessments alone do not provide a definitive diagnosis and are best used as complementary tools rather than standalone solutions.

Despite these advancements, no single method has been completely reliable for diagnosing ADHD, highlighting the need for a broader, more integrated approach [24, 25, 26]. AI offers new possibilities by combining multiple

sources of diagnostic information, such as behavioral patterns, cognitive assessments, physiological markers, and patient history. By analyzing large datasets, AI models can detect subtle indicators of ADHD, helping to reduce inconsistencies and improve early identification.

The applications of AI in ADHD detection extend beyond diagnosis. AI-based screening tools could be used in schools and pediatric clinics, making it easier to spot children who may be at risk before symptoms become more pronounced. In telehealth settings, AI assessments could increase access to ADHD evaluations, especially in areas with limited specialist services.

1.4 Review of Available Literature

Conducting a comprehensive literature review is a crucial step in any research that provides a foundational understanding of current works and debates within a specific field, thus placing the research work in the proper framework. This process provides insights and highlights various perspectives on the topic, presenting valuable information and revealing alternative ideas. In the context of AI-enhanced ADHD detection, numerous studies have explored the application of ML and AI in ADHD detection, leveraging various data sources, including neuroimaging, cognitive performance, speech analysis, and behavioral tracking.

1.4.1 AI and Machine Learning in Mental Health Detection

AI and ML have emerged as powerful tools for analyzing mental health data to enhance diagnostic accuracy and early detection. Ameer *et al.* [27] aimed to address the increased mental health challenges caused by social isolation during the COVID-19 pandemic by using AI to detect mental illnesses through social media data. They focused on analyzing unstructured posts from Reddit to classify five common mental disorders, depression, anxiety, bipolar, ADHD, and PTSD, using traditional ML, DL, and advanced transfer learning models. Their experiments showed that the RoBERTa transfer learning model performed best, achieving accuracy and an F1-score of 83%.

Kim *et al.* [28] developed a deep learning model to detect users' mental states based on their social media posts. They gathered data from mental health-related communities on Reddit and developed a model to categorize posts

according to various mental disorders, such as depression, anxiety, bipolar disorder, borderline personality disorder, schizophrenia, and autism. Their model aims to assist in identifying potential mental health sufferers by analyzing their written content. Additionally, they discuss the implications of their approach as a supplementary tool for monitoring the mental health of frequent social media users. Their results show that the CNN models generally outperformed XGBoost models in accuracy across all subreddits, with the highest accuracy of 96.96% for r/autism. However, class imbalance affected performance for some disorders, leading to lower F1-scores for minority classes. Among the more balanced datasets, the depression subreddit achieved strong performance, with an F1-score of 79.49%. The authors conclude that expanding the dataset could help address imbalance issues and further improve classification results.

Dey and Desai [29] proposed a neural network-based approach for early detection of mental health issues by classifying counseling chat messages. They utilize an LSTM model, with and without GloVe embeddings, to extract sequential patterns from natural language text and categorize mental health conditions. The dataset includes text descriptions labeled with ground truth categories, which undergo NLP preprocessing before training the model. Their approach achieves 84% accuracy before overfitting, demonstrating the feasibility of such a system. They highlight the potential for improved performance with a larger dataset, emphasizing the role of AI in early mental health diagnosis.

Tom *et al.* [30] proposed a unified framework for multi-disorder classification using Reddit posts across various mental health subreddits. Their model combined fine-tuned, domain-specific transformer models with recurrent neural networks such as Long Short-Term Memory (LSTM) and Bidirectional LSTM (Bi-LSTM). This composite approach achieved an F1-score exceeding 88%, highlighting the effectiveness of hybrid architectures in detecting nuanced mental health signals from diverse sources.

Dristy *et al.* [31] focused on detecting mental health conditions using text messages, recognizing the prevalence of digital communication in everyday life. They evaluated five supervised ML classifiers, which are Logistic

Regression (LR), Bernoulli Naive Bayes (BernoulliNB), Random Forest (RF), and Decision Tree (DT) to identify conditions such as depression, anxiety, panic disorder, and Post-Traumatic Stress Disorder (PTSD). LR achieved the highest accuracy (93%), demonstrating the viability of lightweight models for mental health screening in mobile and real-time applications.

Qayyum *et al.* [32] investigated the use of Reddit as a source for detecting multiple mental health conditions, including depression, anxiety, bipolar disorder, schizophrenia, autism, and a general mental health category. They applied a range of models, from CNNs, LSTMs, GRUs, and Bidirectional GRUs (Bi-GRUs) to transformer-based architectures like Bidirectional Encoder Representations from Transformers (BERT), and Robustly Optimized BERT Pre-training Approach (RoBERTa). Their final system employed a hierarchical classification framework, with RoBERTa at its core, achieving an average test accuracy of 84%. This study emphasizes the utility of domain-specific model fine-tuning and multi-tier classification strategies.

Similarly, Pandey and Kumar [33] applied NLP and ML techniques to predict mental health and stress conditions from Reddit posts. Their methodology included text preprocessing, feature extraction using Term Frequency - Inverse Document Frequency (TF-IDF) and n-grams, and classification using transformer models such as BERT, RoBERTa, and A Light BERT (ALBERT). Among these, RoBERTa demonstrated superior performance, reinforcing its efficacy in identifying psychological patterns from informal social media text.

Abdullah and Negied [34] extended existing work by exploring both current and predictive mental health classification. They analyzed Reddit posts to detect not only existing conditions, such as ADHD, anxiety, depression, and bipolar disorder, but also to forecast potential future diagnoses. Nineteen models were evaluated, including traditional classifiers, ensemble methods (Light Gradient-Boosting Machine (LightGBM), Extreme Gradient Boosting (XGBoost), Bagging Estimator), and large language models (LLMs) like RoBERTa and OpenAI's Generative Pre-training Transformer (GPT). The LLMs outperformed other models in both detection and prediction tasks, achieving an F1-score of 80.0% on clinical data and 43.0% in future disorder

prediction, marking a significant step forward in temporal mental health modelling.

In another significant contribution, Kabir *et al.* [35] focused specifically on early detection of depression using social media text. Their study used a balanced dataset of posts from both depressed and non-depressed individuals and applied a variety of NLP techniques including tokenization, stemming, n-grams, CountVectorizer, and TF-IDF to extract meaningful linguistic features. A wide range of ML classifiers, such as stochastic gradient descent (SGD), naïve Bayes (NB), decision tree (DT), random forest (RF), SVM, and multilayer perceptron (MLP), were trained and evaluated using metrics such as accuracy, precision, recall, and F1-score. The results demonstrated consistently strong model performance, affirming the value of combining feature engineering with traditional ML in depression detection tasks. Specifically, for dataset-1, the SVM, RF, MLP, and DT achieved the highest accuracies of 97.00%, 91.00%, 90.00%, and 88.00%, respectively. For dataset-2, the top-performing models were the RF (94.28%), MLP (93.00%), Stochastic Gradient Descent (92.00%), and SVM (88.00%), outperforming the other classifiers on the same dataset.

1.4.2 AI Models for ADHD Detection Using Behavioral and Cognitive Data

Building upon general mental health detection research, several studies have focused specifically on ADHD diagnosis using behavioral, cognitive, and psychometric data. In this vein, Mengi and Malhotra [36] explored the potential of neural networks for ADHD classification as an alternative to traditional standardized diagnostic tests, which are often time-consuming and costly. Using the ADHD-200 dataset, they train neural network models on Personal Characteristic Information (PCI) attributes to differentiate ADHD patients from typically developing (TD) children. They experiment with different neural network configurations and find that a model with four hidden layers (1024-512-256-2) achieves the highest accuracy of 75%, outperforming other known methods for the dataset. The study highlights the effectiveness of PCI attributes for ADHD classification and suggests that incorporating

additional discriminative features, such as neuroimaging, could further improve diagnostic accuracy in the future.

S. and Anandaraju [37] proposed a ML-based approach for classifying ADHD subtypes using Conners data and EEG signals. They employ supervised learning algorithms, including K-Nearest Neighbors, Decision Tree, and Random Forest, to analyze ADHD-related patterns. The performance of their model was assessed using statistical metrics such as recall, precision, accuracy, and F1-score, resulting in a classification accuracy of 79.51%. The study highlights the effectiveness of Conners data in ADHD categorization, emphasizing the potential of machine learning in enhancing early diagnosis and subtype differentiation for improved treatment strategies.

1.4.3 Neuroimaging-Based ADHD Detection

In addition to behavioral data, neuroimaging modalities such as EEG and fMRI have been extensively investigated for objective ADHD diagnosis using AI and deep learning approaches. In this vein, García-Ponsoda *et al.* [38] investigated the role of preprocessing and temporal segmentation of EEG signals in improving ADHD diagnosis. They applied filtering, Artifact Subspace Reconstruction (ASR), and Independent Component Analysis (ICA) to remove noise and artifacts from EEG data collected from ADHD and neurotypical children. Features were extracted from segmented EEG recordings and selected based on statistical significance, with classification performed using support vector machine (SVM), K-Nearest Neighbors (KNN), and XGBoost models. Their results showed that later EEG segments yielded higher classification accuracy, suggesting that cognitive fatigue may influence the ability to distinguish ADHD. The highest performance of 86.1% accuracy was achieved using data from the P3, P4, and C3 channels, with features such as kurtosis, Katz fractal dimension, and power spectra in Delta, Theta, and Alpha bands driving classification. Their study highlights the significance of EEG preprocessing and segmentation in improving diagnostic reliability and suggests that cognitive fatigue may be a promising avenue for future ADHD research.

Elkhodary and Youssef [39] presented a deep learning-based approach for diagnosing ADHD using resting-state functional Magnetic Resonance

Imaging (rs-fMRI) data. Utilizing the ADHD-200 dataset, they employ Nilearn's CanICA, an Independent Component Analysis method, to identify relevant brain regions of interest (ROIs). A Gated Recurrent Unit (GRU) model is then applied to capture temporal dependencies in the rs-fMRI time series, achieving an accuracy of 87.9%. Their study highlights the potential of deep learning in ADHD diagnosis, demonstrating how integrating advanced techniques with informative neuroimaging features enhances both model interpretability and diagnostic accuracy.

Abdolmaleki and Abadeh [40] proposed an AI-driven approach for ADHD diagnosis using non-invasive brain imaging techniques to enhance reliability over traditional psychological tests. They develop a deep multi-modal 3D CNN trained on structural gray matter (GM) and functional MRI data using the fractional amplitude of low-frequency fluctuations (fALFF) from the ADHD-200 dataset. The extracted features are classified using SVM, KNN, and Linear Discriminant Analysis (LDA), with the latter achieving the best accuracy of 74.93%. Additionally, integrating personal characteristics with early and late fusion techniques improves classification accuracy by 5.84%. The study underscores the effectiveness of combining neuroimaging with machine learning to enhance the reliability of ADHD diagnosis and clinical decision support systems.

The remainder of this thesis is outlined as follows. Chapter 2 summarizes the background material that is used in this work. Chapter 3 is devoted to the methodology and algorithms developed for the early detection of ADHD. In Chapter 4, the performance of the proposed methods is assessed, highlighting the model of choice for the data at hand. Finally, Chapter 5 summarizes this work and opens future points of investigation.

Chapter 2. Background Material

As AI continues to transform the field of medical diagnosis, its application in ADHD detection presents a promising avenue for improving accuracy, efficiency, and accessibility. This chapter provides the necessary background knowledge needed to understand normal vs. ADHD lifestyle, the computational tools, ML methodologies, and DL frameworks that support AI-based ADHD classification as well as the techniques that will be used throughout this research.

2.1 Brain Structure

The brain is an extraordinarily complex organ responsible for perception, cognition, emotion, and behavior. It is composed of billions of neurons that communicate through electrical and chemical signals. Moreover, the brain is divided into distinct regions, each responsible for specialized functions. For instance, the cerebral cortex (cerebrum) governs consciousness, the cerebellum controls balance and coordination, and the brainstem connects the cerebrum to the spinal cord [41].

Studies have shown that the brain structure of individuals with ADHD differs from those of neurotypical individuals. A large MRI study done by Poulton and Nanan [42], analyzed brain scans from 1,713 individuals with ADHD and 1,529 controls (people without ADHD), revealing significant variations in five key areas, suggesting that ADHD is linked to differences in brain structure and function. So, the brain areas that are different are as follows [42];

1. **Basal Ganglia (Caudate and Putamen):** These regions play a crucial role in controlling emotions, cognition, and voluntary movement.
2. **Nucleus Accumbens:** This brain area is involved in motivation and action, playing a vital role in reward-seeking behavior, including eating, sexual behavior, and stress responses.
3. **Amygdala:** This is known for processing emotions like fear and stress. The amygdala is often hyperactive or dysregulated in individuals with ADHD.

4. **Hippocampus:** This region is critical for memory formation and goal-directed behavior, both of which are often impacted by ADHD.

Apart from the affected areas, patients with ADHD exhibit reduced brain volume, particularly in regions essential for attention, organization, control, and future planning, such as the frontal lobes [43]. To visually summarize these findings, a schematic illustration of the brain regions associated with ADHD-related volume reduction is presented in Fig. 2.1.

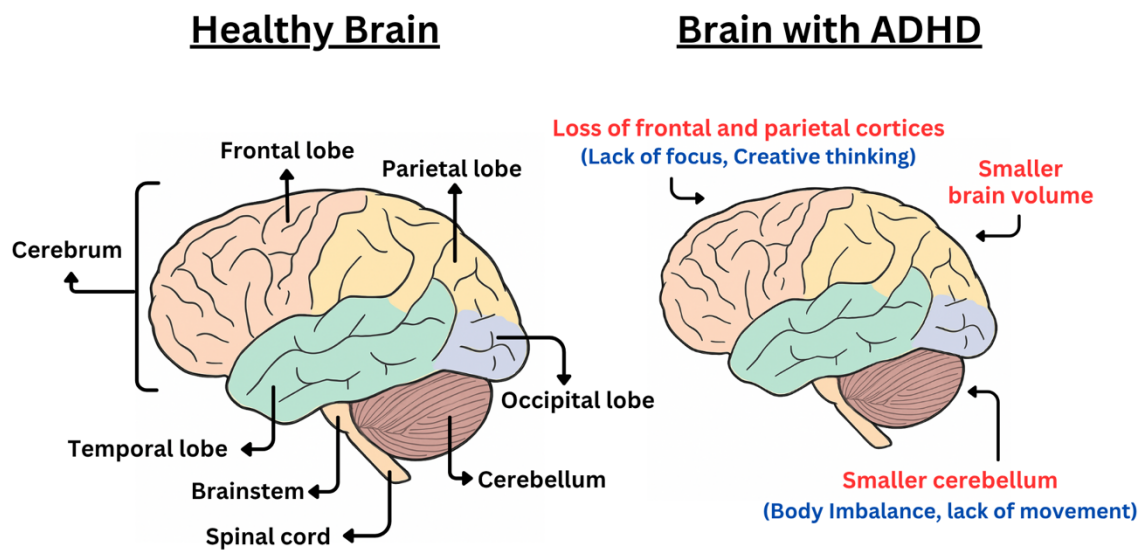


Figure 2.1 Difference between a Healthy Brain and a Brain with ADHD

2.2 How ADHD Affects Productivity and Organization

People with ADHD often encounter unique challenges related to productivity and organization, so it is essential to learn and develop effective methods that support them to focus on their strengths and manage challenges. As individuals with ADHD age, their symptoms transform. Adults often struggle with restlessness, impatience, distractibility, and challenges in organization, time management, and memory, due to which they often struggle with work productivity and regular daily activities [44].

Many individuals with ADHD struggle with time blindness, which refers to difficulty in accurately estimating the passage of time or the duration required to complete tasks. As a result, procrastination and last-minute rushing are common. Additionally, difficulties with working memory, which is the ability to hold and manipulate information in the short term, can lead to forgetfulness, misplaced items, and incomplete tasks. It also affects impulse control, causing individuals to become easily distracted by their environment or shifting thoughts, making it hard to sustain attention on a single task. This can result in frequent interruptions, unfinished projects, and difficulty following multi-step instructions [45]. Organizational challenges often extend beyond work and school into daily life, affecting everything from maintaining a clean-living space to managing appointments and responsibilities. However, strategies such as structured routines, visual reminders, and external accountability can help individuals with ADHD improve productivity and organization.

2.3 Functional Brain Imaging Modalities

Functional brain imaging represents a revolutionary approach to understanding the human brain's complex operations, allowing researchers and clinicians to observe brain activity in real time. Unlike structural imaging techniques, which provide static anatomical views, functional imaging captures the dynamic processes of neural activity, offering insights into cognitive, emotional, and physiological mechanisms. Some of the key functional brain imaging techniques include:

1. Functional Magnetic Resonance Imaging (fMRI)

Functional Magnetic Resonance Imaging (fMRI) is among the most utilized techniques for functional neuroimaging [46]. It relies on the Blood Oxygen Level Dependent (BOLD) signal to measure changes in blood oxygenation and flow associated with neural activity. The need for oxygen increases as neurons are activated, causing an increase in blood flow to certain regions of the brain. One of the key advantages of fMRI is its non-invasive nature, as it does not expose patients to radiation. Additionally, it offers high spatial resolution, making it ideal for mapping brain activity across multiple cognitive domains, such as memory, language processing, and decision-making.

2. Positron Emission Tomography (PET)

Positron Emission Tomography (PET) is a medical imaging technique that reveals the metabolic and biochemical functions of tissues and organs [47]. It involves injecting a small amount of a radioactive tracer to measure metabolic activity, neurotransmitter function, and molecular interactions, especially in the brain. A major advantage of PET is its ability to directly visualize molecular-level brain processes, making it highly valuable for studying neurodegenerative diseases, psychiatric disorders, and cancer. Moreover, PET scans can be tailored to track specific neurotransmitter systems by using different tracers, allowing researchers to study dopamine, serotonin, and other important neurochemicals.

3. Electroencephalography (EEG)

Electroencephalography (EEG) is a method for recording the brain's electrical activity by placing electrodes on the scalp, allowing the detection of neuronal firing activity [47]. It provides a high temporal resolution, capturing brain activity within milliseconds, which is crucial for studying rapid cognitive processes such as perception, attention, and language. EEG is also relatively low-cost, portable, and adaptable to a variety of research settings, making it an accessible tool for both clinical and experimental studies. Furthermore, it is widely used in the study of neural oscillations and brain connectivity, providing valuable insights into conditions such as epilepsy, sleep disorders, and cognitive impairments.

4. Magnetoencephalography (MEG)

Magnetoencephalography (MEG) is a brain imaging method that measures the magnetic fields produced by neural electrical activity [47]. Capturing these subtle magnetic signals enables researchers to map brain activity with high accuracy. One of the primary advantages of MEG is its superior spatial and temporal resolution compared to traditional EEG, which allows for more accurate tracking of neural activity over time. Additionally, MEG is less susceptible to signal distortion caused by the skull and scalp, which can affect EEG recordings. This makes MEG particularly useful for pinpointing the exact location of neural activity, making it an important

tool in epilepsy research, brain-computer interfaces, and cognitive neuroscience.

2.4 Machine Learning Models

ML has become a critical tool in many domains, including healthcare, where it is increasingly being used to enhance diagnosis, predict outcomes, and identify patterns. In the context of mental health, ML models are utilized to analyze diverse types of data, including text, images, and sensor information. This results in advanced systems that can be extensively integrated across all areas of medical science, ranging from drug discovery to clinical decision-making, significantly transforming the way medicine is practiced [48]. As Nasteski [49] mentions, ML models can be categorized into the following groups:

- **Supervised Learning:** The algorithms create a function that maps inputs to desired outputs based on labeled data samples, often used for classification tasks.
- **Unsupervised Learning:** The algorithms analyze unlabeled data with the aim of identifying patterns or clusters.
- **Semi-Supervised Learning:** Combines both labeled and unlabeled data to improve the accuracy of the function or classifier.
- **Reinforcement Learning:** The algorithm learns to make decisions by interacting with the environment, receiving feedback that guides its actions.

Supervised ML algorithms operate with predetermined classes, where a finite set of labels is assigned by experts to segments of data. The algorithm's task is to identify patterns and construct mathematical models, which are evaluated based on their predictive capacity and variance in the data [49]. Supervised learning is primarily divided into algorithms that perform classification and regression tasks. Classification techniques assign input data to specific predefined categories or class (for example: ADHD vs. Healthy). Meanwhile, regression models predict continuous numerical values by learning the relationship between input and a target value (for example: A continuous score indicating the severity of ADHD symptoms on a scale from 0 to 100). On the other hand, unsupervised learning does not depend on predefined

labels but instead detects patterns within data to create clusters. This clustering approach helps uncover hidden structures and relationships.

2.4.1 Traditional Machine Learning Models

The development of ML models typically follows a structured process that involves data preparation, model training, and performance evaluation. One of the first steps in building an ML model is dividing the dataset into training and testing sets. The training set is used to teach the model by providing it with labeled data (in the case of supervised ML), where each data point consists of input features along with its corresponding class label. The model processes the data to identify patterns between inputs and outputs by updating its parameters to reduce the discrepancy between its predictions and the true class labels.

After training, the model is evaluated on a test set, which contains unseen data. This step assesses its generalization capability to new inputs. Performance is commonly quantified using metrics such as accuracy, precision, recall, F1-score, and, depending on the task, the Receiver Operating Characteristic (ROC) curve or the Area Under the Curve (AUC). However, simply training and testing a model is not enough; further refinement is often necessary to achieve optimal performance. Some techniques such as hyperparameter tuning and cross-validation are commonly used to fine-tune the model [50, 51]. Hyperparameter tuning involves adjusting model-specific settings (such as the learning rate or number of decision tree depths), while cross-validation ensures that the model performs well across different data splits.

A key challenge in machine learning is managing the bias-variance tradeoff. High bias, or underfitting, arises when a model is overly simple and cannot capture the underlying patterns in the data. On the other hand, high variance, or overfitting, occurs when a model fits the training data too closely, including noise and irrelevant details, which leads to reduced performance on unseen data. To strike the right balance, techniques such as regularization (L1/L2 penalties), dropout (for neural networks), and ensemble learning can be used to improve generalization [51].

Several supervised learning algorithms that are widely used in various applications are as follows:

2.4.1.1 Logistic Regression

Logistic Regression (LR) is a supervised learning method commonly used for classification tasks. Despite the term “regression,” it models the probability of discrete categorical outcomes rather than continuous variables. It is especially effective in binary classification scenarios, where the outcome variable has two distinct categories, such as the presence or absence of a disease. LR predicts the likelihood that an input belongs to a specific class by applying the logistic (sigmoid) function, which converts a linear combination of features into a probability ranging from 0 to 1 for binary classification problems [52]. It can be categorized into three types based on the number of outcome classes:

- **Binary Logistic Regression:** It is used when the dependent (target) variable has only two categories, such as “yes” or “no”.
- **Multinomial Logistic Regression:** It extends the model to handle three or more unordered categories, such as predicting different types of diseases.
- **Ordinal Logistic Regression:** It is applied when the dependent (target) variable has three or more categories with a natural order, such as rating customer satisfaction as low, medium, or high.

In the context of binary classification, logistic regression models the relationship between input variables and the probability of a specific outcome using a hypothesis function [49]:

$$h_{\theta}(\mathbf{x}) = \frac{1}{1 + e^{-\theta^T \mathbf{x}}} \quad (2.1)$$

where,

- $h_{\theta}(\mathbf{x})$ represents the predicted probability that the output $y = 1$,
- θ is the vector of model parameters (weights),
- $\mathbf{x}$ is the input feature vector,
- ^{T} denotes the transpose operator,
- The function $\frac{1}{1+e^{-()}}$ is known as the sigmoid function.

The sigmoid function maps any real-valued input into a range between 0 and 1, making it ideal for probability-based classification. If the probability is above a certain threshold (usually 0.5), the input data ($\mathbf{x}$) is classified as class 1; otherwise, it is classified as class 0. This is called a threshold-based decision, which defines a decision boundary that separates the two classes in the feature space.

The performance of the logistic regression model is evaluated using a cost function derived from the principle of maximum likelihood estimation. The cost function, also called log loss or binary cross-entropy, is defined as [49]:

$$J(\boldsymbol{\theta}) = -\frac{1}{m} \sum_{i=1}^m \left[y^{(i)} \log(h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)})) + (1 - y^{(i)}) \log(1 - h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)})) \right] \quad (2.2)$$

where,

- m is the number of training samples,
- $y^{(i)}$ is the actual label of the i^{th} training sample that can be either 0 or 1,
- $h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)})$ is the predicted probability for that sample and $\mathbf{x}^{(i)}$ denotes the i^{th} training vector.

2.4.1.2 Support Vector Machine

Support vector machine (SVM) is a supervised learning algorithm widely applied to both classification and regression problems. It is especially effective for high-dimensional data and is frequently applied in fields such as pattern recognition, image processing, text classification, and bioinformatics. As Wang *et al.* [53] mention, SVM aims to identify an optimal hyperplane that best separates different classes. The objective is to maximize the distance between this hyperplane and the nearest data points from each class, referred to as support vectors (margins). These closest or nearest data points play a crucial role in defining the decision boundaries.

SVM operates by mapping the input features to a higher dimension where it identifies the optimum hyperplane which effectively classifies the different classes. The mathematical representation of the hyperplane is [53]:

$$\mathbf{w}^T \mathbf{x} + b = 0 \quad (2.3)$$

where,

- $\mathbf{w}$ is the weight vector,
- $\mathbf{x}$ is the input feature vector,
- b is the bias term.

The goal is to make the margin as large as possible while still classifying all the data points accurately. The optimization problem is formulated as [53]:

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 \quad (2.4)$$

Subject to the constraints for all training example:

$$y^{(i)}(\mathbf{w}^T \mathbf{x}^{(i)} + b) \geq 1, \quad \forall i \quad (2.5)$$

where, $\|\mathbf{w}\|^2$ denotes the norm of vector $\mathbf{w}$. In this case, the labels $y^{(i)}$ belong to $\{-1, +1\}$ instead of $\{0, 1\}$ because the optimization and decision function rely on signed distance from the hyperplane. Using -1 and +1 simplifies the mathematical formulation of constraints. This unifies both classes under a single inequality and naturally defines the margin around the hyperplane.

In cases where data is not linearly separable, SVM uses a slack variable ($\xi_i > 0$) to allow some misclassification or to violate the margin while still maximizing it, as follows [53]:

$$y^{(i)}(\mathbf{w}^T \mathbf{x}^{(i)} + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad (2.6)$$

- $\xi_i = 0$ implies the point is correctly classified and outside the margin,
- $0 < \xi_i < 1$ means the point is inside the margin but correctly classified,
- $\xi_i \geq 1$ indicates a misclassified point.

This leads to the soft-margin SVM, which introduces a regularization parameter C to balance the trade-off between maximizing the margin and minimizing misclassification.

$$\min_{\mathbf{w}, b, \xi} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^m \xi_i \quad (2.7)$$

2.4.1.3 Decision Tree

A Decision Tree (DT) is a supervised learning algorithm used for classification and regression tasks. It mimics human decision-making by breaking down a dataset into smaller subsets while making decisions based on feature conditions. Decision Trees are widely used due to their interpretability and ease of use, making them a popular choice in fields such as credit risk analysis, medical diagnosis, and fraud detection.

It consists of nodes, where each node represents a feature-based decision. The tree starts at the root node and splits data based on feature values, creating branches that lead to leaf nodes, which represent the final classification or regression output [49]. To determine the best split at each node, Decision Trees use splitting criteria such as [54]:

- **Gini Impurity:** Measures the probability of incorrect classification:

$$\text{Gini}(t) = 1 - \sum_{i=1}^N p_i^2 \quad (2.8)$$

where,

- N is the number of classes,
- p_i is the probability of class i at node t .

- **Entropy (Information Gain):** Measures the purity of the dataset:

$$\text{Entropy}(t) = - \sum_{i=1}^N p_i \log_2(p_i) \quad (2.9)$$

The tree grows until all data points are classified, or a stopping condition (such as tree depth or minimum samples per leaf) is met.

2.4.1.4 Random Forest

Random Forest (RF) is an ensemble learning technique that constructs multiple Decision Trees and aggregates their predictions to enhance accuracy and minimize overfitting [55]. It makes predictions by taking a majority vote for classification tasks or the average prediction for regression tasks [55]. It is widely used for classification and regression tasks in applications such as finance, healthcare, and recommendation systems.

2.4.1.5 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a non-parametric, supervised learning algorithm used for classification and regression tasks. It is based on the principle that similar data points are generally located near each other in the feature space [56]. KNN classifies a data point by looking at the K nearest neighbors in the training dataset and assigning the most common class among them. It is a popular algorithm in areas like pattern recognition, anomaly detection, and recommendation systems due to its straightforward approach and strong performance on smaller datasets. It works by following these steps [56]:

Step 1. Select the value of K, which represents the number of nearest neighbors to consider.

Step 2. Calculate the distance between the query point and all training samples. Some of the common distance metrics include:

- **Euclidean Distance:** This metric calculates the straight-line distance between two points in Euclidean space. It is defined mathematically as:

$$d_{\text{Euclidean}}(\mathbf{u}, \mathbf{v}) = \sqrt{\sum_{i=1}^n (\mathbf{u}_i - \mathbf{v}_i)^2} \quad (2.10)$$

where, $\mathbf{u}_i, \mathbf{v}_i$ respectively denote the i^{th} element of vectors $\mathbf{u}$ and $\mathbf{v}$.

- **Manhattan Distance:** This metric measures the distance between two points along axes at right angles. It is defined mathematically as:

$$d_{\text{Manhattan}}(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^N |\mathbf{u}_i - \mathbf{v}_i| \quad (2.11)$$

- **Minkowski Distance (generalized form of Euclidean and Manhattan):** This metric incorporates both Euclidean and Manhattan distances. It is defined mathematically as:

$$d_{\text{Minkowski}}(\mathbf{u}, \mathbf{v}) = \left(\sum_{i=1}^n |\mathbf{u}_i - \mathbf{v}_i|^p \right)^{\frac{1}{p}} \quad (2.12)$$

where, p is the order of the norm, ($p = 1$ corresponds to the Manhattan Distance, while $p = 2$ refers to the Euclidean Distance).

- Step 3.** Identify the K nearest neighbors by selecting the data points with the shortest computed distances.
- Step 4.** Classify the data point using majority voting (for classification) or averaging the values of neighbors (for regression).

2.4.1.6 XGBoost

XGBoost (Extreme Gradient Boosting) is a high-performance, ensemble machine learning algorithm that builds decision trees in a sequential manner while optimizing for both speed and accuracy [57]. It is a variant of Gradient

Boosting Decision Trees (GBDT) and is designed to minimize computational cost while maintaining high predictive power. It is widely used in financial modeling, fraud detection, and recommendation systems due to its efficiency and strong generalization ability.

It builds an ensemble of decision trees using boosting, where each new tree corrects the errors of the previous trees. The key steps involved are [57]:

- Step 1.** Initialize with a weak learner (a shallow decision tree) that makes predictions.
- Step 2.** Calculate residual errors between actual values and predicted values.
- Step 3.** Train the next decision tree on the residual errors to minimize them.
- Step 4.** Continue the process for a set number of iterations or until the model reaches convergence.

2.4.1.7 Multinomial Naïve Bayes

Multinomial Naïve Bayes is a probabilistic ML algorithm based on Baye's theorem, specifically designed for text classification tasks. It assumes that all words in the given sentence, provided as input features, are independent of each other given the class label [49]. Unlike other variants, it takes into account the frequency of words within a document to determine its category. The algorithm calculates the probability of a document belonging to class C based on its word frequencies. Using Baye's theorem, the probability is computed as [49]:

$$P(C | \mathbf{x}) = \frac{P(\mathbf{x} | C) \cdot P(C)}{P(\mathbf{x})} \quad (2.13)$$

where,

- $P(C | \mathbf{x})$ is the posterior probability, the probability that the document belongs to class C given the feature vector $\mathbf{x}$.
- $P(\mathbf{x} | C)$ is the likelihood, the probability of observing the feature vector $\mathbf{x}$ across given that the class is C .
- $P(C)$ is the prior probability of class C .

- $P(\mathbf{x})$ is the evidence, the probability of observing the feature vector $\mathbf{x}$ across all classes

2.4.1.8 Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is a supervised learning approach that functions both as a dimensionality reduction technique and a classification model. Its objective is to find a linear combination of features that optimally separates predefined classes [58]. In contrast to unsupervised methods like Principal Component Analysis (PCA), which focus only on maintaining data variance, LDA uses class labels to enhance class separability.

LDA maps data onto a lower-dimensional space to maximize the separation (variance) between classes while minimizing the variance within each class [58]. It performs optimally when class distributions are Gaussian and share a common covariance matrix. Its interpretability, simplicity, and effectiveness on small datasets make LDA widely used in areas such as pattern recognition, facial recognition, and medical diagnostics [58].

2.4.1.9 Adaptive Boosting (AdaBoost)

AdaBoost (Adaptive Boosting) is an ensemble learning technique that constructs a strong classifier by sequentially combining multiple weak learners, typically decision stumps [59]. It operates by increasing the weights of misclassified samples so that subsequent learners concentrate on the more difficult instances. It has been extensively applied in areas such as face detection, fraud detection, and bioinformatics due to its high accuracy and adaptability.

2.4.1.10 LogitBoost

LogitBoost is a boosting algorithm specifically designed to optimize logistic regression for classification tasks. It is an adaptation of AdaBoost, replacing the weak learner's exponential loss function with a logistic loss function, making it less sensitive to outliers and more stable in noisy datasets [59]. It builds an ensemble of weak classifiers (typically decision stumps) and iteratively adjusts weights to improve classification performance.

2.4.2 Parameters vs Hyperparameters

Parameters and hyperparameters play distinct yet complementary roles in ML models. Parameters are the internal variables of a model that are adjusted and learned during training using optimization algorithms. These variables form the model's representation of data and identify patterns within the training dataset. In iterative training approaches, parameter values undergo continuous updates, with the goal of determining optimal values that accurately fit training data while maintaining effectiveness on unseen data. In neural network contexts, these parameters are commonly labelled as weights and biases, or coefficients in regression models. In contrast, hyperparameters are configuration settings defined prior to training that controls the learning process and establish the fundamental structure and parameters of the learning algorithm [60]. Common hyperparameters include learning rate, dropout probability, neural network architecture specifications (layer count, unit numbers), batch size, and activation function selection.

To define this main difference, the parameters are optimized automatically to minimize the loss function, while hyperparameters require careful tuning through methods like grid search, random search, or Bayesian optimization to achieve optimal model performance. Most importantly, the careful selection of hyperparameter values is important, as these choices significantly affect learning capacity, training efficiency, generalization capabilities, and the model's ultimate performance metrics.

2.4.3 Machine Learning Pipeline

ML pipelines represent structured, end-to-end workflows that systematically organize the sequence of data processing and modelling operations required to transform raw data into deployable ML solutions. These pipelines encapsulate multiple interconnected stages including data acquisition, data preparation and feature engineering, model training, evaluation, and deployment as a cohesive system. Each component within the pipeline performs a specific function, with outputs from one stage serving as inputs to subsequent stages, creating a modular yet integrated framework.

The process begins with data collection, where researchers either receive existing datasets or gather data independently to train, select, and evaluate

models. Next comes data preparation and feature extraction, which is a critical phase where data is cleaned, transformed, and potentially normalized to ensure its suitability for modelling while mitigating errors and biases. The process continues with model type selection, where an appropriate model architecture is chosen based on the specific problem domain and data characteristics. Model training then follows, typically involving the training of multiple model instances with varying hyperparameters rather than a single fixed configuration, as optimal hyperparameter values are rarely known in advance. To identify the best-performing model, we usually split the dataset into 70% as training data, 10% as validation, and 20% as testing data. This validation data allows for performance assessment on unseen examples during the selection process, and the testing data provides evaluation on a separate test set to obtain an unbiased estimate of real-world performance. This pipeline approach ensures reproducibility, facilitates systematic improvement, and provides credible performance metrics essential for research validation and practical application.

2.4.4 Generalization

Generalization is a fundamental concept in ML. It describes a model's capacity to maintain accurate performance on new, unseen data after training on a given dataset. The goal of an ML model is not simply to memorize the training data, but to extract underlying patterns and relationships that can be applied to similar, unseen examples. Effective generalization represents the balance between underfitting, where models fail to capture the essential relationships in the data, and overfitting, where models excessively adapt to training data noise and peculiarities.

Consider, as an example, an NLP task aimed at classifying patient responses into categories such as ADHD, normal mental health, or other mental health conditions. During training, the model identifies and learns to recognize textual features, such as word sentiment, tone, and syntactic structure. However, if it merely memorizes specific phrases seen during training, it will struggle to generalize when encountering semantically similar expressions phrased differently. To generalize well, the model must identify broader linguistic patterns that span both the training and validation datasets. It should

detect sentiment-laden words like “creative,” “disorganized,” “distracted,” or “passionate,” and understand how context affects meaning. For instance, the phrase “not disorganized” conveys a positive trait, despite including a word with a negative connotation.

In essence, strong generalization requires the model to abstract the core patterns underlying the data, rather than fixating on superficial details. As models increase in complexity, particularly with deep learning architectures, understanding and improving generalization behavior becomes increasingly crucial for developing robust systems capable of reliable performance in real-world, dynamic environments where data distributions may shift over time.

2.4.5 Overfitting vs Underfitting

Overfitting and underfitting are fundamental challenges in developing ML models, both of which relate to a model's capacity to generalize effectively from the training dataset to new, unseen data.

Underfitting happens when a model shows poor performance on both the training and validation datasets. This suggests that the model is either overly simplistic or that the dataset's features lack sufficient informativeness, preventing it from capturing meaningful patterns in the data. It is often associated with high bias, where the model consistently fails to recognize important relationships between the input features and the target variable.

Overfitting occurs when a model achieves excellent performance on the training data but performs poorly on validation or test datasets. This happens because the model has memorized the training data, including its noise and outliers, rather than learning generalizable patterns. As a result, it struggles to perform well on new, unseen data. It is often linked to high variance, where the model becomes overly responsive to small fluctuations in the training set.

To overcome underfitting, one might consider increasing model complexity, training for additional epochs, or engineering more informative features. For overfitting, effective strategies include regularization techniques, reducing model complexity, implementing early stopping, or applying dropout in neural networks. The fundamental challenge in model development is identifying the optimal complexity level that minimizes both underfitting and

overfitting errors, thereby maximizing generalization performance. This balance often requires systematic experimentation with various model architectures and hyperparameter configurations to find the most effective solution for a given problem domain. A visual comparison of underfitting, optimal fitting, and overfitting behaviors is shown in Fig. 2.2.

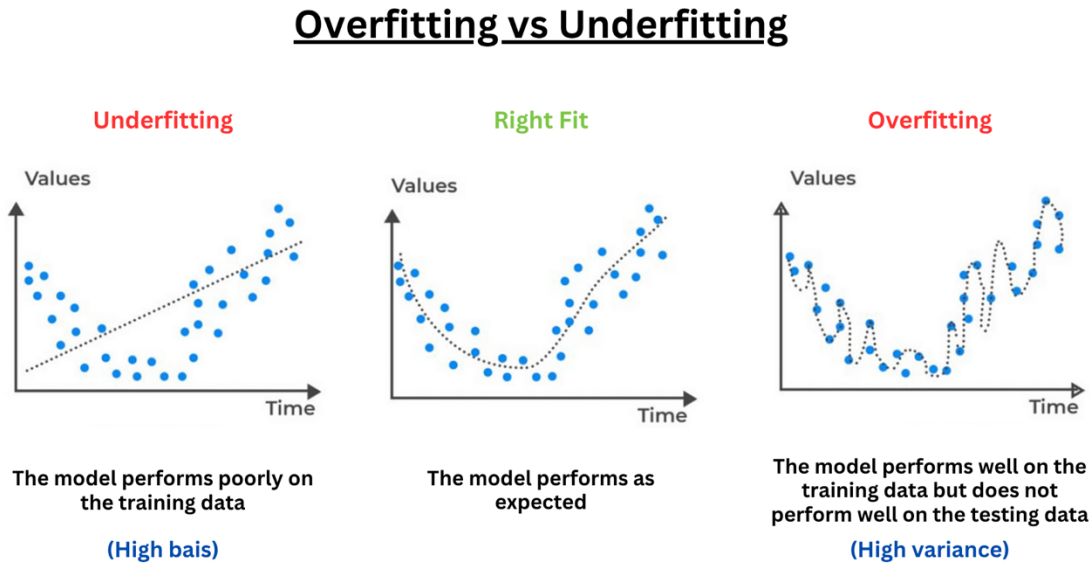


Figure 2.2 Comparison of Model Fitting Behavior in Machine Learning

2.4.6 Performance Measures

Performance measures provide valuable quantitative evaluations of ML models, where multiple measures serve different roles in analysis. In classification tasks, they are usually derived from the confusion matrix, which categorizes predictions into four categories:

- **True Positives (TP):** The predicted class is positive, and the actual class is also positive.
- **False Positives (FP):** The predicted class is positive, but the actual class is negative.
- **False Negatives (FN):** The predicted class is negative, but the actual class is positive.

- **True Negatives (TN):** The predicted class is negative, and the actual class is also negative.

From these foundational elements, we calculate several key performance indicators such as,

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.14)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2.15)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2.16)$$

$$\text{F1 - Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (2.17)$$

Fig 2.3 illustrates the confusion metrics along with the performance metrics.

		PREDICTED CLASS		
		Positive	Negative	
ACTUAL CLASS	Positive	True Positive (TP)	False Negative (FN) Type II Error	$\text{Recall} = \frac{TP}{TP + FN}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	$\text{Precision} = \frac{TP}{TP + FP}$
$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$		$\text{F1 - Score} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$		$\text{Specificity} = \frac{TN}{TN + FP}$

Figure 2.3 Performance Metrics in Machine Learning

2.5 Deep Learning Models

Deep learning (DL) has significantly advanced the field of mental health diagnostics by enabling automatic extraction of complex patterns from data,

surpassing traditional machine learning approaches in both accuracy and efficiency. This advancement is particularly crucial for ADHD classification, where subtle variations in handwriting, neuroimaging, and linguistic patterns must be captured and analyzed. Unlike conventional models that rely on handcrafted features, DL models learn hierarchical representations directly from raw data, making them highly effective for image-based and text-based classification tasks.

2.5.1 Convolutional Neural Networks (CNNs) for Image-Based Analysis

CNNs were initially designed to address the issue of image classification by learning non-linear representations from input data to the target labels. Unlike traditional ML approaches based on hand-designed features, CNNs automatically extract and learn required features from unprocessed input data during training. This capability in feature extraction enables their performance for complex recognition tasks.

The foundation of CNNs was laid by Yann LeCun, who introduced a network for recognizing handwritten digits [61]. This model processed 32×32 pixel images and classified them into ten-digit categories. Despite their potential, early CNN models were constrained by high memory demands, large dataset requirements, and computational limitations. A breakthrough came in 2012 when Alex Krizhevsky developed AlexNet, a CNN model that significantly outperformed previous techniques in the ImageNet classification challenge, which involved training on 1.5 million images across a thousand categories [62]. Following AlexNet, several advanced CNN architectures such as Visual Geometry Group (VGGNet), Residual Neural Network (ResNet), and GoogLeNet (or Inception V1) were developed, further enhancing performance and efficiency in deep learning applications.

CNNs are structured with specialized layers designed for hierarchical feature extraction. These networks are composed of an input layer followed by a series of convolutional and pooling layers, combined with one or more fully connected layers, activation functions, and a final output layer. As Zhao *et al.* [63] mentions, the initial part of the network consists of a series of convolutional and pooling layers arranged alternately to create a feature

extractor. This structure allows the raw input data, which has been pre-processed, to undergo operations that transform it into more abstract and high-level feature representations. Fully connected layers, along with activation functions, are used to perform tasks such as classification or regression on the extracted features. To enhance the performance of the CNN, additional components like batch normalization and dropout are incorporated, along with various mapping functions. The most significant aspects of such a structure are:

- 1. Convolutional layer:** The convolutional layer serves as the fundamental component of a CNN. It applies a mathematical operation called convolution, wherein filters (or kernels) move across the input data to extract relevant features [63]. Each filter is a small matrix of weights that learns to identify specific patterns. For example, in image processing, early convolutional layers might detect simple features like edges or textures. A typical filter might be a 3 or 5×5 matrix that slides across the input image, computing dot products at each position to create a feature map.
- 2. Pooling layer:** Pooling layers are used to reduce the spatial dimensions (height and width) of the input feature maps produced by convolutional layers while preserving valuable information and decreasing computational load [63]. The most common type is max pooling, which takes the maximum value from each patch of the feature map. For instance, a 2×2 max pooling operation with stride 2 would reduce a 28×28 feature map to 14×14 , retaining the most prominent features while discarding 75% of the values. This not only reduces computation but also introduces a form of positional invariance. This function helps in reducing overfitting and improving model generalization.
- 3. Activation Functions:** Activation functions introduce non-linear transformations into the neural network. This allows the model to capture and learn complex patterns. The two commonly used activation functions are:

- a. Rectified Linear Unit (ReLU):** The ReLU activation function is widely used due to its ability to accelerate convergence during training. It is defined as [63]:

$$\text{ReLU}(x) = \max(0, x) \quad (2.18)$$

This function sets negative values to zero, which preserves only the positive activations and reduces computational complexity. The function curves of ReLU can be shown as in Fig. 2.4.,

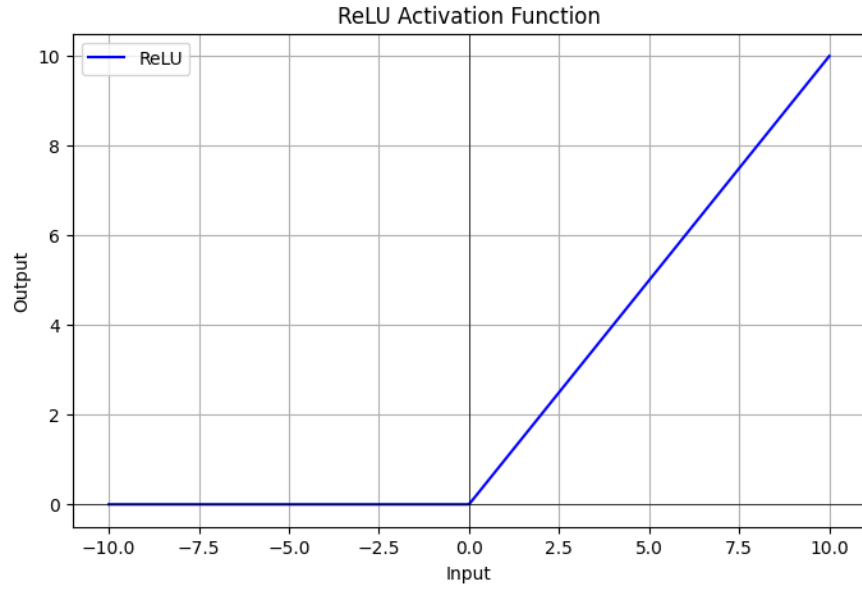


Figure 2.4 Functions curves of ReLU

- b. Leaky Rectified Linear Unit (Leaky ReLU):** To address the “dying ReLU” problem, where neurons can become inactive and output only zeros, the Leaky ReLU introduces a small, non-zero slope for negative values. This slope is controlled by a learnable parameter α , which allows the function to have a small gradient for negative input values [63]:

$$\text{LeakyReLU}(x) = \begin{cases} x, & \text{if } x \geq 0 \\ \alpha x, & \text{if } x < 0 \end{cases} \quad (2.19)$$

If $\alpha = 0$, the function reduces to the standard ReLU. The function curves of Leaky ReLU can be shown as in Fig. 2.5.,

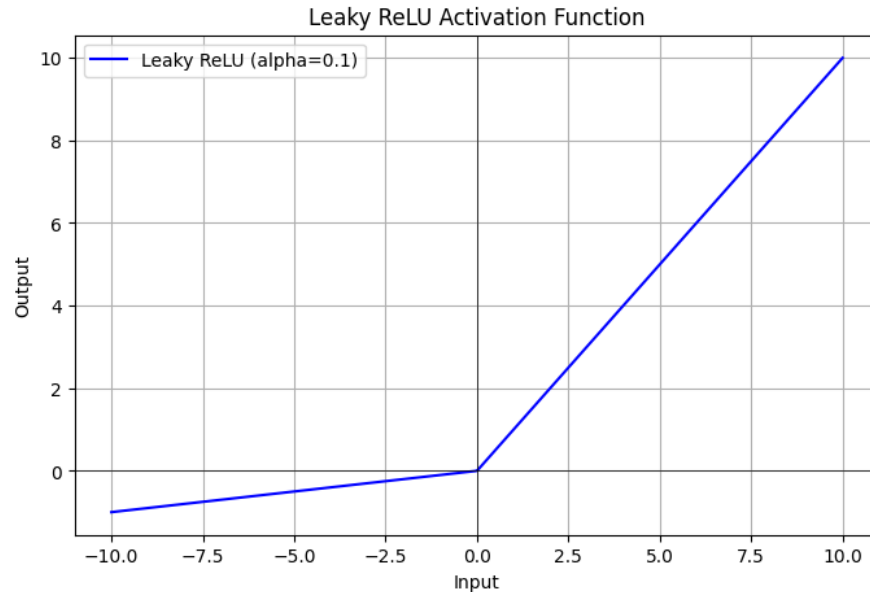


Figure 2.5 Functions curves of Leaky ReLU

- c. Dropout:** Dropout is a regularization method that prevents overfitting by randomly dropping out some fraction of the neurons at each training iteration. This encourages the model to develop more generalized and resilient feature representations.
- d. Fully connected layer:** A Fully Connected Layer, also referred to as a dense layer, is characterized by every neuron being linked to all neurons in the previous and next layers. These layers are typically used in the final stages to perform classification.

The purpose is to combine features learned by previous layers and apply weights to them to make predictions. Each connection has an associated weight that gets adjusted during training to minimize the error between predicted and actual outputs.

2.5.2 Recurrent Neural Networks (RNNs) for NLP-Based Classification

Recurrent Neural Networks (RNNs) are a type of neural network designed to handle sequences by keeping track of past information. Unlike feedforward networks that treat each data point independently, RNNs use recurrent connections to pass information from one step to the next, allowing the network to capture temporal patterns. This makes them particularly useful in domains such as NLP, speech analysis, and time-series prediction. The key element of an RNN is its hidden state, which updates at every time step by integrating the current input with prior states. However, standard RNNs suffer from limitations such as the vanishing gradient problem, making it difficult to learn long-term dependencies. To overcome this, advanced variants like Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs) introduce gating mechanisms that regulate information flow, allowing the network to retain and forget information more effectively.

Since, many individuals with ADHD exhibit distinct linguistic patterns, such as frequent topic-switching, impulsive phrasing, or unusual sentence structures, NLP seems like an attractive technique to capture such characteristics. RNNs enable the model to learn contextual relationships between words and phrases, aiding in the classification of ADHD-related speech and text data.

In this study, we utilize RNN-based models for analyzing ADHD-related text from sources such as self-reported online discussions (Reddit / Twitter datasets).

- LSTM: An advanced variant of RNNs designed to capture long-range dependencies by incorporating memory cells, reducing issues like vanishing gradients in long text sequences.
- Bi-LSTM: A bidirectional version of LSTM that reads text both forwards and backwards, enhancing the ability to detect ADHD-specific linguistic markers, such as repetitive phrasing, impulsivity in writing, and attention shifts.

2.5.3 Hybrid Deep Learning Models

Hybrid deep learning models combine multiple neural network architectures or integrate deep learning with traditional ML techniques to enhance performance across complex tasks. This integration enables the models to capitalize on the strengths of each approach to improve accuracy, generalization, and efficiency. For instance, CNNs can be integrated with Long Short-Term Memory (LSTM):

- **CNN Component:** It extracts spatial features from handwritten samples or fMRI scans, identifying ADHD-related structural irregularities.
- **LSTM Component:** It processes sequential dependencies in text-based datasets, recognizing ADHD-related linguistic patterns.

2.5.4 Transfer Learning

Machine learning models are typically trained using either of the two approaches: traditional learning (“from scratch”) and transfer learning (TL). Each approach has its own distinct strengths and limitations, especially when considering factors such as resource availability and dataset size. An overview of these two approaches is presented in Fig 2.6.

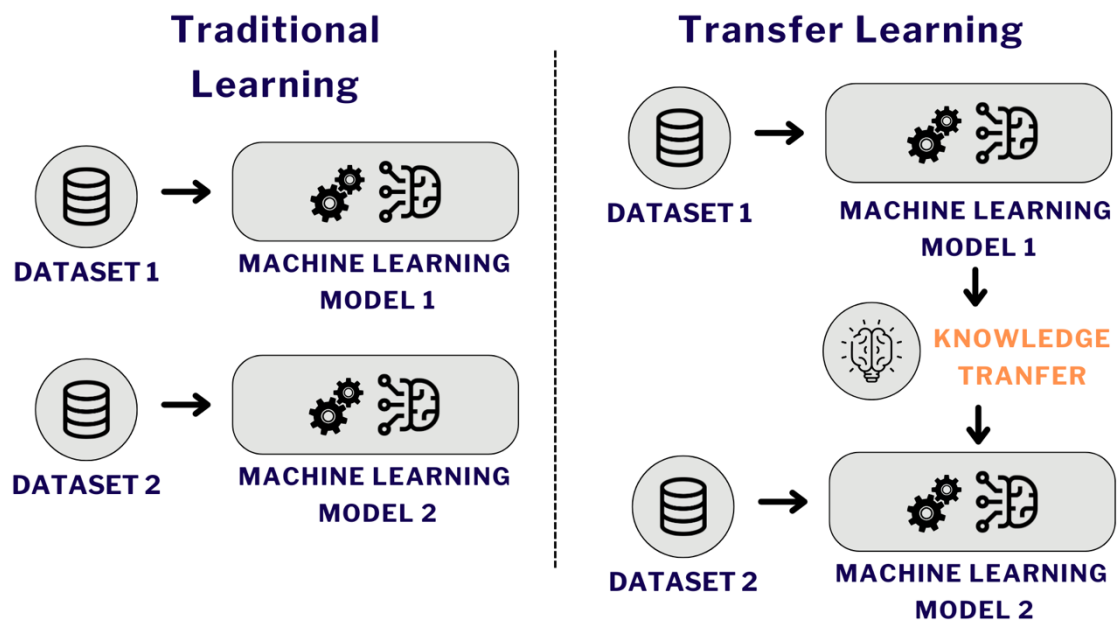


Figure 2.6 Traditional Learning Vs. Transfer Learning

Traditional learning involves training a model from the ground up using a large, labeled dataset. This method requires significant computational resources, such as multiple GPUs or distributed computing systems, and is highly time intensive. While it allows for task-specific customization, it is often impractical for domains like medical imaging, where obtaining large, annotated datasets is costly and time-consuming.

In contrast, transfer learning build upon pre-trained models developed from large and diverse datasets as a starting point. These models are fine-tuned on a smaller, domain-specific dataset, significantly reducing the computational burden and data requirements. Transfer learning is especially beneficial in scenarios where data is scarce, as it allows the model to reuse learned features from similar tasks or domains. For example, in medical imaging, TL has been successfully applied to tasks such as ADHD-related brain MRI analysis, including brain region segmentation and ADHD classification. By utilizing pre-trained models, TL reduces the dependence on large, annotated datasets while maintaining high performance in identifying ADHD-related patterns and abnormalities in brain MRIs. A side-by-side comparison of the key characteristics of traditional learning and transfer learning is summarized in Table 2.1.

Table 2.1 Difference Between Traditional and Transfer learning

Properties	Traditional Learning	Transfer Learning
Training approach	From scratch	Uses pre-trained models
Data requirement	Large dataset needed	Can work with limited data
Training Time	Long	Short
Computational Cost	High	Lower compared to traditional learning

Customization	Highly customizable	Limited to the structure of the pre-trained model
Performance	Depends on the quantity and quality of data	Often superior with limited data and uses prior knowledge
Application Domain	Specific to the dataset used	Effective in similar domains or tasks
Feature extraction	Learning features from scratch &	Uses features learned by pre-trained models
Initial model state	Randomly initialized weights	Pre-trained weights
Fine-tuning	Requires extensive training to reach optimality	Quick adjustments to pre-trained weights

2.6 Ensemble Learning

Ensemble learning refers to the process of integrating multiple models to improve overall accuracy, reliability, and the ability to generalize. Using a variety of models often enables ensemble approaches to outperform single-model predictions in terms of accuracy [64]. The main objective behind merging several models is to lower prediction errors by addressing variance (which leads to overfitting), reducing bias (which causes underfitting), and enhancing predictive capability. According to T. N. Rincy and R. Gupta [65], ensemble approaches are generally categorized into two types: parallel and sequential.

In parallel ensemble approaches, the individual base classifiers are trained independently, and their predictions are combined using techniques such as voting for classification tasks or averaging for regression tasks. These methods aim to reduce variance by aggregating diverse predictions generated from different subsets or variations of the training data. In contrast, sequential ensemble approaches build models one after another, with each new model aiming to address the errors made by its previous one. These methods are

designed to reduce both bias and variance through incremental improvements [64]. However, they may be more prone to overfitting, particularly when applied to noisy or high-variance datasets.

The major types of ensemble learning are:

1. Bagging

Bagging, short for Bootstrap Aggregating, is a technique that trains multiple instances of a base model on different randomly chosen subsets of the training dataset. These subsets are generated through sampling with replacement, known as bootstrap sampling. In regression problems, the final output is calculated by averaging the predictions from all models, while in classification problems, it is determined through majority voting [65]. Notable characteristics of bagging include training models in parallel, allowing each to learn independently from varied subsets, and effectively lowering variance to combat overfitting. A common and well-known example of bagging is the Random Forest algorithm, which integrates multiple decision trees trained on random subsets of both data and features to produce a stable and high performing ensemble model.

2. Boosting

Boosting is a type of ensemble method that builds models in a sequential manner, where each new model is trained to address the errors made by its predecessors [64]. In this approach, greater importance is given to data points that were previously misclassified, enabling the weak learners to improve and enhance the ensemble's overall predictive performance. Some of the key features of boosting include its step-by-step learning process, emphasis on correcting mistakes, and ability to reduce both bias and variance. Nonetheless, boosting methods can be prone to overfitting, particularly in the presence of noise or outliers, unless regularization techniques are applied. Well-known examples of boosting algorithms include AdaBoost (Adaptive Boosting), Gradient Boosting Machines (GBM), XGBoost, LightGBM, and CatBoost.

3. Stacking

Stacking, also known as Stacked Generalization, is an ensemble approach that merges the outputs of multiple base models through the use of a meta-learner trained to combine these predictions effectively [64]. In this method, the base models are first fitted on the training data, and their predictions serve as input features for the meta-model. It allows us to combine various types of models, including DTs, SVMs, and neural networks, allowing it to capture complex patterns in the outputs of the base models. To reduce the risk of overfitting, cross-validation is frequently applied when generating inputs for the meta-learner. This strategy often leads to substantial improvements in predictive accuracy.

4. Voting

Voting is a simple ensemble method, which is sometimes referred to as Simplified Stacking Variant, combines predictions from multiple models using aggregation techniques such as majority voting (for classification) or averaging (for regression). There are three common types of voting [66]:

- **Hard Voting:** Every model votes for a specific class, and the class with the highest number of votes becomes the final prediction.
- **Soft Voting:** Models output class probabilities instead of direct class labels, and these probabilities are averaged to select the most probable class.
- **Weighted Voting:** Predictions from each model are assigned weights according to the model's performance, which allows higher-performing models to have greater influence on the final decision.

2.7 Natural Language Processing

Natural Language Processing (NLP) is a branch of ML that focuses on the interaction between computers and human language with the focus on enabling machines to understand, interpret, and generate language meaningfully. It involves computational techniques and algorithms to process, analyze, and produce both written and spoken language. This field is a highly interdisciplinary field that integrates principles from computer science, linguistics, mathematics, psychology, and other related disciplines.

Modern NLP systems have evolved from rule-based approaches to advanced deep learning architectures. These sophisticated models can capture complex linguistic patterns and semantic relationships that form the foundation for applications ranging from machine translation and sentiment analysis to question answering systems and conversational agents.

2.7.1 Data Cleaning

Data cleaning and preprocessing are crucial steps in NLP machine learning projects. They involve transforming raw text into a clean and structured format suitable for ML algorithms. Key motivations for these steps include improving model performance by reducing noise and handling missing or inconsistent data. Standard techniques include stopwords removal, which eliminates common words like “the”, “and”, and “of” that contribute minimal semantic value. Stemming and lemmatization are used to reduce inflected words to their base or root forms: stemming mechanically strips suffixes (e.g., “jumping” becomes “jump”), while lemmatization maps words to their canonical forms (e.g., “were” becomes “be”). Special character and punctuation removal can further improve model robustness, although over-cleaning risks losing important contextual cues or introducing bias. Collectively, these preprocessing steps lay the foundation for more effective and reliable NLP models.

2.7.2 Tokenization

Tokenization is the process of splitting text into individual words or subwords units, forming a fundamental step in NLP pipelines. Various tokenization strategies exist, ranging from simple whitespace splitting to more sophisticated subword approaches. WordPiece tokenization, a method similar to Byte Pair Encoding (BPE), constructs a vocabulary of frequent subwords by iteratively merging adjacent tokens based on their frequency. This technique enables handling out-of-vocabulary words by decomposing them into familiar subwords, improving robustness to rare terms, misspellings, and morphologically rich languages. Subword tokenization also ensures a more compact and generalizable representation of textual data, particularly beneficial in modern transformer-based models.

2.7.3 Word Embeddings

Word embeddings represent words as dense, low-dimensional vectors that capture semantic and syntactic relationships, in contrast to sparse one-hot encodings. In these vector spaces, words with similar meanings tend to cluster together, enabling operations such as cosine similarity to measure word relationships. Pre-trained word embeddings, such as Word2Vec or GloVe, allow for transfer learning across diverse NLP tasks by leveraging knowledge from large text corpora. However, traditional embeddings are context-independent, meaning they assign the same vector to a word regardless of its usage in different sentences. To address this limitation, contextualized embeddings like those used in BERT dynamically generate word representations based on surrounding text, capturing nuanced, context-specific meaning

2.8 Generative Adversarial Network

A Generative Adversarial Network (GAN) generates realistic new data based on the learning patterns of existing samples. In this study, we are utilizing the Wasserstein Generative Adversarial Network (Wasserstein GAN or WGAN), a variant of the standard Generative Adversarial Network (GAN) designed to address some of the limitations of traditional GANs.

The original GAN framework, introduced by Ian Goodfellow and colleagues in 2014, is a machine learning algorithm using two neural networks, a generator and a discriminator, that compete with one another [67]. The generator is responsible for generating synthetic data, while the discriminator evaluates whether the data is real or synthetically generated. Through this competition, GANs can simulate complex data distributions and generate realistic new data such as images, music, or text. Although conventional GANs can generate high quality data, they often suffer from issues such as mode collapse, unstable training, and difficulties in assessing convergence [68, 69, 70].

To overcome these challenges, Arjovsky *et al.* introduced the WGAN in 2017, which modifies how the discriminator called the “critic”. In this context, it evaluates generated data [69]. Instead of relying on binary classification, the

WGAN measures the difference between real and generated data distributions using the Earth Mover's Distance (or Wasserstein distance). This metric yields smoother, more meaningful gradients, leading to more stable and reliable training [69]. An overview of the WGAN architecture, showing the generator (G) and critic (C), is presented in Fig. 2.7, while a comparison of the fundamental differences between GAN and WGAN is summarized in Table 2.2.

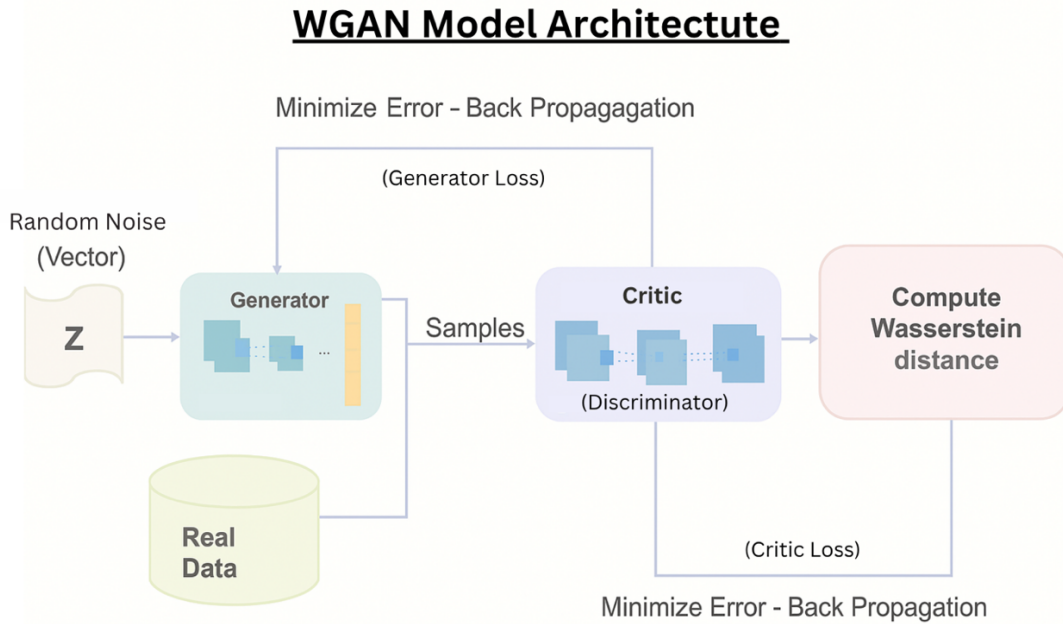


Figure 2.7 Wasserstein GAN Network Architecture with a Generator (G) and a Critic (C)

Table 2.2 Comparison of GAN and WGAN Network

Aspect	GAN	WGAN
Objective Function	Minimax game	Wasserstein distance
Loss Function	Jensen-Shannon Divergence	Wasserstein distance
Discriminator/Critic Role	Binary classifier to distinguish real vs. fake data	Evaluates the Wasserstein distance, ensuring meaningful gradients

Gradient Issue	Suffers from vanishing gradients	Mitigates vanishing gradients
Mode Collapse	Prone to mode collapse	Less prone to mode collapse
Performance	Depends on the quantity and quality of data	Often superior with limited data; leverages prior knowledge
Lipschitz Continuity Enforcement	Not enforced	Enforced via weight clipping or gradient penalty

2.9 Optical Character Recognition

Optical Character Recognition (OCR) has been traditionally used as one of the preprocessing steps that converts visually represented textual information into machine-encoded text. Beyond this traditional use, modern OCR systems can now also extract rich metadata about the character formation, spatial relationships, and typographical properties. These additional features, such as character height-width ratios, stroke densities, baseline variations, and character-to-character or word-to-word spacings, can be structured into numerical feature vectors to be used with ML algorithms. By providing a dense quantitative description of handwriting or typography, OCR-extracted features are discriminative inputs to numerous downstream tasks, including classification (e.g., mental disorder based on handwriting), language identification, writer verification, and semantic content interpretation.

The hierarchical nature of OCR processing, from pixel-level recognition to semantic understanding, mirrors the feature abstraction process in deep learning, making OCR-extracted features particularly compatible with contemporary ML architectures and transfer learning approaches.

Chapter 3. Methodology

This section outlines the methodological framework applied in the study, including data collection, preprocessing methods, model development strategies, training, and testing techniques. The purpose of this work is to present a clear and reproducible approach for enhancing early diagnosis of ADHD using AI techniques.

3.1 Data Collection

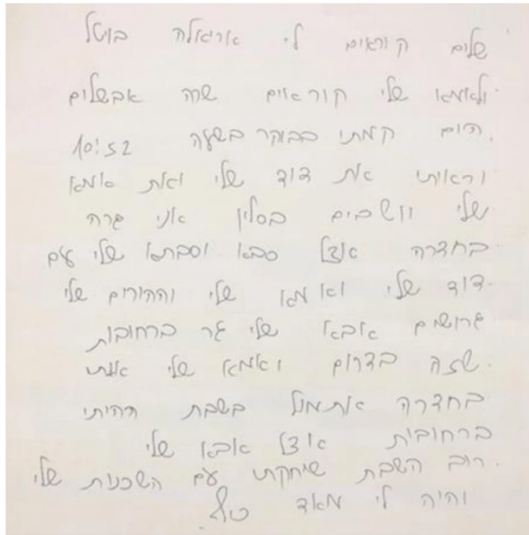
In order to investigate diagnostic methods for the early identification of ADHD, this study integrates five separate datasets. This includes handwriting analysis, drawing analysis, Natural Language Processing of mental health texts, EEG, and neuroimaging.

3.1.1 Handwriting Dataset

The handwriting dataset was collected to explore the potential of handwriting characteristics as indicators of ADHD. The dataset obtained from the study by Cohen *et al.* [71] consisted of handwriting samples from 49 teenagers (ages 13–18 years) attending a tertiary neurology and epilepsy center. Among them, 22 participants had a confirmed DSM-IV/V diagnosis of ADHD, while 27 served as controls. Each participant was instructed to write a 10-12 line story in Hebrew on a blank sheet of paper within a 20-minute period, using a standard blue pen. The collected samples were then analyzed by a licensed, blinded graphologist according to a predefined ADHD handwriting profile. Figure 3.1 illustrates representative handwriting samples without ADHD and with ADHD.

This profile encompassed 15 handwriting characteristics, such as text layout, margins, line direction, spacing between letters and words, handwriting slant, deviation, size, and width of letters, writing speed, and strength of graphism. Each abnormal characteristic identified in a sample earned one point, resulting in a graphology score ranging from 0 to 15. A cutoff score of 9 was used to classify handwriting as an indicative of ADHD [71].

Sample without ADHD



Sample with ADHD

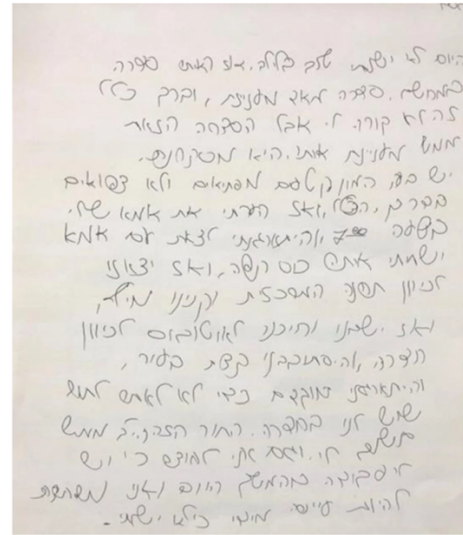


Figure 3.1 Handwriting samples of a girl without ADHD and with ADHD. Adopted from [71]

The dataset revealed that teenagers diagnosed with ADHD had significantly higher mean graphology scores (9.61 ± 3.49) compared to their non-ADHD counterparts (5.79 ± 4.01 , $p = 0.002$). Some noticeable abnormalities included issues with the margins, text layout, and spacing between letters, words, and lines. Additionally, the handwriting showed slant and irregularity, with letters formed in an unconventional style featuring flourishes and a flow not typical of standard Hebrew writing.

3.1.2 Drawing Dataset

The drawing dataset was utilized to investigate if pattern drawing characteristics are predictive indicators of ADHD. The dataset created by Shin *et al.* in Japan [72], consists of pattern samples from 29 Japanese children (14 with ADHD and comorbid ASD, and 15 neurotypical controls). All experiments were conducted under controlled circumstances on a pen tablet. Participants were placed in a seated position on a desk with flat feet and tablet centered, with an even viewing distance of approximately 40 cm from the screen.

Two structured drawing tests were given: zigzag lines (with a top angle of 70° , bottom angle of 80° , and side widths of 2.5 cm) and periodic lines (PL) made up of baseless alternate quadrangles (2 cm wide $\times$ 1.3 cm high) and isosceles triangles (70° vertex). Each of the designs was performed under two conditions: trace condition, where the whole design was revealed, and predict condition, where incomplete information was provided and the subjects were required to fill in the pattern, as illustrated in Fig. 3.2.

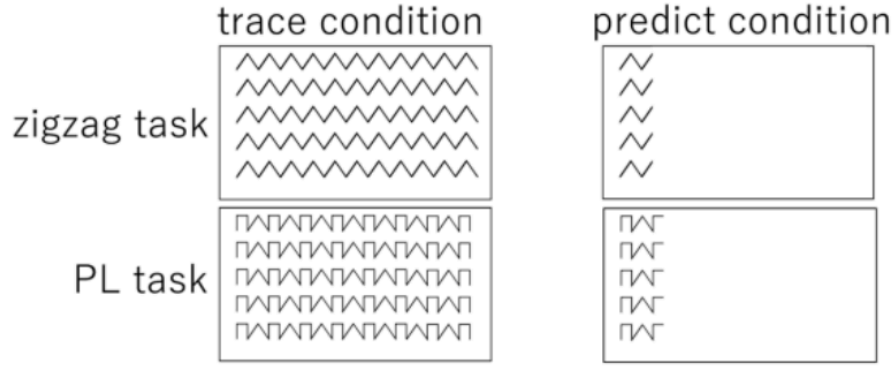


Figure 3.2 Handwritten drawing data collection procedure. Adopted from [72]

The experimental procedure involved three consecutive blocks. Each block included a 20-second rest period, a 30-second trace drawing task, a second 20-second rest, and finally a 30-second predictive drawing task. Each participant completed all tasks three times. A total of 30 features were extracted from the raw drawing data, encompassing spatial metrics (e.g., width, height, stroke length), kinematic variables (e.g., velocity, acceleration), pen pressure, grip angle, and error-based measures. Feature selection was performed using Sequential Forward Floating Search (SFFS) [72] to identify optimal subsets for machine learning classification. For the purpose of this research, only the zigzag trace dataset was utilized.

3.1.3 Reddit based Mental Health Dataset

The dataset used in this study is derived from Reddit, focusing on posts from both mental health-related and general-interest subreddits. The core dataset was originally compiled by Low *et al.* [73] using the Pushshift API, which

provides comprehensive access to historical Reddit data. Our corpus includes content from four mental health-focused subreddits, r/ADHD, r/anxiety, r/BipolarReddit, and r/depression, as well as eight general subreddits, such as r/meditation, r/fitness, r/legaladvice, r/parenting, and r/relationships. These general subreddits were selected to represent the “Non-Mental Health” class. In total, the dataset spans five distinct classes: ADHD, Anxiety, Bipolar, Depression, and Non-Mental Health.

During initial analysis, class imbalance was identified as a limitation, namely affecting the Bipolar classes, which had significantly fewer samples than the others. To rectify this imbalance, two augmentation techniques were tried: (1) back translation as a standard NLP technique by translating to a second language and translating back to the original language [74], and (2) supplementing the dataset with additional data from the original subreddit sources.

Back translation, while effective in increasing sample size, poses risks in this context. Specifically, it stands to undermine the semantic integrity and emotional tone of user-generated mental health text, variables of supreme importance for accurate classification in sensitive domains like psychological well-being. We opted for the latter approach for this reason. We used the dataset compiled by Kim *et al.* [28] to collect additional entries for the underrepresented classes, maintaining linguistic style, context, and subreddit-specific discussion. By doing this, we could keep the authenticity and validity of the dataset while approaching a more balanced class distribution.

Although demographic data are not available for each subreddit individually, Reddit’s overall user base is predominantly American (49.9%), male (67%), and young (22% aged 18–29; 14% aged 30–49) [75, 76]. These demographics should be considered when interpreting the findings of this dataset because it may not fully represent the overall ADHD population.

The inclusion of classes such as anxiety, depression, and bipolar disorder into a textual mental health classification model for ADHD and healthy controls is important since mental health disorders are overlapping and comorbid. ADHD often co-exists with other disorders, such as depression and anxiety, and disorders such as restlessness, lack of ability to focus, or mood swings can be

found in all of these disorders, making them hard to distinguish. By including these additional classes, the model learns to differentiate ADHD-specific patterns from others so that misclassification is less likely to occur. This step mirrors the process of diagnosis in mental health treatment, wherein practitioners must consider several conditions to arrive at an accurate diagnosis. Furthermore, inclusion of additional mental health disorders enhances the generalizability and interpretability of the model with respect to complex text data, which reflects the complexity and interrelatedness of mental health. The model becomes more clinically meaningful, more accurate, and better able to provide insight into an individual's overall mental health status.

The textual data from these communities and various populations provide rich and unique information regarding the lived experiences, conflicts, adaptations, and language practices of those who report living with ADHD. This provides a different perspective than that of the clinical and lab-based data used in this research. This real-world language data can be applied to studying how ADHD is present in ordinary conversation and self-report, potentially revealing less obvious behavioral indicators that are not apparent on extremely structured clinical tests.

3.1.4 EEG Dataset

The EEG dataset was obtained from IEEE DataPort and consisted of EEG recordings from 121 children in the age range 7-12 years [77]. The dataset contains 61 clinically diagnosed ADHD children and 60 control subjects with no psychiatric disorder, epilepsy, or high-risk behavior history. EEG recording was done according to the international 10-20 electrode placement system on 19 channels, with 128 Hz sampling of data to allow adequate capture of pertinent neurophysiological signals. The channels are Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T7, T8, P7, P8, Fz, Cz, and Pz, as illustrated in Fig. 3.3.

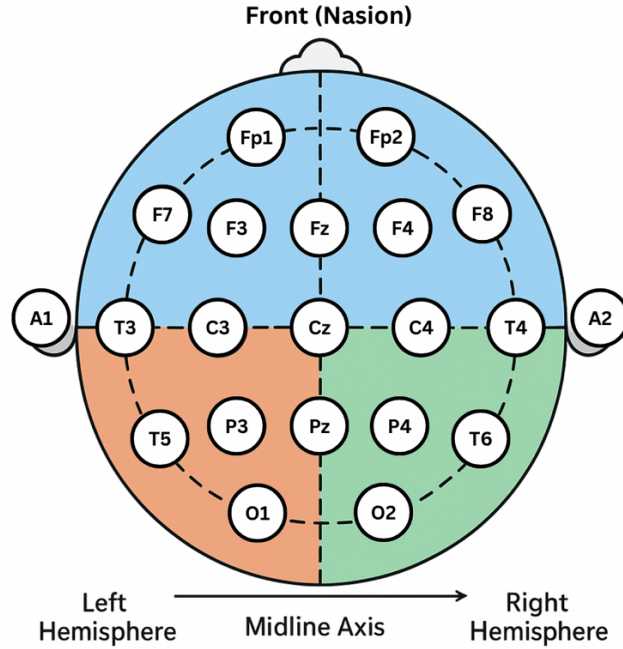


Figure 3.3 Electrode placement for 10-20 system

During the EEG recording trials, participants carried out a standardized visual attention task that comprised counting cartoon characters on an on-screen display. The task was particularly selected to assess sustained attention skills, a cognitive process typically impaired in ADHD. The recording time varied across participants because it followed their task performance, i.e., recordings were terminated when participants provided the response for the counting task. This recording length variable in itself can be employed as a performance measure, possibly reflecting control and ADHD group differences in processing speed and attentional capacity. The dataset also contains raw EEG signals along with task performance measures, enabling extensive analysis of both electrophysiological patterns and behavioral correlates of ADHD.

3.1.5 Neuroimaging Dataset

The neuroimaging dataset utilized in this study is the ADHD-200 Preprocessed Anatomical Dataset, which provides preprocessed three-dimensional structural MRI (sMRI) scans derived from the publicly available ADHD-200 dataset [78, 79]. This multi-site dataset comprises high-resolution sMRI scans of 776 subjects collected across eight imaging sites worldwide,

including 491 typically developing controls and 285 patients diagnosed with ADHD (categorized into hyperactive/impulsive, inattentive, and combined subtypes).

The structural scans were preprocessed using the FSL (Functional Magnetic Resonance Imaging of the Brain (FMRIB) Software Library) neuroimaging toolkit's script, which implements a standardized pipeline consisting of image reorientation to MNI standard space, automatic cropping, bias field correction to address intensity inhomogeneity, linear and non-linear registration, and skull-stripping to isolate brain tissue. However, Tissue-type and subcortical segmentation were not performed in this version.

Each preprocessed scan is provided in Neuroimaging Informatics Technology Initiative (NIfTI) open file format (.nii) and has been resampled to two standard resolutions: $128 \times 128 \times 128$ and $224 \times 224 \times 224$ voxels (we have considered using $224 \times 224 \times 224$ voxel as it ensures input compatibility with the convolutional neural network architecture such as ResNet50). Voxel intensities are normalized to have zero mean and unit standard deviation to ensure consistency across subjects. In addition to the imaging data, the dataset includes extensive phenotypic information such as age (ranging from 7 to 21 years), sex, handedness, IQ, and detailed ADHD symptom scores.

3.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a critical approach to data analysis that involves examining, investigating, and summarizing datasets to understand their main characteristics before applying formal statistical methods or preprocessing for preparing the data for our ML models. Therefore, it is essential for the effective preprocessing of diverse ADHD datasets, as it directly impacts data quality and subsequent analytical outcomes.

3.2.1 Handwriting Dataset

On the handwriting data set, EDA reveals important facts regarding graphology scores distribution to prevent wrongful standardization that could conceal the clinically significant cutoff of 9 distinguishing ADHD cases from non-ADHD cases. Lacking sufficient EDA, preprocessing would normalize

the very pattern that distinguishes the 22 ADHD subjects from the 27 controls, threatening diagnostic accuracy.

The data set contains 49 instances, with each instance consisting of 14 numeric features and one binary classification label indicating ADHD (label 1) or healthy control (label 0). Initial structural inspection revealed no missing values and no duplicate records. The data set consists of 22 ADHD patients and 27 controls, a relatively balanced division (approximately 44.9% ADHD and 55.1% control), as revealed through a pie chart in Fig. 3.4 that confirmed class balance for purposes of avoiding biased training of classification models.

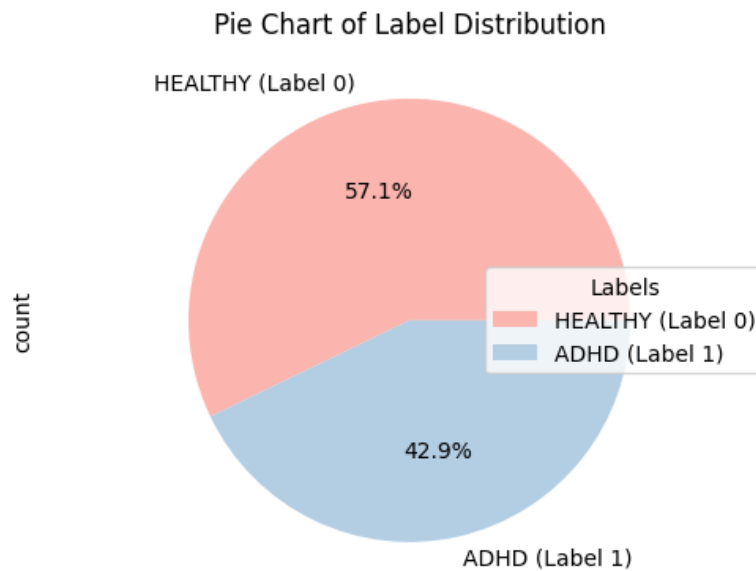


Figure 3.4 Label Distribution in Handwriting Dataset

Feature-wise, all the variables were numeric in nature, and there were no categorical predictors other than the class label. The features included a number of handwriting features including spatial organization (e.g., *text layout, margins*), spacing measurements (e.g., *lines spacing, words spacing, letters spacing*), letter specific features (e.g., *shape of letters, width of letters, letter deviation*), and dynamic measures like *speed of writing* and *strength of graphism*. To explore the relationships among these features, a feature

correlation heatmap was generated, as shown in Fig. 3.5, which illustrates potential dependencies, redundancies, and shared patterns across features.

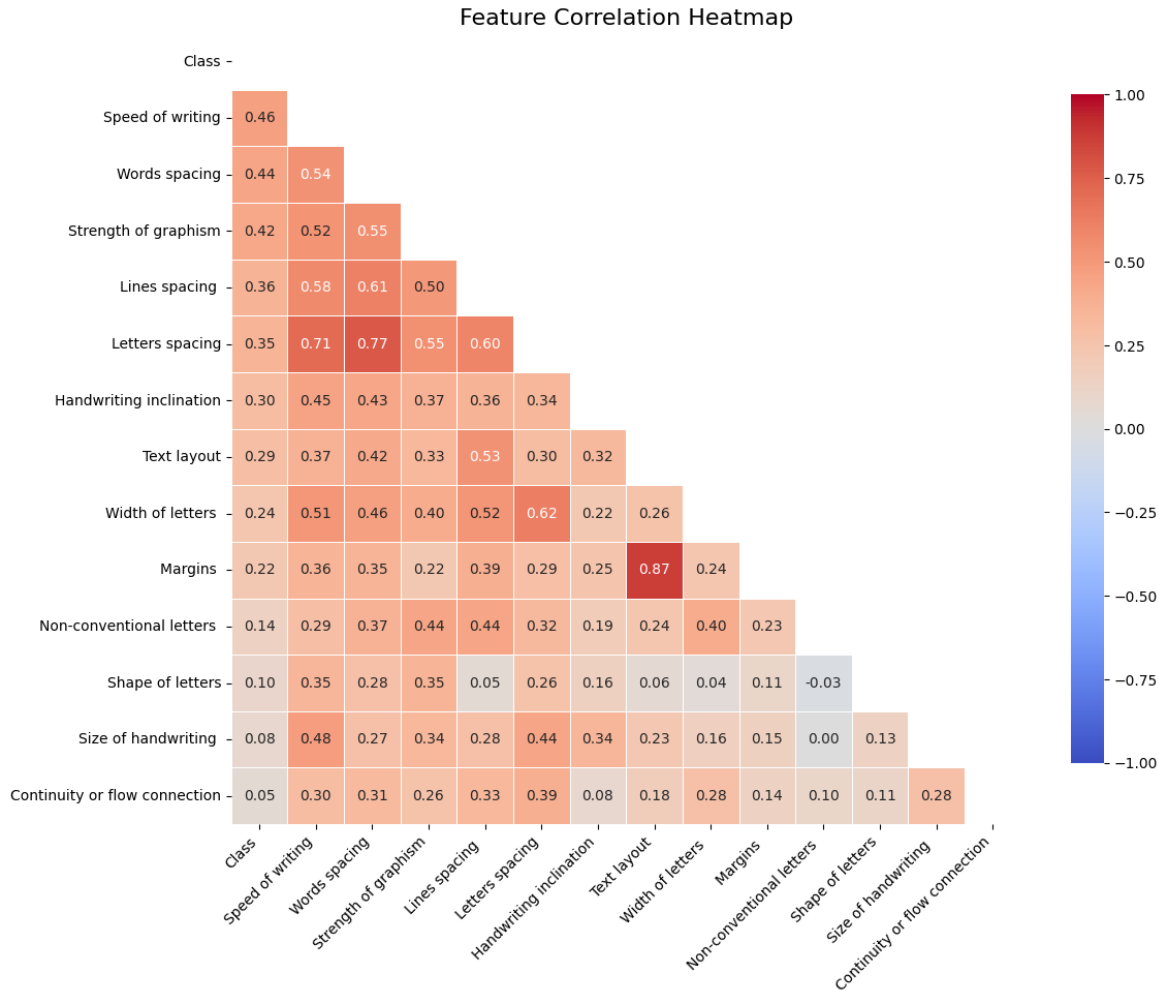


Figure 3.5 Feature Correlation Heatmap of Handwriting Dataset

Multicollinearity analysis using the Variance Inflation Factor (VIF) exposed significant feature redundancy. Specifically, nine features were identified with VIF values exceeding the commonly accepted threshold of 5, suggesting high linear dependence among predictors. The most collinear features included continuity or flow connection (VIF = 15.85), margins (VIF = 15.07), letters spacing (VIF = 14.05), letter deviation (VIF = 12.94), width of letters (VIF = 9.51), words spacing (VIF = 8.91), text layout (VIF = 7.48), speed of writing

(VIF = 6.75), and lines spacing (VIF = 6.62). This level of multicollinearity can hinder the interpretability and generalizability of predictive models, necessitating dimensionality reduction or feature pruning in subsequent preprocessing steps.

To identify the most diagnostically relevant features, a RF classifier was trained on the dataset, and feature importances were extracted based on the Gini impurity metric, as shown in Fig. 3.6. The top ten most influential features were found to be lines spacing (importance = 0.1199), words spacing (0.1076), shape of letters (0.1010), speed of writing (0.0973), handwriting inclination (0.0940), size of handwriting (0.0889), strength of graphism (0.0724), letters spacing (0.0607), width of letters (0.0553), and text layout (0.0507). These features collectively capture the spatial consistency, control, and fluency of handwriting, and cognitive motor dimensions that are known to be impacted in individuals with ADHD. Importantly, these findings offer preliminary validation of the clinical relevance embedded in this handwriting-based dataset.

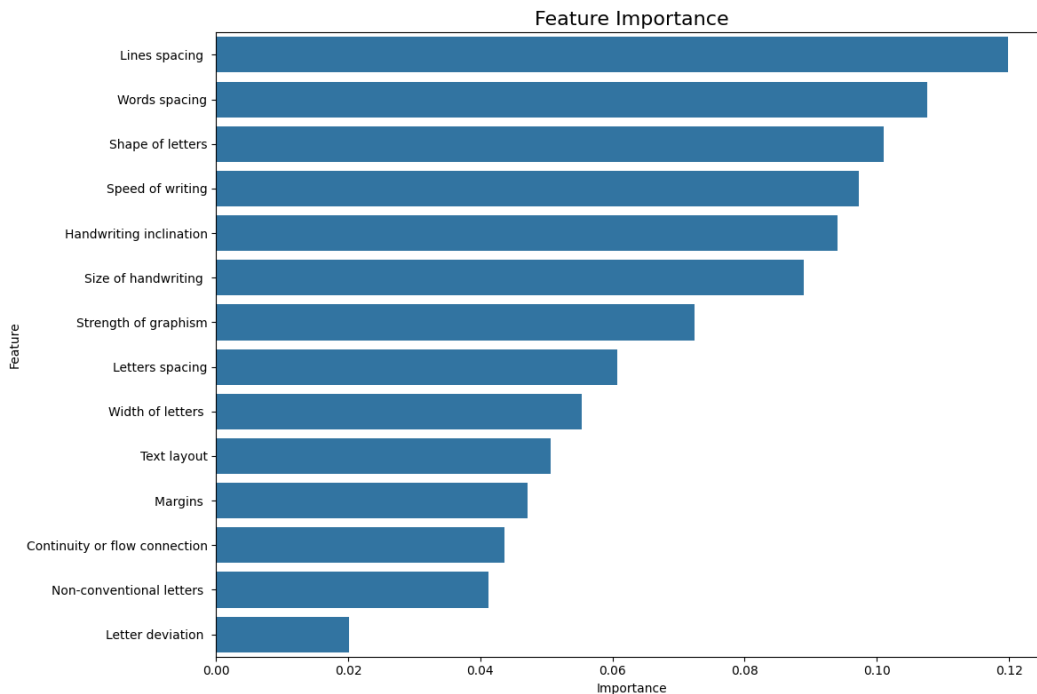


Figure 3.6 Feature Importance Analysis of Handwriting Dataset

To further explore underlying data structure and potential separability between the ADHD and control groups, PCA applied. The first two principal components accounted for 38.25% and 11.44% of the total variance, respectively, totaling approximately 49.69% of explained variance. When projected onto this reduced 2D space, the ADHD and control samples showed moderate visual separation, suggesting latent structure in the data that could be harnessed for classification tasks. PCA visualization, as illustrated in Fig. 3.7, also supports dimensionality reduction efforts, offering a means to capture most of the dataset variability while mitigating the effects of multicollinearity and noise.

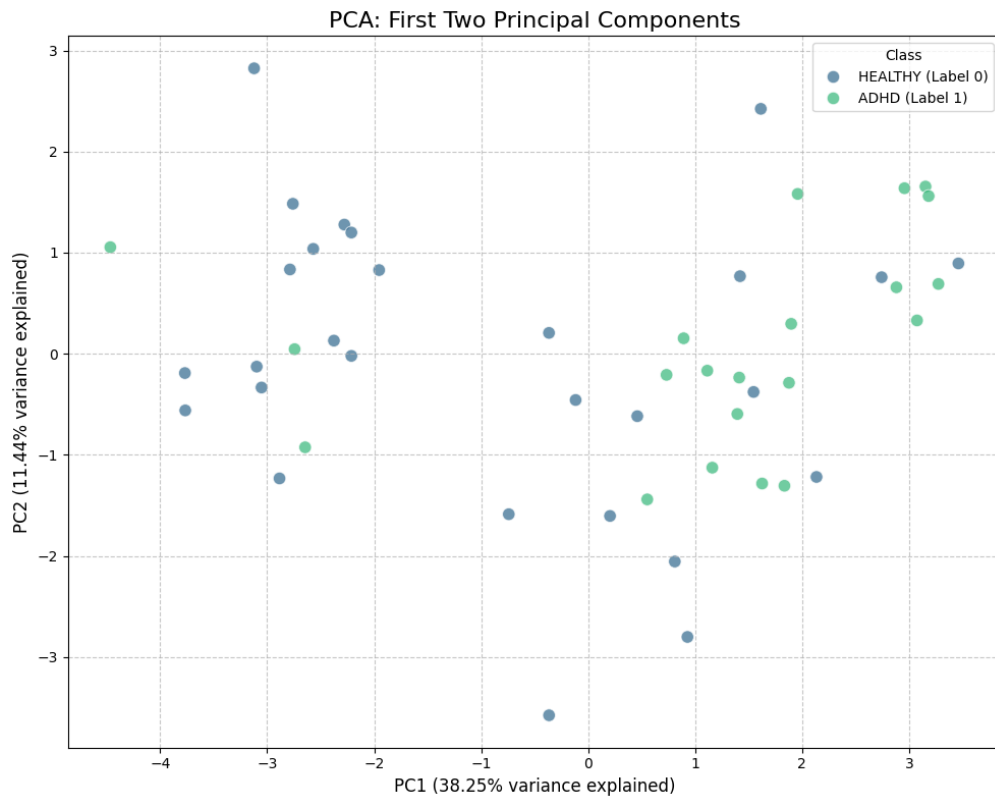


Figure 3.7 Sub-space using the first two Principal Components in the Handwriting Dataset

3.2.2 Drawing Dataset

In the drawing dataset, EDA plays a crucial role in distinguishing meaningful signal from noise across the 30 extracted features. This ensures that preprocessing retains discriminative information while reducing dimensionality and mitigating the risk of overfitting in subsequent models.

An initial EDA of the Zigzag Trace dataset was conducted to examine its structure, assess data quality, and explore the distribution of the target variable. The dataset consists of 87 instances and 31 attributes, including spatial, kinematic, and pressure-based features derived from participants' handwriting trajectories. Of these, 27 attributes are continuous numerical features (float64), representing measures such as velocity, angle variance, mean pressure, and grip angles. The remaining 4 attributes, along with the target class column, are integer variables (int64). Summary statistics and null-value checks confirmed that the dataset is complete, with no missing values, eliminating the need for imputation or additional data cleaning at this stage.

To evaluate the class distribution, frequency analysis of the Class variable was conducted. The result was a relatively balanced dataset with 45 samples labeled as HEALTHY (Class 0) and 42 samples labeled as ADHD (Class 1). This is favorable, as it reduces the likelihood of bias in training classification models. A pie chart, as illustrated in Fig. 3.8, was plotted for visualizing the class distribution, successfully demonstrating the nearly equal counts between both classes.

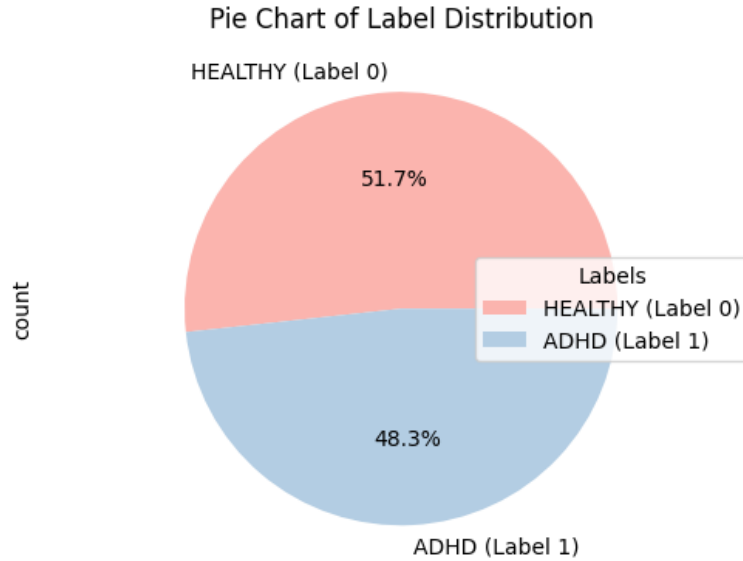


Figure 3.8 Label Distribution in Drawing Dataset

3.2.3 Reddit based Mental Health Dataset

The Reddit text data presents unique challenges that can be mostly addressed by proper EDA. The natural language from online communities contains platform-specific vernacular, abbreviations, and emoticons that require specialized preprocessing approaches. Unless the patterns are detected using EDA, regular NLP preprocessing pipelines could potentially remove the very linguistic features that characterize ADHD communication patterns.

The Reddit Mental Health Dataset consists of posts from four mental health-related subreddits (Anxiety, ADHD, Bipolar, and Depression) along with a control group of non-mental health content. An initial review of the dataset revealed a class imbalance: while the Anxiety, ADHD, and Depression categories each contained around 10,000 records, the Bipolar category included only 6,964 posts. To address this imbalance and ensure more reliable model training, a secondary dataset was incorporated to augment the Bipolar class and balance the overall distribution. The Fig. 3.9 provides with an overview of resulting class distribution.

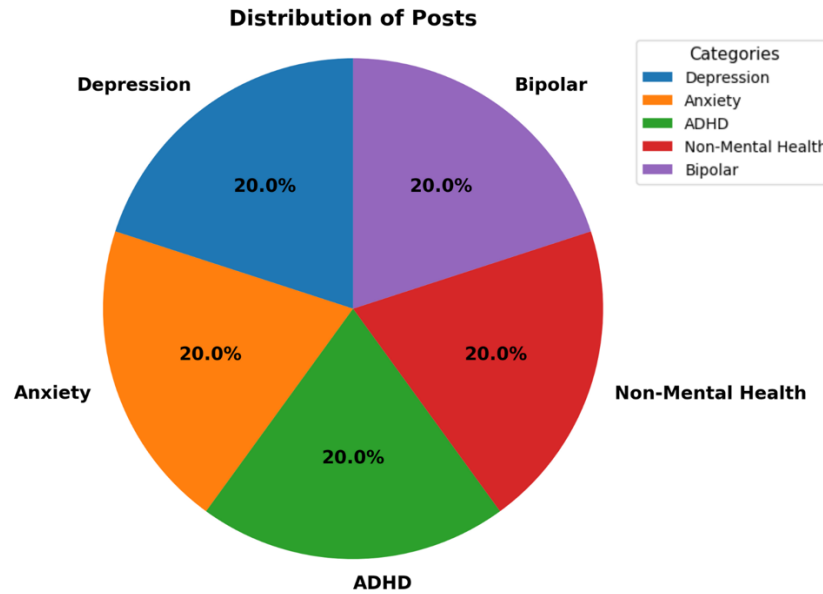


Figure 3.9 Class Distribution of Reddit Based Mental Health Dataset

An analysis of the average word count per post showed that Depression posts were the longest (172.91 words on average), followed closely by ADHD (171.79) and Anxiety (171.73), with Bipolar posts being slightly shorter (156.84). In contrast, posts from the non-mental health subreddit averaged 171.99 words. These results suggest that users posting about mental health conditions, particularly Depression, are more likely to post longer, more elaborate posts, potentially reflective of the complexity and emotional depth of these conditions. This is illustrated in Fig. 3.10, which shows samples from the Reddit-based mental health dataset.

----- Subreddit: Anxiety -----		
	text	subreddit
554	Why living being alone ? Hi guys , i have anxiety , a really bad anxiety , sometimes i'm afraid to walk on the street , talk to strangers , or I don't know what t...	Anxiety
----- Subreddit: Adhd -----		
	text	subreddit
2683	Stood up to my boss today. It did not go well... I work in manufacturing, and our jobs depend on how many plates we can move. I got an "M" on my review last year ...	Adhd
----- Subreddit: Bipolar -----		
	text	subreddit
9813	Musing about self-identity, I ponder the bipolar paradox: did I think/feel/act that way just now because of who I am as a person, or was that a bipolar symptom th...	Bipolar
----- Subreddit: Depression -----		
	text	subreddit
8008	To cut the long story short, I've been depressed for 4 years nearly 5. Over the past 4 weeks, I've become a lot worse, I have anxiety, developed insomnia and I ca...	Depression
----- Subreddit: Non_Mental_Health -----		
	text	subreddit
1247	I went to the Cinema yesterday... ..I was buying popcorn, a drink and some sweets. As I went to pay I said, "I'm ever so sorry, I've only got a £50 note." Th...	Non_Mental_Health

Figure 3.10 Samples of Reddit Based Mental Health Dataset

When analyzing word frequencies, sentiment, and contextual language, several patterns emerged across the mental health subreddits. Word cloud visualizations highlighted the prevalence of emotionally charged and introspective terms such as “feel,” “time,” “know,” “just,” along with mental health-specific words like “Anxiety,” “ADHD,” and “Bipolar.” These findings underscore a consistent focus on personal experiences and emotional states within these communities.

Further analysis using TF-IDF revealed considerable overlap in high-ranking terms across all mental health subreddits. Common words like “know,” “feel,” and “like” appear frequently, indicating shared modes of expression and recurring concerns among users, regardless of their specific mental health condition. Fig. 3.11 illustrates this by showing more frequently occurring words in larger fonts.

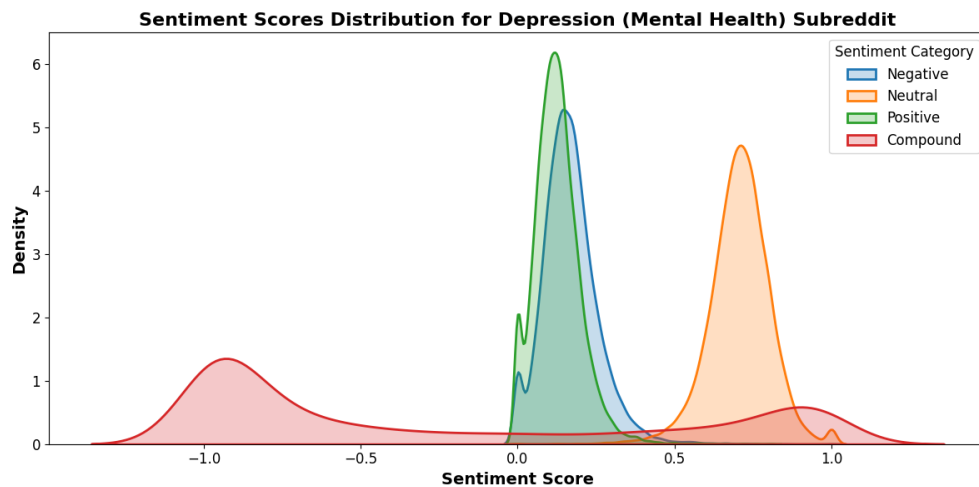
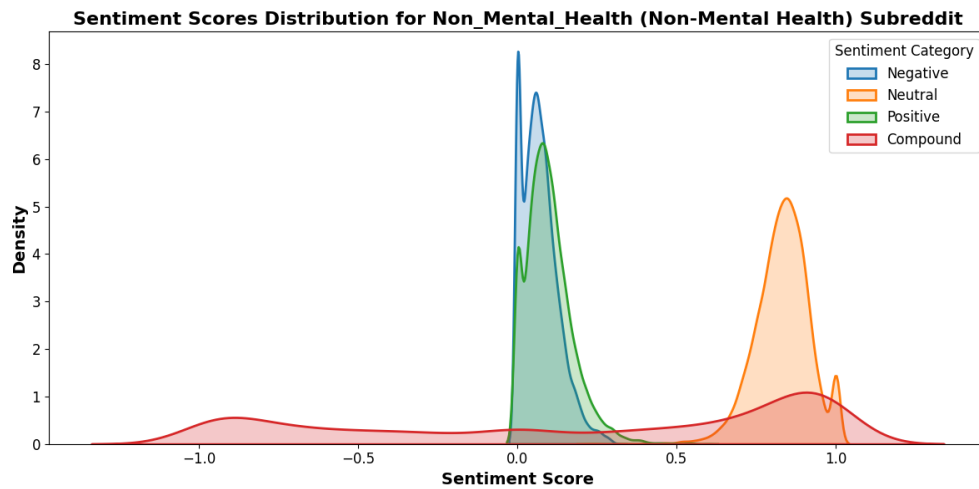


Figure 3.11 Word Cloud of Reddit Based Mental Health Dataset

Sentiment analysis provided additional insight into the emotional tone of these posts. While positive words such as “help,” “friend,” and “love” were present in both mental and non-mental health subreddits, mental health posts more often featured intense negative terms like “anxiety,” “panic,” “depression,” and “die.” This reflects the emotional weight of the topics being discussed and the personal vulnerability often expressed in these spaces. In contrast, the non-mental health subreddit exhibited a more balanced emotional tone with fewer extreme emotional terms.

Sentiment analysis, conducted using VADER, as illustrated in Fig. 3.12, showed that mental health posts were more likely to be of negative or neutral sentiment compared to non-mental health posts. Specifically, the Anxiety

subreddit had the highest average negative sentiment score (0.1536), followed by Depression (0.1672) and Bipolar (0.1197). Compound sentiment scores were also similarly trending, with Anxiety at -0.2807, Depression at -0.2654, and Bipolar at -0.0969, all of which are significantly lower than the mean for the non-mental health subreddit of 0.1845. Neutral sentiment was still relatively high in mental health posts (e.g., Anxiety at 0.7398, ADHD at 0.8038), but non-mental health posts had a more balanced overall tone, with a higher neutral mean (0.8315) and more middling positive sentiment (0.0985).



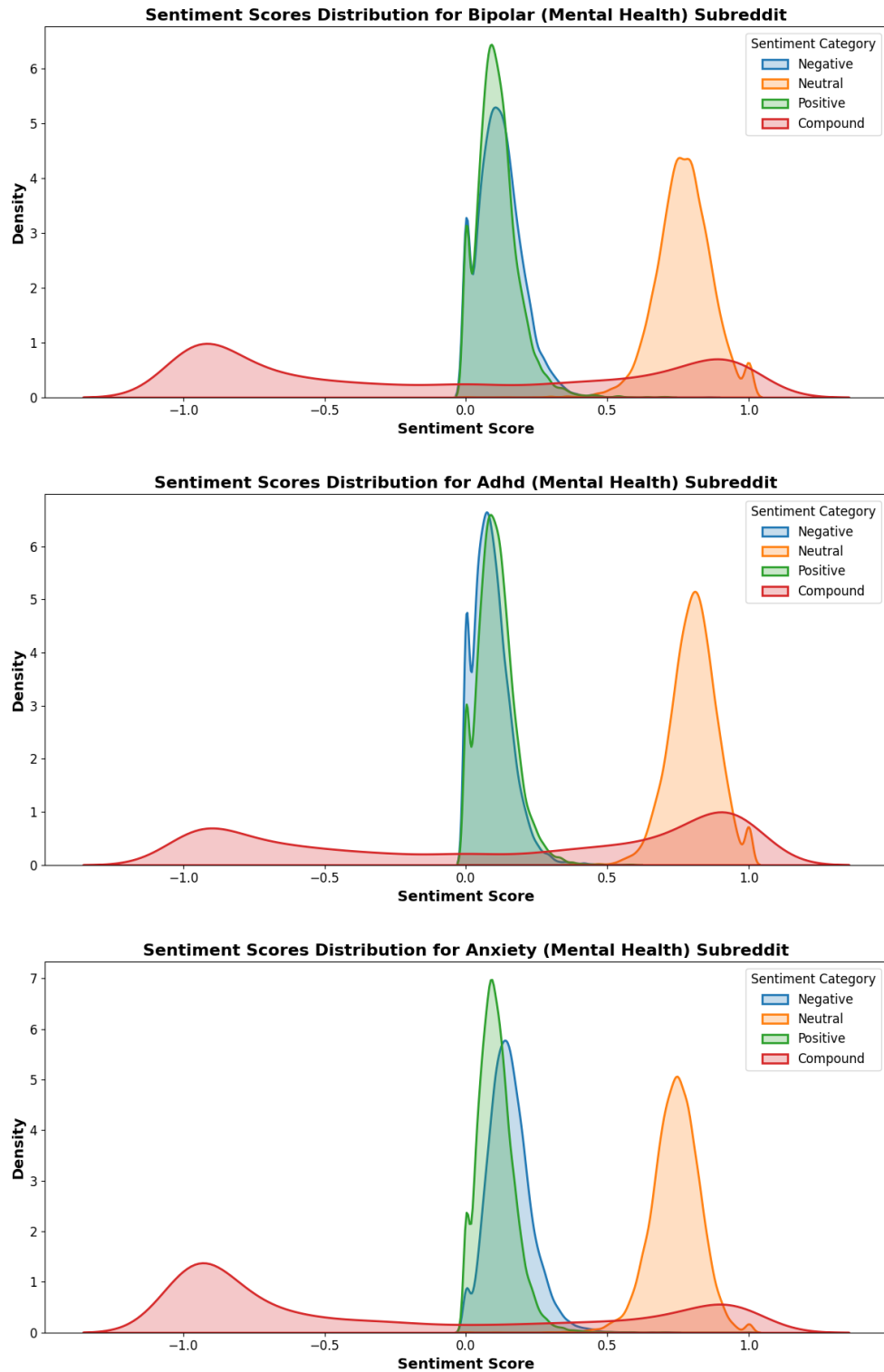
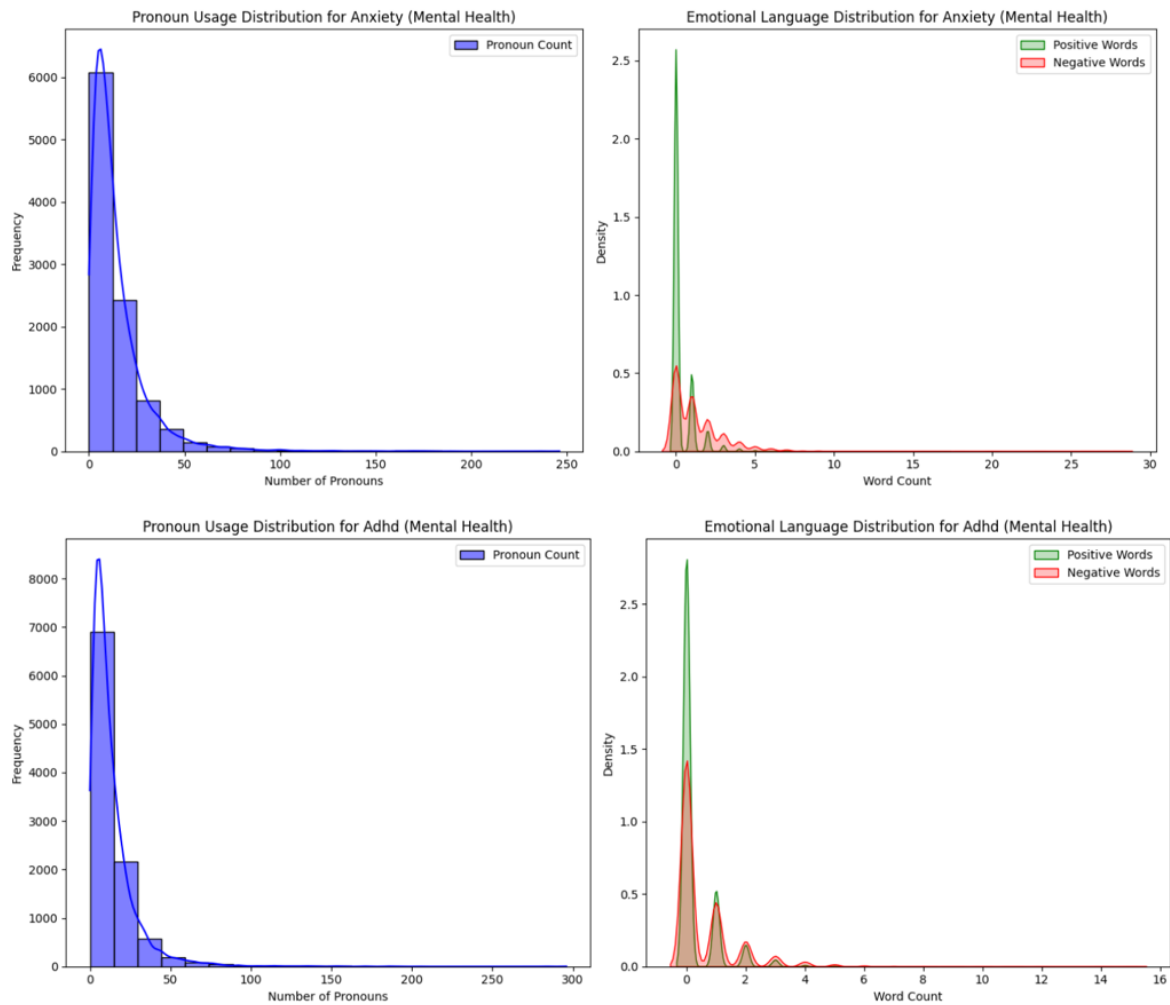


Figure 3.12 Sentiment Analysis of Reddit Based Mental Health Dataset

Readability metrics were calculated using established indices, including Flesch-Kincaid and SMOG. It revealed that posts in mental health communities generally trend toward simpler, more accessible language. This may reflect users’ efforts to clearly articulate emotional experiences or to engage a broad, supportive audience. Finally, linguistic feature analysis focusing on pronoun use and emotional vocabulary revealed that mental health subreddits contained a higher frequency of personal pronouns and negative emotion words. This suggests that content in these communities is not only introspective but also emotionally expressive, aligning with expectations about the vulnerable and supportive nature of discussions within mental health spaces. These patterns are illustrated in Fig. 3.13.



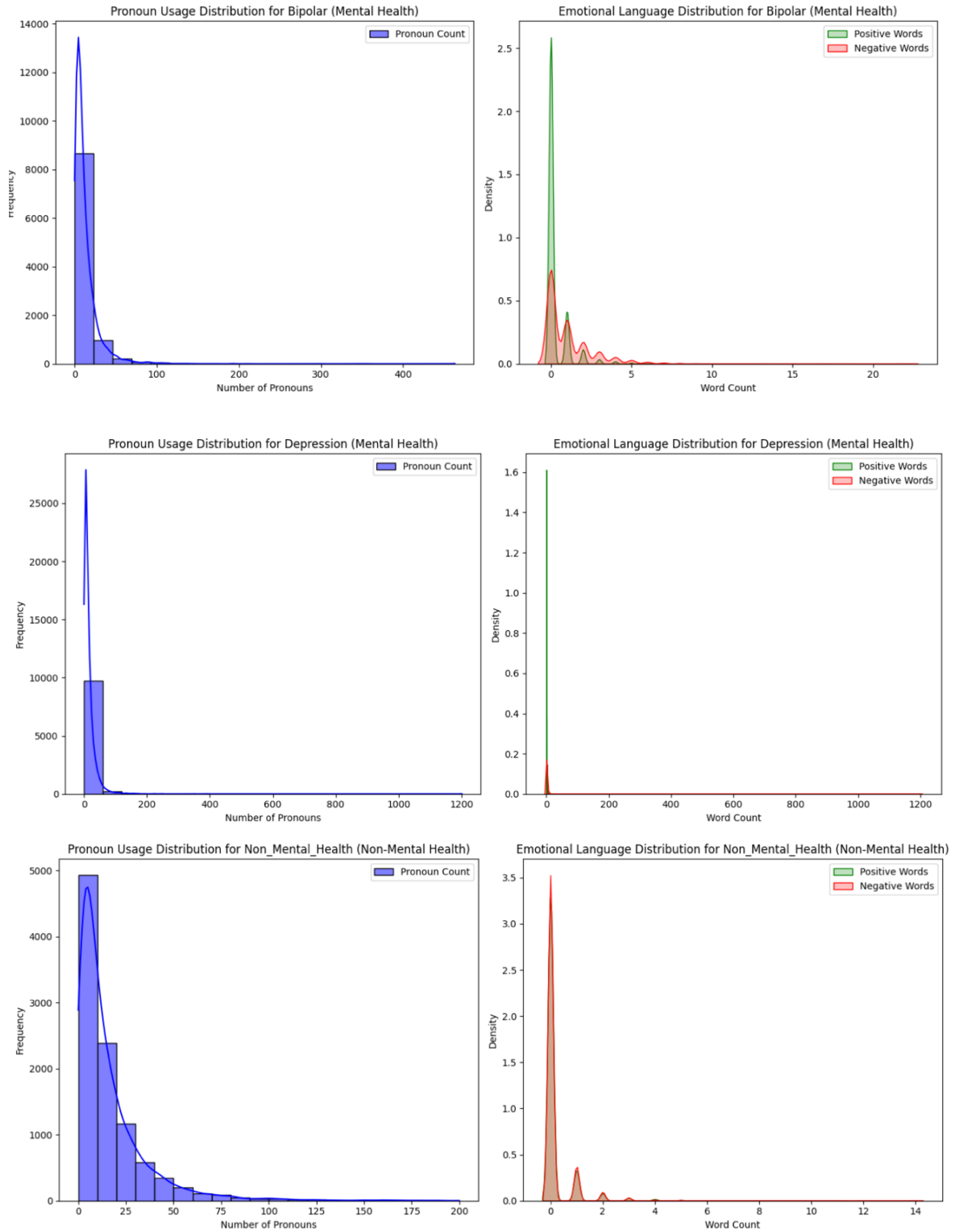


Figure 3.13 Pronoun Usage and Emotional Language Analysis of Reddit Based Mental Health Dataset

3.2.4 EEG Dataset

For the EEG dataset, EDA is necessary for detecting artifacts from movement or equipment issues, as children with ADHD typically exhibit more motor activity during recording sessions. The failure to identify these artifacts would lead to contaminated signal preprocessing and ultimately flawed neurophysiological interpretations.

The EEG dataset consists of 2,166,383 rows and 21 columns, reflecting brain activity recorded on 19 EEG channels at 128 Hz sampling. One row corresponds to one timestamped EEG sample. Raw “.mat” files were cleaned using a custom function that extracted multichannel EEG signals for all participants during the data preparation for analysis. These signals were normalized from raw NumPy arrays to organized pandas DataFrames, with standard channel labels. Each sample was assigned to a group label (0 for Control and 1 for ADHD) and assigned a unique 6-digit Person ID to ensure consistent sample identification. This processing was performed independently to the ADHD and control groups to enable clean integration. By this process, we confirmed the existence of 121 distinct participants, with 61 having ADHD and 60 being classified as healthy controls. The dataset is sufficiently balanced, which is absolutely critical for reducing bias in future machine learning models. Preliminary checks also revealed no significant missing data and equal signal ranges across channels, further confirming the quality of the dataset for statistical inference and model training.

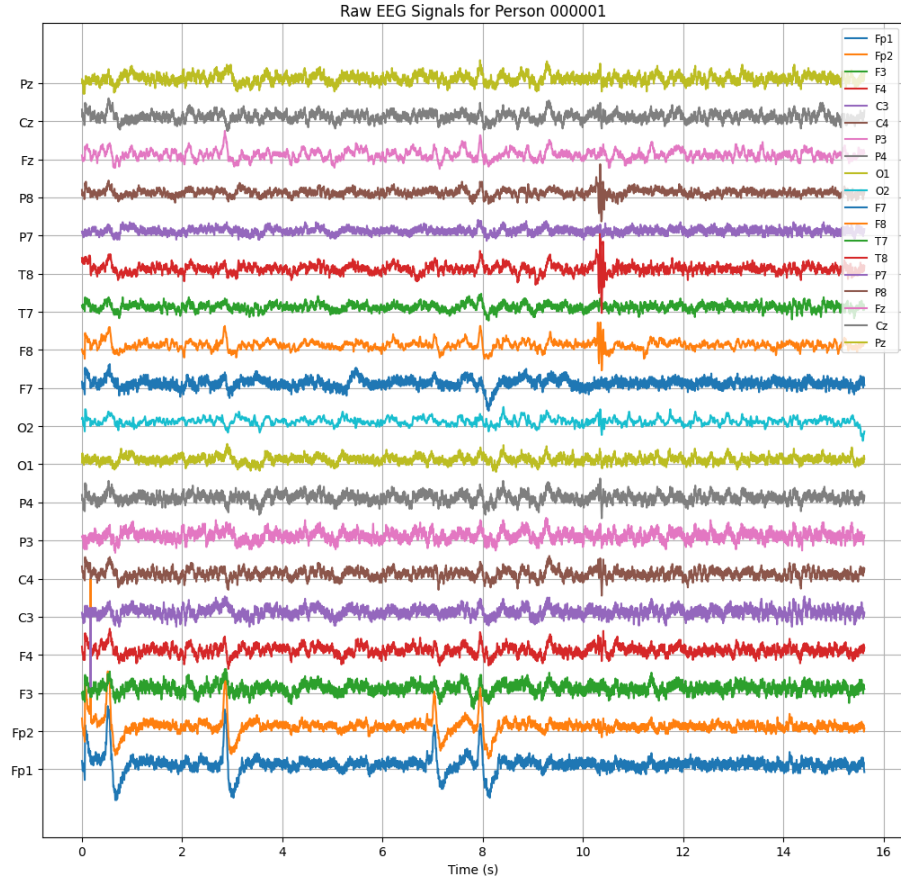


Figure 3.14 Raw EEG Signals for Person ID 000001 (ADHD group)

The Fig. 3.14 above shows the raw EEG signals from a sample (Person ID 000001) across 19 standard electrode positions of the international 10–20 system. Each channel is a plot of voltage as a function of time in one scalp region, in units of microvolts. The horizontal axis is time in seconds (to 16 seconds), and vertical stacking allows channels to be viewed visually. Despite the overall prevalence of biological and environmental noise, certain neural patterns and signal fluctuation can be observed between different brain sites. The Fp1, Fp2, and F7 channels exhibit higher fluctuations, likely corresponding to frontal lobe activity, which is usually of interest in studies on ADHD since it is responsible for attention and executive function. The signals appear synchronized and steady, indicating good quality recording, and are the reference input for the following preprocessing steps such as filtering, artifact rejection, and feature extraction.

3.2.5 Neuroimaging Dataset

The ADHD-200 neuroimaging dataset incorporates data from eight international imaging sites with varying scanning protocols. Without thorough EDA to identify and quantify such site-specific differences, the application of standardized preprocessing procedures can instill systematic bias rather than removing it. So, it is important that through EDA, we uncover the inherent data structure, quality, and nature of each location’s data so that evidence-based preprocessing decisions can be made in comparison to assumptions. This prevents data quality anomalies from worsening during feature extraction and classification, ultimately preserving the integrity and diagnostic quality of conclusions derived from this complex but valuable dataset.

To gain an initial understanding of the ADHD-200 preprocessed anatomical dataset, we first aggregated subject-level phenotypic metadata from four acquisition sites:

- Kennedy Krieger Institute (KKI)
- New York University Child Study Center (NYU)
- NeuroIMAGE Sample (NeuroIMAGE)
- Oregon Health & Science University (OHSU)

For each site, we filtered the original metadata to retain only participants with corresponding scan directories and anatomical images. These cleaned, site-specific metadata were then combined into a single dataset containing only subjects with valid anatomical scans. We also added diagnostic labels by mapping the original DX variable into two interpretable formats. One categorical label was added (Typically Developing or ADHD), and the other was a binary encoding (0 for Typically Developing, 1 for ADHD). This ensured consistent and clear diagnostic grouping across all sites. After this, we examined the distribution of subjects by diagnostic group and acquisition site. There were 214 counts of Typically Developing (group 0) and 198 counts of ADHD (group 1) participants in the combined dataset.

To visually inspect the anatomical scans, representative mid-sagittal, coronal, and axial slices were extracted from selected participants to verify the

structural integrity of the preprocessed brain images. All slices have been bias-corrected, resampled, and intensity-normalized to enhance visual consistency across individuals and sites. These views, as illustrated in Fig. 3.15 and 3.16, allow for an intuitive inspection of brain tissue boundaries, alignment, and artifact removal, which confirms that the preprocessing pipeline preserves essential anatomical features.

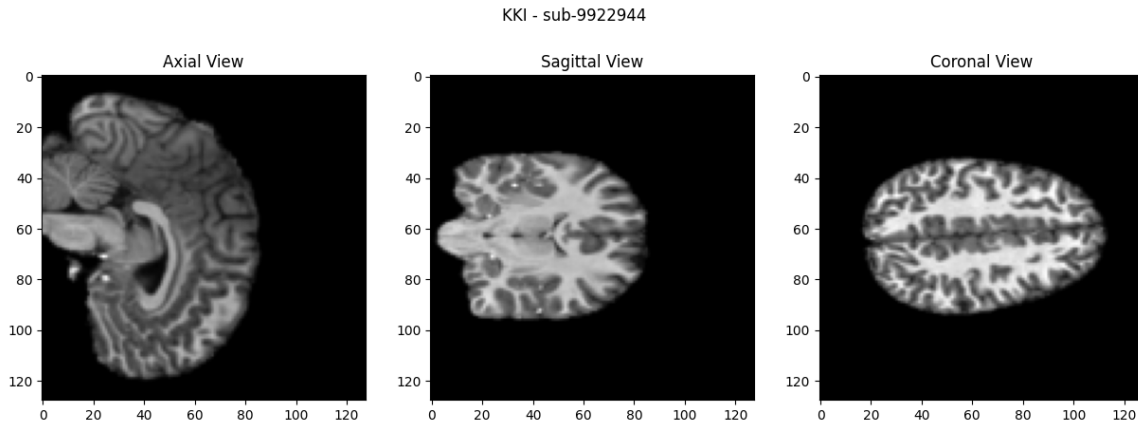


Figure 3.15 MRI Samples of participants from Typically Developing Group

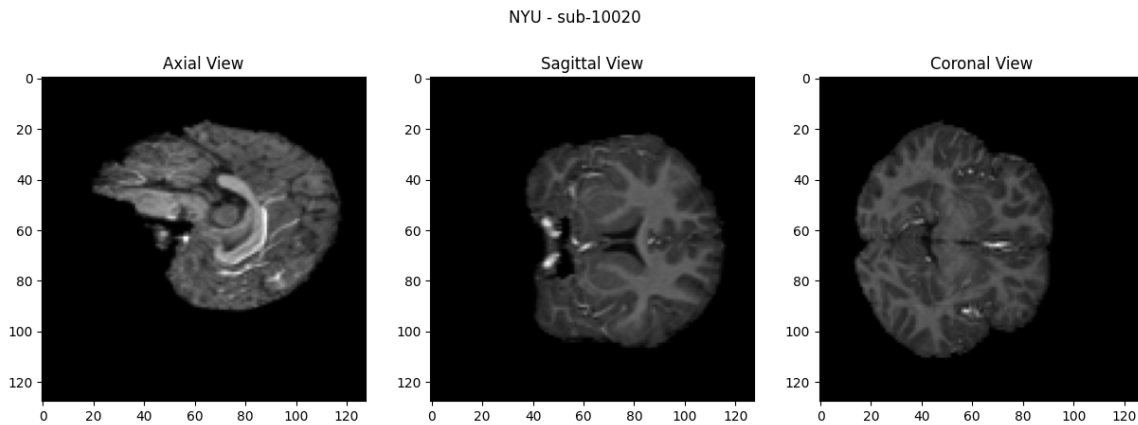


Figure 3.16 MRI Samples of participants from ADHD Group

In addition, we provide five equally spaced slices from each anatomical plane (axial, sagittal, and coronal) of a single bias-corrected and normalized brain scan, as shown in Figs. 3.17 and 3.18, which enables close examination of tissue continuity, anatomical landmarks, and image quality throughout the

whole brain volume and ensures that no significant regions are lost or deformed due to resampling.

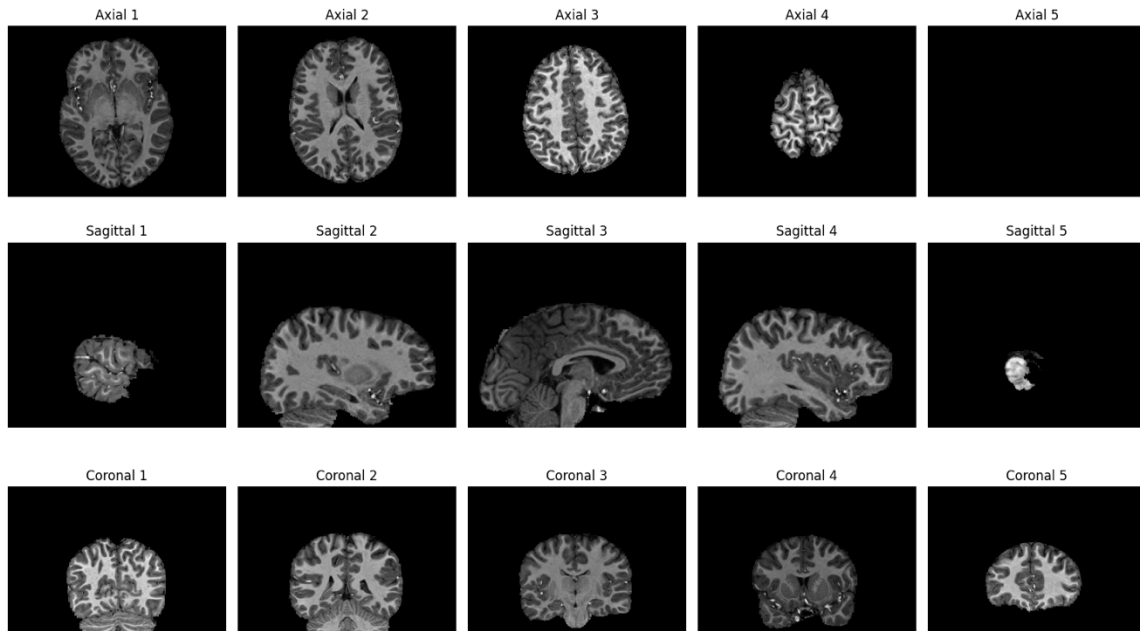


Figure 3.17 Five evenly spaced Axial, sagittal, and coronal mid-plane slices from Participant of Typically Developing group in ADHD200 Preprocessed Dataset

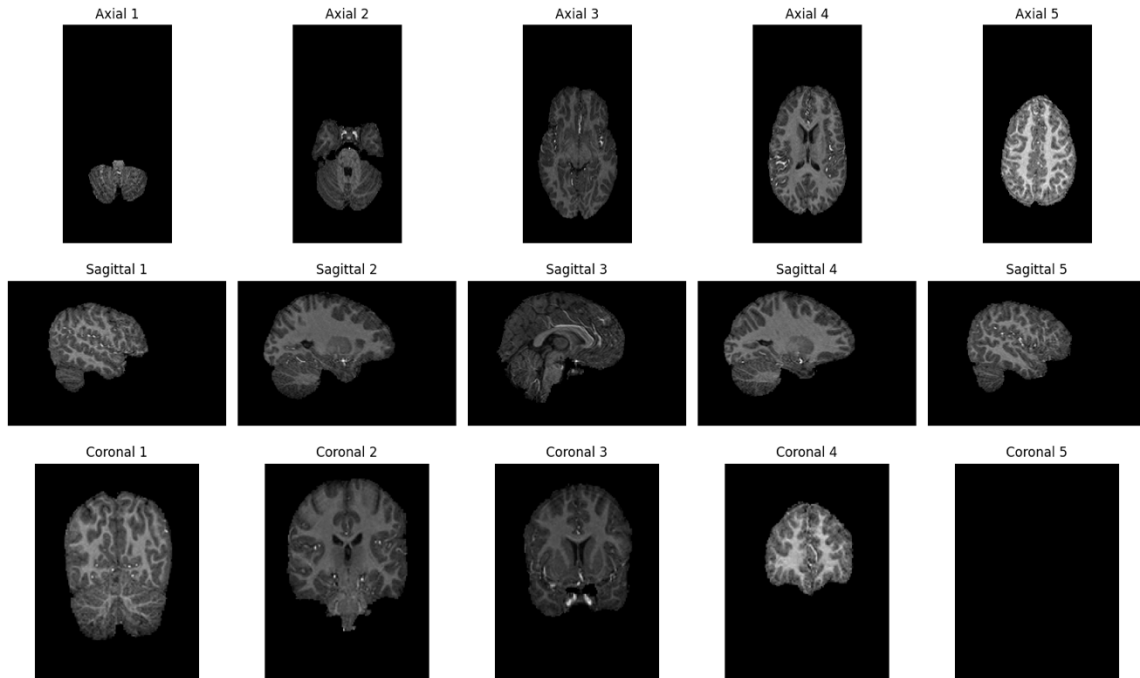


Figure 3.18 Five evenly spaced Axial, sagittal, and coronal mid-plane slices from Participant of ADHD group in ADHD200 Preprocessed Dataset

Following EDA and quality checks, we compiled a curated dataset of sagittal view slices from each subject's bias-corrected anatomical scan, sampling five equally spaced slices to provide a representative cross-section of structural brain anatomy; a sample is shown in Fig. 3.19. The sagittal view was chosen for its clear visualization of midline structures such as the corpus callosum, cerebellum, and brainstem, which are often of particular interest in neurodevelopmental studies.

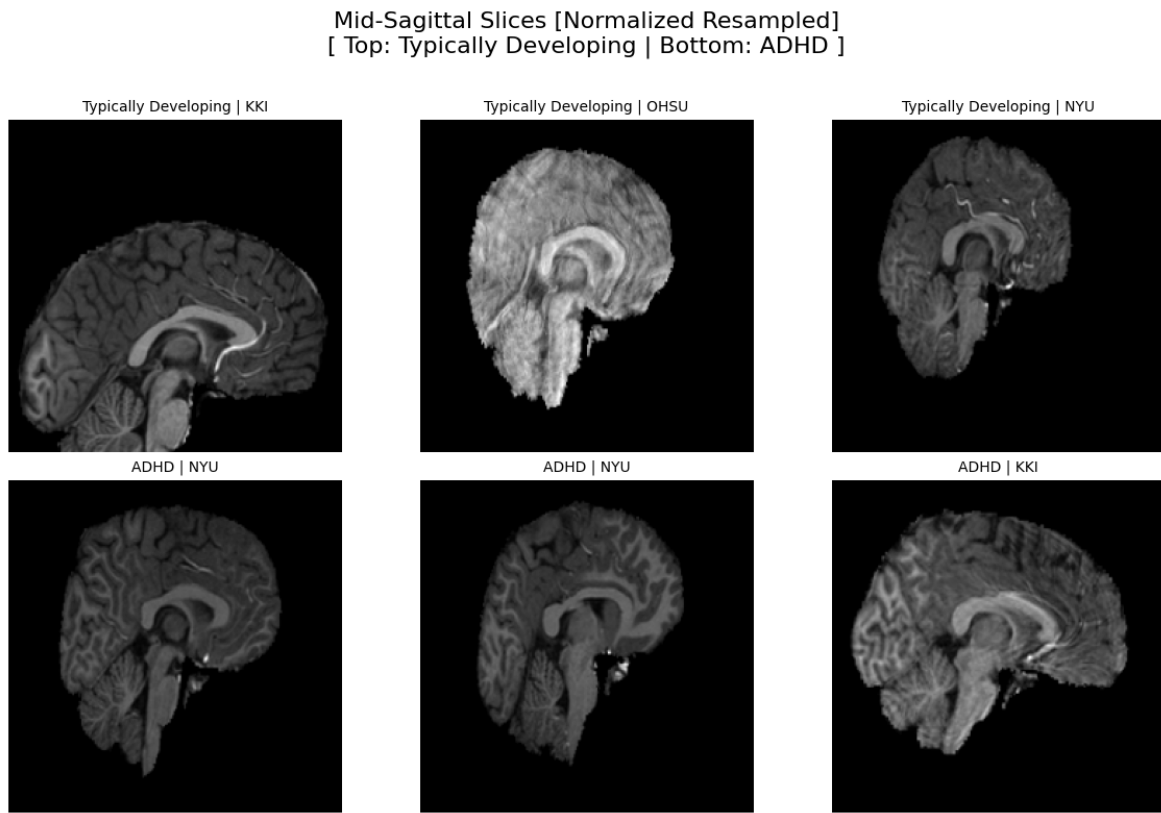


Figure 3.19 MRI Samples from Curated Dataset of Sagittal View Slices

This view also gives equal alignment across subjects, which is well-paired for downstream modeling. Overall, the resulting dataset is N sagittal slices from M participants ($5 \text{ slices} \times M \text{ participants}$), providing an anatomically meaningful and standardized image set for classification and analysis tasks.

3.3 Pre-Processing

Preprocessing is a crucial step in training any ML model, as it prepares raw data into a suitable format for analysis. However, it becomes more critical when dealing with biomedical data that may suffer from imbalances, noise, and small sample sizes. This step involves cleaning the data by handling missing values, removing duplicates, and correcting inconsistencies, which ensures that the model learns from accurate and reliable input features. The transformation techniques such as normalization, scaling, and encoding are applied to convert data into numerical forms that algorithms can effectively process. Moreover, by reducing noise and irrelevant information, preprocessing not only speeds up training but also increases model accuracy and generalizability.

3.3.1 Handwriting Dataset

A structured preprocessing pipeline was developed to prepare the handwriting dataset for ML based ADHD detection. The pipeline was designed to address the challenges of limited data size, class imbalance, feature redundancy, and high dimensionality.

The original dataset consisted of 49 samples and 14 categorical handwriting attributes, along with the binary target variable Class, where Healthy was represented by 0 and ADHD by 1. An early analysis indicated a moderate class imbalance, with 28 samples labeled as Healthy (Class 0) and 21 samples labeled as ADHD (Class 1). Importantly, there were no missing values, allowing the pipeline to proceed without imputation. The recorded handwriting attributes included text layout, margins, line spacing, word spacing, letter spacing, handwriting inclination, letter deviation, size of handwriting, use of non-conventional letters, width of letters, continuity or flow connection, shape of letters, speed of writing, and strength of graphism.

To reduce feature redundancy, Pearson correlation coefficients were calculated for all feature pairs. The correlation map is illustrated in Fig. 3.5. Features with correlation coefficients greater than 0.8 were considered highly correlated. As a result, the feature “Margins” was removed due to its strong correlation ($r = 0.869$) with Text layout, reducing the feature space to 13

independent variables. This step helped improve model interpretability and reduce the risk of multicollinearity during training.

Since the dataset was imbalanced, the Synthetic Minority Oversampling Technique (SMOTE) was applied to generate synthetic samples of the minority class (ADHD). This procedure resulted in a perfectly balanced dataset of 56 records, with 28 samples in each class. To identify the most informative predictors, feature selection was performed using the Chi-Square test, which measures the statistical dependency between each feature and the target variable. A significance threshold of $p < 0.05$ was applied, meaning that features with p-values below this level were considered significantly associated with the target class. In this dataset, all 13 features satisfied the threshold and were retained, confirming their overall relevance to the classification task. Importantly, some features exhibited particularly strong discriminative power, suggesting a greater ability to distinguish ADHD cases from healthy controls. The results of the chi-square test are presented in Fig. 3.20 and top features include:

- Words spacing ($\chi^2 = 11.80$, $p < 0.001$)
- Handwriting inclination ($\chi^2 = 6.53$, $p \approx 0.011$)
- Lines spacing, Letters spacing, Speed of writing

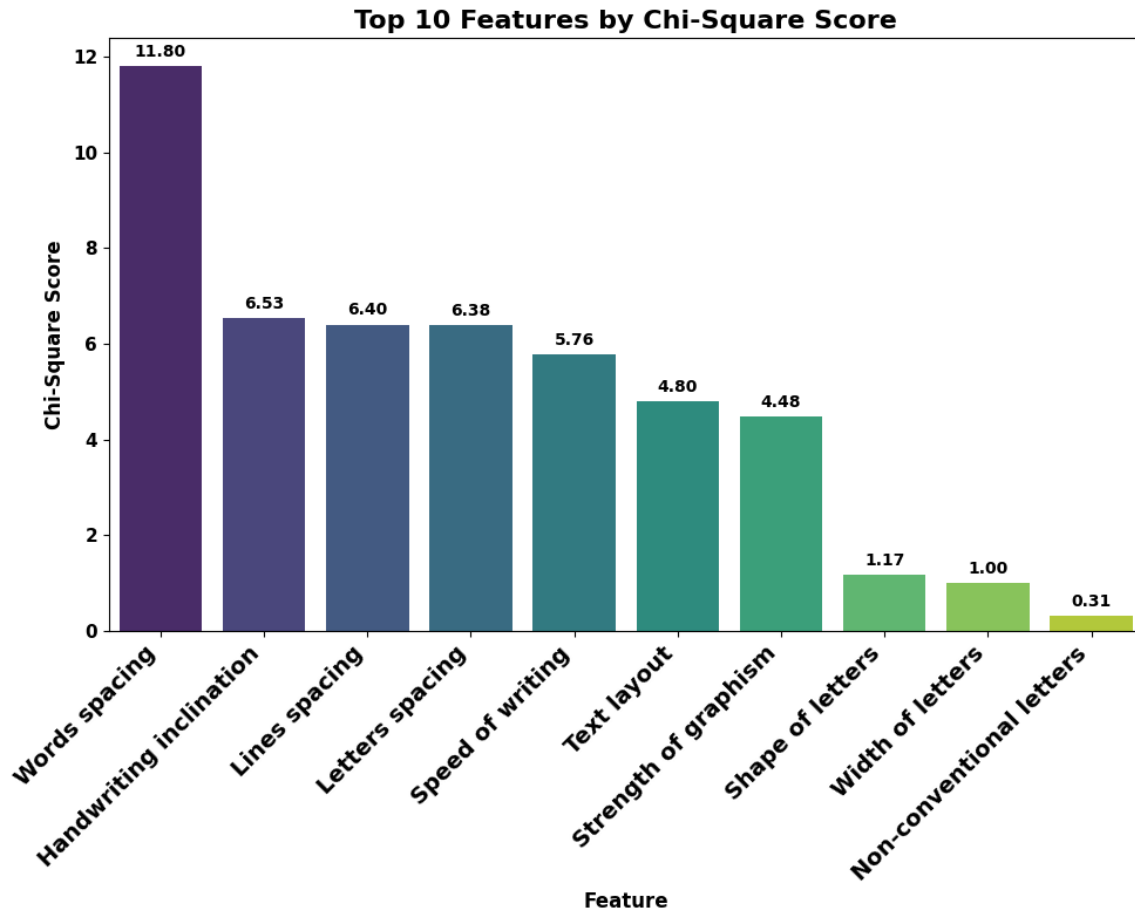


Figure 3.20 Top Features by Chi-Square Score in Handwriting Dataset

The balanced dataset was then divided into training and testing subsets using stratified sampling to preserve class proportions. An 80/20 split was applied, producing 44 training samples and 12 testing samples. Before dimensionality reduction, z-score standardization was performed to normalize feature scales, ensuring that all features contributed equally during subsequent analysis. Finally, Principal Component Analysis (PCA) was applied to reduce the dimensionality of the dataset while retaining most of the information. As shown in Figure 3.21, the cumulative explained variance indicates that the first 10 principal components account for approximately 95% of the total variance. Therefore, the original feature space of 13 features was effectively reduced to 10 components, compressing the data and simplifying the subsequent ML pipeline without significant loss of information.

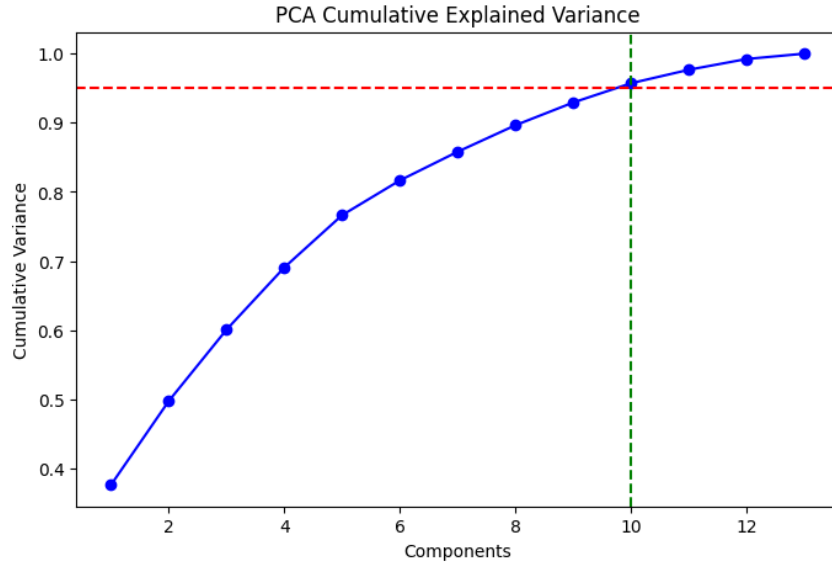


Figure 3.21 Cumulative Explained Variance of Principal Components in Handwriting Dataset

3.3.2 Drawing Dataset

A comprehensive preprocessing strategy was implemented to enhance the model performance and ensure robustness in ADHD detection based on drawing-derived features. The dataset includes 30 statistical features designed to differentiate the handwriting patterns of children with ADHD and co-occurring ASD from those of healthy children. These features were derived from six raw measurements: x and y coordinates (in pixels), drawing speed, pen pressure, and pen tilt (both horizontal and vertical angles) [72].

It is composed of 87 samples (45 Healthy and 42 ADHD), was partitioned into feature vectors ($\mathbf{X}_i$) and target labels (y_i), revealing a relatively balanced yet small dataset. To address this, synthetic data was generated through a custom augmentation method: continuous features were perturbed with uniform random noise within the range of $[-0.5, 0.5]$, while categorical features were resampled from empirical distributions, resulting in 20 new synthetic samples.

Despite this augmentation, slight class imbalance persisted, which was resolved using the SMOTE, equalizing the number of samples in both classes

to 55. Following augmentation and balancing, numerical features were standardized to ensure zero mean and unit variance, a crucial step for scale-sensitive algorithms.

The original drawing dataset consisted of 87 records, with 45 belonging to the stable (majority, Class 0 - Healthy) group and 42 to the unstable (minority, Class 1 - ADHD) group. After applying synthetic data generation, the dataset increased to 107 records, comprising 55 stable and 52 unstable cases. Finally, following the application of the SMOTE technique, the dataset reached a balanced distribution of 110 records, with 55 stable and 55 unstable instances.

3.3.3 Reddit based Mental Health Dataset

The preprocessing phase involved a comprehensive pipeline designed to prepare the Reddit-based Mental Health Dataset for downstream ML tasks. The process began by defining five target categories based on Reddit subreddit sources, namely ADHD, Anxiety, Bipolar, Depression, and Non-Mental Health. Each class was assigned a unique integer label to facilitate categorical representation of textual data, where Non-Mental Health was encoded as 0, ADHD as 1, Anxiety as 2, Bipolar as 3, and Depression as 4.

To create a unified dataset for modeling, data collected from various mental health-related subreddits were first consolidated into a single data frame. This process involved resolving inconsistencies in subreddit naming conventions, for instance, replacing variations like “bipolarreddit” with standardized labels such as “bipolar.” Structural uniformity was ensured by renaming the subreddit column to label across all component data frames. Additionally, some rows were removed that had word length below 3, are empty rows, or do not carry too much insight. Once harmonized, the individual class-specific datasets were concatenated, and the combined data frame was shuffled using a fixed random seed to preserve reproducibility. The resulting dataset comprised 49,869 entries with an approximately balanced distribution of classes.

To address the linguistic variability and noise inherent in user-generated Reddit posts, a text cleaning pipeline was implemented. The pipeline began with the replacement of all Unicode emojis using regular expressions and converting them into their corresponding short code descriptions. This helps

preserve semantic meanings of emojis as textual representations. Next, a comprehensive noise removal step was executed to eliminate line breaks, carriage returns, URLs, user mentions, HTML tags, non-ASCII characters, numerically embedded words, special characters (such as “Ã”, “ñ”, “â”), and punctuation, thereby retaining only relevant and clean textual content. Hashtags were selectively retained if found within sentences but removed when located at the end, while additional symbols like \$ and & were filtered out. White space normalization followed, wherein multiple consecutive spaces were reduced to a single space using regular expressions, enhancing consistency across the corpus.

For stop word removal, two techniques were assessed, a static list of English stop words and spaCy’s NLP filtering. The latter was adopted for its context aware processing capabilities. We used SpaCy’s “en_core_web_sm” model, which is a small English pipeline trained on written web text (blogs, news, comments), that includes vocabulary, syntax, and entities. It effectively removed both common stopwords and punctuation, further improving the signal-to-noise ratio in the text. Lemmatization was subsequently performed using Natural Language Toolkit (NLTK), which standardized inflected words to their base forms. This step enhanced the generalization and semantic compactness of the data. To ensure quality, the length of each cleaned text was analyzed. Posts containing fewer than three words were excluded. A histogram of word counts revealed that 95% of posts were under 132 words, while the longest entry spanned 2,052 words.

Based on all the analysis, a truncation limit of 132 tokens was established during vectorization using TensorFlow’s TextVectorization layer. This layer tokenized the text into sequences of integers. The vocabulary learned dynamically from the training set consisted of 51,428 unique tokens. Among these, the five most frequent tokens were “”, [UNK], *nt*, *m*, and *like*, while the five least frequent tokens were identified and reviewed for quality control.

At the end, the dataset was split using a two-stage stratified sampling with a replacement method to ensure an equal class representation in subsets. The split allocated 70% of data for training, 10% for validation, and 20% for testing. The training dataset had 34,908 samples, the validation dataset 4,987

samples, and the testing dataset 9,974 samples, all of which had an equal number of features and labels. Furthermore, the distribution of text length within the training set was also explored further, validating the maximal sequence length to be 2,052 words and revalidating the 95th percentile threshold.

3.3.4 EEG Dataset

Raw EEG recordings often contain noise, artifacts, and irrelevant variability that obscure the underlying neural patterns. To address this, a comprehensive preprocessing pipeline was developed to transform the raw data into clean, structured, and analyzable signals. The main steps included loading and structuring subject-level data, applying filtering and artifact rejection, segmenting the signals into epochs, and extracting meaningful neural features for analysis.

The EEG data, collected from both ADHD and control subjects, were provided in MATLAB's “.mat” format. A custom Python function was implemented to import these files, organizing each subject's recordings into dictionaries where filenames served as keys and multidimensional EEG arrays as values. These datasets were converted into unified pandas’ data frames, where each row represented a time point across all EEG channels. Metadata fields were appended to support stratified analysis, including a binary “Group” indicator (1 for ADHD, 0 for Control) and a unique “PersonID” identifier for each subject.

For downstream processing and visualization, the EEG signals were reshaped into MNE RawArray objects, a standard format in the MNE (Magnetoencephalography and Electroencephalography) library that makes EEG/MEG analysis and visualization easier. A standard 10 - 20 montage was applied to maintain anatomical accuracy of electrode placement. In order to focus the analysis on the appropriate frequency ranges, a bandpass filter of 1 to 40 Hz was applied. It includes fundamental neural rhythms of delta, theta, alpha, beta, and low gamma and excludes DC drift and high-frequency artifacts.

The continuous signals were segmented into three-second epochs with 50% overlap, a configuration that increased the number of training samples while

preserving temporal resolution. Automated artifact removal was then performed using the AutoReject algorithm, which adaptively determines sensor-specific thresholds and applies interpolation or trial-based rejection [80, 81]. This approach minimized human bias and substantially improved the quality of the retained data.

Each clean epoch was stored as a row in the final preprocessed data frame, with EEG channel values as features and metadata fields (e.g., “Epoch,” “PersonID,” and “Group”) for subject-level tracking. The final dataset consisted of over 2.16 million epochs across 21 columns, with approximately 1,342,880 epochs from ADHD subjects and 999,075 from controls. Figure 3.22 shows the contrast between raw and preprocessed EEG samples. Moreover, Frequency-domain features were computed using Welch’s method to estimate the power spectral density (PSD) across canonical EEG bands as these are widely recognized in neurodiagnostic for their relevance to cognitive and neurodevelopmental functions. Fig. 3.23 compares power spectral density (PSD) estimates for ADHD signals before and after preprocessing.

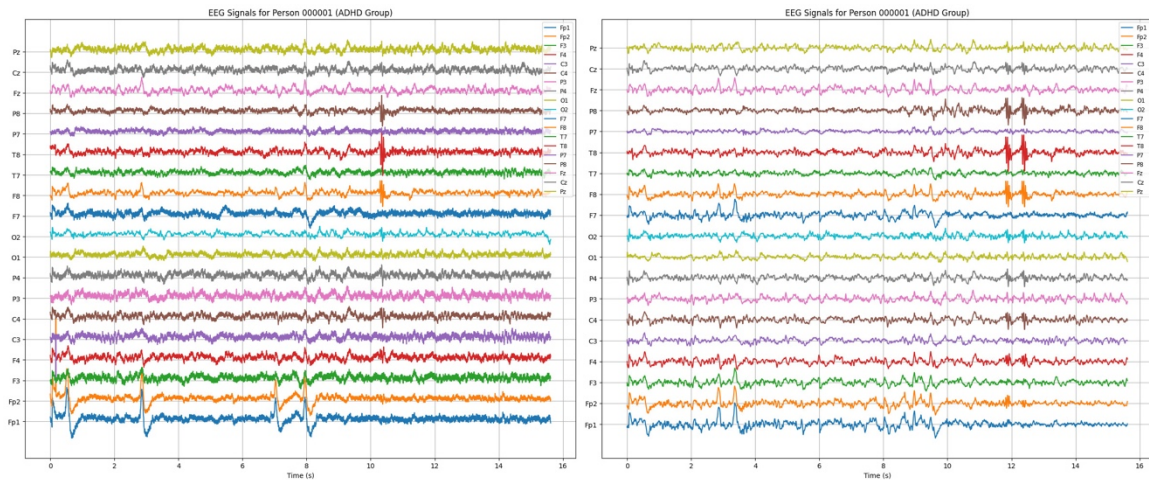


Figure 3.22 Comparison of EEG Signals for ADHD Sample before and after preprocessing

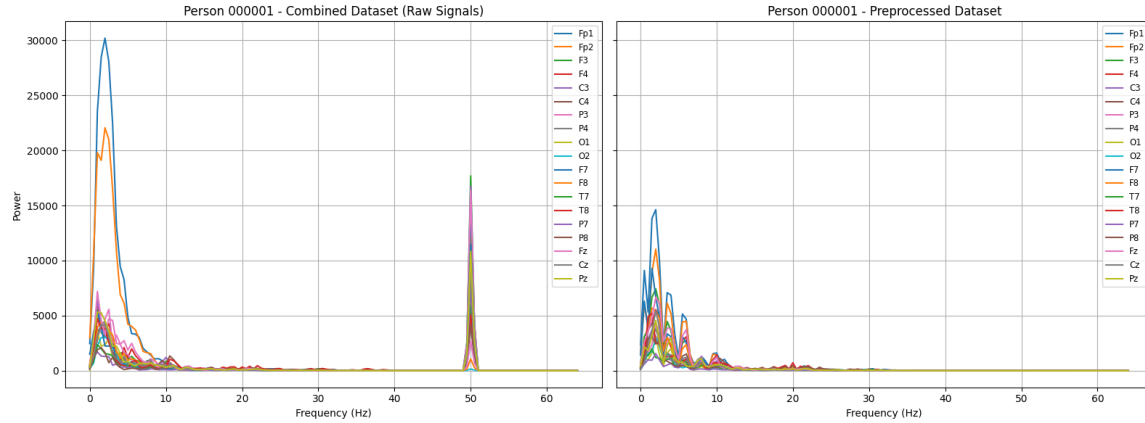


Figure 3.23 Comparison of PSD Signals for ADHD Sample before and after preprocessing

The comparison of Fig. 3.22 and 3.23 confirms that the preprocessing pipeline effectively cleans the EEG signals without removing meaningful neural information. In Fig. 3.22, the raw EEG signals are heavily influenced by artifacts such as baseline drifts, muscle activity, and eye blinks, which obscure the underlying neural activity. After preprocessing, these artifacts are suppressed, and the signals appear cleaner and more stable while preserving the overall waveform patterns. The presence of consistent oscillatory components across channels indicates that essential neural information, particularly within the delta, theta, alpha, and beta frequency bands, remains intact. The Fig. 3.23 further validates this by examining the power spectral density (PSD) of the signals. The raw PSD shows a high concentration of power in low-frequency ranges (0-5 Hz) and a sharp peak around 50 Hz caused by line noise. After preprocessing, these non-neural components are significantly reduced, and the spectral distribution of physiological rhythms becomes clearer. Importantly, the expected neural frequency bands are preserved, demonstrating that the preprocessing removed noise without distorting or eliminating meaningful neural activity.

In addition, the Theta/Beta Ratio (TBR), a common biomarker for ADHD, was calculated for each channel and averaged across frontal electrodes (Fp1, Fp2, F3, F4, Fz) to derive a frontal TBR metric. Complementing the frequency-domain analysis, time-domain statistical descriptors were derived from both the raw EEG signal and its amplitude envelope (obtained using the

Hilbert transform), including the envelope mean and standard deviation, variance, and spectral entropy. These metrics capture spatial-temporal dynamics and signal complexity associated with neural activity.

The feature set for each epoch consisted of descriptors from all 19 EEG channels, producing a high-dimensional space of over 150 features per epoch. To reduce dimensionality while retaining informative features, a two-stage selection strategy was used. First, univariate Analysis of Variance (ANOVA) with FDR correction identified approximately 80-100 features with significant group differences between ADHD and control subjects. The features retained through this ANOVA-based selection are summarized in Fig. 3.24, highlighting the top contributors in the EEG dataset. Then, Lasso regression with L1 regularization and five-fold cross-validation was applied, shrinking non-informative coefficients to zero and yielding a final subset of roughly 60-70 predictive features. The optimal regularization parameter ($\alpha = 0.01$) was determined via grid search.

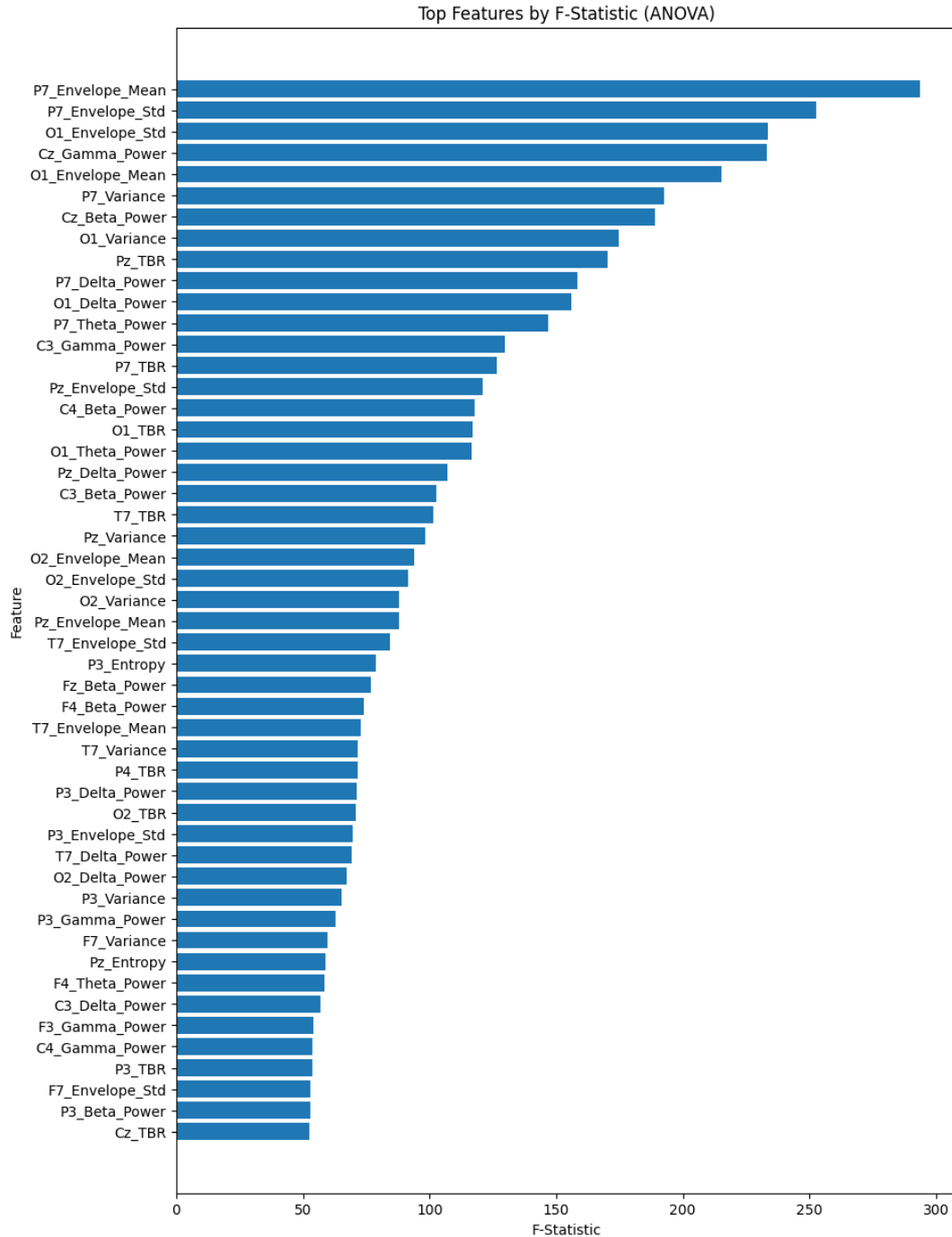


Figure 3.24 Top Features by ANOVA in EEG Dataset

To ensure reproducibility and facilitate downstream modeling, the selected feature sets were saved alongside 80/20 training-testing splits. Person-level feature aggregations were also generated for clustering and group-level analyses.

3.3.5 Neuroimaging Dataset

The ADHD-200 Preprocessed Anatomical Dataset comprises T1-weighted structural MRI scans that were preprocessed in advance using the “Anatomical Processing Script: fsl_anat” pipeline from the FSL neuroimaging toolkit. This standardized anatomical processing script provides an efficient workflow for preparing high-resolution anatomical brain images for analysis. The key preprocessing steps performed by the dataset authors included [78, 79]:

- **Reorientation:** Each scan was reoriented to match the standard Montreal Neurological Institute (MNI) coordinate system that ensures spatial consistency across subjects.
- **Cropping:** Non-informative peripheral regions were automatically removed from each scan to focus subsequent analysis on brain tissue.
- **Bias Field Correction:** Intensity inhomogeneities caused by RF/B1 field variations were corrected to improve signal uniformity throughout the image volume.
- **Registration:** Both linear and non-linear registrations were performed to align each individual scan to the standard MNI152 template.
- **Skull-Stripping:** Non-brain tissues were removed to isolate the brain to improve anatomical clarity for subsequent analyses.
- **Tissue-Type Segmentation:** This step was skipped/not applied in this study, as only whole-brain anatomical structures were analyzed.
- **Subcortical Structure Segmentation:** Similarly, subcortical segmentation was not conducted.

Following the standardized preprocessing, additional post-processing steps were implemented to prepare the data for deep learning by the dataset authors [78, 79]:

- **Resampling:** Each brain scan volume was resampled to a uniform resolution of $224 \times 224 \times 224$ voxels.

- **Normalization:** Voxel intensity values were standardized by applying zero-mean and unit-variance normalization, bringing all scans onto a comparable scale and aiding model convergence during training.

In Parallel to the processing being investigated, demographic and diagnostic metadata were preprocessed and loaded. Subject-level labels were given as ADHD (label = 1) or TD (label = 0). For class balancing, the small class imbalance involved in the original dataset (214 TD vs. 198 ADHD subjects), random down sampling was applied to the larger class, resulting in a balanced dataset with an equal number of subjects per group (198 subjects per class).

3.4 Model Development and Explanation

Across all experiments, extracted features from each dataset were used to train a range of ML and DL models. The objective is to evaluate, compare, and optimize these models for accurate and efficient prediction of ADHD and other mental health conditions.

Unless otherwise specified, models were tuned via grid search with five-fold cross-validation, and performance was assessed using accuracy, precision, recall, F1 score, ROC-AUC, and confusion matrix analysis. For DL models, adaptive learning rate scheduling, early stopping, and model checkpointing were employed to prevent overfitting. Ensemble learning (both hard and soft voting) was explored to integrate the strengths of top-performing models.

3.4.1 Handwriting Dataset

For handwriting-based ADHD detection, features were extracted from digitized handwriting samples to capture both visual-motor and structural characteristics such as text layout, line spacing, word spacing, letter spacing, handwriting inclination, letter deviation, size of handwriting, non-conventional letters, letter width, continuity/flow connection, letter shape, writing speed, and strength of graphism. Following preprocessing and PCA for dimensionality reduction, the dataset was used to train the following supervised ML models. The models selected for this experiment include;

- **Linear Models:** Logistic Regression
- **Margin-Based Models:** Support Vector Machine

- **Tree-Based Models:** Decision Tree, Random Forest, XGBoost
- **Instance-Based Models:** K-Nearest Neighbors
- **Ensemble Models:** Stacking Ensemble (RF + XGB as base learners, LR as meta-learner)

The hyperparameter search space included variations in regularization strength, kernel types, maximum tree depth, number of neighbors, learning rates, and boosting parameters. To check for the possibility of real-time or resource-constrained applications, training time was also recorded for computational efficiency analysis. To further enhance predictive performance, a stacking ensemble was implemented using RF and XGBoost as base learners and LR as the meta-learner. This approach allows the meta-learner to learn how to optimally combine the predictions of base learners, improving stability and accuracy over individual classifiers. The end-to-end ML workflow for ADHD detection using a handwriting-based dataset is illustrated in Fig 3.25.

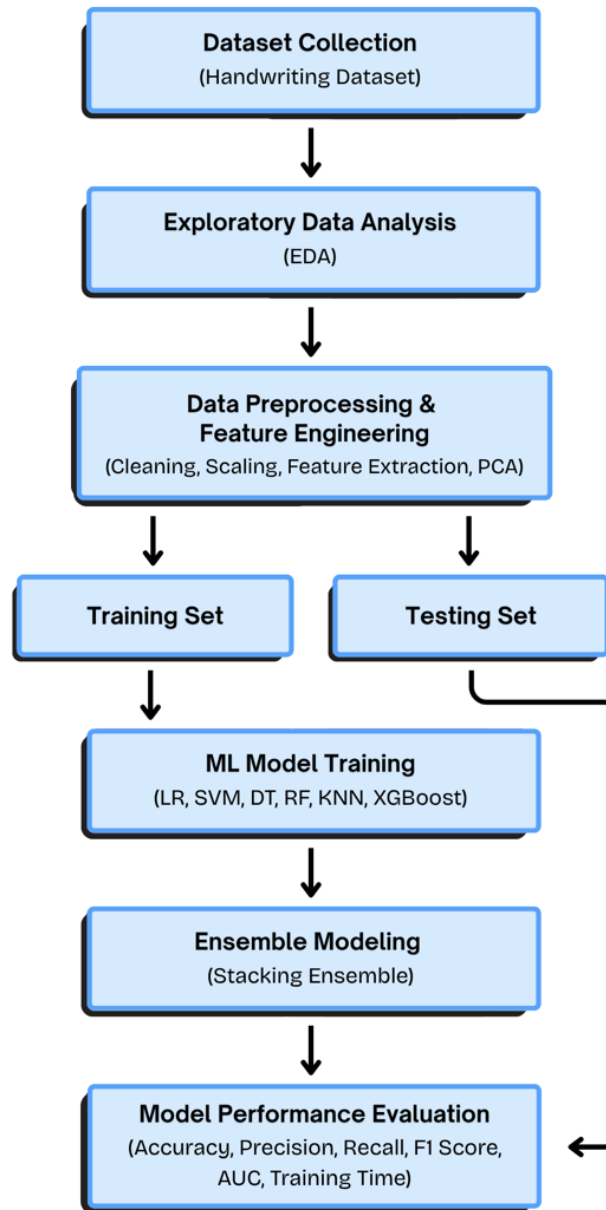


Figure 3.25 End-to-End ML Workflow for ADHD Detection using a Handwriting-Based Dataset

3.4.2 Drawing Dataset

The drawing-based ADHD detection task shares some structural similarities with handwriting but exhibits higher variability in spatial features and a smaller dataset size. Six supervised ML models were implemented;

- **Linear Models:** Logistic Regression

- **Margin-Based Models:** Support Vector Machine
- **Tree-Based Models:** Decision Tree, Random Forest, XGBoost
- **Instance-Based Models:** K-Nearest Neighbors

Each model underwent hyperparameter tuning via grid search with five-fold cross-validation. Performance was compared numerically and visually using bar plots for metric comparisons and ROC curves for class separation visualization. For tree-based models, feature importance scores were analyzed to identify the most influential drawing features. The end-to-end ML workflow for ADHD detection using a drawing-based dataset is illustrated in Fig 3.26.

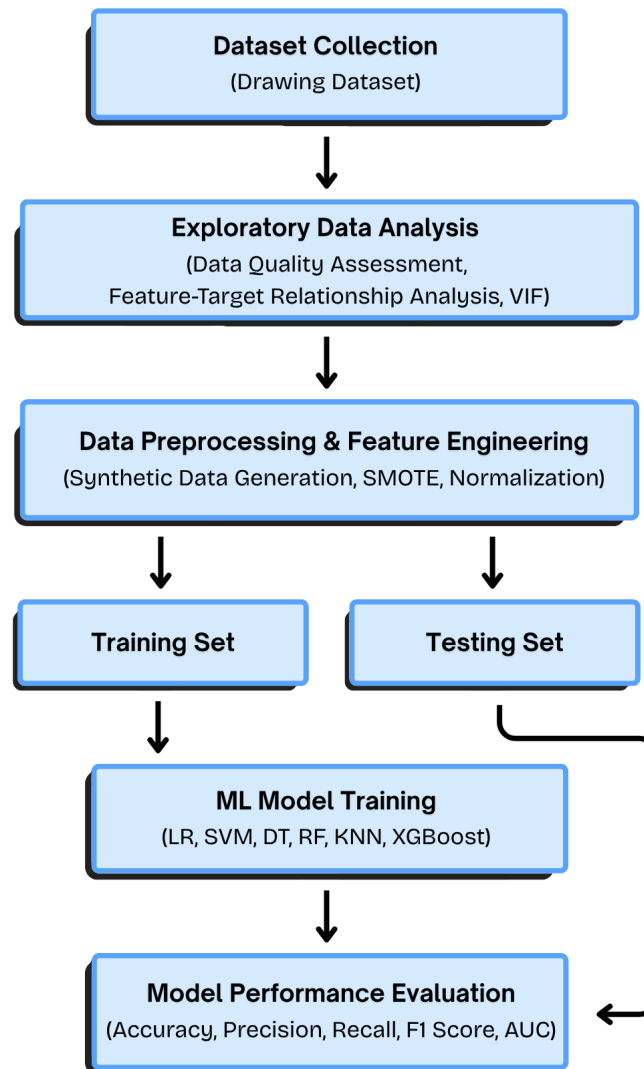


Figure 3.26 End-to-End ML Workflow for ADHD Detection using a Drawing-Based Dataset

3.4.3 Reddit based Mental Health Dataset

To classify mental health textual data into mental health-related (ADHD, Anxiety, Bipolar Disorder, Depression) and non-mental health categories, we utilize a diverse set of models spanning from classical ML to advanced transformer-based architectures and hybrid designs. This approach allows us to compare simple statistical baselines against state-of-the-art language models and sequence-based neural architectures. This ensures a comprehensive evaluation for this experiment. These models include;

- **Baseline Statistical Model:**
 - Naïve Bayes (Baseline) - A classic probabilistic classifier that utilizes Bag-of-Words and TF-IDF representations, serving as a benchmark for model performance.
- **Transformer Models:**
 - BERT - A bidirectional transformer model fine-tuned for five-class classification.
 - MentalBERT - A specialized version of BERT pre-trained on mental health-related corpora, enabling improved contextual understanding of psychological language.
 - RoBERTa - A transformer-based model that builds on BERT with refined pre-training methods for better performance.
 - MentalRoBERTa - A domain-adapted RoBERTa model trained on mental health text datasets to increase relevance for clinical and psychological contexts.
- **Sequence Models:**
 - BiLSTM with Attention - A bidirectional LSTM network enhanced with an attention mechanism to emphasize critical tokens in a sequence.
- **Convolutional Models:**
 - CNN for Text Classification - A convolutional neural network designed to detect local n-gram features in textual data.

- **Hybrid / Ensemble Model:** A hybrid architecture that combines contextual embeddings from best performing pre-trained model with recurrent layers for deeper sequential representation learning.

All models are trained using the balanced Reddit Mental Health Dataset. However, the training methodology differs according to model type. Model performance is evaluated on the test dataset using multiple metrics to provide both quantitative and qualitative insights. These include overall accuracy, weighted F1-score to account for any class imbalance, precision and recall per class, and a confusion matrix to visualize misclassifications. In addition, loss curves are plotted to monitor overfitting or underfitting during training, and random sample predictions are reviewed to assess qualitative prediction behavior. The best-performing model is selected based on the highest validation weighted F1-score. The end-to-end ML workflow for ADHD detection using a Reddit-based mental health dataset is illustrated in Fig 3.27.

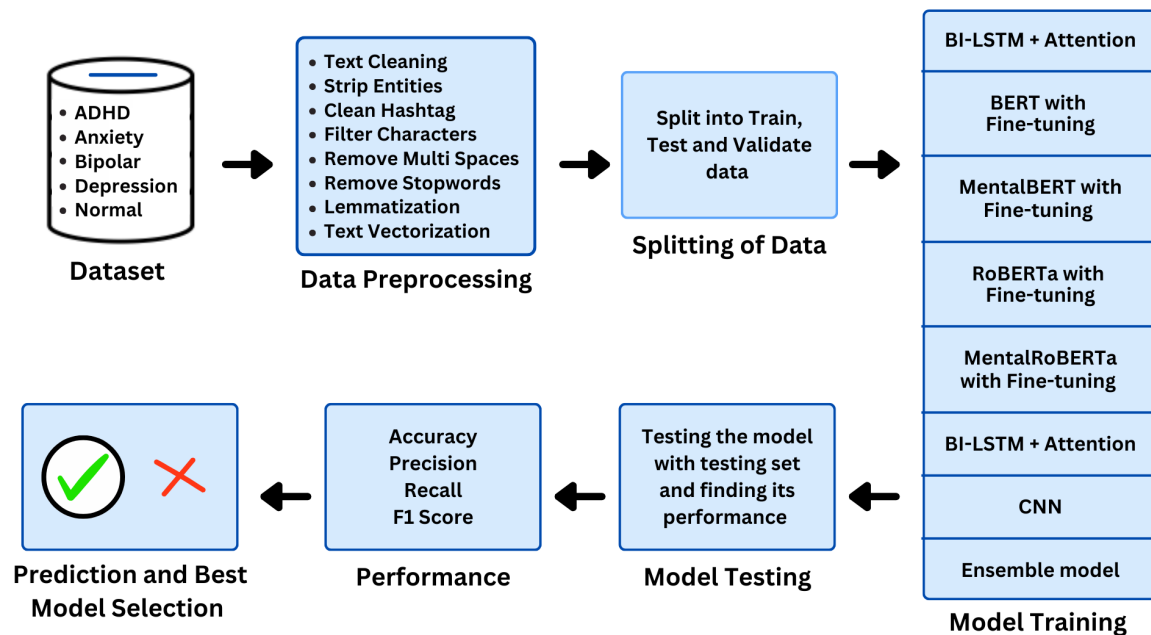


Figure 3.27 End-to-End ML Workflow for ADHD Detection using a Reddit based Mental Health Dataset

3.4.4 EEG Dataset

To classify ADHD and control groups based on EEG data, once signal preprocessing is complete and statistically significant EEG features are extracted, the next step is to train and evaluate a diverse range of ML and DL models. At this stage, the preprocessed and feature-selected datasets are split into training, validation, and testing subsets. These subsets are then used to train multiple classification algorithms; each selected for its distinct learning approach and proven effectiveness in biomedical signal classification tasks.

- **Traditional ML Models:**

- Logistic Regression (LR) - A linear baseline model effective for binary classification with balanced interpretability.
- Linear Discriminant Analysis (LDA) - A projection-based technique that maximizes class separability.
- Support Vector Machine (SVM) - A margin-based classifier capable of handling high-dimensional feature spaces
- K-Nearest Neighbors (KNN) - A non-parametric method relying on feature similarity for classification.
- Decision Tree (DT) - A simple yet interpretable tree-based learner for decision boundary visualization.
- Random Forest Classifier (RF) - An ensemble method based on decision trees, robust to noise and non-linear relationships.
- Logit Boost - A boosting algorithm using logistic regression as a base learner to enhance classification performance.
- AdaBoost - An adaptive boosting method combining multiple weak learners to form a strong classifier

- **Ensemble Methods:**

- Ensemble of Top 3 ML Models - A soft voting ensemble combining the three best-performing traditional ML algorithms, designed to utilize complementary strengths and achieve superior predictive performance.

- **Deep Learning Models:**

- Convolutional Neural Network (CNN) - Captures spatially local patterns in EEG feature sequences via convolutional layers.

- CNN-LSTM Hybrid - Combines convolutional feature extraction with recurrent temporal modeling for enhanced spatiotemporal learning.

For classical models, hyperparameter optimization is performed using GridSearchCV with k-fold cross-validation. For DL models, performance is refined through adaptive learning rate scheduling, early stopping, model checkpointing, and learning rate reduction on plateau. Evaluation metrics include accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix analysis, with ROC and learning curves generated to visualize classification capability and generalization behavior.

Finally, ensemble learning is used to combine the strengths of top-performing models. Soft voting ensembles (probability averaging) are implemented, which enables strong decision-making by aggregating model outputs. This multi-model, multi-paradigm strategy ensures a comprehensive evaluation across different learning frameworks and facilitates the identification of the most effective approach for reliable ADHD vs. control classification using EEG-derived biomarkers. The end-to-end ML workflow for ADHD detection using an EEG dataset is illustrated in Fig 3.28.

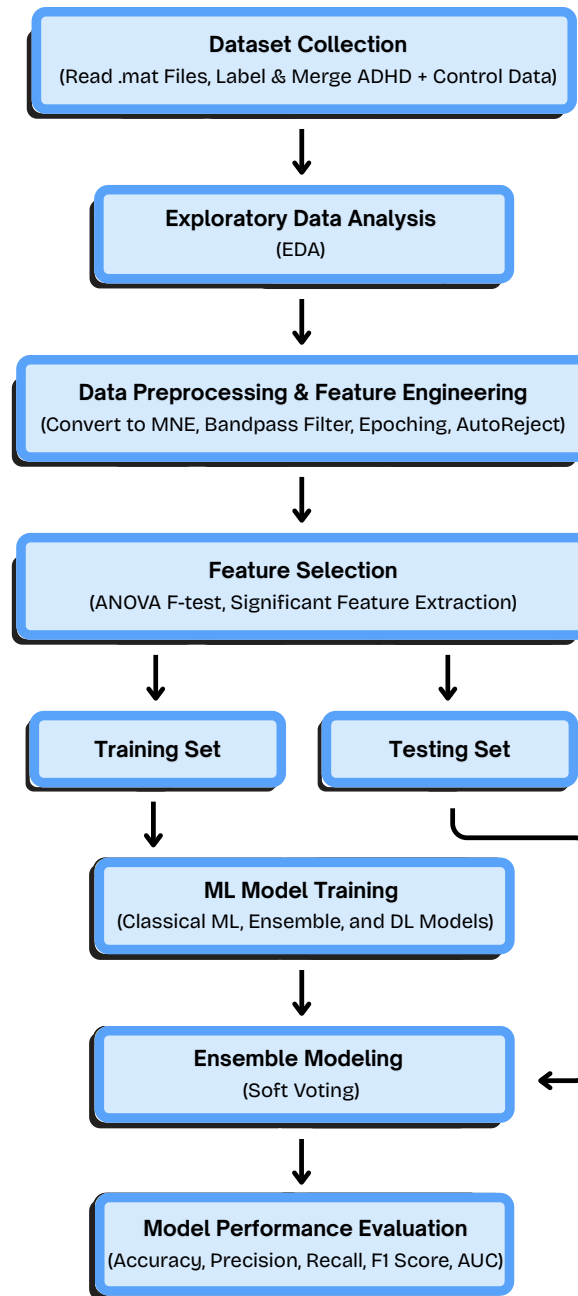


Figure 3.28 End-to-End ML Workflow for ADHD Detection using an EEG Dataset

3.4.5 Neuroimaging Dataset

For neuroimaging-based ADHD detection, DL models were implemented using transfer learning with ImageNet, which are pretrained convolutional neural networks. The selected models include:

- **ResNet50V2:** A residual architecture using skip connections to enable stable training of deep networks and effective feature extraction from MRI slices.
- **DenseNet121:** A densely connected convolutional neural network in which each layer receives input from all preceding layers. This facilitates feature reuse and improves gradient flow, enabling the capture of subtle neuroanatomical differences.

ResNet50 and DenseNet121 have been selected due to their performance on medical image tasks and complementary architectural advantages [82, 83, 84, 85]. ResNet50, using its residual learning architecture, is efficient in managing the vanishing gradient problem and is well-suited to train deep networks capable of learning high-level hierarchical features from MRI imaging. DenseNet121, on the other hand, uses dense connectivity to facilitate feature reuse, gradient flow improvement, and learning efficiency enhancement, making it highly effective for identifying fine-grained structural variations in brain images.

Training was carried out in two stages. First, the convolutional base was frozen to preserve general image representations. Then, higher layers were fine-tuned to adapt to domain-specific neuroimaging patterns. Optimization used the Adam optimizer with adaptive learning rate scheduling. To mitigate overfitting and class imbalance, dropout, L2 regularization, and class weighting were applied. Additional safeguards included early stopping, checkpointing, and learning rate reduction on plateau.

Model performance was evaluated using accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix analysis. Random test sample predictions were also visualized to illustrate correct classifications and common error cases. The end-to-end ML workflow for ADHD detection using a neuroimaging dataset is illustrated in Fig 3.29.

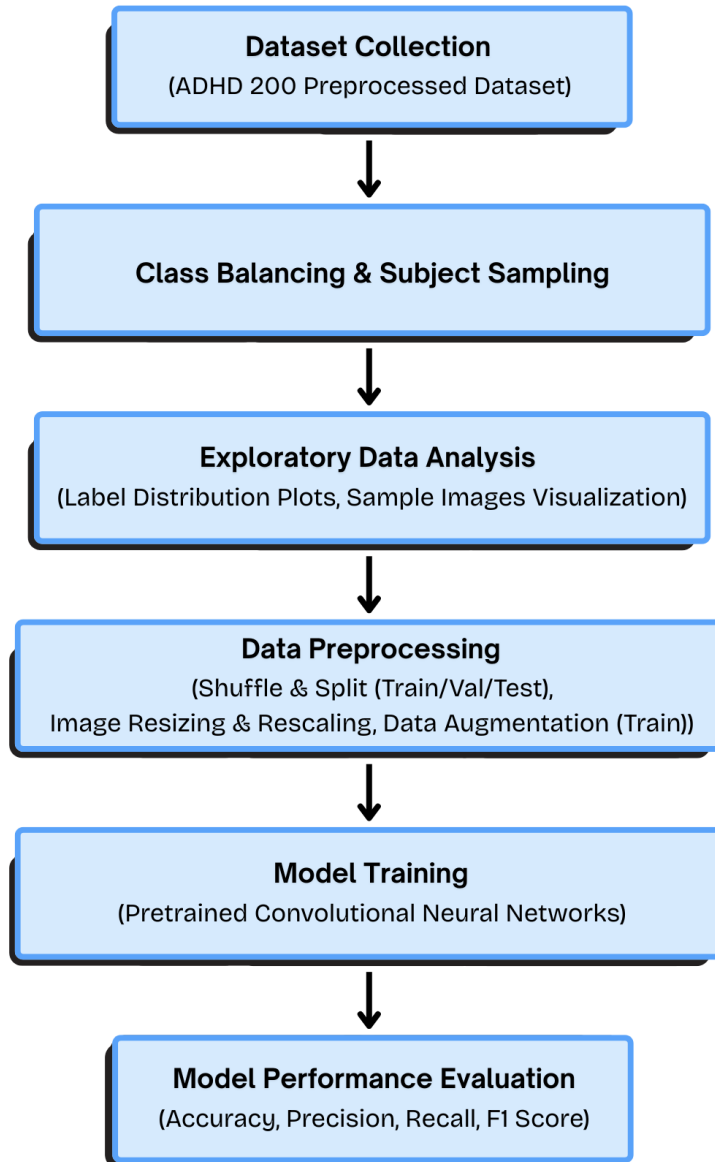


Figure 3.29 End-to-End ML Workflow for ADHD Detection using Neuroimaging Dataset

3.5 Experimental Setup

All experiments in this study were conducted in a unified computational environment to ensure consistency, reproducibility, and comparability across all datasets and modeling approaches. The implementation was carried out using Python on the Google Colab platform, with a NVIDIA Tesla T4 GPU (16 GB VRAM) enabled for accelerating the training of DL architectures.

Standard scientific computing and machine learning libraries, including NumPy, Pandas, Matplotlib, and Seaborn, were used for data handling, visualization, and exploratory analysis. Traditional ML models were implemented using Scikit-learn, while DL models were developed using TensorFlow and Keras. Gradient boosting models relied on the XGBoost library. Hyperparameter optimization was performed primarily through GridSearchCV, applied consistently across the datasets to ensure fair model comparison. All preprocessing steps, model training procedures, and evaluation metrics were executed within this environment following a unified workflow, which supported controlled experimentation and facilitated the reliable assessment of each model's performance.

Chapter 4. Experimental Results and Performance Evaluation

This chapter presents the results of the experiments conducted across all datasets, namely handwriting, drawing, Reddit-based mental health text, EEG, and neuroimaging. Each dataset was analyzed using traditional ML, DL, and, where applicable, ensemble models to evaluate their performance in detecting ADHD or other related mental health conditions.

This chapter is structured to present results dataset by dataset, followed by a comparative analysis and a discussion of findings that highlight the strengths, limitations, and implications of each modeling approach.

4.1 Handwriting Dataset Results

The handwriting dataset was evaluated using a range of traditional ML algorithms, a gradient boosting approach, and a stacking ensemble. The models were fine-tuned using GridSearchCV with a 5-fold cross-validation strategy to achieve effective hyperparameter optimization. Performance was evaluated based on accuracy, precision, recall, F1-score, ROC-AUC, and training duration. Additionally, confusion matrices and ROC curves were plotted to visualize classification performance and discriminative capability.

The test sample is very small. Consequently, all performance metrics are reported with conservative rounding to avoid overprecision. Accuracy is rounded to the nearest 10%, while precision, recall, F1-score, and AUC are rounded to one decimal place. Training times are rounded to whole seconds. Due to the limited number (12) of test cases, the reported metrics should be interpreted cautiously, with confusion matrices providing the most reliable insight into model behavior.

4.1.1 Traditional Machine Learning Models

Each traditional model was optimized over its key hyperparameters, such as solver and penalty terms for LR, kernel and regularization strength for SVM, depth and splitting criteria for DT, and neighbors/metrics for KNN. SVM emerged as the best-performing traditional method, achieving approximately

90% accuracy, 0.9 F1-score, and an AUC of 0.8, indicating strong discriminative capability. Logistic Regression and Random Forest achieved moderate but stable results (80% accuracy), while KNN offered high sensitivity to ADHD cases (Recall = 1.0) but lower overall accuracy. Decision Tree performed the weakest, with 60% accuracy and low discriminative ability (AUC = 0.6). Figure 4.1 illustrates the confusion matrices of top-performing models (SVM and Random Forest).

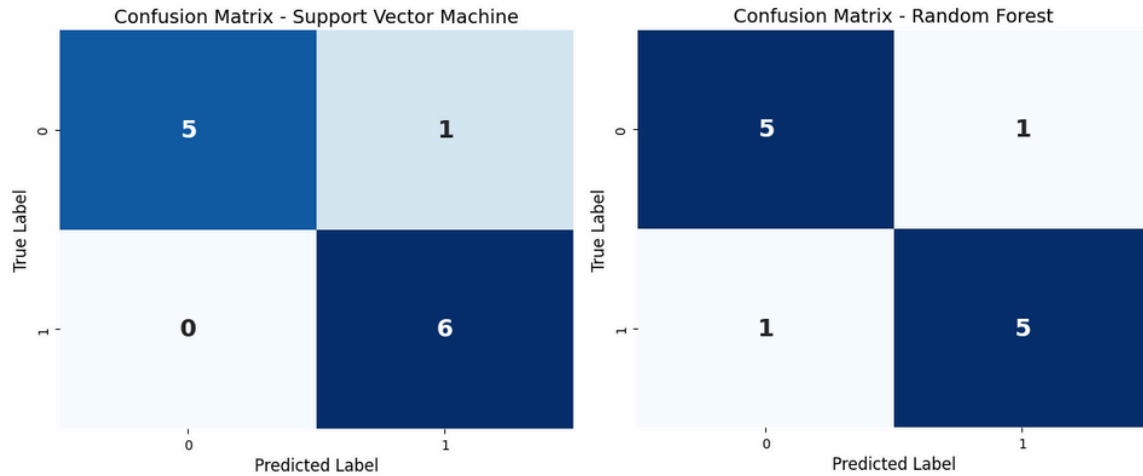


Figure 4.1 Confusion Matrices of top-performing models (SVM and Random Forest) on Handwriting Dataset

4.1.2 Gradient Boosting (XGBoost)

The XGBoost classifier was extensively tuned across parameters such as learning rate, maximum depth, subsampling, and column sampling ratios. Despite high computational costs (~450 seconds), XGBoost achieved performance equivalent to SVM, with 90% accuracy, 0.9 F1-score, and an AUC of 0.8. Its recall of 1.0 shows that it successfully captured ADHD cases without false negatives, though at the expense of efficiency. Figure 4.2 illustrates the confusion matrix of XGBoost.

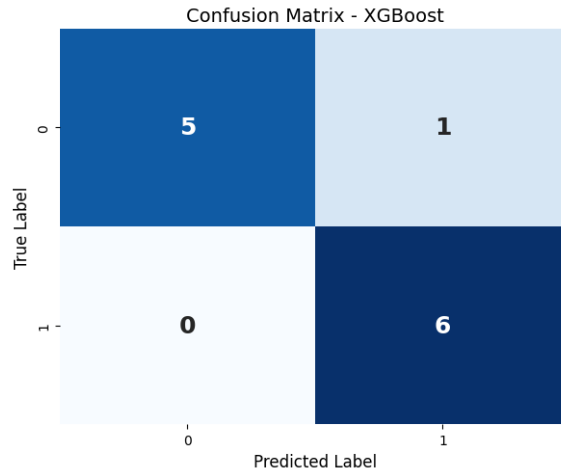


Figure 4.2 Confusion Matrix of XGBoost on Handwriting Dataset

4.1.3 Stacking Ensemble

To further enhance predictive performance, two stacking ensembles were implemented. In both cases, LR served as the meta-learner, while the base learners varied. The first ensemble combined XGBoost and SVM to utilize the strengths of boosting and margin-based classification. This model achieved the same predictive performance as the strongest individual classifiers, with an accuracy of 90%, F1-score of 0.9, and AUC of 0.8. Crucially, it maintained recall (1.0), ensuring that no ADHD cases were missed. However, this approach proved computationally expensive, requiring roughly 5,000 seconds for training due to the complexity of integrating both models within the stacking framework.

The second ensemble combined RF and XGBoost as base learners. Like the first approach, it achieved identical predictive metrics (accuracy = 90%, F1-score = 0.9, AUC = 0.8, recall = 1.0), but it was markedly more efficient, completing training in only 5 seconds. This demonstrates that while both stacking ensembles preserved the strong predictive capabilities of XGBoost and SVM, the RF + XGBoost combination offered a far more practical trade-off between accuracy and efficiency. Since the predictive outcomes were identical across both stacking strategies and the strongest individual models, the confusion matrices of the stacking classifiers are equivalent to those of XGBoost and SVM (Fig. 4.2).

4.1.4 Comparative Summary

The comparative results reveal several important insights. SVM, XGBoost, and the Stacking Ensemble demonstrated identical predictive performance, each achieving 90% accuracy, an F1-score of 0.9, and an AUC of 0.8. These models consistently identified ADHD cases with perfect recall (1.0), thereby minimizing false negatives. However, differences in efficiency clearly emerged. XGBoost, despite its strong performance, required a substantially longer training time of approximately 450 seconds. In contrast, the Stacking Ensemble achieved the same level of accuracy in approximately 5 seconds, underscoring its practical advantage. Similarly, SVM offered an excellent balance, delivering equivalent predictive power with minimal computational overhead of less than 1 second. Traditional models such as Logistic Regression and Random Forest served as stable baselines, each achieving approximately 80% accuracy, though they lacked the discriminative capacity of the top-performing methods. KNN demonstrated higher sensitivity but suffered from reduced precision, while the Decision Tree classifier underperformed overall. Clinically, the perfect recall of the top-performing models is especially significant, as it ensures that ADHD cases are not missed—an essential factor in early detection and intervention. Equally significant, the computational efficiency of the Stacking Ensemble highlights its potential for deployment in practical clinical decision support systems. These trends are summarized visually in Fig. 4.3, which presents a comparative bar chart of predictive metrics, and Fig. 4.4, which highlights ROC curves of all models. A summary of the results from all applied models is provided in Table 4.1. This shows that using individual SVM seems to be the best option for this dataset, avoiding the extra computational cost of the ensemble model.

Table 4.1 Performance of all ML Models on Handwriting Dataset

Model	Accuracy	Precision	Recall	F1-Score	AUC	Training Time (s)
SVM	90 %	0.9	1.0	0.9	0.8	< 1
XGBoost	90 %	0.9	1.0	0.9	0.8	450

Stacking Ensemble (XGBoost + SVM)	90 %	0.9	1.0	0.9	0.8	5000
Stacking Ensemble (RF + XGBoost)	90 %	0.9	1.0	0.9	0.8	5
LR	80 %	0.8	0.8	0.8	0.7	3
RF	80 %	0.8	0.8	0.8	0.8	87
KNN	70 %	0.7	1.0	0.8	0.6	< 1
DT	60 %	0.6	0.5	0.5	0.6	1

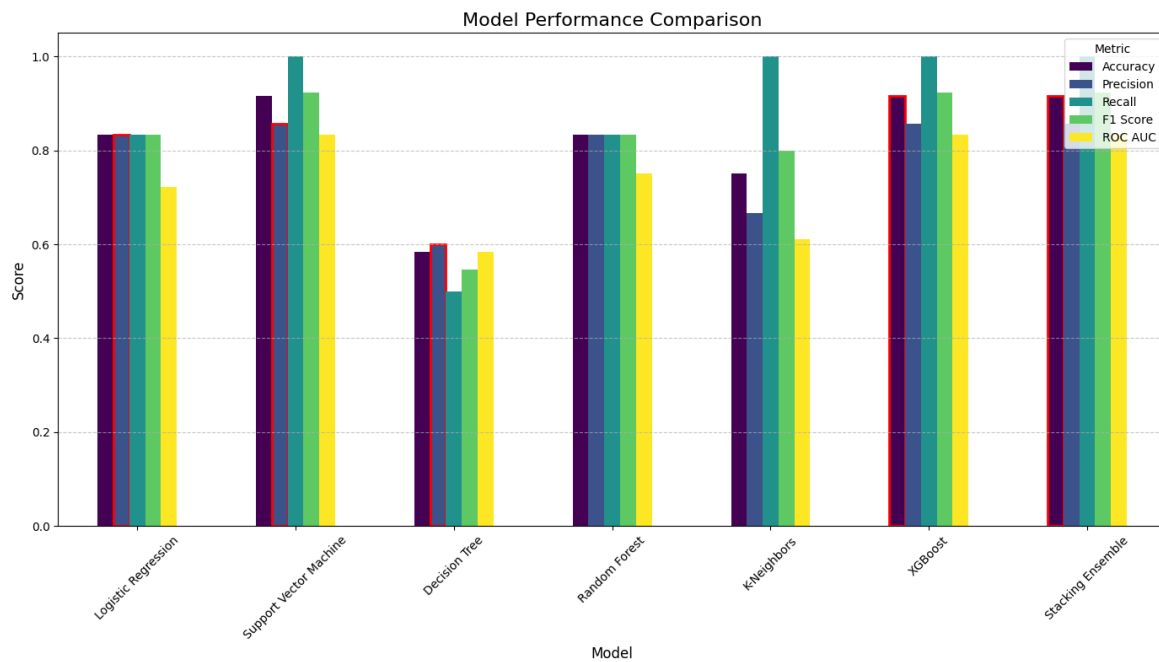


Figure 4.3 Model Performance of all Models on Handwriting Dataset

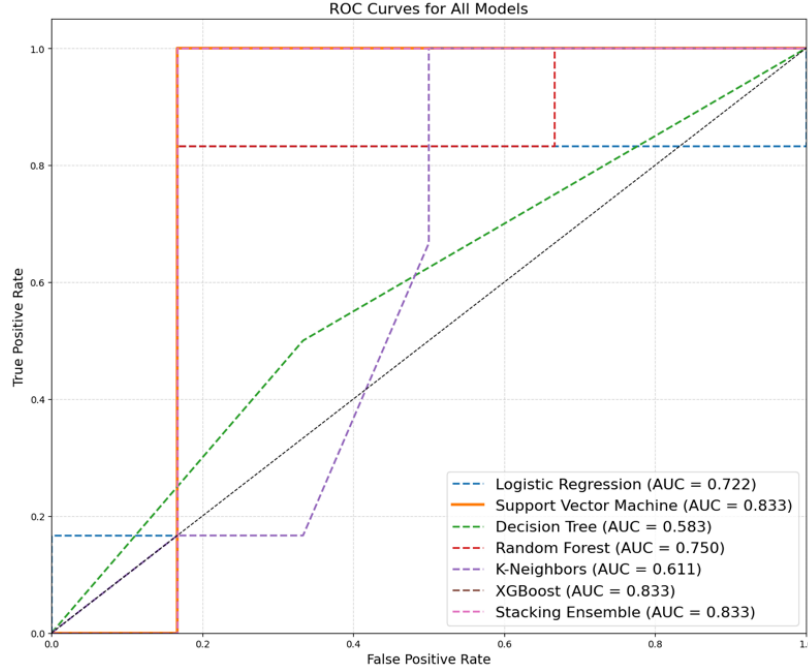


Figure 4.4 Comparison of ROC curves of all Models on Handwriting Dataset

4.1.5 Optical Character Recognition and Graphology Analysis

Apart from ML simulations, while the original handwriting dataset was manually labeled with the help of professional graphologists [71], the OCR pipeline can provide an automatic approach to extracting graphology-motivated features. This tool could potentially be used to create new handwriting datasets with little or no reliance on expert graphologists, enabling researchers to collect larger datasets more efficiently and reliably. Moreover, it would help to further standardize the feature extraction process and allow for scalable research in ADHD detection from handwriting.

To complement the handwriting classification models, an OCR based handwriting analysis pipeline was developed to capture handwriting specific characteristics beyond raw classification. This approach focused on clinically relevant features such as text layout, spacing, inclination, size, strength, and writing speed, which are often associated with ADHD-related patterns [71]. The pipeline was designed with modular functions to process handwriting samples, extract structured features, and provide both numerical scores and semantic descriptors.

The framework analyzed multiple aspects of handwriting, including margins and text layout, line and word spacing, letter inclination, size and intensity, as well as deviations in letter formation and writing flow. Writing speed was indirectly estimated from stroke smoothness and contour connectivity. A complete wrapper function consolidated these modules, producing interpretable results such as “straight lines,” “connected letters,” or “variable slant,” while visualization outputs displayed the original, preprocessed, and contour-enhanced images (Fig. 4.5).

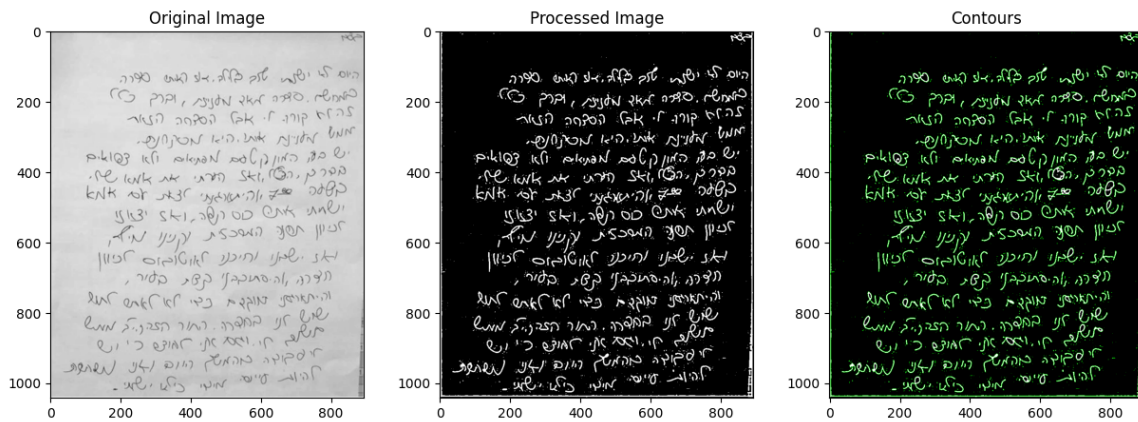


Figure 4.5 OCR based Graphology Analysis Visualization

The analysis identified distinct qualitative and quantitative characteristics that are not typically captured by conventional ML or DL classifiers but may serve as informative inputs for such models. Notably, children and adolescents with ADHD often displayed irregular line spacing, variations in letter size, disrupted flow of writing, inconsistent inclinations, and unconventional letter formations [71]. These results are consistent with clinical findings on dysgraphia associated with ADHD.

4.2 Drawing Dataset Results

Experiments on the drawing dataset were conducted to evaluate the discriminative power of trajectory-based features in classifying participants into ADHD and healthy (control) groups. The models underwent evaluation using a structured methodology that combined grid search-based

hyperparameter tuning, cross-validation, and bootstrapped ensemble predictions. Their performance was measured in terms of accuracy, precision, recall, and F1-score, with additional insights provided by confusion matrices and ROC-AUC curves.

4.2.1 Traditional Machine Learning Models

A diverse set of traditional classifiers were implemented, including LR, SVM, DT, RF, KNN, and XGBoost. Each algorithm was paired with a parameter grid for hyperparameter tuning via GridSearchCV to ensure that optimal configurations were selected across bootstrapped datasets. For instance, LR optimization focused on the regularization constant C and solver type, while tree-based methods tuned depth and split thresholds.

For this particular dataset, the optimized hyperparameters are presented in Table 4.2. Notably, LR achieved its best performance with a large regularization constant ($C = 1000$) and class balancing, while SVM demonstrated strong results with a linear kernel ($C = 10$). KNN favored a distance-weighted configuration with 7 neighbors, and XGBoost converged on a learning rate of 0.1 and 200 estimators.

Table 4.2 Optimized Hyperparameters for ML Models on Drawing Dataset

Model	Key Parameters
LR	solver="lbfgs," $C=1000$, class_weight="balanced"
SVM	$C=10$, kernel='linear', probability=True
KNN	n_neighbors=7, weights='distance'
RF	n_estimators=10, max_depth=5
XGBoost	learning_rate=0.1, n_estimators=200
DT	max_depth=10, min_samples_split=2

The evaluation results are summarized in Table 4.3. KNN emerged as the top-performing model, achieving perfect performance across all metrics (Accuracy = 100%, Precision = 1.0, Recall = 1.0, F1-score = 1.0). SVM also

performed exceptionally well, with an accuracy of 95.45% and an F1-score of 0.96, reflecting its strong linear separability of features. Tree-based methods (DT, RF, XGBoost) consistently achieved around 91% accuracy with balanced precision and recall, demonstrating robust but slightly lower performance compared to SVM and KNN. LR lagged slightly behind, with an accuracy of 86.36% but a relatively high recall (0.92), suggesting its utility as a stable baseline model. Figures 4.6 and 4.7 present the confusion matrices and ROC curves of the top two models (KNN and SVM), respectively.

Table 4.3 Performance of all ML Models on Drawing Dataset

Model	Accuracy	Precision	Recall	F1-Score
KNN	100.000 %	1.00	1.00	1.00
SVM	95.450 %	1.00	0.92	0.96
RF	90.910 %	0.92	0.92	0.92
XGBoost	90.910 %	0.92	0.92	0.92
DT	90.910 %	0.92	0.92	0.92
LR	86.360 %	0.86	0.92	0.89

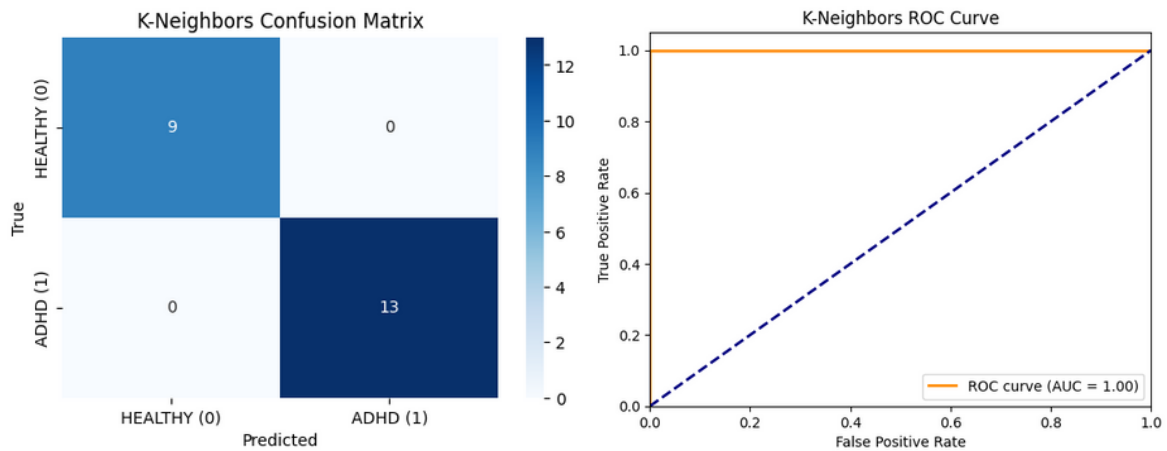


Figure 4.6 Confusion Matrix and ROC Curves of the KNN Models

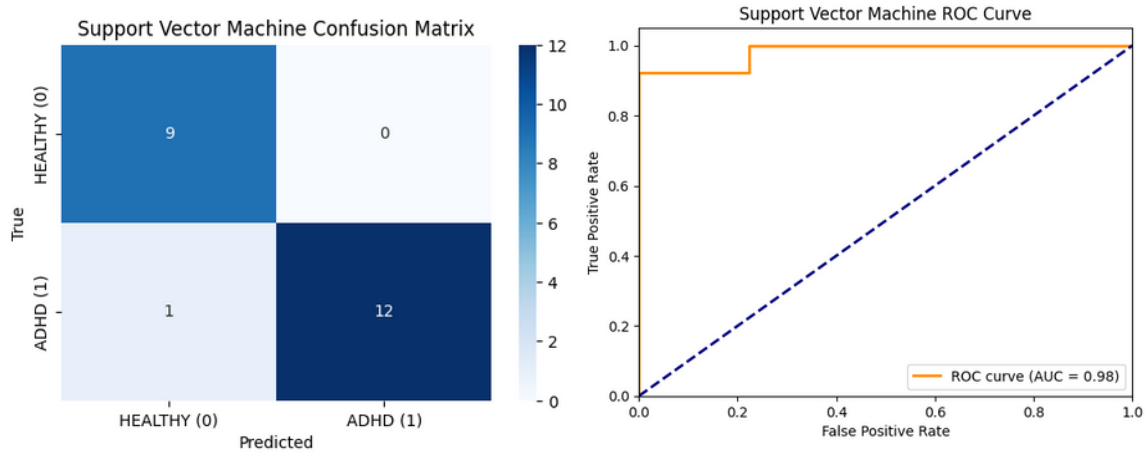


Figure 4.7 Confusion Matrix and ROC Curves of the SVM Models

4.3 Reddit-Based Mental Health Dataset Results

The Reddit dataset was investigated with DL and ensemble-based methods to classify user posts into the respective mental health categories. To ensure strong performance and reproducibility, all the experiments followed an identical pipeline with standardized data preprocessing, stratified train-test split, hyperparameter tuning, and evaluation based on Accuracy, Precision, Recall, F1-score, and ROC-AUC.

4.3.1 Baseline Statistical Model

As a baseline, a Multinomial Naive Bayes (MNB) classifier was implemented, given its established effectiveness in text classification [86]. The pipeline combined a Count Vectorizer (1-2 n-grams, 10,000 features, optimized frequency thresholds) with SMOTE oversampling to address class imbalance. Hyperparameter tuning identified an optimal smoothing parameter of $\alpha = 0.1$.

The MNB baseline achieved an accuracy of 78.2%, with Precision (0.789), Recall (0.782), and F1-score (0.784) closely aligned, indicating balanced predictive performance across categories. While it was limited in modeling deeper semantic dependencies, the model served as a reasonable reference point for comparison with more advanced approaches. The confusion matrix (Fig. 4.8) illustrates category-level misclassifications, while the ROC curves

(Fig. 4.9) show that class-level discriminative performance remained consistently above chance.

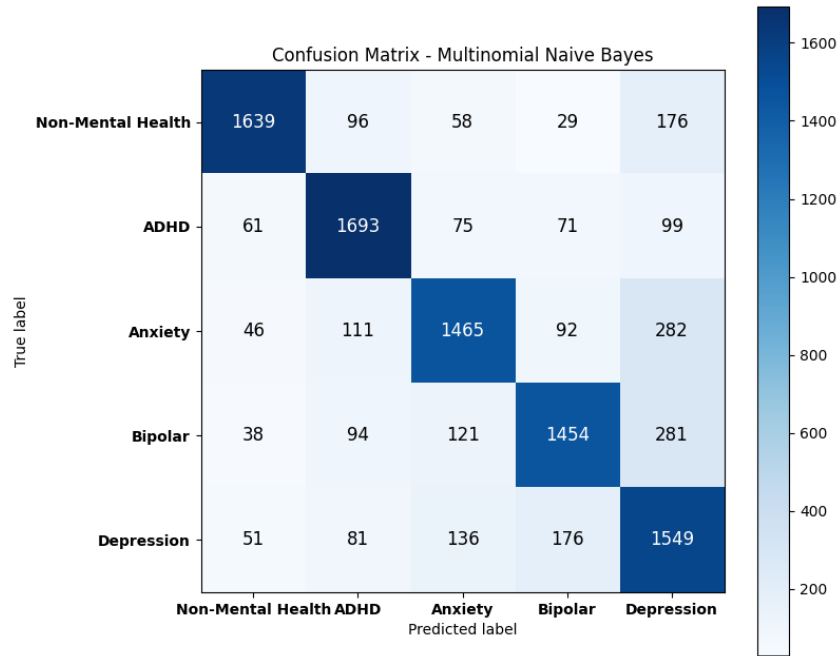


Figure 4.8 Confusion Matrix of MNB Classifier on Mental Health Dataset

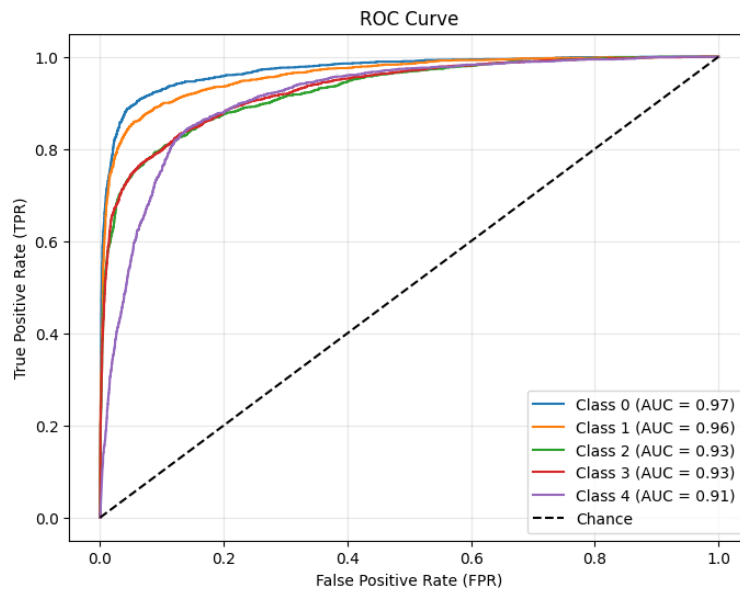


Figure 4.9 ROC Curve of MNB Classifier on Mental Health Dataset

4.3.2 Transformer Models

To capture contextual and semantic regularities in Reddit posts regarding mental health issues. Four models were tested and fine-tuned: BERT, MentalBERT, RoBERTa, and MentalRoBERTa. They were trained for five-label multi-classification with one label for each mental health category, utilizing a standard pipeline with dropout regularization, cosine learning rate schedule, and early stopping to prevent overfitting.

The BERT baseline (bert-base-uncased) achieved an accuracy of 86.55% and a weighted F1-score of 0.866. Its performance was particularly strong in detecting Non-Mental Health (F1 = 0.94) and ADHD (F1 = 0.89), while it struggled to some extent with categories involving more subtle or overlapping symptoms, such as Depression (F1 = 0.80) and Anxiety (F1 = 0.84). This indicates that although BERT effectively captures general contextual information, its ability to differentiate between conditions with nuanced symptomology remains limited. The confusion matrix for BERT is shown in Fig. 4.10.

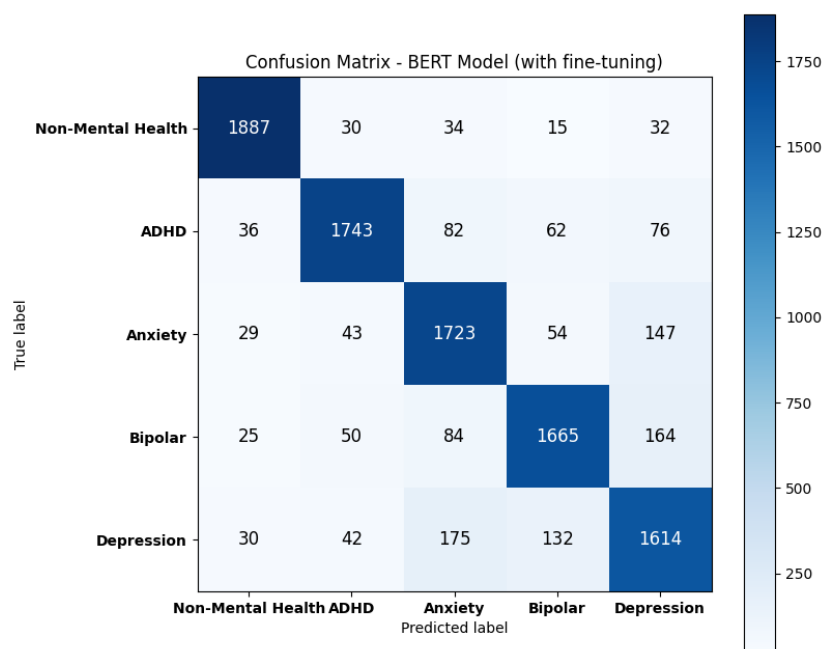


Figure 4.10 Confusion Matrix of BERT on Mental Health Dataset

MentalBERT (mental/mental-bert-base-uncased), being trained on mental health-related corpora, was expected to perform better by learning domain-specific vocabulary and context. True enough, it performed better slightly with 86.89% accuracy and weighted F1-score of 0.870. Improvements were largest in Depression (F1 = 0.81) and Anxiety (F1 = 0.85) classes, which demonstrated better sensitivity to categories on the basis of subtle, context-dependent language patterns. Figure 4.11 illustrates the confusion matrix for MentalBERT.

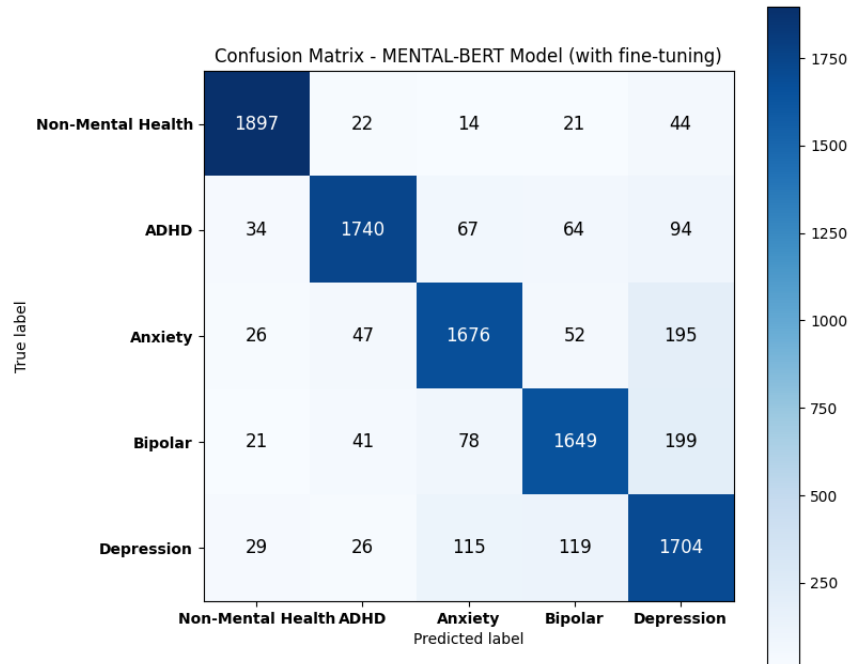


Figure 4.11 Confusion Matrix of Mental BERT on Mental Health Dataset

RoBERTa model (FacebookAI/roberta-base), trained with dynamic masking and optimized pretraining strategies, was also evaluated under the same fine-tuning conditions. It achieved an accuracy of 86.98% and a weighted F1-score of 0.870, making it highly competitive with both BERT and MentalBERT. RoBERTa was especially strong in classifying Non-Mental Health (F1 = 0.95) and ADHD (F1 = 0.91), but like the other general-purpose models, it demonstrated weaker results for Depression (F1 = 0.80) and Anxiety (F1 = 0.84). These results suggest that while RoBERTa's architecture captured general linguistic features effectively, it still struggled to fully resolve

overlapping linguistic patterns in clinically relevant categories. The corresponding confusion matrix is shown in Fig. 4.12.

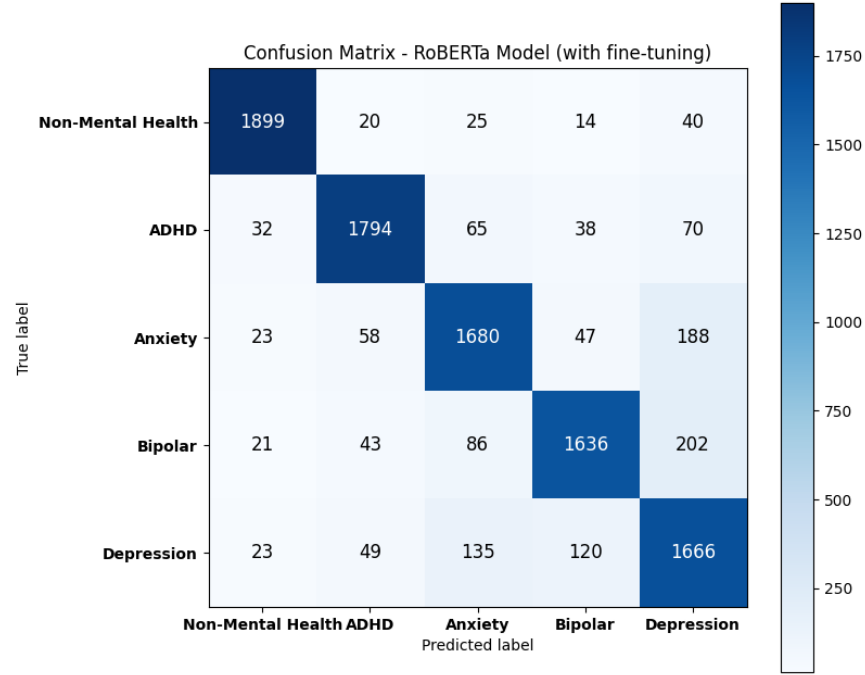


Figure 4.12 Confusion Matrix of RoBERTa on Mental Health Dataset

Finally, the MentalRoBERTa model (mental/mental-roberta-base) performed the best across all. Utilizing pretraining on general corpora over domain-specific ones, it surpassed others in most classes, achieving 87.22% accuracy and a weighted F1-score of 0.873. Performances were especially enhanced for bipolar disorder ($F1 = 0.86$) and Depression ($F1 = 0.81$), where the domain-specific training was able to capture linguistic signals and subtle indications that general-purpose models did not. The respective confusion matrix is presented in Fig. 4.13.

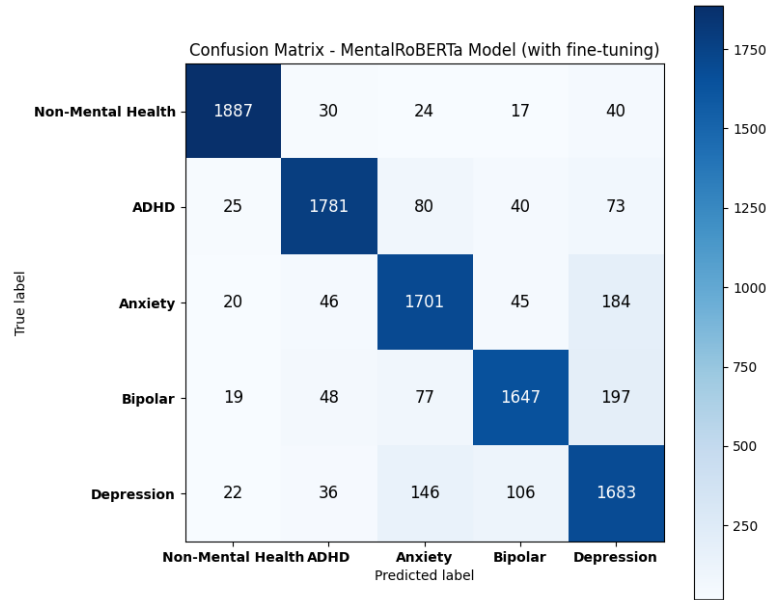


Figure 4.13 Confusion Matrix of Mental RoBERTa on Mental Health Dataset

These results confirm the advantage of domain-specific pretraining for transformer models in mental health text classification. While BERT and RoBERTa provided strong baselines, MentalBERT and MentalRoBERTa generalized better to nuanced categories.

4.3.3 Sequence Models

To model sequential dependencies, a Bidirectional LSTM with attention (BiLSTM-Attention) was implemented. The architecture used a 128-dimensional trainable embedding, a BiLSTM with 64 hidden units, an attention layer, and dense layers with dropout and batch normalization. Training used Adam with sparse categorical cross-entropy, early stopping, and checkpointing.

The BiLSTM-Attention achieved 83% accuracy, with high performance for Non-Mental Health ($F1 = 0.90$), ADHD ($F1 = 0.87$), and Anxiety ($F1 = 0.83$). Lower results were observed for Bipolar ($F1 = 0.80$) and Depression ($F1 = 0.74$), reflecting difficulty in separating categories with overlapping language. The confusion matrix (Fig. 4.14) shows frequent misclassifications between Bipolar and Depression, though the attention mechanism aided Anxiety classification compared to simpler RNNs.

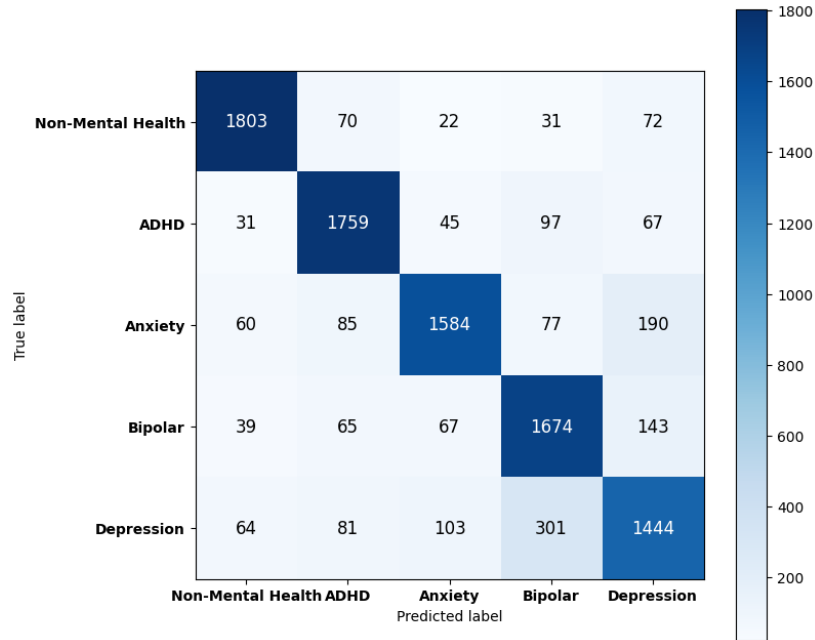


Figure 4.14 Confusion Matrix of BiLSTM-Attention Model on Reddit Dataset

4.3.4 Convolutional Models

A 1D CNN with pre-trained GloVe embeddings was applied to capture local n-gram features. The model embedded tokens into 300-dimensional vectors, applied Conv1D layers with kernel sizes 2-4, followed by batch normalization, global max pooling, concatenation, and dense layers with dropout. Training used Adam ($\text{lr} = 0.0005$) with early stopping, learning rate scheduling, and checkpointing.

The CNN achieved 82% accuracy. It showed strong precision for ADHD (0.88) and Bipolar (0.90), but lower recall for Bipolar (0.71) and Depression (0.85), indicating challenges in separating overlapping linguistic cues. The confusion matrix (Fig. 4.15) highlights strong classification of Non-Mental Health and ADHD posts, but frequent confusion between Bipolar and Depression.

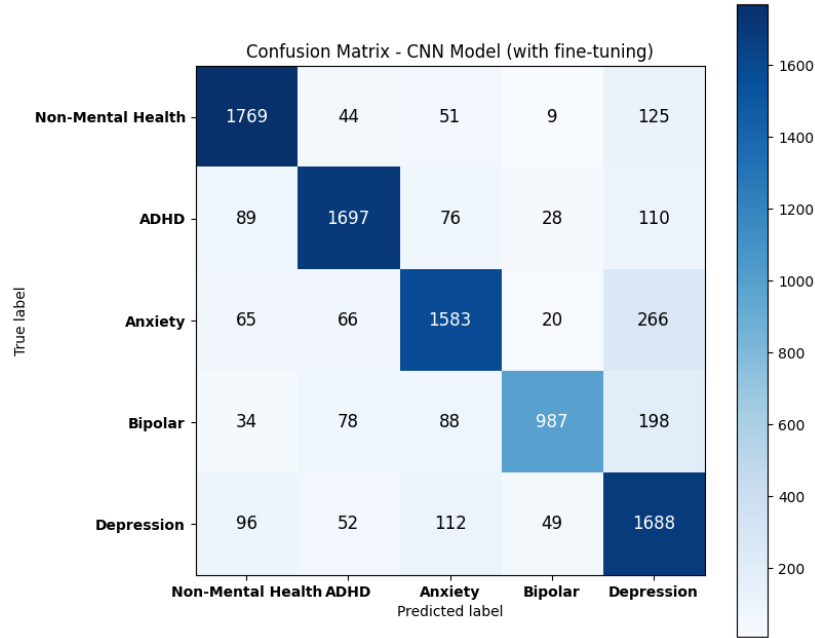


Figure 4.15 Confusion Matrix of CNN Model on Reddit Dataset

4.3.5 Ensemble Models

For combining the strengths of MentalRoBERTa and Bidirectional LSTM, weighted soft-voting ensemble and meta-classification methods were used. The aim of these techniques is to improve classification accuracy through the aggregation of predictions made by models with different architectures and feature spaces.

In the first method (Weighted Soft-Voting Ensemble), the predictions of MentalRoBERTa and BiLSTM were converted to class probabilities. Grid search overweight combinations were performed over the validation set to identify the best combination. The best combination was found to be 0.6 for MentalRoBERTa and 0.4 for BiLSTM, which achieved the best weighted F1-score of 0.870. Using this weighting over the test set produced an F1-score of 0.874, showing how utilization of contextual embeddings taken from transformers in conjunction with sequential representations from LSTM can improve overall predictive ability.

The second ensemble method (Meta-Classification Ensemble) used a meta-classifier with training on predictions of the base models. Logistic Regression

and Random Forest were used as meta-classifiers, using validation predictions of BiLSTM and MentalRoBERTa as input features. Two aggregation processes were experimented on the test set: majority voting and averaging probabilities. Using majority voting, the meta-classifier ensemble achieved an accuracy of 0.872 and a weighted F1-score of 0.872. When averaging probabilities were applied, performance slightly improved, reaching an accuracy of 0.874 and a weighted F1-score of 0.875. The averaging probabilities method worked best at test, and this indicates that the probabilistic output of several classifiers together better captures complementary information than voting.

4.3.6 Comparative Summary

The results demonstrate a clear performance hierarchy across statistical, sequential, convolutional, and transformer-based models. The Multinomial Naive Bayes baseline provided a good foundation, achieving 78.2% accuracy with balanced precision and recall. While limited in modeling deeper semantics, its stable and interpretable performance confirmed its value as a benchmark.

Among deep learning models, sequence-based (BiLSTM-Attention) and convolutional (1D CNN) architectures performed competitively, with accuracies of 83% and 82%, respectively. Both models were effective at classifying Non-Mental Health and ADHD posts but struggled with Bipolar and Depression, where linguistic overlap created challenges. The BiLSTM's attention mechanism provided some improvements for Anxiety detection, while the CNN's convolutional filters captured strong n-gram patterns but missed longer dependencies.

The transformer-based models consistently outperformed traditional and sequential approaches. General-purpose BERT and RoBERTa models performed strongly, each exceeding 86.5% accuracy, but domain-adaptive pretraining provided an additional advantage. MentalBERT and particularly MentalRoBERTa achieved the best standalone results, with MentalRoBERTa reaching 87.22% accuracy and an F1-score of 0.873. These improvements were most notable in Depression and Bipolar categories, where subtle linguistic cues benefited from domain-specific pretraining.

Finally, ensemble approaches, Combining MentalRoBERTa and BiLSTM-Attention through weighted soft voting or meta-classification slightly but consistently improved performance. The best ensemble, using probability averaging, achieved 87.4% accuracy and an F1-score of 0.875, demonstrating that hybrid strategies can leverage the complementary strengths of transformer-based contextual embeddings and sequential dependency modeling. Overall, the findings highlight three key insights: (1) transformer architectures dominate in contextual understanding, (2) domain-adaptive pretraining enhances sensitivity to clinically nuanced categories, and (3) ensembles offer incremental yet meaningful improvements in robustness and generalization. Relative performance of all models is listed in Table 4.4.

Table 4.4 Performance of all Models on Mental Health Dataset

Model	Accuracy	Precision	Recall	F1-Score
Ensemble (Weighted voting - MentalRoBERTa + BiLSTM)	87.400 %	0.874	0.874	0.874
Ensemble (Meta, Averaging)	87.400 %	0.876	0.875	0.875
MentalRoBERTa	87.220 %	0.875	0.872	0.873
RoBERTa	86.980 %	0.872	0.870	0.870
MentalBERT	86.890 %	0.872	0.869	0.870
BERT	86.550 %	0.867	0.865	0.866
BiLSTM - Attention	83.000 %	0.834	0.830	0.832
CNN	82.000 %	0.828	0.820	0.823
Multinomial Naive Bayes	78.200 %	0.789	0.782	0.784

4.4 EEG Dataset Results

Experiments on the EEG dataset were performed using a combination of traditional ML algorithms, ensemble models, and DL architectures. The

primary objective was to classify subjects into ADHD and Control groups using features derived from the preprocessed EEG signals. To ensure fairness and reproducibility, all models were trained and evaluated using a consistent pipeline, including standardized data splits, hyperparameter tuning, and evaluation metrics. Performance was assessed using Accuracy, Precision, Recall, F1-score, and AUC to provide a comprehensive view of each model's effectiveness.

4.4.1 Traditional Machine Learning Models

Traditional ML models were trained on a feature-engineered dataset that integrated statistical, spectral, and temporal descriptors to capture the multi-dimensional characteristics of the EEG signals. Hyperparameter optimization was conducted using GridSearchCV with 5-fold cross-validation, which ensures robust parameter selection while minimizing the risk of overfitting.

For each algorithm, a parameter grids were designed to tune the most influential hyperparameters. For example, the regularization strength C and kernel type for SVM, or the number of trees and maximum depth for Random Forest. For this particular dataset, the optimized configurations are summarized in Table 4.5.

Table 4.5 Optimized Hyperparameters for ML Models on EEG Dataset

Model	Key Parameters
LR	solver='sag', penalty='l2', C=100, max_iter=10000, class_weight="balanced"
LDA	solver='lsqr', shrinkage='auto'
SVM	C=10, kernel='rbf', class_weight='balanced', probability=True
KNN	n_neighbors=5, weights='distance', p=2 (Euclidean), with StandardScaler
DT	max_depth=20, min_samples_split=2, min_samples_leaf=1, class_weight='balanced'
RF	n_estimators=200, max_depth=20, min_samples_split=2, min_samples_leaf=1
Logit Boost	base_estimator=LogisticRegression(max_iter=1000), n_estimators=100, learning_rate=0.1
AdaBoost	base_estimator=DecisionTreeClassifier(max_depth=1), n_estimators=100, learning_rate=0.1

Among the algorithms, KNN and SVM demonstrated the highest classification performance with both achieving AUC values of 0.96, indicating excellent discriminative capability between ADHD and control groups. The strong performance of SVM and KNN highlights the presence of non-linear patterns in the EEG feature space, which these models capture effectively. Methods like Random Forest also performed competitively (AUC = 0.96), benefitting from aggregated decision trees that reduce variance and improve generalization. Logistic Regression and LDA, while being less complex, provided stable and interpretable baselines, with moderate accuracy (~77%) but lower AUC (~0.83-0.85), suggesting their limited ability to capture complex feature interactions. The Fig. 4.16 illustrates ROC-AUC curves comparing the performance of all traditional ML models, while Fig. 4.17 illustrates the confusion matrices of top-performing models (SVM, KNN, and Random Forest).

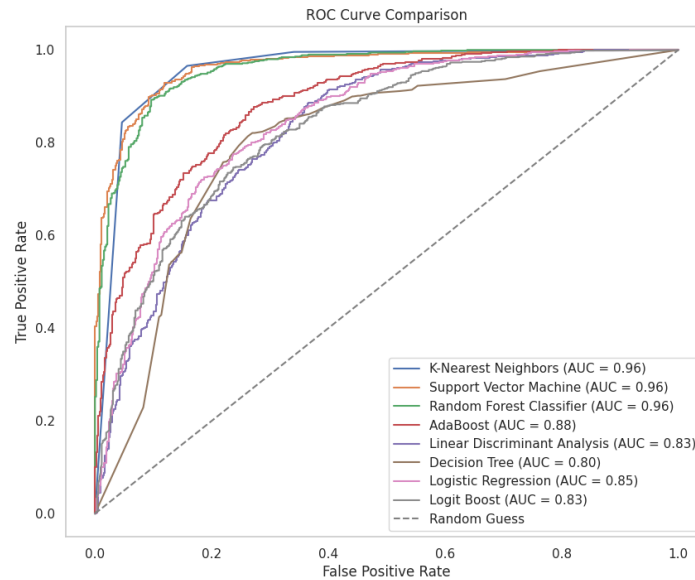


Figure 4.16 ROC-AUC curves of all traditional ML models on EEG Dataset

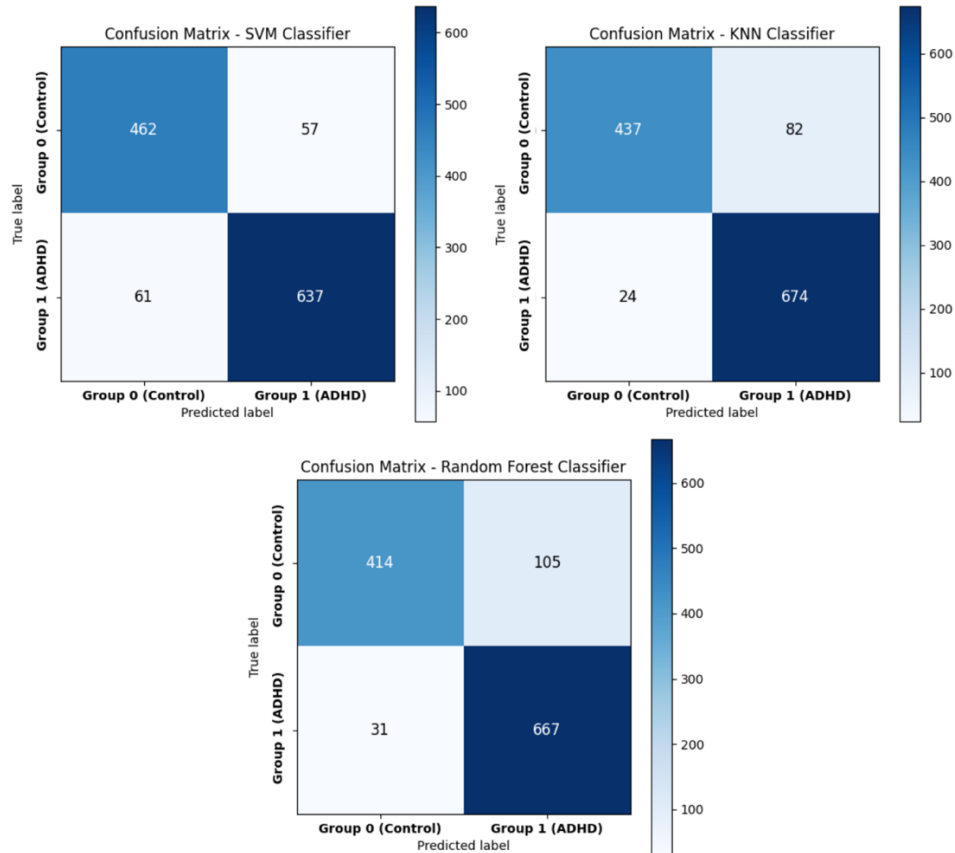


Figure 4.17 Confusion matrices of top-performing models (SVM, KNN, and Random Forest) on EEG Dataset

4.4.2 Ensemble Models

To further enhance predictive accuracy and reduce model-specific biases, an ensemble learning approach using a soft-voting classifier was implemented. The methodology focused on combining the strengths of the top-performing models in the dataset. The top three models selected based on their validation accuracies were KNN, SVM, and Random Forest Classifier. This approach allows the ensemble to account for the confidence levels of each classifier, often resulting in more balanced and accurate predictions compared to hard majority voting.

The soft voting ensemble demonstrated strong predictive performance. It achieved an overall accuracy of 93.26% on the test set, outperforming most individual models. The classification report indicated consistent performance across both classes;

- Class 0 (Control) achieving a precision of 0.95, recall of 0.88, and F1-score of 0.92,
- Class 1 (ADHD) achieving a precision of 0.92, recall of 0.97, and F1-score of 0.94.

Figure. 4.18 illustrates the confusion matrices and ROC-AUC curves of soft-voting ensemble.

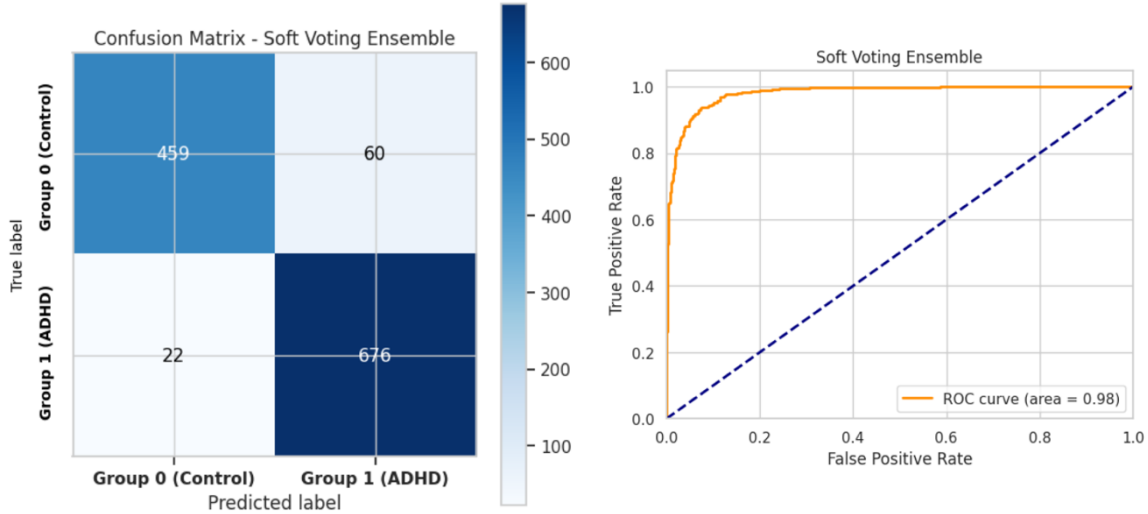


Figure 4.18 confusion matrices and ROC-AUC curves of soft-voting ensemble

The high recall for ADHD suggests the ensemble was particularly effective at identifying ADHD cases, reducing the likelihood of false negatives, which is a critical factor in clinical settings.

4.4.3 Deep Learning Architectures

Two deep learning models were implemented to evaluate their ability to learn discriminative patterns directly from the EEG data, CNN, and Hybrid CNN-LSTM architecture.

Both models were trained for up to 32 epochs with a batch size of 32, using early stopping, checkpoint saving, and learning rate reduction to optimize performance, prevent overfitting, and optimize convergence. The CNN model outperformed the hybrid CNN-LSTM, achieving 87% accuracy and a balanced performance across both classes. It achieved a recall of 91% for

ADHD and 82% for controls, indicating strong generalization while slightly favoring the ADHD class. The CNN-LSTM model achieved a lower overall accuracy of $\sim 80\%$, with relatively high recall for ADHD (87%) but weaker performance for controls (70% recall), suggesting it tended to overfit to ADHD patterns. Figure 4.19 illustrates the confusion matrices of DL models.

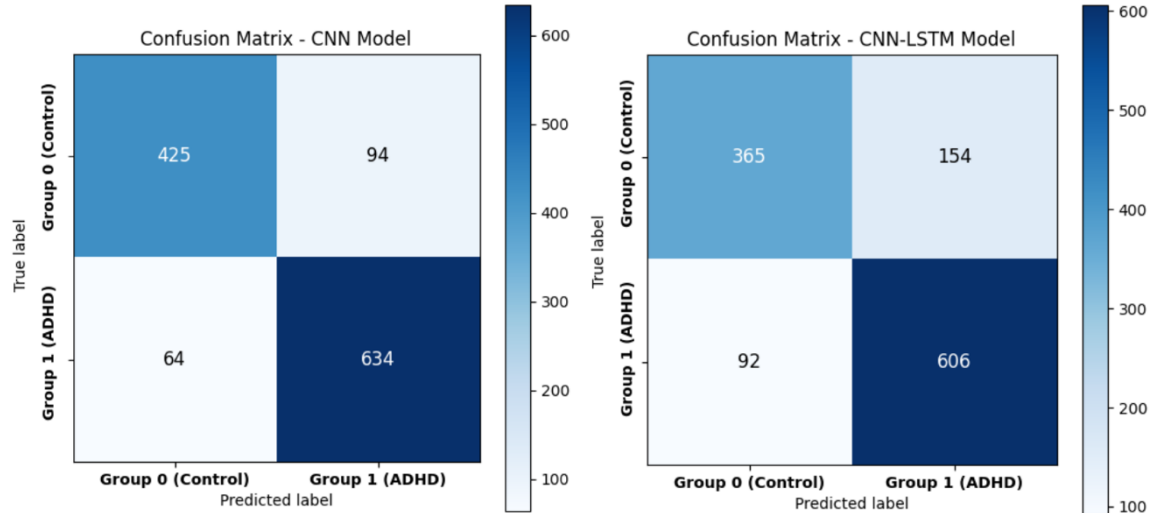


Figure 4.19 Confusion matrices of CNN and CNN-LSTM models on EEG Dataset

4.4.4 Comparative Summary

The comparative results highlight several important trends. The soft-voting ensemble model emerged as the best performing approach, achieving 93.26% accuracy with balanced precision and recall across classes. This demonstrates that integrating the complementary strengths of KNN, SVM, and RF models yields superior generalization compared to using any single model. When comparing traditional ML methods with DL approaches, optimized traditional models, particularly KNN, SVM, and RF, consistently outperformed the tested deep learning architectures. This suggests that the feature-engineered representations extracted from EEG data effectively captured the underlying patterns, thereby reducing the added benefit of the automatic feature extraction typically provided by DL models. Although the CNN model achieved competitive performance with 87% accuracy, it still lagged behind

the ensemble and top-performing ML models, this indicates that deeper architectures, larger datasets, or advanced regularization strategies may be required for DL to reach its full potential in this context. From a clinical perspective, the ensemble's high recall for ADHD cases (0.97) is particularly significant, as it minimizes false negatives, which is an essential factor for early detection and timely intervention in clinical screening scenarios. Relative performance of all models is listed in Table 4.6.

Table 4.6 Performance of all Models on EEG Dataset

Model	Accuracy	Precision	Recall	F1-Score	AUC
Ensemble Model (KNN + SVM + RF)	93.260 %	0.94	0.93	0.93	-
KNN	91.290 %	0.92	0.91	0.91	0.96
SVM	90.304 %	0.90	0.90	0.90	0.96
RF	88.825 %	0.90	0.88	0.89	0.96
CNN	87.017 %	0.87	0.87	0.87	-
AdaBoost	80.772 %	0.81	0.81	0.81	0.88
CNN-LSTM	79.786 %	0.80	0.80	0.80	-
LDA	77.814 %	0.78	0.77	0.77	0.83
DT	77.650 %	0.78	0.78	0.78	0.80
LR	76.910 %	0.76	0.76	0.77	0.85
Logit Boost	75.514 %	0.76	0.76	0.76	0.83

4.5 Neuroimaging Dataset Results

The neuroimaging dataset experiments were designed to classify participants into ADHD and TD groups using features extracted from both structural and functional MRI scans. The primary objective was to use DL architectures

capable of capturing complex spatial and contextual information directly from neuroimaging data, thereby reducing the need for extensive handcrafted feature engineering. Two advanced convolutional neural network (CNN) architectures, DenseNet121 and ResNet50V2, were implemented and systematically compared. Both models utilized transfer learning with ImageNet-pretrained weights and were trained in two phases:

1. A warm-up phase, during which the convolutional layers were frozen and only custom dense layers were trained.
2. A fine-tuning phase, where selected convolutional layers were unfrozen to allow deeper optimization.

Moreover, Key hyperparameters such as learning rate, batch size, and optimizer configurations were carefully tuned to maximize performance while minimizing overfitting.

4.5.1 Deep Learning Architectures

To ensure a fair and standardized comparison, both models were trained with identical configurations. The input images were resized to $224 \times 224 \times 3$, and a batch size of 16 was used. The models were optimized using the Adam optimizer with a learning rate of $1e-3$ during warm-up and $1e-5$ during fine-tuning, guided by the categorical cross-entropy loss function. Regularization techniques, including dropout and L2 weight regularization, were applied to prevent overfitting. Class imbalance in the training set was addressed by calculating class weights proportional to the class distribution, ensuring balanced learning for ADHD and TD samples.

The DenseNet121 architecture incorporated a global average pooling layer followed by two dense layers with batch normalization and dropout, culminating in a SoftMax output layer for classification. Despite the benefits of transfer learning, DenseNet121 showed moderate performance, with an overall accuracy of 70%, precision of 0.70, recall of 0.70, and F1-score of 0.70. The confusion matrix (Fig. 4.20) and sample predictions (Fig. 4.21) provide a detailed view of its classification performance, showing relatively balanced yet less discriminative predictions across both classes. This outcome suggests that DenseNet121 struggled to capture subtle discriminative spatial features in the neuroimaging data. These limitations could potentially be

mitigated by extending the fine-tuning phase, applying more advanced data augmentation techniques, or incorporating multi-modal learning strategies.

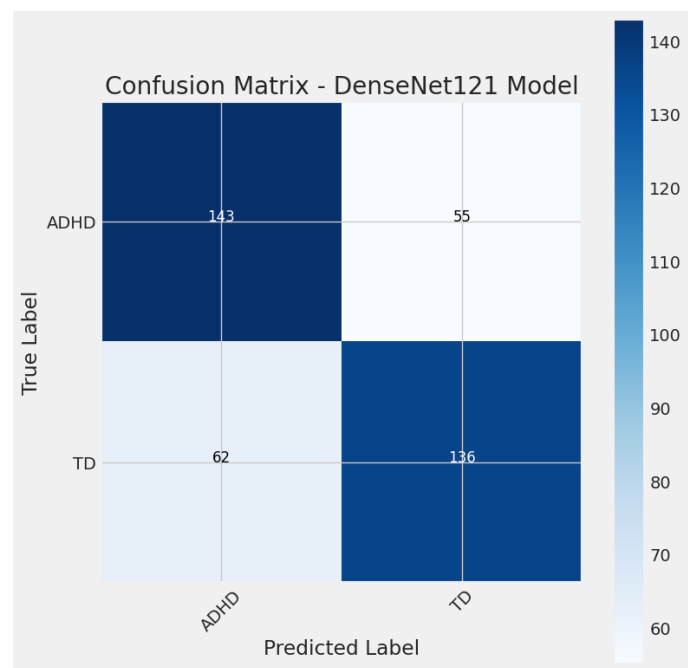


Figure 4.20 Confusion Matrix of DenseNet121 Model on Neuroimaging Dataset

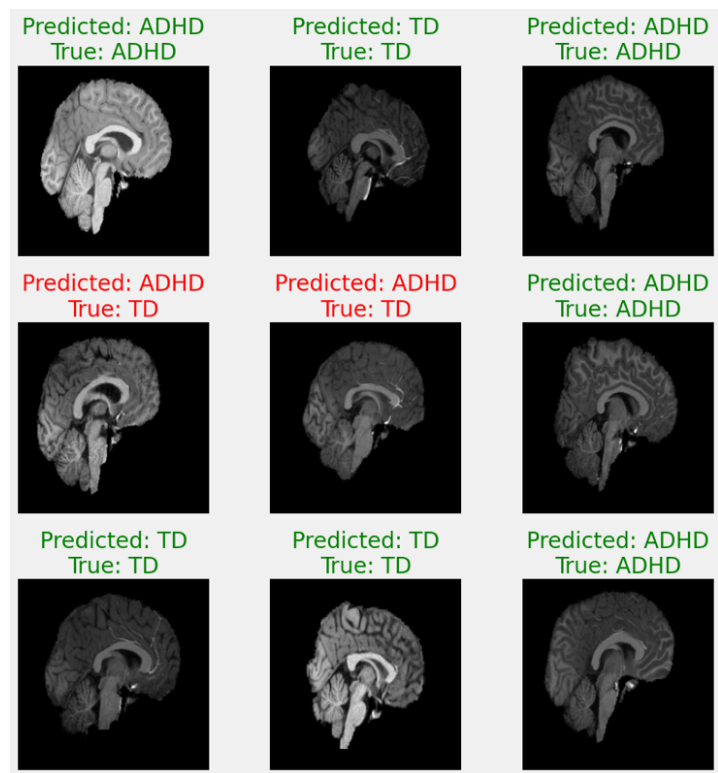


Figure 4.21 Predictions made by DenseNet121 Model on Neuroimaging Dataset

By contrast, the ResNet50V2 model leveraged residual connections, which facilitated deeper feature extraction and improved gradient propagation. This architectural advantage enabled the model to capture complex spatial representations of neuroimaging data more effectively. The ResNet50V2 architecture significantly outperformed DenseNet121, achieving 83% accuracy, along with precision, recall, and F1-score values of 0.83 each. The confusion matrix (Fig. 4.22) and sample predictions (Fig. 4.23) confirm its superior predictive capability, with better sensitivity and specificity across both ADHD and TD classes.

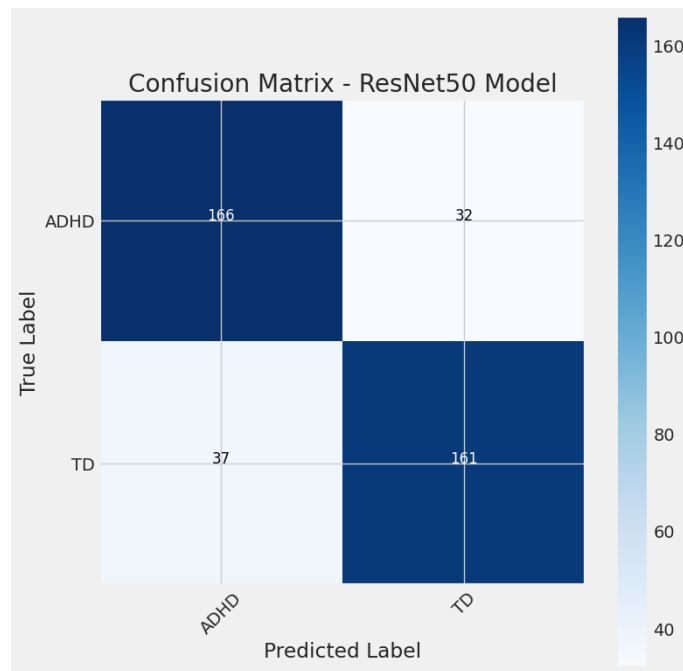


Figure 4.22 Confusion Matrix of ResNet50 Model on Neuroimaging Dataset

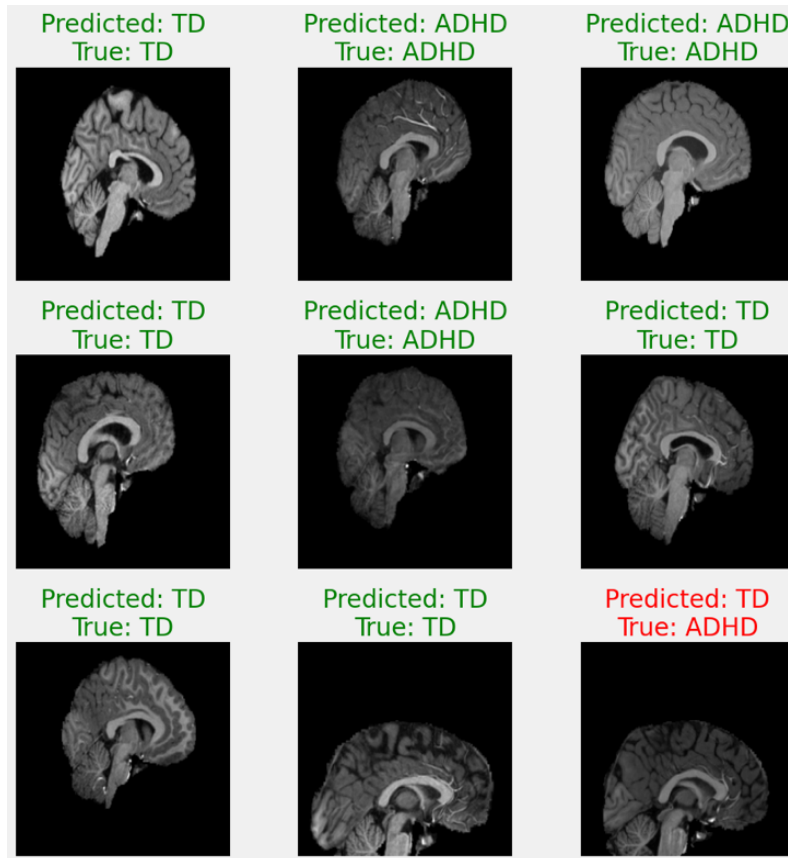


Figure 4.23 Sample Predictions made by ResNet50 Model on Neuroimaging Dataset

4.5.2 Comparative Summary

ResNet50V2 consistently outperformed DenseNet121, achieving an accuracy of 83% with balanced class performance. Its residual architecture enabled more effective feature propagation, making it particularly well-suited for the given dataset. In comparison, DenseNet121 achieved 70% accuracy across all evaluation metrics, whereas ResNet50V2 demonstrated a 13% improvement in accuracy, along with consistently higher precision, recall, and F1-score (all at 0.83). These results highlight the advantage of deeper architectures with residual learning, which appear more effective at leveraging subtle spatial cues present in MRI data for ADHD classification. Key insights from this analysis suggest that transfer learning is highly effective, even with limited datasets, by leveraging generalized image representations. Furthermore, ResNet50V2 proves better suited for neuroimaging classification tasks due to its residual learning mechanism and deeper architecture. For future directions,

multi-modal learning approaches could be explored, such as integrating EEG features with neuroimaging data or combining deep feature extraction with ensemble traditional machine learning classifiers to further enhance performance. A summary of the quantitative results for both models is presented in Table 4.7.

Table 4.7 Performance of DL Models on Neuroimaging Dataset

Model	Accuracy	Precision	Recall	F1-Score
DenseNet121	70.0 %	0.70	0.70	0.70
ResNet50 V2	83.0 %	0.83	0.83	0.83

4.6 Recommendations

The experimental findings across handwriting, drawing, textual, EEG, and neuroimaging modalities provide a basis for prioritizing ADHD detection strategies in children according to their feasibility, diagnostic value, and deployment considerations. Based on model performance, training efficiency, and clinical usability, we recommend a staged diagnostic pathway that begins with low-cost, non-invasive approaches and progressively advances to more resource-intensive modalities only when necessary.

The first line of assessment should rely on handwriting and drawing tasks. These modalities are inexpensive, child-friendly, and easy to administer in both schools and primary care. In our experiments, handwriting models achieved strong balanced performance, while drawing-based models demonstrated near-perfect detection, particularly with KNN, although this result requires careful validation to exclude overfitting. SVMs emerged as the most practical algorithm of choice, balancing accuracy, training speed, and ease of deployment. RFs offer an interpretable alternative, while lightweight stacking ensembles of RF and XGBoost can be employed if computational resources are abundant. For real-world use, these behavioral tasks hold particular clinical value because they capture motor control and attention patterns in a naturalistic and scalable way.

As a second line of assessment, digital textual data, such as caregiver reports, clinical questionnaires, or ethically collected online interactions, can provide a valuable supplementary layer of information. Transformer-based models, particularly MentalRoBERTa, achieved the highest accuracy (87.22%) and sensitivity to nuanced ADHD-related language patterns. While these models are computationally intensive, they can be distilled or quantized for lighter deployment in digital health platforms. Ensemble models could theoretically improve accuracy slightly (87.40%), but the marginal gain does not justify the significantly higher computational cost, longer inference time, and increased complexity for deployment in clinical or school settings. In contexts with limited resources, simpler models such as MNB can serve effectively as initial screening tools. The primary advantage of text-based analysis lies in its potential for continuous or population-scale monitoring, supporting early identification and follow-up while requiring strict attention to privacy and ethical safeguards.

When considering NLP-based detection, it is important to recognize the limitations of binary classification. When ADHD detection is framed simply as Normal versus ADHD, the model often achieves deceptively high accuracy because it only distinguishes ADHD-related language from all other forms of text. This ignores clinically important overlaps with other disorders. For instance, both ADHD and Depression posts may describe “difficulty concentrating” or “trouble focusing,” so a binary Normal vs. ADHD model could mistakenly classify a Depression post as ADHD. In contrast, when the model is trained on a five-class setup (Normal, ADHD, Anxiety, Bipolar, Depression), it faces a more challenging problem, which naturally lowers overall accuracy. However, this added complexity forces the model to learn finer-grained distinctions between ADHD and other disorders that share overlapping symptoms and language markers.

The third line of assessment should involve EEG, which provides objective neurophysiological evidence of ADHD. In our experiments, ensemble methods combining SVM, KNN, and Random Forest delivered the best results, achieving 93.26% accuracy and high recall. However, for practical deployment in clinics, SVM (90.30% accuracy) or RF (88.83% accuracy) models may be preferable due to their faster inference and interpretability.

Although KNN achieved 91.29% accuracy, it is less practical for clinical deployment due to slow inference, higher memory demands, and limited interpretability. EEG requires specialized equipment and trained professionals but remains notably less expensive and invasive than most neuroimaging techniques. It serves as a valuable confirmatory tool, especially when initial behavioral or linguistic assessments produce inconclusive results.

Finally, neuroimaging should be reserved as the last line of assessment, primarily for confirmatory purposes or in complex clinical cases. Although ResNet50V2 achieved the strongest accuracy among MRI-based models in our study, neuroimaging remains expensive, time-consuming, and logistically challenging, particularly in pediatric populations. Its clinical utility lies in providing structural or functional evidence that can support multidisciplinary decision-making rather than serving as a first-line diagnostic tool.

It is also crucial to situate these AI-driven approaches alongside traditional diagnostic methods, such as the DSM-5 ADHD Symptom Checklist. These tools remain essential because they capture self-reported symptoms and clinician-rated observations that no ML model can yet replace. In practice, traditional methods should form the foundation of any ADHD evaluation, with AI-based tools serving as adjuncts that enhance objectivity, scalability, and early screening capacity. For example, handwriting or drawing analysis could flag children at risk in a school setting, prompting administration of the DSM-5 checklist, followed by confirmatory EEG or neuroimaging in more complex cases.

Taken together, these findings support a multistage diagnostic workflow such as starting with initial behavioral screening through handwriting and drawing, followed by digital text monitoring where available, then objective EEG testing in clinical settings, and finally neuroimaging for complex or unresolved cases. Traditional diagnostic inventories should be interwoven throughout this workflow, providing clinical grounding and ensuring that AI outputs are interpreted within the established diagnostic framework. This staged pathway balances accessibility, diagnostic precision, and clinical practicality. Importantly, for early screening, models with high recall should be prioritized to minimize false negatives, as the cost of misdiagnosing a child

with ADHD outweighs the risks of false positives. Moreover, deployment should emphasize explainability, clinician oversight, and ethical data governance, particularly for text-based monitoring. By doing so and adopting this structured approach, clinicians and educators can integrate AI-assisted ADHD detection tools in a manner that maximizes early detection, minimizes unnecessary burdens, and ensures reliability across diverse real-world settings.

Chapter 5. Summary and Conclusion

5.1 Summary

This research focused on using AI techniques to enhance the early detection of mental health disorders, in particular, Attention Deficit Hyperactivity Disorder (ADHD). This is a highly prevalent neurodevelopmental condition characterized by persistent patterns of inattention, hyperactivity, and impulsivity. Despite its prevalence, ADHD remains largely underestimated in society and is often confused for naughtiness in children's attitude. This is aggravated by the fact that ADHD diagnosis is still largely dependent on subjective clinical assessments, which often result in delayed diagnosis, misdiagnosis, and untreated symptoms. This study investigated the use of various AI techniques including traditional ML, DL, and ensemble methods to enhance the early detection of ADHD, especially in children. This was accomplished through a comparative evaluation of different data modalities such as handwriting, drawing trajectories, social media posts, EEG signals, and neuroimaging.

The diagnosis of ADHD remains one of the most complex challenges in neuropsychiatry, relying significantly on subjective clinical interviews and behavioral checks. Misdiagnosis rates remain significant, and many individuals, especially children, experience delayed or missed identification, leading to academic, occupational, and social difficulties. This research was motivated by the need for objective, data-driven approaches to support clinicians, reduce diagnostic delays, and facilitate earlier intervention. Through intense comparison among different modalities of data from handwriting and drawing analysis to neuroimaging and language models, this study aspired to better map the way to make use of ML for diagnosing ADHD.

The objective of this study was to identify which modality, and algorithmic approach could offer the most accurate, accessible, and clinically meaningful solution for ADHD classification. Handwriting classification appears to be a promising screening tool for ADHD. SVM and XGBoost achieved comparable performance, each reaching approximately 90% accuracy with

perfect recall (1.0) and F1-scores of 0.9. The stacking ensemble combining XGBoost and SVM did not provide additional predictive gains, while requiring substantially longer training time (roughly 5,000 seconds). RF and LR achieved moderate accuracy (~80%), KNN reached 70% accuracy, and DT performed the weakest at 60% accuracy. These results indicate that margin-based classifiers (SVM) and gradient boosting algorithms (XGBoost) are the most robust approaches for handwriting-based ADHD detection, offering both high recall and efficient training.

Drawing-based classification performed even better, with KNN boasting a 100% F1-score and accuracy, showing perfect discrimination between control and ADHD samples in the dataset. SVM performed equally well (95.45% accuracy, $F1 = 0.96$), then came RF, XGBoost, and DT (90.91% accuracy, $F1 = 0.92$ all of them), and LR being marginally lower at 86.36% accuracy. Although these results highlight the significant discriminative capability of drawing-based features, the best KNN score may indicate overfitting and should be validated on larger, independent datasets.

Transformer language models performed best in classifying ADHD (textual) posts. MentalRoBERTa was the highest performer at 87.22% accuracy ($F1 = 0.873$), followed closely by vanilla RoBERTa (86.98%), MentalBERT (86.89%), and BERT (86.55%). Ensemble techniques provided a further modest improvement, with a weighted vote ensemble of MentalRoBERTa and BiLSTM giving 87.4% accuracy. Traditional methods, such as Multinomial Naive Bayes, lagged (78.2% accuracy), demonstrating the utility of contextual embeddings in capturing subtle linguistic markers of ADHD in social media text.

EEG classification with KNN, SVM, and RF ensemble model achieved the best accuracy (93.26%) and F1-score (0.93). KNN and SVM also performed well (91.29% and 90.30% accuracy, respectively), both achieving AUC values of 0.96, reflecting high discriminative power. CNN and CNN-LSTM were slightly worse off (87.02% and 79.79% accuracy, respectively), highlighting that conventional ML approaches still have the edge in well-structured, feature-rich EEG datasets. DL algorithms on MRI data were variably successful. The best was ResNet50 V2 with 83% accuracy ($F1 =$

0.83), while DenseNet121 was less successful with 70% accuracy. While encouraging, these are still computationally demanding and are best left to confirmatory diagnosis rather than large-scale screening.

From the comparative evaluation of modalities and models, we recommend a staged and pragmatic pipeline for future ADHD screening. The first stage should use handwriting and drawing tasks, collected in schools or clinics, and analyzed with margin-based classifiers such as SVM or lightweight gradient boosting methods like XGBoost. These tasks are quick to administer, inexpensive to process, child-friendly, and demonstrated the highest balanced accuracy and recall in our experiments. This makes them ideal for early detection. In parallel, brief textual samples such as caregiver questionnaires, structured child narratives, or ethically collected clinical notes can be processed using a distilled or quantized MentalRoBERTa pipeline or transformed into embeddings classified with MLP, to capture linguistic markers with minimal inference latency. For confirmatory assessment, EEG models based on SVM or RF ensembles using standardized 10-20 montage datasets may be applied when initial results are inconclusive, as they offer strong accuracy despite higher resource demands. Neuroimaging should remain reserved for complex or unresolved cases given its cost and logistical challenges.

Hopefully, this work will add value to the growing body of research on AI-assisted mental health diagnostics and bring more awareness to society about this widespread yet overlooked problem. Ultimately, we hope that this work would constitute a modest contribution towards applying AI to better serve society and alleviate the suffering caused by ADHD, a stealthy illness.

5.2 Future Points of Expansion

The results presented in this study exhibit significant potential for ML and DL models for diagnosing ADHD. However, there are many avenues that remain open for further development to improve robustness, generalizability, and clinical impact. Some of these points include:

1. **Validation on Larger and More Diverse Datasets:** The optimal performance of KNN on the drawing dataset (100% accuracy and F1-

score) suggests possible overfitting due to the relatively small sample size. Future studies must cross-validate these models in larger, independent, and demographically diverse samples to ascertain generalizability and rule out dataset-specific biases.

- 2. Multi-Modal Fusion Approaches:** Since handwriting, drawing, EEG, and text-based models all had strong but distinct predictive capacities, the next research direction could involve exploring multi-modal fusion methods (early fusion, late fusion, or hybrid architectures) to combine complementary information. As a suggested experiment, one might combine EEG biomarkers (93.26% ensemble accuracy) with text-based transformer embeddings (MentalRoBERTa $F1 = 0.873$) to increase sensitivity and reduce false negatives for borderline cases. Moreover, If datasets are diverse and drawn from multiple sources, comparative studies across modalities can yield more generalizable and robust results. To achieve this, future work could incorporate data from multiple institutions, age groups, and demographic backgrounds, or augment datasets to capture broader variability.
- 3. Explainability and Interpretability:** Clinical deployment of ML systems must not only be precise but also interpretable. Techniques such as Shapley Additive exPlanations (SHAP) or Local Interpretable Model-agnostic Explanations (LIME) for attributing features, transformers' attention heatmaps for CNN-based neuroimaging models could uncover which features, brain regions, or linguistic indicators affect model predictions [87, 88]. This can make predictions more transparent and trustworthy for clinicians.
- 4. Temporal and Longitudinal Analysis:** Most experiments in this study were cross-sectional, meaning the data were collected at a single point in time. This provides a snapshot rather than capturing changes over time. In contrast, longitudinal studies track the same individuals across multiple time points, allowing researchers to observe the progression of symptoms, developmental trajectories, or responses to interventions. Future studies should incorporate longitudinal datasets to assess if these models could be used to predict ADHD risk patterns over time, detect

symptom development, or quantify treatment response. Furthermore, the integration of cross-sectional and longitudinal data may increase model generalizability and yield more clinically useful information.

- 5. Model Optimization for Deployment:** Some models, such as the XGBoost + SVM stacking ensemble on handwriting data (training time > 4600 seconds), were expensive to compute and barely showed performance improvements. Future research needs to tackle model compression, pruning, and knowledge distillation for producing slimmer, faster models deployable in real-time on edge devices or clinical decision-support systems.
- 6. Ethical and Privacy Considerations:** Particularly in the case of text-based data gathered from social media, future research needs to explore privacy-preserving ML approaches such as federated learning and differential privacy that facilitate user data protection without compromising efficient model training.
- 7. Clinical Trials and Regulatory Pathways:** To translate from research to clinical practice, these models need to be proven in prospective clinical trials and verified under regulatory standards (e.g., Food and Drug Administration (FDA), Conformité Européenne (CE)). Moreover, having standard benchmarks and protocols for the AI detection of ADHD will accelerate clinician and policymaker adoption.
- 8. Data Augmentation with Generative Models:** To mitigate the small-sample limitation, future work could utilize generative models, such as WGANs, to synthesize realistic MRI images. WGANs improve training stability over traditional GANs and produce high-quality synthetic samples, which could help balance underrepresented classes and enhance the robustness of classifiers [69, 70].

The above improvements should hopefully make progress towards creating more robust, interpretable, and scalable ADHD detection systems that can be used as first-line screening tools and assist clinicians in making diagnostic recommendations.

Reference

- [1] G. Kumar, B. Sharma, R. K. Tiwari, R. Kumar, and M. K. Singh, "A review on ADHD disorder, diagnosis and therapeutic approach in children and adolescents," *Brain-Apparatus Communication: A Journal of Bacomics*, vol. 3, no. 1, Sep. 2024. doi:10.1080/27706710.2024.2400070
- [2] T. E. Wilens and T. J. Spencer, "Understanding attention-deficit/hyperactivity disorder from childhood to adulthood," *Postgraduate Medicine*, vol. 122, no. 5, pp. 97–109, Sep. 2010. doi:10.3810/pgm.2010.09.2206
- [3] N. A. P. Association, *Diagnostic and Statistical Manual of Mental Disorders*. 2013. doi: 10.1176/appi.books.9780890425596.
- [4] A. Niina, O. Eyre, R. Wootton, E. Stergiakouli, A. Thapar, and L. Riglin, "Exploring ADHD Symptoms and Associated Impairment across Development," *Journal of Attention Disorders*, vol. 26, no. 6, pp. 822–830, Jun. 2021, doi: 10.1177/10870547211025612.
- [5] D. Singh, A. Kumar, R. Kaur, S. Mandal, and S. Ku-Mar, "A short review on Attention Deficit Hyperactivity Disorder (ADHD): Types, Symptoms, Prevention and Treatment," *International Journal for Pharmaceutical Research Scholars*, vol. 7, pp. 1202-1204, 2022, doi: 10.35629/7781-070112021204.
- [6] S. Young *et al.*, "Guidance for identification and treatment of individuals with attention deficit/hyperactivity disorder and autism spectrum disorder based upon expert consensus," *BMC Medicine*, vol. 18, no. 1, May 2020, doi: 10.1186/s12916-020-01585-y.
- [7] S. Lee, L. Wang, and C. Yen, "Identification of diagnostic and therapeutic biomarkers for attention-deficit/hyperactivity disorder," *The Kaohsiung Journal of Medical Sciences*, Jan. 2025, doi: 10.1002/kjm2.12931.
- [8] R. Knott *et al.*, "Age at diagnosis and diagnostic delay across attention-deficit hyperactivity and autism spectrums," *Australian & New Zealand Journal of Psychiatry*, vol. 58, no. 2, pp. 142–151, Oct. 2023, doi: 10.1177/00048674231206997.
- [9] N. Website, "Attention deficit hyperactivity disorder (ADHD)," *nhs.uk*, Mar. 13, 2024. <https://www.nhs.uk/conditions/attention-deficit-hyperactivity-disorder-adhd/>
- [10] "Data and statistics on ADHD," *Attention-Deficit / Hyperactivity Disorder (ADHD)*, Nov. 19, 2024. <https://www.cdc.gov/adhd/data/index.html>
- [11] K. Bruchmüller, J. Margraf, and S. Schneider, "Is ADHD diagnosed in accord with diagnostic criteria? Overdiagnosis and influence of client

- gender on diagnosis.,” *Journal of Consulting and Clinical Psychology*, vol. 80, no. 1, pp. 128–138, Dec. 2011, doi: 10.1037/a0026582.
- [12] WV ACC Guidelines, “ADHD Guidelines - WV ACC Guidelines,” *WV ACC Guidelines*, Jan. 15, 2025. <https://wvadhd.org/guidelines/>
- [13] G. Russell and T. Ford, “The costs and benefits of diagnosis of ADHD: Commentary on Holden et al.,” *Child and Adolescent Psychiatry and Mental Health*, vol. 8, no. 1, p. 7, 2014. doi:10.1186/1753-2000-8-7
- [14] S. E. Holden *et al.*, “The prevalence and incidence, resource use and financial costs of treating people with attention deficit/hyperactivity disorder (ADHD) in the United Kingdom (1998 to 2010),” *Child and Adolescent Psychiatry and Mental Health*, vol. 7, no. 1, Oct. 2013. doi:10.1186/1753-2000-7-34
- [15] L. Hasemann *et al.*, “Health care costs of incident ADHD in children and adolescents in Germany - A claims data analysis within the framework of the consortium project INTEGRATE-ADHD.,” *PubMed*, vol. 9, no. 3, p. e12303, Sep. 2024, doi: 10.25646/12303.
- [16] J. Thome *et al.*, “Biomarkers for attention-deficit/hyperactivity disorder (ADHD). A consensus report of the WFSBP Task Force on Biological Markers and the World Federation of ADHD,” *The World Journal of Biological Psychiatry*, vol. 13, no. 5, pp. 379–400, Jul. 2012. doi:10.3109/15622975.2012.690535
- [17] “Adult ADHD Self-Report Scale (ASRSV1.1) - Psychology Tools,” *Psychology Tools*, 2001. <https://psychology-tools.com/test/adult-adhd-self-report-scale>
- [18] “Copeland Symptom Checklist Template & Example | Free PDF download.” <https://www.carepatron.com/templates/copeland-symptom-checklist>
- [19] NeuroLaunch.com, “Understanding SASI ADHD: A Comprehensive Guide to Screening and assessment,” *NeuroLaunch.com*, Aug. 04, 2024. https://neurolaunch.com/sasi-adhd/#google_vignette
- [20] J. N. Kiro, T. Kundu, and Mohan Kumar Dehury, “Detecting ADHD Among Children Using EEG Signals,” *Communications in computer and information science*, pp. 203–217, Jan. 2024, doi: https://doi.org/10.1007/978-3-031-64067-4_14.
- [21] G. Coppersmith, M. Dredze, C. Harman, and K. Hollingshead, “From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses,” *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2015. doi:10.3115/v1/w15-1201

- [22] M. A. Teruel, J. Sanchis, N. Ruiz-Robledillo, N. Albaladejo-Blázquez, R. Ferrer-Cascales, and J. Trujillo, "Measuring attention of ADHD patients by means of a computer game featuring biometrical data gathering," *Heliyon*, vol. 10, no. 5, p. e26555, Feb. 2024, doi: 10.1016/j.heliyon.2024.e26555.
- [23] H. Roebuck, C. Freigang, and J. G. Barry, "Continuous performance tasks: not just about sustaining attention," *Journal of Speech Language and Hearing Research*, vol. 59, no. 3, pp. 501–510, Apr. 2016, doi: 10.1044/2015_jslhr-l-15-0068.
- [24] X. Wang, Q. Jia, L. Liang, W. Zhou, W. Yang, and J. Mu, "Artificial intelligence in ADHD: a global perspective on research hotspots, trends and clinical applications," *Frontiers in Human Neuroscience*, vol. 19, Apr. 2025, doi: 10.3389/fnhum.2025.1577585.
- [25] L. Caselles-Pina, A. Quesada-López, A. Sújar, E. M. G. Hernández, and D. Delgado-Gómez, "A systematic review on the application of machine learning models in psychometric questionnaires for the diagnosis of attention deficit hyperactivity disorder," *European Journal of Neuroscience*, vol. 60, no. 3, pp. 4115–4127, Feb. 2024, doi: 10.1111/ejn.16288.
- [26] M. M. Rahman, "AI for ADHD: Opportunities and Challenges," *Journal of Attention Disorders*, vol. 27, no. 8, pp. 797–799, Apr. 2023, doi: 10.1177/10870547231167608.
- [27] I. Ameer, M. Arif, G. Sidorov, H. Gómez-Adorno, and A. Gelbukh, "Mental Illness Classification on Social Media Texts using Deep Learning and Transfer Learning," *arXiv (Cornell University)*, Jan. 2022, doi: 10.48550/arxiv.2207.01012.
- [28] J. Kim, J. Lee, E. Park, and J. Han, "A deep learning model for detecting mental illness from user content on social media," *Scientific Reports*, vol. 10, no. 1, Jul. 2020. doi:10.1038/s41598-020-68764-y
- [29] J. Dey and D. Desai, "NLP based approach for classification of Mental Health Issues Using LSTM and glove embeddings," *International Journal of Advanced Research in Science, Communication and Technology*, pp. 347–354, Jan. 2022. doi:10.48175/ijarsct-2296
- [30] A. Tom et al., "Detecting Mental Health Disorders Using NLP," 2023 IEEE Pune Section International Conference (PuneCon), Pune, India, 2023, pp. 1-7, doi: 10.1109/PuneCon58714.2023.10450025.
- [31] I. J. Dristy, A. M. Saad and A. A. Rasel, "Mental Health Status Prediction Using ML Classifiers with NLP-Based Approaches," 2022 International Conference on Recent Progresses in Science, Engineering

- and Technology (ICRPSET), Rajshahi, Bangladesh, 2022, pp. 1-6, doi: 10.1109/ICRPSET57982.2022.10188544.
- [32] R. Qayyum, H. Afzal, K. Mahmood and N. Iltaf, "Detection and Analysis of Mental Health Illness using Social Media," 2023 International Conference on Communication Technologies (ComTech), Rawalpindi, Pakistan, 2023, pp. 34-41, doi: 10.1109/ComTech57708.2023.10165143.
- [33] A. Pandey and S. Kumar, "Mental Health and Stress Prediction Using NLP and Transformer-Based Techniques," 2024 IEEE Symposium on Wireless Technology & Applications (ISWTA), Kuala Lumpur, Malaysia, 2024, pp. 61-66, doi: 10.1109/ISWTA62130.2024.10651843.
- [34] M. Abdullah and N. Negied, "Detection and prediction of future mental disorder from social media data using machine learning, ensemble learning, and large language models," *IEEE Access*, vol. 12, pp. 120553–120569, Jan. 2024, doi: 10.1109/access.2024.3406469.
- [35] F. Kabir, M. A. Hossain, A. F. M. Minhazur Rahman and S. Z. Mishu, "Depression Detection From Social Media Textual Data Using Natural Language Processing and Machine Learning Techniques," 2023 26th International Conference on Computer and Information Technology (ICCIT), Cox's Bazar, Bangladesh, 2023, pp. 1-6, doi: 10.1109/ICCIT60459.2023.10441612.
- [36] M. Mengi and D. Malhotra, "A Study of the ADHD-200 Repository for Attention Deficit Categorization Using Deep Neural Networks," 2023 *IEEE Fifth International Conference on Advances in Electronics, Computers and Communications (ICAIECC)*, Bengaluru, India, 2023, pp. 1-5, doi: 10.1109/ICAIECC59324.2023.10560312.
- [37] S. S and M. B. Anandaraju, "Machine Learning Approach to Predict ADHD Types using EEG Signal Data," 2023 International Conference on Recent Advances in Science and Engineering Technology (ICRASET), B G NAGARA, India, 2023, pp. 1-4, doi: 10.1109/ICRASET59632.2023.10420342.
- [38] S. García-Ponsoda, A. Maté, and J. Trujillo, "Refining ADHD diagnosis with EEG: The impact of preprocessing and temporal segmentation on classification accuracy," *Computers in Biology and Medicine*, vol. 183, p. 109305, Nov. 2024, doi: 10.1016/j.combiomed.2024.109305.
- [39] H. O. Elkhodary and S. M. Youssef, "An Improved Model for The Diagnosis of Attention Deficit/Hyperactivity Disorder (ADHD) Using Resting-state fMRI Data," 2024 International Conference on Machine Intelligence and Smart Innovation (ICMISI), Alexandria, Egypt, 2024, pp. 60-63, doi: 10.1109/ICMISI61517.2024.10580368.

- [40] S. Abdolmaleki and M. S. Abadeh, "Brain MR Image Classification for ADHD Diagnosis Using Deep Neural Networks," 2020 International Conference on Machine Vision and Image Processing (MVIP), Iran, 2020, pp. 1-5, doi: 10.1109/MVIP49855.2020.9116877.
- [41] "Brain anatomy and how the brain works," *Johns Hopkins Medicine*, Jul. 14, 2021. <https://www.hopkinsmedicine.org/health/conditions-and-diseases/anatomy-of-the-brain>
- [42] A. Poulton and R. Nanan, "Subcortical brain volume differences in participants with attention deficit hyperactivity disorder in children and adults," *The Lancet Psychiatry*, vol. 4, no. 6, pp. 437–438, Jun. 2017, doi: [https://doi.org/10.1016/s2215-0366\(17\)30105-0](https://doi.org/10.1016/s2215-0366(17)30105-0).
- [43] M. A. Mooney *et al.*, "Smaller total brain volume but not subcortical structure volume related to common genetic risk for ADHD," *Psychological Medicine*, pp. 1–10, Jan. 2020, doi: <https://doi.org/10.1017/s0033291719004148>.
- [44] A. Joseph, C. E. Kosmas, C. Patel, H. Doll, and P. Asherson, "Health-Related Quality of Life and Work Productivity of Adults With ADHD: A U.K. Web-Based Cross-Sectional Survey," *Journal of Attention Disorders*, vol. 23, no. 13, pp. 1610–1623, Sep. 2018, doi: 10.1177/1087054718799367.
- [45] A. B. M. Fuermaier, L. Tucha, M. Butzbach, M. Weisbrod, S. Aschenbrenner, and O. Tucha, "ADHD at the workplace: ADHD symptoms, diagnostic status, and work-related functioning," *Journal of Neural Transmission*, Feb. 2021, doi: 10.1007/s00702-021-02309-z.
- [46] L. Whitten, "Functional Magnetic resonance Imaging (fMRI): an invaluable tool in translational neuroscience," Dec. 2012. doi: 10.3768/rtpress.2012.op.0010.1212.
- [47] C. Yen, C.-L. Lin, and M.-C. Chiang, "Exploring the Frontiers of Neuroimaging: A review of recent advances in understanding brain functioning and disorders," *Life*, vol. 13, no. 7, p. 1472, Jun. 2023, doi: 10.3390/life13071472.
- [48] B. M. Rashed and N. Popescu, "Machine Learning Techniques for Medical Image Processing," 2021 International Conference on e-Health and Bioengineering (EHB), Iasi, Romania, 2021, pp. 1-4, doi: 10.1109/EHB52898.2021.9657673.
- [49] V. Nasteski, "An overview of the supervised machine learning methods," *HORIZONS B*, vol. 4, pp. 51–62, Dec. 2017, doi: 10.20544/horizons.b.04.1.17.p05.

- [50] J. A. Ilemobayo *et al.*, “Hyperparameter tuning in Machine Learning: A Comprehensive review,” *Journal of Engineering Research and Reports*, vol. 26, no. 6, pp. 388–395, Jun. 2024, doi: 10.9734/jerr/2024/v26i61188.
- [51] E. Bartz, T. Bartz-Beielstein, M. Zaefferer, and O. Mersmann, *Hyperparameter Tuning for Machine and Deep Learning with R*. 2023. doi: 10.1007/978-981-19-5170-1.
- [52] K. Qu, “Research on linear regression algorithm,” *MATEC Web of Conferences*, vol. 395, p. 01046, Jan. 2024, doi: 10.1051/mateconf/202439501046.
- [53] H. Wang, J. Xiong, Z. Yao, M. Lin, and J. Ren, “Research Survey on Support Vector Machine,” Dec. 2017, doi: <https://doi.org/10.4108/eai.13-7-2017.2270596>.
- [54] I. D. Mienye and N. Jere, "A Survey of Decision Trees: Concepts, Algorithms, and Applications," in *IEEE Access*, vol. 12, pp. 86716-86727, 2024, doi: 10.1109/ACCESS.2024.3416838.
- [55] G. Louppe, “Understanding random forests: From Theory to practice,” *arXiv.org*, Jul. 28, 2014. <https://arxiv.org/abs/1407.7502>
- [56] K. Taunk, S. De, S. Verma and A. Swetapadma, "A Brief Review of Nearest Neighbor Algorithm for Learning and Classification," 2019 International Conference on Intelligent Computing and Control Systems (ICCS), Madurai, India, 2019, pp. 1255-1260, doi: 10.1109/ICCS45141.2019.9065747.
- [57] T. Chen and C. Guestrin, “XGBoost: a Scalable Tree Boosting System,” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, vol. 1, no. 1, pp. 785–794, Aug. 2016, doi: <https://doi.org/10.1145/2939672.2939785>.
- [58] IBM, “Linear Discriminant Analysis,” *Ibm.com*, Nov. 27, 2023. <https://www.ibm.com/think/topics/linear-discriminant-analysis>
- [59] S. S. Rambhatla, M. Jones, and R. Chellappa, “To Boost or not to Boost: On the Limits of Boosted Neural Networks,” *arXiv.org*, Jul. 28, 2021. <https://arxiv.org/abs/2107.13600>
- [60] L. Yang and A. Shami, “On hyperparameter optimization of machine learning algorithms: Theory and practice,” *Neurocomputing*, vol. 415, pp. 295–316, Jul. 2020, doi: 10.1016/j.neucom.2020.07.061.
- [61] L. Alzubaidi *et al.*, “Review of deep learning: concepts, CNN architectures, challenges, applications, future directions,” *Journal of Big Data*, vol. 8, no. 1, Mar. 2021, doi: 10.1186/s40537-021-00444-8.
- [62] A. Krizhevsky, I. Sutskever, and G. E. Hinton “ImageNet Classification with Deep Convolutional Neural Networks,” *Neural Information*

- Processing Systems*, vol. 25, pp. 1097–1105, Dec. 2012, [Online]. Available: http://books.nips.cc/papers/files/nips25/NIPS2012_0534.pdf
- [63] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, and M. Parmar, “A review of convolutional neural networks in computer vision,” *Artificial Intelligence Review*, vol. 57, no. 4, Mar. 2024, doi: 10.1007/s10462-024-10721-6.
- [64] I. D. Mienye and Y. Sun, “A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects,” in *IEEE Access*, vol. 10, pp. 99129–99149, 2022, doi: 10.1109/ACCESS.2022.3207287.
- [65] T. N. Rincy and R. Gupta, “Ensemble Learning Techniques and its Efficiency in Machine Learning: A Survey,” *2nd International Conference on Data, Engineering and Applications (IDEA)*, Bhopal, India, 2020, pp. 1–6, doi: 10.1109/IDEA49133.2020.9170675.
- [66] A. Mohammed and R. Kora, “A comprehensive review on ensemble deep learning: Opportunities and challenges,” *Journal of King Saud University - Computer and Information Sciences*, vol. 35, no. 2, pp. 757–774, Feb. 2023, doi: 10.1016/j.jksuci.2023.01.014.
- [67] Karthika. S and M. Durgadevi, “Generative Adversarial Network (GAN): a general review on different variants of GAN and applications,” *2022 7th International Conference on Communication and Electronics Systems (ICCES)*, pp. 1–8, Jul. 2021, doi: 10.1109/icces51350.2021.9489160.
- [68] A. Borji, “Pros and cons of GAN evaluation measures,” *arXiv (Cornell University)*, Jan. 2018, doi: 10.48550/arxiv.1802.03446.
- [69] M. Arjovsky, S. Chintala, and L. Bottou, “Wasserstein GAN,” *arXiv (Cornell University)*, Jan. 2017, doi: 10.48550/arxiv.1701.07875.
- [70] L. Weng, “From GAN to WGAN,” *arXiv.org*, Apr. 18, 2019. <https://arxiv.org/abs/1904.08994>
- [71] R. Cohen, B. Cohen-Kroitoru, A. Halevy, S. Aharoni, I. Aizenberg, and A. Shuper, “Handwriting in children with Attention Deficient Hyperactive Disorder: role of graphology,” *BMC Pediatrics*, vol. 19, no. 1, Dec. 2019, doi: <https://doi.org/10.1186/s12887-019-1854-3>.
- [72] J. Shin, Md. Maniruzzaman, Y. Uchida, Md. A. M. Hasan, A. Megumi, and A. Yasumura, “Handwriting-Based ADHD detection for children having ASD using machine learning approaches,” *IEEE Access*, vol. 11, pp. 84974–84984, Jan. 2023, doi: 10.1109/access.2023.3302903.
- [73] D. M. Low, L. Rumker, T. Talkar, J. Torous, G. Cecchi, and S. S. Ghosh, “Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on Reddit during COVID-

- 19: Observational study,” *Journal of Medical Internet Research*, vol. 22, no. 10, p. e22635, Sep. 2020, doi: 10.2196/22635.
- [74] D. R. Beddiar, S. Jahan, and M. Ouassalah, “Data expansion using back translation and paraphrasing for hate speech detection,” *Online Social Networks and Media*, vol. 24, p. 100153, Jun. 2021, doi: 10.1016/j.osnem.2021.100153.
- [75] Statista, “Distribution of Reddit.com traffic 2024, by country,” *Statista*, Nov. 11, 2024. <https://www.statista.com/statistics/325144/reddit-global-active-user-distribution/>
- [76] M. Barthel and M. Barthel, “1. Reddit news users more likely to be male, young and digital in their news preferences,” *Pew Research Center*, Apr. 14, 2024. <https://www.pewresearch.org/journalism/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/>
- [77] Ali Motie Nasrabadi, Armin Allahverdy, Mehdi Samavati, Mohammad Reza Mohammadi, June 10, 2020, “EEG data for ADHD / Control children”, IEEE Dataport, doi: <https://dx.doi.org/10.21227/rzfh-zn36>.
- [78] P. Bellec, C. Chu, F. Chouinard-Decorte, Y. Benhajali, D. S. Margulies, and R. C. Craddock, “The Neuro Bureau ADHD-200 Preprocessed repository,” *NeuroImage*, vol. 144, pp. 275–286, Jul. 2016, doi: 10.1016/j.neuroimage.2016.06.034.
- [79] “ADHD200 Preprocessed Anatomical Dataset,” Dec. 02, 2024. <https://www.kaggle.com/datasets/purnimakumarr/adhd200-preprocessed-anatomical-dataset>
- [80] M. Jas, D. A. Engemann, Y. Bekhti, F. Raimondo, and A. Gramfort, “Autoreject: Automated artifact rejection for MEG and EEG data,” *NeuroImage*, vol. 159, pp. 417–429, Jun. 2017, doi: 10.1016/j.neuroimage.2017.06.030.
- [81] “autoreject - autoreject 0.4.2 documentation.” <https://autoreject.github.io/stable/index.html>
- [82] W. Xu, Y.-L. Fu, and D. Zhu, “ResNet and its application to medical image processing: Research progress and challenges,” *Computer Methods and Programs in Biomedicine*, vol. 240, p. 107660, Jun. 2023, doi: 10.1016/j.cmpb.2023.107660.
- [83] T. Zhou, X. Ye, H. Lu, X. Zheng, S. Qiu, and Y. Liu, “Dense convolutional network and its application in medical image analysis,” *BioMed Research International*, vol. 2022, pp. 1–22, Apr. 2022, doi: 10.1155/2022/2384830.

- [84] A. W. Salehi *et al.*, “A study of CNN and transfer learning in Medical imaging: Advantages, challenges, future scope,” *Sustainability*, vol. 15, no. 7, p. 5930, Mar. 2023, doi: 10.3390/su15075930.
- [85] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, “Transfer learning for medical image classification: a literature review,” *BMC Medical Imaging*, vol. 22, no. 1, Apr. 2022, doi: 10.1186/s12880-022-00793-7.
- [86] H. Allam, L. Makubvure, B. Gyamfi, K. N. Graham, and K. Akinwolere, “Text classification: How machine learning is revolutionizing text categorization,” *Information*, vol. 16, no. 2, p. 130, Feb. 2025, doi: 10.3390/info16020130.
- [87] A. M. Salih *et al.*, “A perspective on explainable artificial intelligence methods: SHAP and LIME,” *Advanced Intelligent Systems*, Jun. 2024, doi: 10.1002/aisy.202400304.
- [88] M. Chung, J. B. Won, G. Kim, Y. Kim, and U. Ozbulak, “Evaluating visual explanations of attention maps for Transformer-Based medical Imaging,” in *Lecture notes in computer science*, 2025, pp. 110–120. doi: 10.1007/978-3-031-77610-6_11.

الملخص

اضطراب نقص الانتباه وفرط النشاط (ADHD) هو حالة اضطراب في النمو العصبي شائعة للغاية تتميز بأنماط مستمرة من عدم الانتباه وفرط النشاط والاندفاع. على الرغم من انتشاره، لا يزال اضطراب نقص الانتباه وفرط النشاط (ADHD) يُستخف به إلى حد كبير في المجتمع وغالبًا ما يُخلط بينه وبين سوء سلوك الأطفال. ويزداد الأمر سوءًا بسبب حقيقة أن تشخيص اضطراب نقص الانتباه وفرط النشاط (ADHD) لا يزال يعتمد إلى حد كبير على التقييمات السريرية الذاتية، مما يؤدي غالبًا إلى تأخير التشخيص والتشخيص الخاطئ والأعراض غير المعالجة. تبحث هذه الدراسة في الاستخدام المحتمل لتقنيات الذكاء الاصطناعي (AI) المختلفة، بما في ذلك التعلم الآلي التقليدي (ML) والتعلم العميق (DL) وطرق التجميع لتعزيز الكشف المبكر عن ADHD، خاصة عند الأطفال. يتم تحقيق ذلك من خلال تقييم مقارن لطرق البيانات المختلفة مثل الكتابة اليدوية ومسارات الرسم ومنشورات وسائل التواصل الاجتماعي وإشارات تخطيط كهربية الدماغ (EEG) والتصوير العصبي.

قام الإطار التجريبي بتقييم كل طريقة بشكل منهجي من خلال معالجة مسبقة دقيقة، وتحسين المعلمات الفائقة، والتحقق المتبادل. على بيانات الكتابة اليدوية، حققت آلة الدعم المتجه (SVM) و XGBoost ومجموعة التراص أداءً متساويًا (دقة $\approx 90\%$ ، درجة $F1 = 0.9$ ، استرجاع $= 1.0$)، مع توفير SVM التوازن الأمثل بين الدقة والكفاءة الحسابية. في التصنيف القائم على الرسم، حققت K-Nearest Neighbors (KNN) أعلى أداء تمييزي، مع نتائج مثالية في جميع المقاييس (الدقة $= 100\%$ ، درجة $F1 = 1.0$)، تليها SVM بنسبة 95.45% ، مما يشير إلى إمكانية عالية للفحص القائم على المهام السلوكية. بالنسبة للبيانات النصية، تفوقت نماذج اللغة القائمة على المحولات بشكل كبير على المعايير الإحصائية والتسلسلية. حقق MentalRoBERTa دقة بنسبة 87.22% (درجة $F1 = 0.873$)، مع استراتيجيات التجميع التي توفر تحسينات طفيفة ولكنها متسقة. فضل التصنيف القائم على EEG طرق المجموعة التي تدمج KNN و SVM و Random Forest، وحقق دقة 93.26% واسترجاعًا شبه مثالي (0.97)، مما يجعلها مناسبة للفحص التأكيد في البيئات السريرية. أظهر تصنيف التصوير العصبي باستخدام التعلم النقلي أن ResNet50V2 حقق دقة 83% ، مما يسلط الضوء على الفعالية المحدودة للكشف عن ADHD القائم على التصوير بالرنين المغناطيسي نظرًا للتحديات في جمع بيانات التصوير بالرنين المغناطيسي.

تدعم النتائج بشكل جماعي مسارًا تشخيصيًا مرحليًا يبدأ بتحليل خط اليد والرسم من أجل فحص غير جراحي ومنخفض التكلفة، ويشمل تحليل النص الرقمي للمراقبة الطولية، ويستخدم تخطيط كهربية الدماغ كطريقة تأكيدية موضوعية، ويحتفظ أخيرًا بالتصوير العصبي لحالات الاضطرابات العقلية المعقدة. يتم التركيز على النماذج ذات الاسترجاع العالي لتقليل النتائج السلبية الخاطئة ودعم التدخل المبكر. تسلط الدراسة الضوء أيضًا على قابلية التفسير والكفاءة الحسابية والأخلاقيات والأبحاث المستقبلية في مجال الاندماج متعدد الوسائط والتعلم الذي يحافظ على الخصوصية. نأمل أن يساهم هذا العمل في توسيع نطاق الأبحاث المتعلقة بتشخيص الصحة العقلية بمساعدة الذكاء الاصطناعي، وزيادة وعي المجتمع بهذه المشكلة المنتشرة ولكن المهملة.

جامعة البحرين

قسم الهندسة الكهربائية
والإلكترونية



كلية الهندسة

تحسين الأداء في الكشف المبكر عن اضطراب نقص الانتباه وفرط النشاط باستخدام الذكاء الاصطناعي

أطروحة مقدمة كجزء من متطلبات الحصول على درجة الماجستير
في نظم الذكاء الاصطناعي

إعداد

محمد ظفر اقبال بدر

الرقم الجامعي: ٢٠١٨٠٩١٠

إشراف

د. زهير البحري

أستاذ مشارك

جامعة البحرين

مملكة البحرين

نوفمبر، ٢٠٢٥ م