CAPITAL UNIVERSITY OF SCIENCE AND TECHNOLOGY, ISLAMABAD



A Multimodal Generative Adversarial Network Based Explainable Framework for Parkinson's Disease Progression Subtype Prediction

by

Amna Ayub

A thesis submitted in partial fulfillment for the degree of Master of Science

in the

Faculty of Computing
Department of Software Engineering

2026

Copyright © 2026 by Amna Ayub

All rights reserved. No part of this thesis may be reproduced, distributed, or transmitted in any form or by any means, including photocopying, recording, or other electronic or mechanical methods, by any information storage and retrieval system without the prior written permission of the author.



CERTIFICATE OF APPROVAL

A Multimodal Generative Adversarial Network Based Explainable Framework For Parkinson's Disease Progression Subtype Prediction

by

Amna Ayub

(MAI243001)

THESIS EXAMINING COMMITTEE

S. No.	Examiner	Name	Organization
(a)	External Examiner	Dr. Iram Matloob	FJWU, Rawalpindi
(b)	Internal Examiner	Dr. Farah Haneef	CUST, Islamabad

Dr. Sabeen Masood
Thesis Supervisor
October , 2026

Dr. Nadeem Anjum
Head
Dept. of Software Engineering
October, 2026

Dr. M. Abdul Qadir
Dean
Faculty of Computing
October, 2026

Author's Declaration

I,**Amna Ayub** hereby state that my MS thesis titled "**A Multimodal Generative Adversarial Network Based Explainable Framework for Parkinson's Disease Progression Subtype Prediction**" is my own work and has not been submitted previously by me for taking any degree from Capital University of Science and Technology, Islamabad or anywhere else in the country/abroad.

At any time if my statement is found to be incorrect even after my graduation, the University has the right to withdraw my MS Degree.



(Amna Ayub)
Registration No: MAI243001

Plagiarism Undertaking

I solemnly declare that research work presented in this thesis titled "**A Multimodal Generative Adversarial Network Based Explainable Framework for Parkinson's Disease Progression Subtype Prediction**" is solely my research work with no significant contribution from any other person. Small contribution/help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero tolerance policy of the HEC and Capital University of Science and Technology towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS Degree, the University reserves the right to withdraw/revoke my MS degree and that HEC and the University have the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized work.



(Amna Ayub)
Registration No: MAI243001

Acknowledgement

First and foremost, I would like to express my sincere gratitude to Almighty Allah for granting me the strength, patience, wisdom and perseverance to complete this research successfully. Without His countless blessings, this achievement would not have been possible.I would like to express my deepest gratitude to my respected supervisor, Dr. Sabeen Masood, for her continuous guidance, valuable suggestions, constructive feedback and unwavering support throughout the completion of this research. Her encouragement and expert supervision greatly contributed to the successful completion of this thesis.

I am also grateful to the faculty members of the Department of Software Engineering, Capital University of Science and Technology (CUST), Islamabad, for providing an excellent academic environment and for sharing their valuable knowledge during my graduate studies.

I would like to acknowledge the Parkinson's Progression Markers Initiative (PPMI) for providing the publicly available dataset used in this research. Their valuable contribution to Parkinson's disease research has made this study possible.

My heartfelt thanks go to my husband, parents and family for their unconditional love, endless prayers, encouragement and continuous support throughout my academic journey. Their sacrifices and motivation have always inspired me to achieve my goals.

Finally, I would like to thank my friends and colleagues who directly or indirectly supported me during this research. Their encouragement, discussions and cooperation were greatly appreciated.May Allah bless everyone who contributed to the successful completion of this work.



(Amna Ayub)

Abstract

Parkinson's disease (PD) is a progressive neurodegenerative disease with considerable heterogeneity in symptom manifestation and in disease progression. Early prognosis and prediction of the disease in terms of subtypes are difficult due to the different patterns of disease progression observed in different patients, which makes effective timely interventions and personalized treatment strategies more difficult. The traditional diagnostics methods are mostly based on clinical assessment and rating scales, which are subjective and do not fully reflect the interaction between various disease features. While there have been recent successes with AI techniques for better diagnosis and prognosis for disease, many current studies rely on only a single modality, lack temporal modeling abilities, are under-represented in certain classes and have limited interpretability.

To overcome these drawbacks, this thesis suggests an explainable framework that combines multiple modalities and builds upon the ideas of GANs for prediction of Parkinson's disease progression subtypes. The framework proposed here combines heterogeneous information from clinical evaluations, gait characteristics, longitudinal time-series data, images of spiral handwriting, and MRI and DaTscan imaging modalities. A Random Forest classifier is used to analyze clinical and gait features, while temporal progression patterns are modeled with Long Short-Term Memory (LSTM) networks. The analysis of spiral handwriting and neuroimaging data uses convolutional neural networks and transfer learning techniques. To tackle the class imbalance issue, Conditional Tabular Generative Adversarial Networks (Conditional Tabular GANs or CTGANs) are used to generate synthetic examples and augment data diversity. In addition, SHapley Additive exPlanations (SHAP) are included to increase the interpretability and transparency of the proposed framework by highlighting the most impacting clinical variables that drive disease progression.

Additionally, the prediction results of each machine learning model and deep learning model are fused at the decision level, which uses complementary information from various modalities. The system is intended to categorize the patients into

three progression groups: slow, moderate and fast progressors. The experimental evaluation is carried out on publicly available datasets of Parkinson's disease, such as a Parkinson's Progression Markers Initiative (PPMI) data set and image-based data sets of spiral handwriting and MRI/DaTscan images. To assess performance, accuracy, precision, recall and F1-score are used.

The results of the experiments show that the model constructed in the LSTM architecture achieves the best performance in the task of prediction of progression subtypes emphasizing the crucial role of temporal information in the prediction of disease. Random Forest models are able to capture clinically relevant features, while image-based models have moderate accuracy due to low intra-class image variance and data size. Easier clinical interpretation and increased transparency of the model through the integration of XAI. Overall, the proposed multimodal framework shows that the integration of machine learning, deep learning, generative learning, XAI and late fusion methods may enable for powerful prediction of Parkinson's disease progression subtypes. The framework developed could help assist with the early prognosis, personalised treatment planning and better clinical decision making for people with Parkinson's disease.

Contents

Author's Declaration iii

Plagiarism Undertaking iv

Acknowledgement v

Abstract vi

List of Figures xiii

List of Tables xv

Abbreviations xvii

Symbols xix

1 Introduction 1

1.1 Background . . . 4
1.2 Symptoms of Parkinson's Disease . . . 5
1.2.1 Motor Symptoms . . . 5
1.2.2 Non-motor Symptoms . . . 6
1.3 Parkinson's Disease Subtypes . . . 7
1.4 Disease Progression Subtypes . . . 8
1.5 Diagnostic Approaches for Parkinson's Disease . . . 10
1.6 Motivation . . . 12
1.7 Problem Statement . . . 14
1.8 Research Gap . . . 15
1.9 Aim of the study . . . 16
1.10 Research Objectives . . . 17
1.11 Research Questions . . . 17
1.12 Major Contributions . . . 17
1.13 Significance of Research . . . 18
1.14 Scope and Limitations . . . 19
1.15 Thesis Organization . . . 20
1.16 Chapter Summary . . . 20

2 Literature Review **22**
2.1 Parkinson's Disease and Artificial Intelligence . . . 24
2.2 Machine Learning Approach for Parkinson's Disease . . . 26
2.2.1 Support Vector Machine . . . 27
2.2.2 Random Forest . . . 28
2.2.3 Logistic Regression . . . 28
2.2.4 Extreme Gradient Boosting . . . 28
2.3 Deep Learning Approaches . . . 29
2.3.1 Convolutional Neural Networks . . . 30
2.3.2 Recurrent Neural Networks . . . 30
2.3.3 Long Short-term Memory Networks . . . 31
2.3.4 Bidirectional Long Short-term Memory Networks . . . 32
2.3.5 Transformer Architectures . . . 32
2.4 Imaging-Based Approaches . . . 33
2.4.1 Magnetic Resonance Imaging . . . 34
2.4.2 Positron Emission Tomography . . . 34
2.4.3 Dopamine Transporter Imaging . . . 35
2.5 Speech-Based Approaches . . . 35
2.6 Handwriting and Spiral Drawing Analysis . . . 36
2.7 Wearable Sensor and Gait Analysis . . . 37
2.8 longitudinal and Time-series Approaches . . . 38
2.9 Explainable Artificial Intelligence . . . 39
2.10 Generative Learning Approaches . . . 40
2.11 Multimodal Learning Approaches . . . 40
2.12 Machine Learning and Deep learning Approaches . . . 41
2.13 Time-series and Longitudinal data Approaches . . . 42
2.14 Explainable Artificial Intelligence . . . 43
2.15 Generative Learning and class imbalance . . . 43
2.16 Research Gaps in Existing Literature . . . 44
2.17 Chapter Summary . . . 45

3 Research Methodology **47**
3.1 Proposed FrameWork . . . 48
3.2 Dataset Description . . . 51
3.2.1 Parkinson's Progression Markers Initiative . . . 53
3.2.1.1 Clinical Dataset for Motor Subtype Classification . 54
3.2.1.2 Longitudinal Dataset for Progression Prediction . . 55
3.2.2 Structural MRI Dataset . . . 56
3.2.3 Spiral Handwriting Dataset . . . 58
3.2.4 Dataset Selection for Multimodal Fusion . . . 60
3.3 Data Preprocessing . . . 61
3.3.1 Clinical Data Preprocessing . . . 62
3.3.1.1 Patient Identification and Baseline Record Selection 62
3.3.1.2 Motor Subtype Label Generation . . . 62
3.3.1.3 Prevention of Target Leakage and Feature Selection 63

3.3.1.4 Missing and Invalid Value Handling . . . 64
3.3.1.5 Clinical Train-test Partitioning . . . 65
3.3.1.6 Feature Scaling . . . 65
3.3.2 CTGAN Data Pre-processing and Minority-Class Augmentation . . . 66
3.3.2.1 CTGAN Input Preparation . . . 66
3.3.2.2 Synthetic Minority-class Generation . . . 66
3.3.2.3 Synthetic Data Validity Control . . . 67
3.3.3 Longitudinal Data Preprocessing . . . 68
3.3.3.1 Temporal Separation of Input and Target . . . 69
3.3.3.2 Future Progression Label Generation . . . 69
3.3.3.3 Sequence Construction . . . 70
3.3.3.4 Longitudinal Feature Scaling . . . 71
3.3.4 History-based Future-visit Motor Subtype Preprocessing . . . 71
3.3.5 Structural MRI Preprocessing . . . 72
3.3.5.1 DICOM Quality Control and Volume Selection . . . 72
3.3.5.2 Patient-level MRI Partitioning . . . 73
3.3.5.3 MRI Slice Selection . . . 74
3.3.5.4 MRI Intensity Processing and Resizing . . . 74
3.3.5.5 MRI Data Augmentation . . . 75
3.3.6 Spiral Handwriting Image Preprocessing . . . 76
3.3.6.1 Subject-level Partitioning . . . 76
3.3.6.2 Image Resizing and Input Preparation . . . 77
3.3.6.3 Training-only Image Augmentation . . . 77
3.4 Model Development and Architecture . . . 78
3.4.1 Clinical Machine-learning Models . . . 79
3.4.1.1 Logistic Regression . . . 79
3.4.1.2 Support Vector Machine . . . 80
3.4.1.3 Random Forest . . . 81
3.4.2 Conditional Tabular Generative Adversarial Network . . . 82
3.4.2.1 Generator and Discriminator . . . 82
3.4.2.2 CTGAN Training Configuration . . . 83
3.4.2.3 Downstream CTGAN evaluation design . . . 84
3.4.3 Long Short-term Memory Network . . . 85
3.4.3.1 LSTM Internal Architecture . . . 85
3.4.3.2 LSTM Model Architecture . . . 86
3.4.4 Bidirectional LSTM . . . 87
3.4.5 Single visit Logistic Regression Baseline . . . 88
3.4.6 Structural MRI Classification with MobileNetV2 . . . 88
3.4.6.1 MobileNetV2 Feature Extraction . . . 89
3.4.6.2 MRI Model Architecture . . . 90
3.4.6.3 Transfer learning and fine-tuning . . . 90
3.4.6.4 MRI Class-imbalance Handling . . . 91
3.4.6.5 Patient-level MRI Prediction . . . 92

3.4.7 Spiral Handwriting Classification Using MobileNetV2 92
3.4.7.1 Spiral Model Training 93
3.4.7.2 Subject-level Spiral Prediction 94
3.4.8 Explainable Artificial Intelligence 95
3.4.8.1 SHapley Additive exPlanations 95
3.4.8.2 Gradient-weighted Class Activation Mapping . . . 96
3.4.9 Decision-Level Multimodal Late Fusion 96
3.4.9.1 Management of leakage in Fusion 97
3.4.10 Model Evaluation Metrics 98
3.4.11 Experimental Environment 99
3.4.12 Summary of Model Development 100

4 Results and Discussion **101**
4.1 Introduction 101
4.2 Clinical Motor-subtype Classification Results 102
4.2.1 Clinical Logistic Regression Performance 103
4.2.2 SHAP-based Clinical Explainability 105
4.3 CTGAN-based Class-imbalance Results 107
4.3.1 Logistic Regression with and without CTGAN 108
4.3.2 Random Forest with and without CTGAN 109
4.3.3 Discussion of CTGAN Findings 109
4.4 Longitudinal Progression Prediction Results 111
4.4.1 Discussion of Longitudinal Modeling 113
4.5 History-based Future-visit Motor Subtype Prediction Results 113
4.5.1 Development and Held-out Evaluation 114
4.5.2 Class-wise Motor Subtype Performance 114
4.5.3 Confusion Matrix 115
4.5.4 Discussion of History-based Prediction 115
4.6 Structural MRI Classification Results 115
4.6.1 MRI Model Selection 116
4.6.2 Final Patient-level MRI Results 116
4.6.3 Discussion of MRI Results 118
4.6.4 MRI Grad-CAM Results 119
4.7 Spiral Handwriting Classification Results 120
4.7.1 Image-level Spiral Results 120
4.7.2 Subject-Level Spiral Results 121
4.7.3 Spiral Grad-CAM Results 122
4.8 Clinical MRI Decision Level Late Fusion Results 123
4.8.1 Same-cohort Comparison 124
4.8.2 Discussion of Fusion Results 127
4.9 Comparison with Recent Studies 128
4.9.1 Clinical vs imaging information 131
4.9.2 Effect of Longitudinal Information 131
4.9.3 Effect of Generative Augmentation 132
4.9.4 Role of Explainability 132

4.9.5 Multimodal Integration . . . 133
4.10 Key Findings . . . 133
4.10.1 Clinical explainability: . . . 133
4.10.2 CTGAN augmentation: . . . 133
4.10.3 Longitudinal modeling: . . . 134
4.10.4 Structural MRI: . . . 134
4.10.5 Spiral handwriting: . . . 134
4.10.6 Explainable image modeling: . . . 134
4.10.7 Multimodal fusion: . . . 134
4.11 Discussion of Overall Findings . . . 135
4.12 Chapter Summary . . . 135

5 Conclusion and Future Work **137**
5.1 Introduction . . . 137
5.2 Conclusion . . . 138
5.3 Major Contributions of the study . . . 142
5.4 Limitations . . . 143
5.5 Future Directions . . . 144
5.6 Overall Implications . . . 146
5.7 Chapter Summary . . . 147

Bibliography **148**

List of Figures

1.1 A comprehensive overview of PD symptoms, causes, subtypes of PD, PD progress and management strategies . . . 2
1.2 Different stages involved in Parkinson's disease progression from healthy condition to advanced disease stage. . . . 5
1.3 Major motor symptoms associated with Parkinson's disease. . . . 6
1.4 Major subtypes of Parkinson's disease based on symptom predominance 8
1.5 Different progression subtypes of Parkinson's disease . . . 9

2.1 Timeline of Parkinson's Disease AI Evolution . . . 24
2.2 Common machine learning techniques used for Parkinson's disease analysis . . . 26
2.3 Major imaging modalities used in Parkinson's disease diagnosis and progression analysis . . . 33

3.1 Framework of the proposed explainable multimodal Parkinson's disease progression prediction model . . . 50
3.2 Overview of the datasets and modality-specific cohorts used in the study. . . . 52
3.3 Clinical Data Preprocessing. . . . 55
3.4 Temporal organization of early longitudinal observations and future progression-label generation. . . . 56
3.5 Representative preprocessed axial slices from a PPMI 3D T1-weighted structural MRI volume used in the imaging experiment. 58
3.6 Representative spiral handwriting samples from healthy-control and Parkinson's disease subjects in the NewHandPD dataset. . . . 59
3.7 CTGAN-based minority-class augmentation and evaluation pipeline. 68
3.8 Representative stages of structural MRI preprocessing prior to MobileNetV2 input. . . . 75
3.9 Spiral handwriting preprocessing and training-only augmentation procedure. . . . 77
3.10 Overall model-development, explainability and multimodal-fusion strategy used in the proposed framework. . . . 78
3.11 Conceptual architecture of CTGAN for conditional generation of minority-class clinical records. . . . 83
3.12 Internal gating mechanism of the LSTM unit used for longitudinal sequence modeling. . . . 86

3.13 MobileNetV2-based architecture used for structural MRI motor-subtype classification. 89
3.14 Decision-level late fusion of clinical and structural MRI class probabilities. 98

4.1 Confusion matrix of the final Random Forest model for clinical motor-subtype classification on the locked independent test set (n = 962). 103
4.2 Confusion matrix of the Logistic Regression model for clinical motor-subtype classification. 105
4.3 SHAP-based global feature-importance analysis of the final Random Forest model for clinical motor-subtype classification. 107
4.4 Comparison of real-only and CTGAN-augmented clinical classifiers. 110
4.5 Comparative performance of single-visit and recurrent models for future progression prediction. 112
4.6 Patient-level confusion matrix of the final structural MRI MobileNetV2 model. 118
4.7 Representative Grad-CAM visualization for a correctly classified structural MRI test case. 119
4.8 Representative Grad-CAM visualization for an incorrectly classified structural MRI test case. 120
4.9 Subject-level confusion matrix of the spiral handwriting MobileNetV2 model. 122
4.10 Representative Grad-CAM visualization for a healthy-control spiral handwriting image. 123
4.11 Representative Grad-CAM visualization for a Parkinson's disease spiral handwriting image. 123
4.12 Same-cohort comparison of the clinical Random Forest, structural MRI and decision-level late-fusion models. 125
4.13 Full-dataset patient-level confusion matrix of the Clinical–MRI late fusion framework (N = 238). 126

List of Tables

1.1 Research Gaps Identified in Existing Studies and Their Consideration in the Proposed Framework . . . 16

3.1 Summary of Datasets, Modalities and Prediction Tasks . . . 52
3.2 Feature Groups and Variables Used for Clinical Classification . . . 54
3.3 Dataset Processing and Final Data Split . . . 58
3.4 Subject-level distribution of the NewHandPD spiral handwriting dataset . . . 60
3.5 Distribution of Real and Synthetic Training Samples Across Motor Subtypes . . . 67
3.6 Logistic Regression configuration . . . 80
3.7 Random Forest Configuration . . . 82
3.8 CTGAN Configuration and Synthetic Data Generation . . . 84
3.9 LSTM Model Architecture and Training Configuration . . . 87
3.10 Class Weights for Motor Subtype Classification . . . 91
3.11 MobileNetV2 Model Architecture and Training Configuration . . . 94
3.12 Software Tools and Their Applications . . . 100

4.1 Five-fold cross-validation performance of the optimized clinical machine-learning models on the development set . . . 102
4.2 Final clinical Logistic Regression test performance . . . 104
4.3 Highest-ranked clinical features according to mean absolute SHAP value . . . 106
4.4 Logistic Regression performance before and after CTGAN augmentation . . . 108
4.5 Random Forest performance before and after CTGAN augmentation 109
4.6 Longitudinal progression prediction results . . . 111
4.7 Performance of the history-based Random Forest and persistence baseline . . . 114
4.8 Class-wise held-out performance of the history-based Random Forest 114
4.9 Held-out confusion matrix of the history-based Random Forest . . . 115
4.10 MRI validation performance during model selection . . . 116
4.11 Final patient-level structural MRI performance . . . 117
4.12 Class-wise MRI test performance . . . 117
4.13 Image-level spiral handwriting test performance . . . 120
4.14 Subject-level spiral handwriting test performance . . . 121
4.15 Clinical, MRI and late-fusion performance on the same 36 subjects . 124

4.16 Class-Wise Classification Performance 124
4.17 Comparison of recent Parkinson's disease studies with the proposed framework 128
4.18 Prediction Tasks and Evaluation Details of Experimental Branches . 130

Abbreviations

Adam	Adaptive Moment Estimation
AI	Artificial Intelligence
AUC	Area Under the Curve
BiLSTM	Bidirectional Long Short-Term Memory
CAM	Class Activation Map
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CSV	Comma-Separated Values
CTGAN	Conditional Tabular Generative Adversarial Network
CV	Cross Validation
CWT	Continuous Wavelet Transform
DaTscan	Dopamine Transporter Scan
DL	Deep Learning
DT	Decision Tree
ECG	Electrocardiogram
EEG	Electroencephalogram
FN	False Negative
FP	False Positive
GAN	Generative Adversarial Network
GNN	Graph Neural Network
GPU	Graphics Processing Unit
Grad-CAM++	Gradient-weighted Class Activation Mapping++
K-Fold CV	K-Fold Cross Validation
LR	Logistic Regression

LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MDS-UPDRS	Movement Disorder Society-Unified Parkinson's Disease Rating Scale
ML	Machine Learning
MRI	Magnetic Resonance Imaging
PCA	Principal Component Analysis
PDBP	Parkinson's Disease Biomarkers Program
PD	Parkinson's Disease
PD/Control	Parkinson's Disease / Healthy Control
PET	Positron Emission Tomography
PIGD	Postural Instability and Gait Difficulty
PPMI	Parkinson's Progression Markers Initiative
ReLU	Rectified Linear Unit
RF	Random Forest
RMSE	Root Mean Square Error
RNN	Recurrent Neural Network
ROC	Receiver Operating Characteristic
R^2	Coefficient of Determination
SHAP	SHapley Additive exPlanations
STFT	Short-Time Fourier Transform
SVM	Support Vector Machine
TD	Tremor Dominant
TF	TensorFlow
TN	True Negative
TP	True Positive
UPDRS	Unified Parkinson's Disease Rating Scale
VAE	Variational Autoencoder
XAI	Explainable Artificial Intelligence

Symbols

X	Input dataset
x	Input feature vector
y	True class label
$\hat{y}$	Predicted class label
n	Number of samples
m	Number of input features
t	Time step
T	Total number of time steps
h_t	Hidden state of LSTM at time step t
c_t	Cell state of LSTM at time step t
i_t	Input gate of LSTM
f_t	Forget gate of LSTM
o_t	Output gate of LSTM
$\tilde{c}_t$	Candidate cell state
W	Weight matrix
b	Bias vector
σ	Sigmoid activation function
tanh	Hyperbolic tangent activation function
L	Loss function
α	Learning rate
θ	Model parameters
G	Generator network (CTGAN)
D	Discriminator network (CTGAN)
z	Random noise vector

p_{data}	Real data distribution
p_z	Noise data distribution
ϕ_i	SHAP value of feature i
$f(x)$	Prediction function
Acc	Classification accuracy
P	Precision
R	Recall
F_1	F1-score
AUC	Area Under the ROC Curve

Chapter 1

Introduction

Parkinson's disease (PD) is a chronic and a progressive neurodegenerative disorder that affects millions of people throughout the world and is the second most prevalent neurodegenerative disorder after Alzheimer's disease [1]. The disease is mainly caused by the death of dopaminergic neurons in the substantia nigra pars compacta area of the brain, which causes a decrease in the levels of the chemical dopamine and disrupts normal motor functions [2]. Patients gradually develop problems of mobility, posture, balance and coordination as the disease progresses, greatly impacting on the quality of their life.

The disease of Parkinson's is traditionally known as a movement disorder, but it is now known that there are many other motor and non-motor symptoms. Resting tremor, bradykinesia, muscular rigidity and postural instability are commonly observed motor symptoms, while depression, anxiety, cognitive impairment, sleep disturbances, fatigue and autonomic dysfunction are non-motor symptoms [3]. Symptoms can present years before a diagnosis is made and can be very different depending on the patient.

Heterogeneity is one of the major features of PD. Patients with Parkinson's do not always progress in the same way and have varying degrees of severity. In some cases, disease progression is slow and lasts for a long duration, while in others it is rapid with a gradual decline of motor and cognitive function [4]. Likewise,

various clinical phenotypes, such as Tremor Dominant (TD), Postural Instability and Gait Difficulty (PIGD) and Mixed subtype have been described, reflecting, once again, the complexity and variability of the disease [5]. This variability is a difficult problem to overcome for disease monitoring and prediction of disease progression.

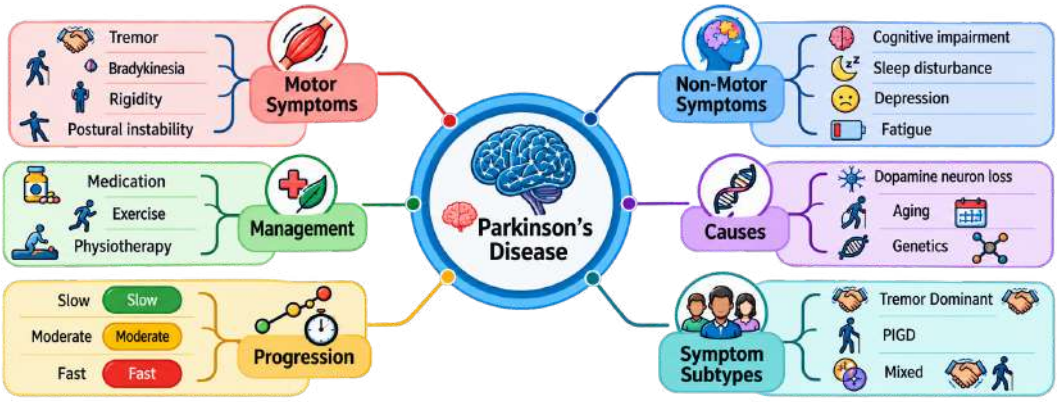


FIGURE 1.1: A comprehensive overview of PD symptoms, causes, subtypes of PD, PD progress and management strategies

Figure 1.1 provides a detailed account of the symptoms, causes, progression and management of Parkinson's disease [6]. The disease is highly heterogeneous, making it difficult to monitor it accurately and to predict its subtypes.

Traditionally, Parkinson's disease diagnosis and progression assessment mainly rely on clinical evaluations and rating scales such as the Movement Disorder Society-Unified Parkinson's Disease Rating Scale (MDS-UPDRS).

These clinical evaluations are regarded as the traditional method, but they are subjective and do not fully address the multiple interactions between various clinical manifestations of the disease [7]. Moreover, traditional diagnostic approaches are not very effective in predicting future disease course or in offering individualized treatment.

In recent years, artificial intelligence (AI) and machine learning (ML) techniques have been increasingly explored for the diagnosis and prognosis of Parkinson's disease. Various computational approaches have been employed using different data modalities, including clinical assessments, neuroimaging, speech signals, gait analysis, wearable sensor data, handwriting patterns and longitudinal time-series in-

formation [8]. These techniques have shown potential success in pattern recognition of diseased conditions and in aiding clinicians in decision making for disease.

There are still some issues in the existing literature regarding the analysis of PD with AI. Most studies are on single data modalities or disease diagnosis, but it is difficult to combine heterogeneous information to analyze disease subtypes and prognosis. Moreover, small-scale longitudinal modeling, class imbalance and limited interpretability of the models can limit the success of AI-driven methods for disease analysis [9].

This study explores a multimodal and explainable framework based on clinical assessments, longitudinal clinical records, structural MRI and spiral handwriting images to address these challenges.

Motor-subtype classification is made using clinical and structural MRI data, while future progression is predicted using only longitudinal clinical data.

The image of the handwriting is evaluated independently for Parkinson's disease and healthy-control classification using the spiral hand-writing images. Furthermore, class-imbalance analysis and interpretation of models are added using generative learning techniques and explainable artificial intelligence techniques.

In this study, a framework of Explainable Artificial Intelligence (XAI) to classify Parkinson's disease subtypes and analyze its progression using heterogeneous and longitudinal data is developed and evaluated. The framework is based on modality-specific machine-learning and deep-learning models to explore various information sources such as clinical assessment, longitudinal clinical records, structural MRI and spiral handwriting images. The motor-subtype classification is performed using clinical and structural MRI data, while subsequently, longitudinal clinical records are analysed separately for predicting disease progression patterns over time. The images of handwriting are assessed independently for the Parkinson's disease and healthy control classification. To integrate multimodal information, clinical and structural MRI branches provide class-probability predictions that are fused in the decision level with late fusion. This design will allow for the study to examine

the contribution of individual modalities and will also understand the effects of additional clinical and imaging data on motor-subtype prediction.

1.1 Background

Parkinson's disease is a progressive disorder of the nervous system, which involves the progressive destruction of a group of nerve cells in the brain that produce a chemical (dopamine). The deficiency of dopamine causes a failure of normal communication between the basal ganglia and leads to abnormal movements, posture and coordination. With its rising prevalence and lack of a definitive cure, Parkinson's disease is regarded as one of the most prevalent neurodegenerative disorders and a significant healthcare problem worldwide [10].

Disease involves both motor and non-motor functions and is highly clinically variable. In addition to aging, other risk factors associated with PD include genetic susceptibility, oxidative stress, mitochondrial dysfunction, environmental toxins and abnormal protein aggregation. Pathological features that are observed in neuronal degeneration include the accumulation of α-synuclein proteins and the development of Lewy bodies [11].

Due to population ageing and the improved life expectancy over the past ten years, the proportion of people with Parkinson's disease has risen markedly. Recent epidemiological studies have shown that the number of people with Parkinson's disease will rise significantly in the next few years, highlighting the critical importance of early diagnosis and good disease monitoring.

Parkinson's is a progressive and heterogeneous disease and its diagnosis and prognosis are particularly difficult. Symptoms and the course of the disease, as well as the response to treatment, may vary from patient to patient.

This means knowledge of the mechanisms and characteristics of the disease is of paramount importance in order to develop personalized and effective management strategies [12].

The progression of Parkinson's disease is commonly described using clinical stages

that reflect increasing motor impairment and disability [13], as illustrated in Figure 1.2. However, as the disease advances, the motor and non-motor symptoms worsen as the dopaminergic neurons slowly die.It is important to recognise these stages for early diagnosis and efficient monitoring of progression.

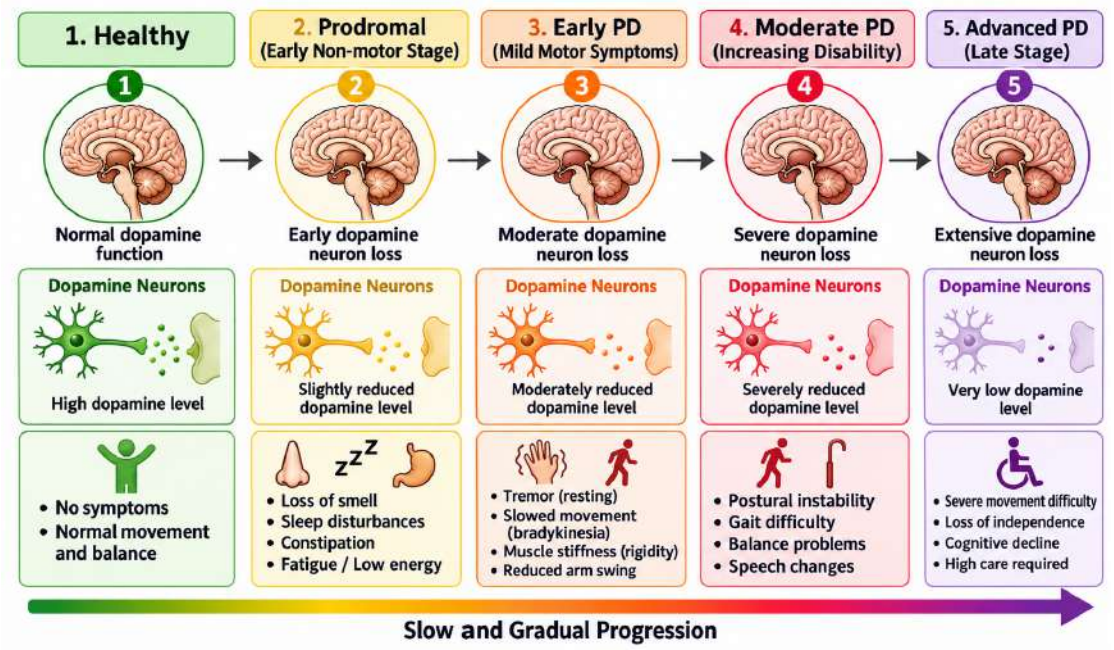


FIGURE 1.2: Different stages involved in Parkinson's disease progression from healthy condition to advanced disease stage.

1.2 Symptoms of Parkinson's Disease

The signs and symptoms of Parkinson's disease can be divided into motor and non-motor signs and symptoms. Symptoms are different in each patient and can evolve during the course of the disease. In many instances non-motor symptoms occur many years before motor [14].

1.2.1 Motor Symptoms

The most obvious signs of Parkinson's disease are the motor symptoms, which are primarily linked to the lack of dopamine. Resting tremor is one of the earliest and most characteristic symptoms. A patient's hand, fingers or legs often shake involuntarily. The slowness of movement (bradykinesia) seriously interferes with tasks and is considered one of the cardinal signs of Parkinson's disease [15].

Muscular rigidity results in stiffness and resistance to movement and postural instability is responsible for imbalance and coordination issues. The disease can also cause patients to freeze, have poor posture, reduced arm swing and show abnormal gait, making them more susceptible to falls and injuries[15]

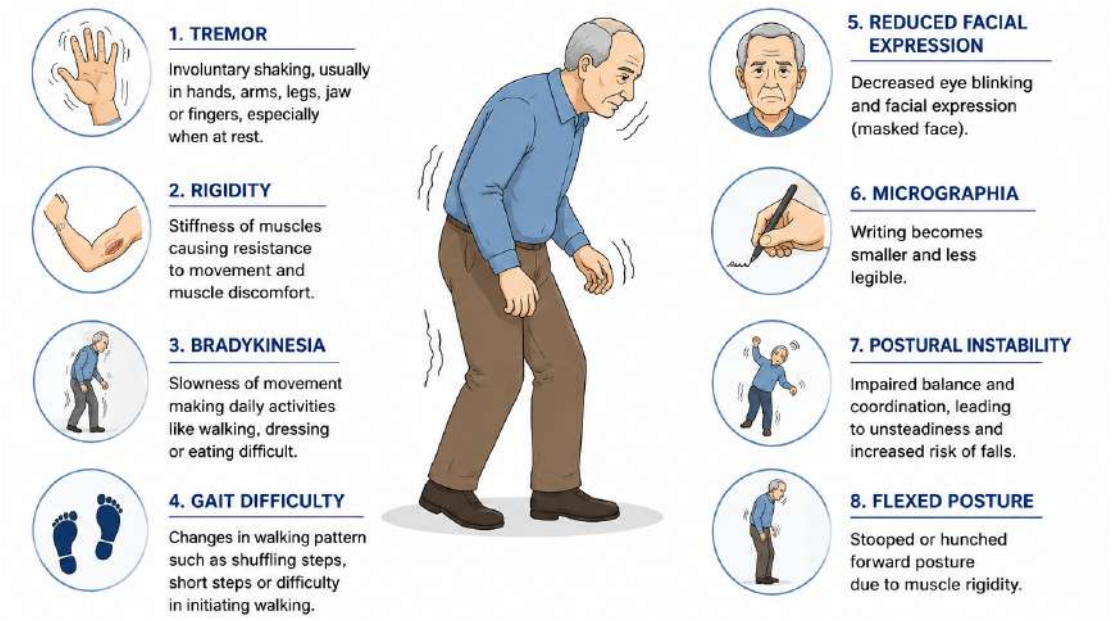


FIGURE 1.3: Major motor symptoms associated with Parkinson's disease.

The major motor symptoms of Parkinson's disease are shown in Figure 1.3. These include tremor, rigidity, bradykinesia, gait abnormalities, postural instability, reduced facial expression, micrographia and flexed posture [16]. The severity of these motor impairments generally increases with disease progression and can substantially affect patients' mobility and quality of life.

1.2.2 Non-motor Symptoms

Parkinson's disease has many non-motor symptoms along with its motor symptoms. Patients often experience depression, anxiety, sleep problems, intellectual dysfunction, constipation, weakness, loss of smell and autonomic dysfunction.[17].

Recent research studies show that although significant clinically, non-motor symptoms are underdiagnosed. Therefore, thorough assessment of motor and non-motor symptoms is crucial to proper disease evaluation and individualized treatment planning [17].

These symptoms include sleep disturbances, mood disorders, fatigue, cognitive impairment, autonomic dysfunction, pain, olfactory dysfunction, gastrointestinal complications, urinary problems, sensory abnormalities and dizziness.

Although these symptoms are not primarily related to motor function, they can significantly affect patients' health status and quality of life.

1.3 Parkinson's Disease Subtypes

Parkinson's disease is a heterogeneous disorder and patients have different symptom profiles and/or symptom progression. Several subtypes have been suggested on the basis of the predominance of clinical symptoms.

These are the most commonly studied phenotypes, Tremor Dominant (TD), Postural Instability and Gait Difficulty (PIGD) and an Indeterminate subtype [18].

Tremor-Dominant (TD) involves more tremor symptoms than gait and postural impairment. This phenotype has been linked to clinical and disease course differences to other motor phenotypes.

The Postural Instability and Gait Difficulty (PIGD) subtype is associated with increased symptoms of gait difficulty, freezing, balance problems and postural instability. In the literature this phenotype has been linked to an increased level of gait related disability and to variations in disease course. [19].

The Indeterminate subtype is defined as patients whose relative tremor and PIGD symptom scores fall outside of the set criteria for TD or PIGD.

For the clinical classification experiment and clinical vs. structural MRI classification experiment, the three motor-subtype classes used are TD, PIGD and Indeterminate [20].

Figure 1.4 illustrates the major motor subtypes of Parkinson's disease based on predominant motor symptoms. These include tremor-dominant (TD), postural instability and gait difficulty (PIGD) and an indeterminate or mixed subtype [21]. These motor phenotypes reflect differences in the predominant clinical manifesta-

tions of Parkinson's disease.

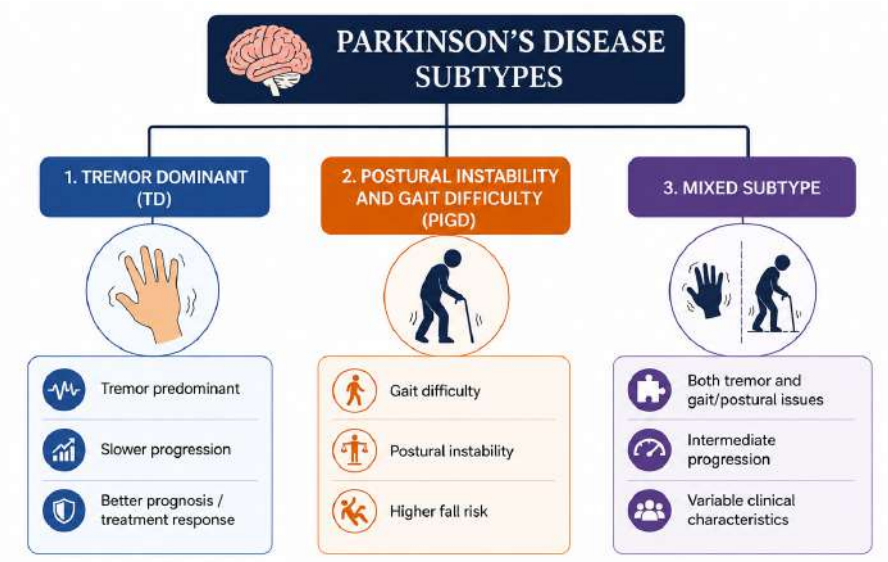


FIGURE 1.4: Major subtypes of Parkinson's disease based on symptom predominance

1.4 Disease Progression Subtypes

Parkinson's disease is a progressive neurological disorder that occurs gradually, although patients progress at different rates. In some people the deterioration happens over a number of years, while in others it can happen more quickly and they have severe motor and cognitive impairment. The heterogeneity of disease course underscores the need for classification of disease progression subtypes and for an understanding of underlying patterns of disease course [22].

Patients can be divided into slow, moderate and fast progressors based on the severity of the symptoms and the evolution of the symptoms over time. Slow progressors have relatively mild deterioration of symptoms over time and have longer periods of functional independence. Medications tend to be effective in these patients, and they have fewer complications early in the disease [22].

Moderate progressors show intermediate progression patterns and slowly experience motor and non-motor impairments. The disease progression in these patients is more rapid than in slow progressors, but slower than in fast progressors. Long-term monitoring and early interventions may need to be implemented to maintain quality

of life and avoid complications [23].

In the case of fast progressors, symptoms deteriorate in a short period of time and can include significant impairment in gait, postural stability, cognition and responsiveness to treatment. The management of these patients is typically complex and there is a higher risk of them having disabilities earlier in the disease process.

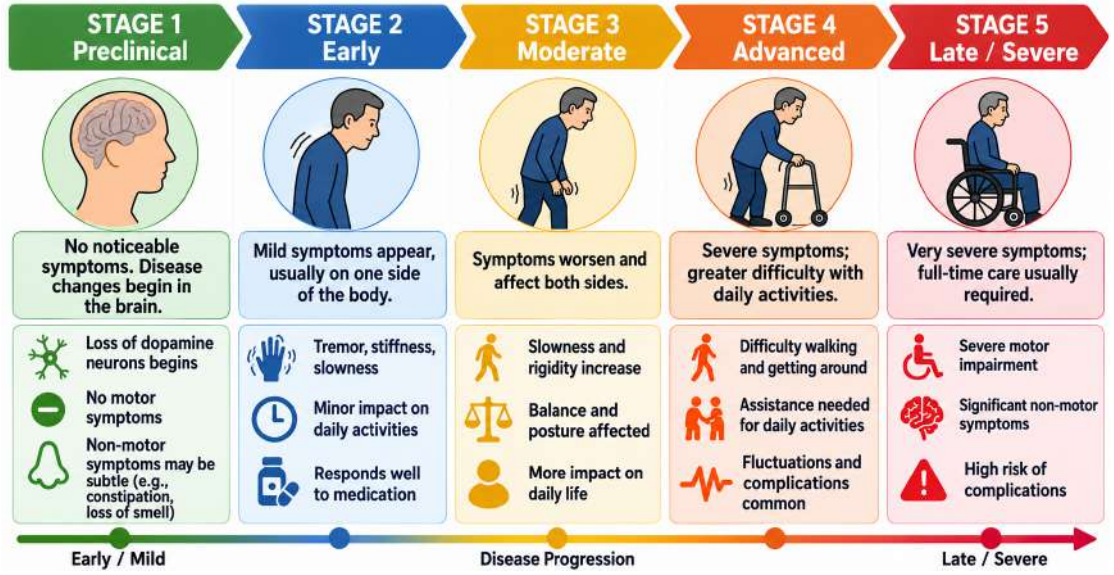


FIGURE 1.5: Different progression subtypes of Parkinson's disease

The progression patterns in Parkinson's disease are shown in Figure 1.5. Symptoms and functional abilities may deteriorate slowly, moderately or rapidly across patients. Longitudinal studies have demonstrated substantial heterogeneity in Parkinson's disease progression and have identified distinct slow, moderate and fast progression groups [24]. Understanding these progression patterns is important for characterizing individual differences in disease course and for developing predictive models of disease progression. Through longitudinal clinical assessments and AI methods, there has been significant potential identified to predict disease course and support early intervention strategies. Hence, the prediction of progression subtypes has become a tempting avenue towards precision medicine in PD. Longitudinal progression categories should be separated from the motor subtypes that were included in this study.

Motor-subtype classification is based on Tremor-Dominant (TD), Postural Instability and Gait Difficulty (PIGD) classes, while the longitudinal progression prediction

is based on Slow, Moderate and Fast progression categories. These are considered as individual prediction problems in the experimental setting.

1.5 Diagnostic Approaches for Parkinson's Disease

Traditionally, the diagnosis and prognosis of Parkinson's disease has been based on clinical assessments done by a neurologist. Commonly used assessments to determine disease severity and monitor disease progression include the Movement Disorder Society- Unified Parkinson's Disease Rating Scale (MDS-UPDRS) and the Hoehn and Yahr staging system. All are useful, but may have subjectivity and do not fully reflect the complexity and heterogeneity of PD [25].

The use of artificial intelligence and machine learning approaches for the diagnosis and prediction of Parkinson's disease (PD) has made great strides in the past few years. Several types of data have been investigated such as neuroimaging, speech signals, handwriting characteristics, gait characteristics, wearable sensor measurements and longitudinal clinical information

Medical imaging techniques that can give information about the structure and function of the brain, like magnetic resonance imaging (MRI), positron emission tomography (PET) and dopamine transporter imaging (DaTscan), have been widely employed for disease diagnosis. These imaging techniques help clinicians detect any abnormalities linked to dopamine deficiency and neuronal degeneration.

In addition, since vocal changes are a common symptom in patients with PD, speech analysis has also been shown to be a useful tool for detecting PD. In the same way, handwriting and spiral drawing patterns have proved to be helpful in the diagnosis of motor abnormalities related to tremors and bradykinesia. Wearable sensors and gait analysis methods allow real-time tracking of patients' gait and objective measurement of the severity of the disease.

Moreover, there has been a growth of interest in analysing the clinical history of a disease, where patient visits are recorded longitudinally. These datasets can

provide time-series data that can be used to study symptom evolution and to forecast the disease trajectory.

Temporal dependency and progression modeling has been demonstrated to be successful with deep learning and recurrent neural networks.Although these developments have come a long way, a large number of current methods are unimodal and are not able to leverage complementary information from multiple data sources.

Model explainability, limited longitudinal modeling, poor generalization ability, and class imbalance are other important challenges that limit the clinical use of available models for predicting PD. Most DL models are black box systems that give predictions without transparency regarding individual clinical variables, imaging features, or other key patient-specific factors that drive the decision. The lack of transparency makes it hard for clinicians to understand, verify and believe the output of the model in the real health care environment.

Besides, the symptoms and severity of Parkinson's disease also evolve over time. However, some current methods are based on cross-sectional or single visit data, and cannot adequately represent temporal variations or patterns of progression across multiple visits. Another limitation to be noted is class imbalance, especially when the patients are sorted into different progression subtypes. Some progression groups might be quite small this can lead to predictive models being overly optimistic about the majority class and cause them to miss the smaller underrepresented groups. Additionally, models created and tested in one patient cohort or on a limited patient population may not perform as well in an independent patient cohort, because of variations in demographics, clinical evaluations, acquisition protocols and/or data distributions. The above constrain the strength, interpretation and applicability of the current methods and still prevent their use in clinical practice [26].

The limitations underscore the need to develop more comprehensive, powerful and meaningful multimodal approaches to analysing Parkinson's disease progression. An efficient framework needs to be able to seamlessly combine all the available information from various data modalities including longitudinal clinical assessments, neuroimaging, handwriting patterns, gait measures and other relevant biomarkers

in order to leverage complementary information across different modalities.

In addition, through the inclusion of the longitudinal modeling, the framework can account for changes of condition over repeated visits, as opposed to only static observations. Proper measures to address class imbalance are also needed for better representation and prediction of minority progression groups, and explainable artificial intelligence methods can enable more transparency of what factors are driving model predictions. Such a multi-layered approach could offer a more complete picture of the complexity of PD and facilitate better and more meaningful prediction of PD disease subtypes. Thus the need for multimodal approaches that consider these challenges together and not separately as is done in [27]: data heterogeneity, temporal progression, class imbalance, interpretability and generalization of models.

1.6 Motivation

Parkinson's disease is a heterogeneous disorder of the nervous system with a high level of inter-individual variability. Some patients will experience different clinical symptoms, symptom severity, progression of symptoms and reactions to treatment than others. Some patients will have relatively slow deterioration over several years while others may be more rapidly deteriorating with increased motor and non-motor impairments. The diversity of these makes characterization and monitoring of disease progression difficult [28].

Traditional analysis of Parkinson's disease is mainly based on clinical and neurological evaluations and standardized rating scales. While these are necessary to measure disease severity, clinical assessment can contain some subjectivity and may not be able to capture the diverse and multi-faceted nature of disease progression. These restrictions have spurred a growing interest in computational strategies that could be used to better analyze Parkinson's disease in an objective and data-driven manner [29].

There has been a growing interest in the use of artificial intelligence (AI) and machine learning (ML) technologies for the diagnosis, classification and prognosis

of Parkinson's disease.

Various information sources such as clinical assessment, neuroimaging, gait characteristics, handwriting pattern, speech signals, wearable sensor measures and other biomarkers have been previously investigated. But, many studies examine single modalities and multi-modal studies show how useful a combination of complementary information from different modalities can be. Thus, the task of effective incorporation of such information is still a significant research challenge.

One more aspect that is crucial: the longitudinal nature of Parkinson's disease. Disease progression takes time and data taken at one clinical assessment point may not capture disease changes. Multiple visits can be used to get information on the temporal changes and progression trajectories. Thus, taking into account the longitudinal data and the proper temporal modelling is relevant for the prediction of disease progression patterns.

Class imbalance is another challenge as the number of patients/observations in different subtypes or progression groups can vary significantly. Models trained on imbalanced data can be biased toward the majority classes and have a poor recognition rate for the minority classes. Moreover, many machine learning and deep learning models lack interpretability with respect to their predictions. Such uninterpretability may limit their applicability in health care-related applications, where knowing the rationale behind the model's prediction is essential.

The study of Parkinson's disease diagnosis and classification has been the focus of many existing studies, but predicting disease subtypes and patterns of progression over time remains difficult, as well as integrating multiple modalities, dealing with class imbalance and providing interpretable predictions.

These restrictions drive the need to develop an integrated framework that can make use of complementary multi-modal and longitudinal information for Parkinson's disease analysis.

Hence, the present study introduces a multimodal and longitudinal explainable AI system for prediction of Parkinson's disease subtype and progression. The

framework examines how to combine heterogeneous data modalities, temporal modeling of repeated clinical observations, how to work with class imbalance and how to use explainable AI to understand what factors are driving model predictions. Performance metrics (quantitative) and comparison of individual and integrated modeling approaches are used to assess the proposed approach.

1.7 Problem Statement

There is substantial clinical heterogeneity around Parkinson's disease and complex progression patterns which make prediction of disease trajectory difficult. Current diagnostic and prognostic methods are mainly based on subjective clinical evaluation and on a single source of information, such as a single modality.

However, approaches based on a single modality may capture only a limited aspect of the heterogeneous manifestations of Parkinson's disease and may not fully utilize the complementary information available across different data sources [30].

In the field of Parkinson's diagnosis, a variety of machine learning and deep learning approaches have been proposed based on speech analysis, gait data, neuroimaging, handwriting analysis and wearable sensor measurements.

These methods have yielded satisfactory results, but most studies have been geared toward the detection of disease rather than prediction of disease progression subtypes. Moreover, most of the current approaches take a single modality approach and hence overlook the complementary information that can be gained from multiple heterogeneous data sources [31].

Yet another drawback of the current methods is the lack of use of the longitudinal information. Parkinson's is a progressive condition and symptoms vary over time. However, many studies use cross-sectional data sets and do not account for temporal dependence between multiple patient visits. Thus, the ability to predict disease progression and to identify subtypes of disease progression is limited [32].

Moreover, class imbalance is often encountered in Parkinson's disease datasets which may lead to biased predictions and poor performance for minority classes.

The models thus far developed also lack interpretability, which hinders their use in clinical settings. For physicians and healthcare professionals, it is desirable to have a robust and easily explainable model that yields information about what factors contribute to the prediction [33, 34].

So, an interpretable AI model that can explore complementary clinical, longitudinal and image-based data, simultaneously addressing the diverse prediction tasks of Parkinson's disease is required. Such a framework should be used for motor-subtype classification, for predicting the progression in time, for analysing class-imbalances, for analysing model interpretability, and for multimodal integration, when modalities, subjects and prediction targets are compatible.

1.8 Research Gap

There are a number of studies that have explored using machine learning and deep learning methods for the diagnosis and prognosis of Parkinson's Disease. Many improvements have been made, but there are still some limitations in the existing literature [35].

The current state-of-the-art solutions are mostly based on one signature modality (e.g., speech data, gait, neuroimaging or clinical evaluation). Thus, these approaches cannot take advantage of the complementary information that can be provided by disparate data sources. Furthermore, numerous studies are targeted at disease diagnosis and do not provide much attention to progression subtype prediction, which is crucial in personalized disease [36].

A second limitation which is of importance is the lack of proper use of longitudinal information. The severity of the symptoms of Parkinson's disease fluctuates over time and is a progressive disorder. However, most of current studies rely on cross-sectional data sets and thus cannot account for temporal dependencies of disease progression [37].

In addition, class imbalance and limited interpretability remain as great challenges. It is difficult to use current models in clinical settings because there are no explain-

able AI mechanisms. Hence, a powerful and interpretable multimodal framework that is able to combine heterogeneous information and the temporal properties is needed to make accurate prediction of progression subtypes. Several machine learning and deep learning approaches have been proposed for Parkinson's disease analysis; however, important research challenges remain.

These include limited integration of heterogeneous data modalities, underutilization of longitudinal information, class imbalance, limited interpretability and comparatively limited investigation of progression subtype prediction. Table 1.1 summarizes these research gaps and shows how they are considered in the proposed framework.

Table 1.1: Research Gaps Identified in Existing Studies and Their Consideration in the Proposed Framework

Research Gap	**Existing Studies**	**Proposed Framework**
Predominant use of single-modality approaches	Frequently reported	Multimodal integration
Limited focus on progression subtype prediction	Limited investigation	Subtype prediction
Limited longitudinal modeling	Often underutilized	Longitudinal modeling
Class imbalance	Common challenge	Generative data augmentation
Limited model interpretability	Frequently reported	Explainable AI (XAI)
Limited integration of heterogeneous data sources	Limited integration	Decision-level multimodal fusion of compatible clinical and structure MRI predictions

1.9 Aim of the study

This study proposes and tests a multimodal, longitudinal, generative and explainable artificial intelligence framework for the analysis of Parkinson's disease subtypes and progression based on clinical assessments, longitudinal clinical records, structural MRI and writing data captured with the spiral hand-writer. The frame-

work explores temporal modelling, dealing with class-imbalance, explainability and decision level late fusion of compatible clinical and structural MRI predictions.

1.10 Research Objectives

The following are the key objectives of this research:

i. To build a multimodal AI framework for analysis of Parkinson's disease subtypes and progression based on clinical assessments, longitudinal clinical records, structural MRI and spiral handwriting data.
ii. To improve prediction reliability by modeling temporal disease progression and addressing class imbalance using appropriate machine learning, deep learning and generative learning techniques.
iii. To integrate multimodal information using late fusion and incorporate explainable AI techniques to interpret model predictions and identify influential features.

1.11 Research Questions

This study will address the following research questions:

i. How can clinical, longitudinal, structural MRI and spiral handwriting data be used for Parkinson's disease subtypes and progression analysis?
ii. How can temporal modeling and class imbalance handling contribute to the prediction of Parkinson's disease progression patterns?
iii. How can decision-level late fusion and explainable AI be used to integrate compatiable clinical and structural and interpret the factors contributing to model predictions?

1.12 Major Contributions

The major contributions of this research are summarized as follows:

i. An integrated multimodal and longitudinal AI framework is developed for Parkinson's disease subtype and progression prediction by utilizing complementary information from clinical, longitudinal and imaging data.
ii. Temporal modeling and class-imbalance handling are investigated for longitudinal and clinical data respectively, while decision-level late fusion is employed to combine compatible clinical and structural MRI motor-subtype predictions.
iii. Explainable Artificial Intelligence (XAI) is incorporated to provide interpretable model predictions and identify the features contributing to the predicted Parkinson's disease subtypes and progression patterns.

1.13 Significance of Research

Parkinson's is one of the fastest growing neurological diseases in the world; due to its progressive nature and heterogeneous manifestations, it is a significant healthcare challenge. Appropriate forecasting of disease development is fundamental for better care management and optimum medication treatment.

This research is significant in that it explores an intelligent framework that is able to analyze information from various modalities of data. The study compares complementary sources of information (clinic, longitudinal data, handwriting patterns, structural MRI) for clinical classification of Parkinson's disease subtypes and analysis of disease progression.

Moreover, the integration of XAI techniques is a crucial element of this research. In healthcare contexts, interpretability is crucial since physicians need to know the predictions of AI systems are accurate and easy to understand before using them in clinical settings. As such, the framework's clinical relevance and trustworthiness are strengthened with the incorporation of explainability mechanisms.

Moreover, the use of temporal information and generative learning methods could

enhance the robustness of the model and the accuracy of predictions. The results of this study can be used in precision medicine for making personalized treatment plans, early intervention, and providing patients with a better quality of life who suffer from Parkinson's disease [38].

1.14 Scope and Limitations

This study focus to develop and assess an explainable AI framework using clinical assessments, longitudinal clinical records, structural MRI and spiral handwriting images. Three-class motor-subtype classification based on clinical and structural MRI data, which includes Tremor-Dominant (TD), Postural Instability and Gait Difficulty (PIGD) and Indeterminate subtypes.

Future progression patterns (Slow, Moderate and Fast) are investigated using independently collected longitudinal MDS-UPDRs Part III. Data from spiral handwriting are analyzed as a separate binary classification task of Parkinson's disease vs. healthy controls.

Conditional Tabular Generative Adversarial Networks (CTGAN) are explored as a tabular clinical data-augmentation approach for class-imbalance issues and model interpretation is further enhanced using SHAP and Grad-CAM. Decision-level late fusion is only possible for clinical and structural MRI predictions due to the common test subjects and the compatible motor-subtype target.

Although some exploration was done with gait data, the data did not significantly overlap with the structural MRI cohort and were not included in the final multimodal fusion.

Sample sizes are relatively small for some of the modalities, especially the image-based experiments (with the exception of the auditory modality); only a small number of subjects participate in multiple modalities; class imbalance; heterogenic data and lack of external clinical validation. Thus, the results are understood in the context of evaluated datasets/experimental setup.

1.15 Thesis Organization

The thesis consists of five chapters. This book addresses each component of the research and collectively develops and evaluates the proposed framework for predicting PD progression subtypes.

The introduction and background of Parkinson's disease is given in Chapter 1. It discusses the problem motivation, problem statement, research gap, objectives, research questions, significance of the study, scope and limitations, and the major contributions of the proposed work.

Chapter 2 reviews the literature in detail to look for literature on the diagnosis and prediction of Parkinson's disease. This provides a summary of recent machine learning and deep learning strategies, discusses their limitations, and explains the research gaps that justify the proposed study.

The research methodology used in this study is described in Chapter 3. It introduces the datasets, preprocessing methods, generation methods for subtypes, multimodal feature fusion, machine learning and deep learning models, explainable AI techniques and evaluation methods throughout the study.

The experimental results and discussion are given in chapter 4. It compares the performance of various models and give comparative evaluation using different performance metrics. In addition, explainability methods and modalities of images are discussed with regard to their effectiveness.

The thesis is concluded in Chapter 5, with the analysis of main findings and contributions of the study. It also covers the scope of the present study and suggests some future research avenues.

1.16 Chapter Summary

This chapter provided a detailed overview of Parkinson's disease (PD), its definition, prevalence, pathological features, clinical symptoms, disease progression

and the major diagnostic methods. Different evaluation methods including clinical evaluation, neuroimaging, speech analysis, handwriting analysis, gait analysis and sensor-based methods were presented and discussed and their benefits and limitations.

The chapter also identified some major obstacles in the diagnosis of PD and prediction of PD progression, such as data heterogeneity, class imbalance, multimodal data integration and the lack of interpretability of many AI models.

In addition, gaps in the current literature were identified in order to motivate the research. The majority of the previous studies used single-modality datasets and cross-sectional analysis of data, with a limited focus on multimodal data integration, longitudinal time-series modeling, explainable AI and how to better deal with imbalanced progression classes.

These restrictions are enough motivation to develop an intelligent and clinically interpretable framework that can accurately predict the progression subtypes of Parkinson's disease with heterogeneous data sources.

Lastly, the chapter stated the aims, research questions, research significance, scope, contribution of the research and organization of the thesis.

This research proposed a multimodal explainable framework, where machine learning, deep learning, longitudinal analysis, Conditional Tabular Generative Adversarial Networks and Explainable Artificial Intelligence techniques are combined. Chapter two offers a thorough survey of the current state of the literature, outlines the gaps in research and lays a foundation for the proposed framework.

Chapter 2

Literature Review

The growing prevalence of parkinson's disease and its complex clinical characteristics have made it an important area of neurodegenerative disease research. Over past ten decade, significant progress has been made in the application of artifical intelligence and machine learning techniques for disease diagnosis, prediction, and monitoring . The increasing availability of multi-modal datasets and improvements in computational approaches have allowed researchers to study disease-related patterns based on clinical measurements, neuro-imaging, speech signals, gait measurements, handwriting, wearable sensors and longitudinal time-series measurements [39].

Traditional Parkinson's disease diagnosis methods were based primarily on clinical evaluations and neurological examinations. The advent of machine learning and deep learning algorithms, however, has made it much easier to spot more intricate patterns between a collection of features that vary widely. These methods have shown to be successful in disease detection and disease prediction. However, some challenges such as limited interpret-ability, class imbalance, poor generalization capability, and lack of temporal modeling still exist and limit the use of these models in the healthcare sector [40].

Deep learning architectures such as convolutional neural networks, recurrent neural networks, long short-term memory networks, transformers and generative models

have gained significant popularity over the past few years due to their automatic learning capacity of meaningful representations from high-dimensional data.

Furthermore, explainable artificial intelligence methods are being incrementally adopted to augment model explainability and to boost the trust of clinicians in the intelligent systems [41].

Despite these developments some of the current studies are primarily aimed at disease diagnosis and have not paid attention to prediction of disease progression subtype. Besides, many methods are dependent upon one modality and do not make use of complementary information that can be derived from several data sources.

As a result, multimodal learning frameworks have been proposed as a promising approach to enhance prediction and offer customized disease management [40].

This chapter is a detailed review of the recent studies of diagnosis and progression prediction in Parkinson's disease. The different types of machine learning and deep learning are explained, then imaging based approaches, time-series approaches, explainable machine learning and deep learning approaches, generative learning approaches and multimodal approaches are discussed.

Furthermore, a critical analysis of recent studies in the period of 2022 to 2026 is conducted to highlight strengths, weaknesses and gaps in the existing studies that motivate the proposed framework [42].

Figure 2.1 shows the developments of AI methods used in the analysis of Parkinson's disease between 2019 and 2026.

The timeline demonstrates the transition from conventional deep learning approaches, including Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, towards more advanced paradigms such as transformers, graph learning, multimodal learning, and explainable artificial intelligence. These techniques have evolved over time, and have contributed greatly to advancing the ability to diagnose diseases, to model their course and to guide clinical decisions. Explainable multimodal models have been proposed as potential solutions in recent

years since they can effectively combine data from different modalities and deliver interpretable predictions that are appropriate for healthcare applications.

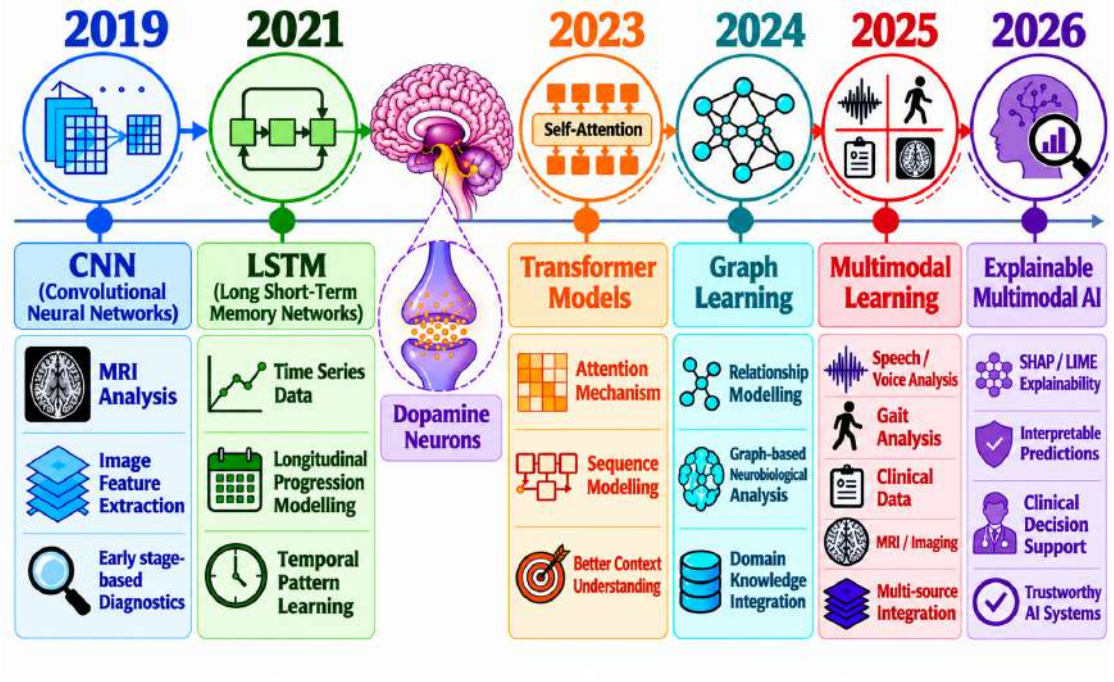


FIGURE 2.1: Timeline of Parkinson's Disease AI Evolution

Although multimodal learning and explainable AI techniques have already been investigated in Parkinson's disease research, challenges remain in jointly utilizing heterogeneous modalities, longitudinal disease information, class-imbalance handling and interpretable prediction within a unified framework. Therefore, the contribution of this study lies in the integration and systematic evaluation of these components for Parkinson's disease subtype and progression prediction, rather than in proposing multimodal learning or explainable AI as entirely new techniques.

2.1 Parkinson's Disease and Artificial Intelligence

The power of AI to process vast amounts of data and uncover patterns that are hard to see in traditional statistical analyses makes it a crucial piece in today's healthcare infrastructure. The application of AI technologies in Parkinson's disease research holds significant promise for enhancing diagnostic accuracy, disease monitoring,

and prediction of disease progression. With the help of machine learning and deep learning algorithms, complex relationships between different, heterogeneous features can be identified and objective decision support can be offered to the physician [43].

The initial research mostly focused on classical machine learning techniques like Support Vector Machines, Decision Trees, Random Forests, Logistic Regression and Naïve Bayes classifiers. These methods leveraged manual features extracted from speech, gait, handwriting and clinical examination data. While being successful, the results of these approaches largely relied on feature engineering and data preprocessing methods.

The study of PD has been advanced further in recent years by developments in deep learning. Convolutional Neural Networks (CNNs) have been widely used in imaging-based modalities and handwriting images, while recurrent neural networks and long short-term memory networks (LSTMs) have been used to model temporal relationships in longitudinal data. Moreover, transformer based models and attention mechanisms have shown great potential to learn complex representations from heterogeneous data sources.

Multimodal information integration has also emerged as an important research direction in parkinson's disease analysis. Clinical assessments, neuroimaging, speech signals, wearable sensor measurements, gait characteristics and handwriting patterns have been investigated in different combinations to exploit complementart information across modalities. Several studies have reported potential benefits from multimodal integration; however, the extent of improvement depends on the selected modalities, cohort characteristics, prediction task, fusion strategy and validation design [43].

While some of the AI techniques have been very successful, there are some challenges that remain to be addressed. The use of intelligent systems in clinical practice is still limited due to model interpretability, class imbalance, lack of longitudinal analysis and poor generalization ability. Therefore, the study of explainable artificial intelligence and multimodal learning technologies that address these challenges to

enhance clinical application have been gaining momentum recently [44].

2.2 Machine Learning Approach for Parkinson's Disease

The use of machine learning techniques for the diagnosis and prognosis of PD has been widely studied, as they are capable of detecting complex patterns from clinical and behavioral data.

The traditional machine learning algorithms are effective for classification and prediction tasks and have been successfully applied in diverse healthcare applications. It is common for these techniques to need feature engineering and data preprocessing to get useful data out of raw data.

The Support Vector Machines, Random Forests, Logistic Regression, Decision Trees and Extreme Gradient Boosting are some of the classic machine learning methods that have been extensively studied in the field of Parkinson's disease research.

These techniques have been used in the context of speech recordings, gait characteristics, handwriting patterns, wearable sensor readings and clinical data for detection and severity assessment of disease [43].

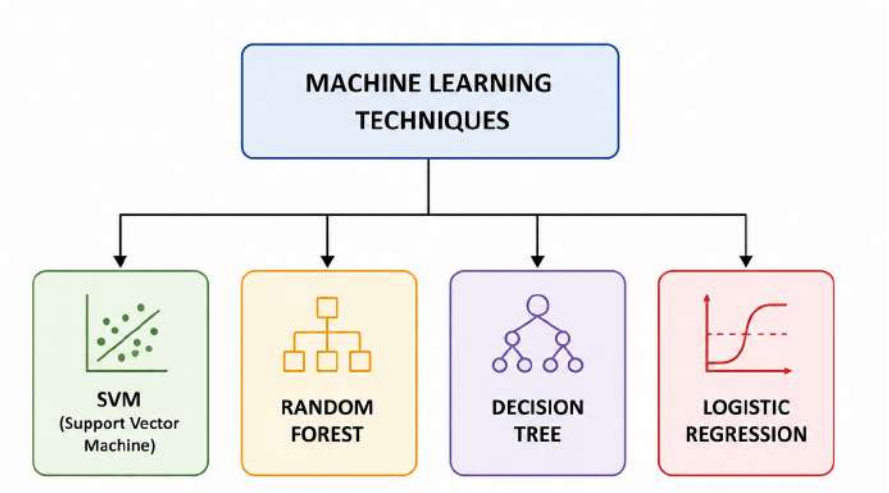


FIGURE 2.2: Common machine learning techniques used for Parkinson's disease analysis

As shown in Figure 2.2, some of the most widely applied machine learning algorithms for Parkinson's disease studies exist. The diagnosis and progression analysis of Parkinson's disease (PD) has been successfully utilized by various machine learning techniques. The techniques can successfully learn useful patterns from clinical and behaviour data and have shown good performance in classification and prediction problems.

2.2.1 Support Vector Machine

One of the most popular supervised learning classifiers used for classification is Support Vector Machine (SVM). SVM builds an optimized hyperplane that separates between classes, with a maximum distance. SVM is widely applied in the diagnosis of Parkinson's disease from speech signals, handwriting images, and gait measurements because of its capability of handling high-dimensional data and nonlinear relationships.

Nevertheless, SVM is proven to be a powerful machine learning algorithm with solid generalization power, especially in the case of relatively small and high dimensional datasets. For medical classification tasks, several studies have reported good overall performance of the SVM because it is able to build an optimal separating hyperplane between different classes with a minimum risk of overfitting. This is one of the reasons for which SVM are well suited for healthcare applications where the availability of large quantities of labeled patient data is often restricted.

Despite these benefits, performance of SVM is very sensitive to the choice of kernel functions (e.g., linear, polynomial, radial basis function (RBF) kernels) and the careful optimization of related hyperparameters. If a kernel and/or parameters aren't set up properly, classification performance can be drastically affected.

Furthermore, SVM might be costly and less scalable when the number of samples and features grows. The restrictions become more important as the size of the data is increased and more modalities are added, such as when handling multimodal or large-scale datasets, where the clinical, imaging, sensor-based, or other sources of information could lead to feature spaces that are high dimensional. Hence,

SVM can be used with relatively small datasets, but its parameter selection and complexity can restrict its utility in complex multimodal learning tasks.

2.2.2 Random Forest

Random Forest is an ensemble learning method which integrates several decision trees to gain both prediction ability and prevent overfitting. It is regarded as one of the most common machine learning techniques due to its strong performance, simplicity and robustness in the presence of noise and high dimensional data[45].

Random Forest has been used in Parkinson's disease research to analyze clinical assessments, gait information, wearable sensor information and speech recording. The algorithm also generates the feature importance measures that help to recognize the variables that are significant in the process of prediction. However, Random Forest can be challenging to use in the presence of high imbalance data or temporal dependencies.

2.2.3 Logistic Regression

Logistic regression is a popular statistical learning method for binary and multi-class classifiers. Logistic Regression has been widely used for medical diagnosis applications because of its simplicity and interpretability. Logistic Regression has been used in a number of studies where clinical and behavioural characteristics were used to differentiate patients from healthy controls [46].

While Logistic Regression offers interpretability, it might not be very effective when dealing with nonlinear relationships and complex interactions between the heterogeneous variables. Hence, more complex methods are generally preferred for multimodal and high-dimensional data.

2.2.4 Extreme Gradient Boosting

An ensemble learning algorithm based on gradient boosting, which builds very precise prediction models, is called Extreme Gradient Boosting (XGBoost). The popularity of XGBoost has increased greatly thanks to its efficiency in computation

and superior prediction performance.

Researchers have used XGBoost to develop a model of Parkinson's disease diagnosis and severity measurement based on clinical features, speech recordings and sensor measurements [47]. Although it is effective, there are some difficulties in the interpretation of the model and optimization of its hyperparameters.

In general, the traditional machine learning techniques have shown potential in predicting Parkinson's disease. These methods, however, are mostly feature engineering-based and are prone to miss a complex temporal dependency and multimodal interaction. This has led to a focus on deep learning methods that aim to provide deeper representation learning and more accurate predictions recently.

2.3 Deep Learning Approaches

The field of medical data analysis is undergoing an incredible revolution with the success of deep learning and the opportunities it creates to improve diagnosis and prediction of Parkinson's disease progression. The advantage of deep learning over traditional machine learning is that the algorithm learns hierarchical representations automatically from raw input without needing to be heavily engineered [48].

These methods have been found to be effective in processing complex and high-dimensional data and have proven to perform better in that regard. CNNs have been widely used for image modalities like MRI, PET, DaTscan, or handwriting images. Lately, Recurrent Neural Networks and Long Short-Term Memory networks have proven to be effective for modeling the longitudinal data and time dependency of disease progression [49].

Transformer architecture and attention mechanisms, more recently, have gained great popularity due to their capability to model long-range dependencies and complex relationships between the various features.

In addition, the surge in the availability of multimodal datasets and the enormous computational power have further spurred the use of deep learning methodologies in the study of Parkinson's disease. These techniques have proven to be highly

effective in the diagnosis, severity, prediction of disease progression and subtyping of disease. However, problems of interpretability, data sparsity and generalization of models have been some of the primary challenges that drive the development of explainable and robust frameworks.

2.3.1 Convolutional Neural Networks

The convolutional neural networks (CNNs) are among the most used deep learning architectures in image analysis and pattern recognition. CNNs learn hierarchical features from raw images automatically, which also do not require manual feature [50]. For their excellent performance, CNNs have been widely used in medical image analysis such as the diagnosis and grading of Parkinson's disease.

CNN architectures have been applied to analyzing MRI scans, PET images, DaTscan images and spiral handwriting images. These models can detect minute irregularities that can't be seen using conventional analysis techniques. CNN-based methods have been shown to yield considerable prediction accuracy in several studies [51].

Although they are effective, CNN models tend to need extensive data sets and high computing power. Furthermore, their black-box effect can make them difficult to interpret and thus not easily used in clinical settings. Therefore, the use of EAI approaches with CNN has grown, to improve transparency of the models.

2.3.2 Recurrent Neural Networks

Recurrent Neural Network (RNN) is a special type of neural network used to analyze sequential and time-dependent data. RNNs have memory mechanisms that make them able to remember past time steps and model temporal relationships, unlike feedforward neural networks [52].

A recent example of using RNNs in Parkinson's disease studies is on speech signals, gait measurements and longitudinal clinical records. They can handle sequential information which is suitable for disease progression studies [52]. Typical RNN

models, however, can experience the vanishing gradient problem, limiting the ability of these models to learn long-term dependency [53].

These constraints have led to more widespread use of architectures such as Long Short-Term Memory (LSTM) and Bidirectional LSTM (Bi-LSTM) for processing temporal data related to disease development.

2.3.3 Long Short-term Memory Networks

Long Short-Term Memory (LSTM) networks are an improved implementation of the Recurrent Neural Network (RNN) to address the vanishing gradient problem. LSTM models have memory cells and gating mechanisms that enable them to hold information for long periods of time and to effectively map out temporal relationships [54].

Longitudinal Parkinson's disease data is well suited for LSTM networks, as it exhibits sequential data characteristics. LSTM architectures can be used effectively to analyze repeated patient visits, symptom progression patterns and clinical assessments collected over time [55].

Previous studies have reported the potential of LSTM networks for forecasting disease progression and severity because they can model temporal dependencies across repeated observations.

Their ability to utilize sequential information makes them suitable for longitudinal Parkinson's disease analysis; however, the magnitude of their performance benefit depends on the dataset, prediction target, sequence construction and validation strategy [55].

But, LSTM networks can be quite demanding in terms of computational power and extensive training set sizes. They can also suffer from poor performance when data is highly imbalanced and/or modalities are different. However, LSTM is still one of the most popular models used for time-series forecasting in healthcare [55].

2.3.4 Bidirectional Long Short-term Memory Networks

The Bidirectional LSTM network (BiLSTM) is an improved version of the usual LSTM networks. The models are able to process sequential information in both forward and backward directions, which allows them to capture more contextual information [56].

BiLSTM models have also been investigated for sequential prediction tasks because they process contextual information in both forward and backward directions. Some studies have reported benefits from this bidirectional processing; however, the performance advantage over conventional LSTM models is task- and dataset-dependent. Increased computational complexity and memory requirements also remain important considerations [56].

Recently, it has been reported that BiLSTM models are superior to traditional LSTM models in a variety of prediction problems. The complexity of computation and memory use are still significant drawbacks of these networks, though [56].

2.3.5 Transformer Architectures

Recently, transformer architectures have been gaining traction as strong deep learning models that can effectively model long-range dependencies using the self-attention mechanism. Transformers process input sequences in parallel, which results in better computational efficiency and scalability as compared to recurrent networks [57].

This has led to an increasing focus on transformer architectures in healthcare for their ability to capture intricate relationships from diverse data sources. Transformer architectures have been used in the context of multimodal learning, speech analysis and longitudinal prediction tasks in the context of Parkinson's disease research.

Transformer architectures that include attention mechanisms allow the identification of salient features that affect prediction results. These architectures offer good

opportunities for enhancing predictions of progression subtypes and multimodal data integration [57]. Although transformer models have impressive capabilities, they typically demand significant amounts of data and compute. Hence, researchers are still trying to find efficient transformer-based architectures that can be used in medical applications [57].

2.4 Imaging-Based Approaches

Medical imaging techniques are an important part of the diagnosis and prognosis of neurological disorders. Imaging modalities in Parkinson's disease include structural and functional analysis of abnormalities that are linked to the neuronal degeneration [58]. The field of medical imaging has also seen significant improvements in the analysis of medical images, with the ability to automatically identify patterns related to diseases, thanks to advances in artificial intelligence and deep learning.

Some of the more common imaging methods used for diagnosing Parkinson's disease include MR imaging (MRI), positron emission tomography (PET) and dopamine transporter imaging (DaTscan) [59]. These modalities have been shown to be useful for deep learning models to extract discriminative features and support clinicians to assess disease state, with specifically convolutional neural networks (CNNs) being promising for these tasks.

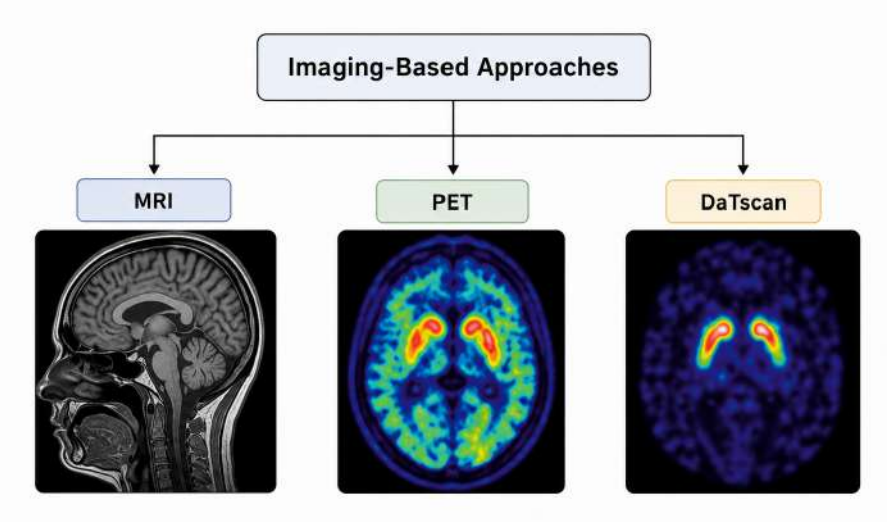


FIGURE 2.3: Major imaging modalities used in Parkinson's disease diagnosis and progression analysis

Medical imaging is important for the diagnosis and progression analysis of Parkinson's. Structural and functional imaging techniques are complementary in providing information on brain anatomy, metabolic activity and dopaminergic degeneration. Figure 2.5 shows the most significant imaging methods used in Parkinson's disease studies.

While encouraging results have been obtained using imaging-based methods, there are important challenges associated with small sample sizes, data variability and computational complexity. However, imaging still plays an important role in disease diagnosis and analysis of disease progression [59].

2.4.1 Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) is a non-invasive imaging technique that provides good anatomical detail from the brain. MRI has been widely used to study neurodegenerative diseases and structural changes related to Parkinson's [60].

There are several studies using machine learning and deep learning techniques to analyze MRI scans. CNNs have demonstrated intriguing features to learn hierarchical representations from MRI images and boost disease classification performance. MRI-derived techniques are specifically useful to distinguish Parkinson's from other neurological disorders and advanced to study the patterns of disease progression [60].

Although such MRI-based models are effective, they need extensive amounts of data and computing power. Furthermore, differences in image acquisition protocols can impact model generalization and prediction accuracy.

2.4.2 Positron Emission Tomography

Positron Emission Tomography (PET) is a functional technique that gives information about brain metabolism and neurotransmitter systems. PET imaging has already been extensively utilised to study neuronal degeneration and disease

progression [61].

Deep learning techniques have been applied to PET images for analysis and patient identification of Parkinson's disease from healthy subjects. Positron emission tomography (PET) imaging gives important insights into dopamine metabolism and allows abnormalities to be identified at an early stage [61].

PET imaging, however, is fairly costly and is subject to radioactive tracer exposure, restricting the use of this imaging technique in the clinic. As a result, other imaging methods and multimodal methods are gaining more and more interest [61].

2.4.3 Dopamine Transporter Imaging

DaTscan imaging is a special type of nuclear imaging, which assesses dopamine transporter activity in the brain. It offers useful data on the degeneration of dopaminergic neurons and has emerged as a key diagnostic aid in Parkinson's disease [62].

In recent years, convolutional neural network and transfer learning models have been used to analyze DaTscan images. These methods have shown to be promising in separating Parkinson patients from healthy controls and to detect subtle abnormalities related to dopamine deficiency [62].

However, DaTscan imaging is costly and not available everywhere because of its effectiveness. Furthermore, imaging results are difficult to interpret and need great expertise, thus the development of automated analysis systems using artificial intelligence [62].

2.5 Speech-Based Approaches

One of the typical symptoms of Parkinson's disease is speech impairment. Patients often have a decreased voice volume, monotony of voice, voice tremors and articulation problems. The abnormalities are helpful in the diagnosis and evaluation of the disease's progress [63].

A variety of techniques from the fields of machine learning and deep learning have been used to decipher speech recordings and to identify the relevant acoustic features. Support Vector Machines, Random Forests, Convolutional Neural Networks and Long Short-Term Memory networks have shown to be effective in the classification of Parkinson's disease patients from healthy controls. Speech based methods have several benefits such as low cost, non-invasive data collection and the possibility of using these methods in remote healthcare applications. Model performance and generalization ability, however, could be affected by environmental noise, recording quality, or speaker variability [63].

The development of intelligent systems that are capable of providing accurate and reliable predictions has improved with recent advances in deep learning and attention mechanisms that are able to better analyze speech.

2.6 Handwriting and Spiral Drawing Analysis

Handwriting problems are viewed as one of the earliest signs of Parkinson's disease and closely linked with motor disturbances. Micrographia is a common symptom of Parkinson's disease, along with tremors, irregular stroke patterns and poor coordination of hand movements. These abnormalities may be evident in handwriting and in drawing spirals and can be used as a source of information for diagnosis and analysis of disease progression [64].

The study of spiral drawing has been a subject of a lot of interest, as it is a simple, non-invasive and inexpensive method for detecting motor abnormalities. Drawing tasks are commonly used in clinical practice to assess the severity of tremor and motor impairment through the use of spiral and wave drawings. Digital tablet and image acquisition technology have also made it easier to collect and analyze handwriting data [64].

Traditional machine learning algorithms such as Support Vector Machines, Random Forests, Decision Trees and Logistic Regression have been used in several studies to classify Parkinson's disease using handwriting features. These techniques are usually based on features that are handcrafted from images or digitized signals.

Various parameters (pen pressure, stroke velocity, acceleration and curvature) have been used to discriminate Parkinson's disease patient from healthy subjects.

Deep learning has contributed to the improvement of handwriting analysis in recent years. CNNs have been shown to have great potential to learn discriminative features in an automatic way when applied to spiral images or handwriting images without any human intervention in the process of feature engineering [65]. CNN-based architectures have been reported to show promising performance for patient–control discrimination in PD.

Moreover, the transfer learning using pre trained models like VGG16, ResNet, MobileNet and EfficientNet has produced promising results in image based Parkinson's disease analysis. These methods enable better feature representation and decrease the training size. By integrating transfer learning, researchers have been able to achieve improved prediction accuracy and better model generalization [66].

Although these developments have been achieved, handwriting-based methods have some drawbacks. Image quality, different writing styles, small data sets and class imbalance can have a negative impact on the performance of the model. Furthermore, the black-box problem of deep learning models makes it hard to interpret and poses clinical questions about applicability. Hence, the use of EAI techniques has become paramount to improve the transparency and establish trust in prediction systems.

In summary, handwriting and drawing analysis through the spiral approach are promising avenues for the diagnosis and monitoring of PD. They are non-invasive, easy to acquire and suitable for deep learning methods, rendering them promising for intelligent healthcare applications [67].

2.7 Wearable Sensor and Gait Analysis

Gait disturbance is a frequent symptom of PD and can be accompanied by balance issues, freezing and postural instability. The use of wearable sensors has emerged as a promising approach to continuously monitor motor symptoms and provide an

objective measure of the severity of the disease. These sensors offer quantitative reading allowing for precise evaluation of movement dysfunction [68].

Accelerometers, gyroscopes, inertial measurement units and pressure sensors are used for gait analysis, among other sensor technologies. These devices allow gait characteristics like stride length, walking speed, cadence and variability to be extracted. These features are useful in understanding the development and progression of the disease and motor dysfunction [68].

Gait features have been used to classify Parkinson's disease patients by several machine learning methods, such as Support Vector Machines, Random Forests and Decision Trees. Recently, deep learning methods have shown to be more effective in automatically learning meaningful representations from sensor data and discovering complex movement patterns.

While wearable sensor-based approaches are beneficial, placement of the sensor, environmental variables and patient variation can impact them. Furthermore, the continuous monitoring process produces vast amounts of data, and the data needs to be analyzed through a sophisticated computational process. However, wearable technologies still offer valuable opportunities for distant healthcare and personalised disease management [68].

2.8 longitudinal and Time-series Approaches

Parkinson's is a progressive neurological disease that has symptoms that wax and wane over time. Therefore, longitudinal study has gained importance to elucidate the trajectory of the disease and the prediction of the disease pattern in the future. Longitudinal data are data from successive visits to the patient and can be useful in generating information about changes over time in clinical manifestations.

Traditional cross sectional methods use data from one point in time only, and are unable to account for temporal dependencies relating to disease progression. Time series methods take advantage of sequential information, allowing more comprehensive modeling of disease evolution, in contrast. These are especially

beneficial for determining progression subtypes and for guiding individual therapeutic decisions [69]. Recurrent Neural Networks and Long Short-Term Memory networks have shown to be very effective models for temporal relations. They have memory mechanisms to store long-term dependencies and effectively analyze repeated observations. Additionally, the bidirectional LSTM networks can be used to capture the sequential information both forward and backward, which improves prediction accuracy [70].

Recent research indicates that the use of longitudinal information greatly outperforms the use of crosssectional information in the prediction of progression. The irregular follow-up intervals, missing values and limited sample sizes, however, remain a challenge in time-series analysis. Thus, there is a need to build advanced architectures and strong preprocessing strategies to maximize the efficiency of longitudinal datasets.

2.9 Explainable Artificial Intelligence

Deep learning techniques have been shown to have excellent predictive power, however they are black box models and are hard to accept in healthcare applications. In order for clinicians to use intelligent systems in clinical practice, the predictions need to be transparent and interpretable. As a result, Explainable Artificial Intelligence (XAI) has become a dynamic field of study [71].

Explainability techniques are designed to give information about the factors involved in the prediction results. SHapley Additive exPlanations (SHAP), Local Interpretable Model-Agnostic Explanations (LIME), and Gradient-weighted Class Activation Mapping (Grad-CAM) are among the most commonly used techniques. Such methods can help to visualize the importance of features and gain insights into model behavior [72].

Global and local explanations are offered, since the values of SHAPs quantify the contribution of individual features to the model predictions [**?**]. LIME is a method for localizing models of complex models and providing explainable answers to individual cases [73]. Grad-CAM techniques are especially valuable for image

analysis and help visualize important regions that can impact the classification given an image.

The incorporation of XAI into healthcare models fosters trust and transparency among professionals, thereby boosting the confidence of patients.By incorporating XAI into healthcare models, transparency is improved and trust is built among professionals, which results in increased trust for patients. Thus, explainability has emerged as an essential part of intelligent healthcare systems of today [74].

2.10 Generative Learning Approaches

Class imbalance is one of the most important challenges faced in medical data sets. Model performance and the generalization ability can be harmed by the limited samples in minority classes. In recent years, generative learning techniques have been developed to solve this problem by generating synthetic data [75].

GANs involve two networks: the generator network and the discriminator network. A generator generates samples, while the discriminator tries to tell the difference between real and generated samples. This competitive learning process allows to obtain realistic data distributions [75].

For tabular data, Conditional Tabular Generative Adversarial Networks (CTGANs) have been tailored to this type of data, and have shown good results in healthcare. The CTGAN models can be used to generate synthetic records without compromising statistical properties of original data. Each of these methods helps in balancing the classes and helps to make the model more robust [76]. Generative models have been shown to be effective but have some drawbacks like train instability and mode collapse. However, they still offer great opportunities to improve prediction accuracy in imbalanced data sets [76].

2.11 Multimodal Learning Approaches

Because of the complexity and diversity of Parkinson's disease, multimodal learning methods have been studied that try to learn information from multiple sources.

Multimodal frameworks take advantage of complementary information and offer a more comprehensive understanding of the disease characteristics [77].Several studies have explored multimodal approaches for Parkinson's disease analysis by combining complementary data sources such as clinical information,neuroimaging, speech, gait, wearable sensor data and handwriting patterns. More recent studies have also investigated deep learning and multimodal fusion approaches to integrate heterogeneous information for Parkinson's disease prediction [78]. Multimodal approaches tend to have improved accuracy and a better generalisation capability than single-modality approaches. However, despite these developments, important challenges remain in Parkinson's disease subtype and progression prediction. Existing approaches do not always jointly utilize complementary multimodal information and longitudinal disease patterns, while class imbalance and limited model interpretability remain important concerns. These limitations motivate the development of an integrated multimodal, longitudinal and explainable framework for Parkinson's disease subtype and progression prediction [79].

In recent years, several significant breakthroughs in multimodal deep learning (MDL) and explainable artificial intelligence (XAI) have created a new demand for the development of strong, clinically relevant prediction systems. Thus, multimodality learning is one of the most promising lines of research in Parkinson's disease diagnosis and prediction of disease subtypes.

Various types of fusion methods can be implemented in multimodal learning frameworks such as early fusion, intermediate fusion and late fusion. This research uses a decision level late fusion technique to fuse the results of machine learning and deep learning models.

2.12 Machine Learning and Deep learning Approaches

Over recent years, significant progress has been made in recent years in the use of AI-based methods to diagnose PD and monitor PD progression. In recent

years, machine learning and deep learning architectures have been used to uncover intricate patterns in diverse data sources and enhance prediction accuracy.

During 2022, a number of research studies were carried out on the traditional machine learning techniques and methods for the classification of PD based on clinical assessment and speech recording. The methods of ensemble learning and SVM showed good performance and yielded interpretable results. However, these techniques were not able to adequately reflect nonlinear relationships or temporal dependence in disease progression [80].

As deep learning has advanced rapidly, CNN and RNN are increasingly popular methods for predicting Parkinson's disease. CNN-based architectures exhibited high performance in medical image analysis and handwriting classification while recursive architectures were used to analyze sequential and longitudinal information [81].

Pre-trained models like VGG16, ResNet50, MobileNet and EfficientNet have also been explored as transfer learning approaches. These methods have proven to be very effective in achieving better classification accuracy and less reliance on huge training sets [82].

2.13 Time-series and Longitudinal data Approaches

The progressive nature of PD has made longitudinal analysis and temporal modelling more of a focus in research. Repeated measures datasets are useful for providing information about the pattern(s) of disease progression and trajectory. In comparison to other models, LSTM networks have shown tremendous success in capturing the temporal or long-term dependency and modeling such relationship. Bidirectional LSTM models additionally process information in the both directions, further improving prediction accuracy. Recurrent neural networks (RNNs) and attention models are reported to be effective for the prediction of progression in several studies [83].

Another attractive alternative is the transformer architecture due to its ability to represent longer-range dependencies and model heterogeneous sequences. Successful applications in health care have been reported and the models are now beginning to be explored in the context of Parkinson's disease progression analysis [84].

2.14 Explainable Artificial Intelligence

Deep learning models possess outstanding prediction capabilities, yet they are black-box models that pose a challenge in medical applications. To overcome this drawback, researchers have been increasingly using explainable AI methods in intelligent prediction systems. There are popular methods for explaining machine learning models such as SHAP and LIME, while Grad-CAM or Grad-CAM++ are heavily used with image-based deep learning architectures. These methods offer insights into the factors that affect prediction results and enhance the credibility of AI systems [85].

The seamless adoption of explainability techniques becomes even more crucial in the field of healthcare, where clinicians need clear and understandable models before they start using artificial intelligence systems in practice [86].

2.15 Generative Learning and class imbalance

The imbalance of class is one of the significant issues faced by Parkinson's datasets. Predictions for minority classes are often biased and the model performance is limited due to limited size of minority classes.

Generative Adversarial Networks and Conditional Tabular Generative Adversarial Networks have proven to be helpful for synthetic data generation and class balancing. These have shown to have a good performance in terms of prediction accuracy and increase in model robustness [87].

Recent works show that augmentation techniques based on CTGAN preserve statistical properties of tabular data and can boost class balancing and diversity of

data significantly [88].

2.16 Research Gaps in Existing Literature

While significant advances have been made in the field of AI-driven Parkinson's disease detection and analysis, there are still a number of challenges in the current literature [89]. Many of the previous studies have focused on single modalities like clinical evaluation, speech, gait analysis, neuroimaging or handwriting. While multimodal approaches have been investigated more and more, the combination of multiple data sources is difficult due to the various data availability, representation, subject overlaps and prediction targets.

Longitudinal information is of special interest for Parkinson's disease because the symptoms evolve and clinical severity over time. However, cross section analysis is still popular and longitudinal analysis must face problems with irregular follow-up periods, missing observations and small populations with adequate numbers of follow-up visits [90].

Class imbalance is also a key methodological challenge that may make it more difficult to identify underrepresented classes. One approach to augmenting minority-class tabular observations is generative, using methods like CTGAN, but the advantage of such methods in the downstream tasks is dependent on the dataset and needs to be tested on independent real data.

Another factor to take into account is model interpretability, especially for complex machine learning and deep learning models. Explainable AI methods like SHAP and Grad-CAM can offer insights into the model's behavior, but outputs should be interpreted with caution and should not be used as proof of causal or disease-specific mechanisms.

Finally, most of the existing literature on Parkinson's disease artificial intelligence is dedicated to diagnosis or disease-state classification, and prediction of disease progression, as well as the longitudinal modelling, multimodal information, class imbalance and explainability remain challenges for future research. These

observations inspire the combined experimental approach explored in this study [91].

2.17 Chapter Summary

In this chapter, the researchers reviewed the literature on Parkinson's disease (PD) diagnosis, progression prediction and the use of artificial intelligence techniques in healthcare in a comprehensive manner. Common machine learning algorithms (SVM, Random Forest, Logistic Regression, Extremely Gradient Boosting) and deep learning architectures (CNN, LSTM, BiLSTM, Transformer based models) were discussed. The chapter also summarized imaging-based approach, speech analysis, handwriting analysis, gait analysis, wearable sensor technologies, multimodal learning, generative learning and Explainable Artificial Intelligence (XAI) and their role in the diagnosis and prediction of PD progression.

A number of significant research gaps were revealed through the analysis of studies. Most current methods are based on one-modality datasets, are mainly concerned with disease diagnosis and only a small number of studies make use of longitudinal patient data. However, issues like class imbalance, multimodal nature of data, difficulty in interpretability and high computational cost are still significant hurdles to clinical deployment. Advanced multimodal learning, transfer learning, CTGAN, and XAI techniques have enhanced the prediction accuracy, but their integration for PD progression subtype prediction is still underutilized.

With the above findings, the chapter set the motivation for the proposed research and validated the development of an intelligent multimodal framework that incorporates clinical assessments, gait features, longitudinal time-series data, spiral handwriting images and MRI/DaTscan images. The proposed framework also includes CTGAN to tackle the issue of class imbalance and XAI techniques to enhance model transparency and clinical trust. The methodology of the research study is presented in the next chapter, including the selection of the datasets, the preprocessing and fusion of the multimodal data, the machine learning and

deep learning models used, synthetic data generation, explainability techniques and the evaluation metrics employed to validate the proposed Parkinson's disease progression subtype prediction framework.

Chapter 3

Research Methodology

In this chapter, the research methodology used for the development and evaluation of the proposed multimodal framework for analysis of Parkinson's Disease using Explainable Artificial Intelligence is presented. The methodology aimed to explore complementary information gathered from clinical evaluation, follow-up patient visits, structural magnetic resonance imaging (MRI) and spiral hand-writing data [92]. Since these modalities are different aspects of PD and not all of the subjects had access to all modalities, modality-specific strategies were used for preprocessing and modeling prior to multimodal integration [93].

The motor subtype classification clinical component was created from baseline clinical data. Parkinson's patients were divided into Tremor-Dominant (TD), Postural Instability and Gait Difficulty (PIGD) and Indeterminate/Mixed groups according to clinical criteria of the Movement Disorders Society (MDS-UPDRS) for the subtypes of Parkinson's. To build a leakage-aware clinical feature set, the variables used to define the clinical subtype labels were dropped.

The clinical subtypes were predicted using Logistic Regression, Support Vector Machine and Random Forest models. To study the problem of class imbalance, Conditional Tabular Generative Adversarial Network (CTGAN) was trained only with the real training partition, while the minority PIGD and Indeterminate classes were synthetically generated.

A separate longitudinal component was developed to model the evolution of the disease over multiple visits to the patients. Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM) networks were tested on longitudinal MDS-UPDRS Part III observations and the future progression was grouped as Slow, Moderate and Fast.

Transfer-learning based MobileNetV2 models were used to analyze structural MRI and handwriting information. The motor subtype classification was evaluated using MRI and the branch of the spiral handwriting was evaluated separately for discrimination between Parkinson's disease and healthy control subjects.

To enhance the interpretability of the model, Explainable Artificial Intelligence (XAI) techniques were added. The selected clinical model was used to quantify the contribution of each individual clinical variable using SHapley Additive exPlanations (SHAP) and Gradient-weighted Class Activation Mapping (Grad-CAM) was used for the image-based models to visualize the regions of the image that contribute to individual predictions.

Finally, the outputs of the class-probability models of the clinical and MRI models were combined in decision-level late fusion for subjects that had both clinical and MRI information. The datasets, pre-processing, classification of subtypes and progression labels, generation of progression labels with the CTGAN, augmentation with CTGAN, model architectures, training strategies, explainability methods and multimodal fusion are described in detail in the following sections, as well as the evaluation protocol.

3.1 Proposed FrameWork

The overall research framework was designed to be complementary modality-specific learning branches and then explainability analysis and multimodal decision-level fusion. The study was based on clinical and longitudinal data from the Parkinson's Progression Markers Initiative (PPMI) and structural MRI data from the PPMI imaging repository, along with handwriting data from the HandPD/NewHandPD resources [92]. Due to the differences in data structure, dimensionality, temporal

characteristics and prediction objectives, each modality had its own dedicated preprocessing pipeline [94].

In the clinical motor-subtype branch, baseline clinical data was merged at the patient level and a leakage-aware dataset of 4,806 subjects with 30 predictor variables was created. Motor subtype labels were created based on subtype criteria using MDS-UPDRs as Tremor-Dominant (TD), PIGD or Indeterminate/Mixed.

A total of 2,055 TD, 1,039 PIGD and 1,712 Indeterminate subjects were included in the resulting clinical dataset. A variety of classifiers were explored, including logistic regression, support vector machine and random forest, and patient-level partitions were used for preprocessing and model evaluation [95].

To handle class imbalance in clinical subtypes dataset, CTGAN was also employed. Before the generative modeling, the patient-level data was split into 3,844 real training subjects and 962 held-out real test subjects. Only real training partition was fitted with CTGAN. The initial training distribution comprised 1,644 TD, 831 PIGD and 1,369 Indeterminate subjects.

Thus, 813 synthetic PIGD records and 275 synthetic Indeterminate records were created, resulting in 1088 synthetic minority class samples. The training data set was then expanded to include 4,932 records, each representing 1,644 samples of one of the subtypes. For synthetic samples, the only learning augmentation used was training augmentation, which were never part of the held-out test set.

Repeated MDS-UPDRs Part III assessments were conducted chronologically for each patient for the purpose of the longitudinal progression model. There were 1,122 eligible subjects that met the longitudinal inclusion criteria. The early observations served as model inputs while the later observations were treated separately to generate Slow, Moderate and Fast progression labels.

The resulting sequences were sequences up to 8 time steps long, each having 34 input features. To explore the temporal dependencies in Parkinson's disease progression, LSTM and BiLSTM networks were tested. Imaging data included structural 3D T1-weighted MRI data. After DICOM-level QC and spatial ordering, removal of

non-volume records and duplicate slices, 238 MRI subjects were included that had motor-subtype labels. A volume of each subject was extracted and 2,380 images were taken from the middle of the volume.

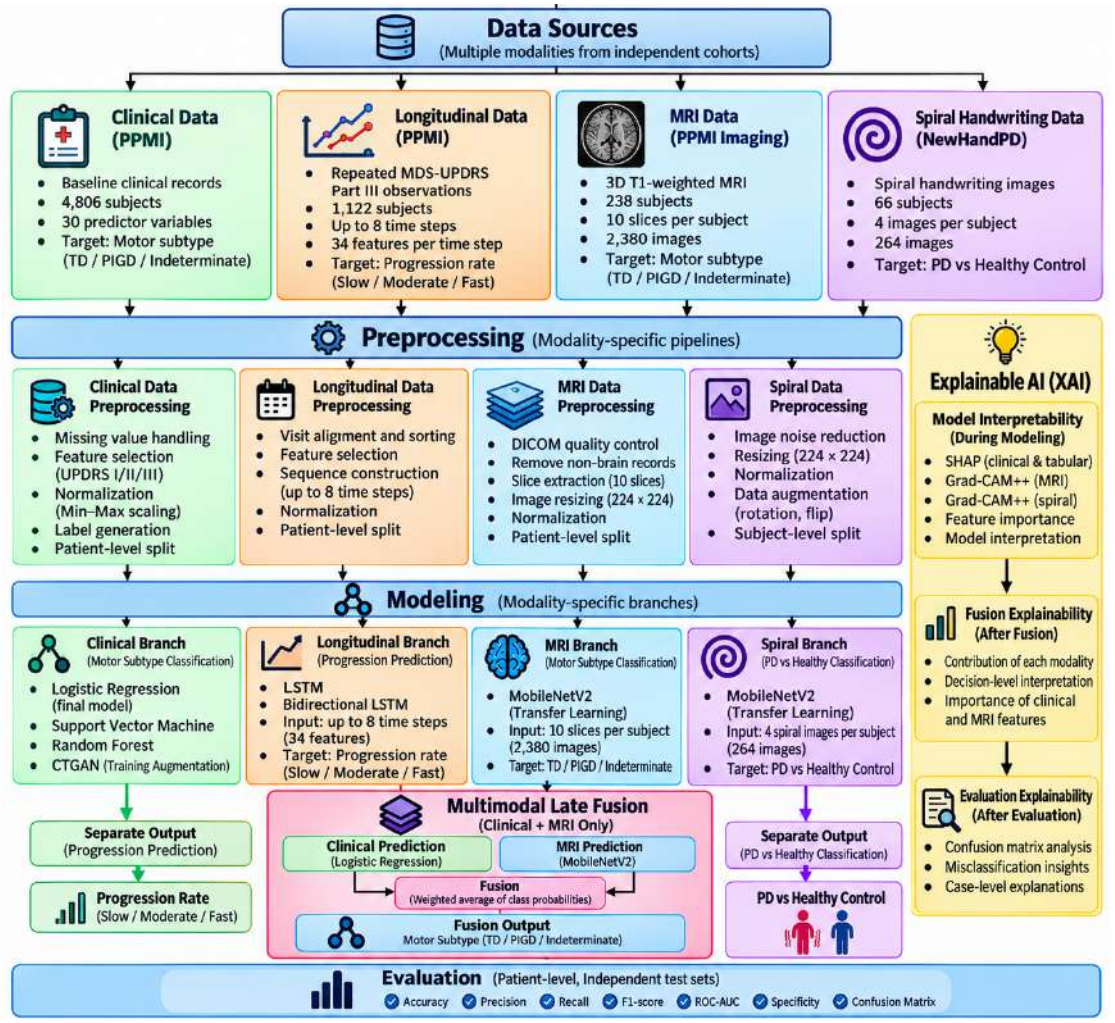


FIGURE 3.1: Framework of the proposed explainable multimodal Parkinson's disease progression prediction model

Three-class motor-subtype classification was carried out using a transfer-learning architecture for MobileNetV2, where patient-level splitting was used prior to the extraction of slices to avoid data leakage. The ten slices of each patient were then used to predict probabilities for the patient and then averaged together.

Images of handwriting in the spiral style were fed into a different branch of transfer-learning, MobileNetV2, for Parkinson's disease versus healthy-control classification. This experiment involved 66 subjects and 264 spiral images in the NewHandPD subset (four images per subject). To avoid images of the same person appearing in

various partitions, subject-level train, validation and test partitions were created before the model was trained.

The interpretability of the clinical model was explored through the use of SHAP while the image-based models were explored through the use of Grad-CAM. When performing multimodal integration, the clinical prediction and MRI prediction were taken together only for those subjects for which both modalities and the same motor-subtype target were present.

To perform the late fusing at the decision level, the probability of each class provided by each of the clinical and MRI models was averaged with fixed weightage of 0.5 for each model. This design enabled the evaluation of the different modalities, the generative augmentation, temporal modeling, explainability and multimodal integration within a common experimental setup.

3.2 Dataset Description

In this study, different aspects of Parkinson's disease were investigated by using heterogeneous data, spanning from clinical, longitudinal studies to neuroimaging and handwriting studies. Clinical evaluations, demographic data and repeated longitudinal visits were collected from the Parkinson's Progression Markers Initiative (PPMI) along with structural magnetic resonance imaging (MRI) data. The images of spiral handwriting were captured independently from the NewHandPD dataset. While some modalities were available for all subjects, a modality specific experimental protocol was used for each data source and a paired multimodal cohort was not assumed for the data.

The clinical data were mainly used for motor subtype classification, i.e. Tremor-Dominant (TD), Postural Instability and Gait Difficulty (PIGD) and Indeterminate/Mixed groups. Observations from the longitudinal MDS-UPDRS part III were used to examine future disease progression separately, which were classified into Slow, Moderate and Fast. Motor subtype classification was performed on structural 3D T1-weighted MRI images while binary classification between PD and healthy controls was performed on spiral handwriting images. These different data types

allowed the proposed approach to evaluate clinical features, temporal evolution, structural MRI data and handwriting motor features independently. Next, in subjects in which both clinical and MRI information and the same motor-subtype target were present, these were combined at the decision-level, in the form of late fusion. The multimodal datasets used in this study are shown in Fig. 3.2.

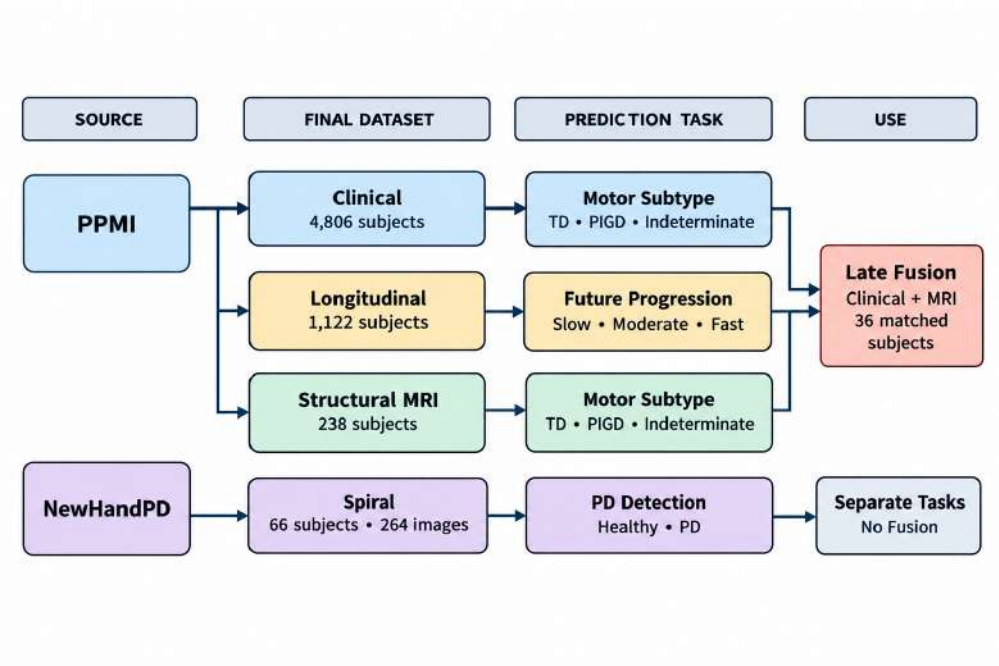


FIGURE 3.2: Overview of the datasets and modality-specific cohorts used in the study.

TABLE 3.1: Summary of Datasets, Modalities and Prediction Tasks

Data Source	**Modality**	**Final Dataset Size**	**Prediction Task**
PPMI	Clinical	4,806 subjects, 30 predictors	TD / PIGD / Indeterminate
PPMI	Longitudinal	1,122 subjects, repeated MDS-UPDRS Part III visits	Slow / Moderate / Fast progression
PPMI Imaging	Structural 3D T1-weighted MRI	238 labeled subjects, 2,380 selected images	TD / PIGD / Indeterminate
NewHandPD	Spiral handwriting	66 subjects, 264 images	PD / Healthy

The use of heterogeneous data allowed the creation of a multimodal approach that can leverage complementary information from different sources. The datasets used in this study are provided in Table 3.1.Reported cohort sizes are the sample numbers after task-specific eligibility criteria, quality control, pre-processing and labeling. They are not the overall size of the source repositories, but rather just the number of entries.

3.2.1 Parkinson's Progression Markers Initiative

Clinical, longitudinal and structural MRI data were the primary source of data in this study from the Parkinson's Progression Markers Initiative (PPMI). PPMI is an ongoing observational study that collects standardised clinical, imaging, biological and demographic data that can be used in Parkinson's disease research. With repeated clinical evaluations, PPMI is especially useful for studies of inter-patient variability and temporal changes in the condition of Parkinson's disease [95].

Several PPMI data files were first reviewed to determine the variables that would be included in the proposed framework. These encompassed demographic data, age at visit, MDS-UPDRS Part I, MDS-UPDRS Part II, MDS-UPDRS Part III, MDS-UPDRS Part IV, Modified Schwab and England Activities of Daily Living, Neuro-QoL mobility assessments, Neuro-QoL upper-extremity assessments and other clinical data available.

All variables downloaded, however, were not always part of all experiments. For clinical motor-subtype classification, however, specific datasets were created for each task, while for modeling longitudinal motor type progression, a specific dataset was created for each motor type.

The MDS-UPDRS evaluations were very significant in the study. Part I provides information about non-motor experiences of daily living; Part II rates motor experiences of daily living; and Part III is a clinician-administered motor examination. These assessments differed in their applications, such as generating subtype labels, building clinical features with leakage-aware and analyzing longitudinal progression.

3.2.1.1 Clinical Dataset for Motor Subtype Classification

To facilitate clinical motor-subtype classification, a patient-level baseline dataset was created from the available clinical records from PPMI. Following integration of the necessary clinical evaluations and application of the motor-subtype labeling procedure, 4806 subjects remained in the clinical experiment.

The subjects were divided into three motor-subtype groups: Tremor-Dominant (TD), PIGD and Indeterminate/Mixed. The resulting clinical dataset contained 2,055 TD subjects (42.76%), 1,039 PIGD subjects (21.62%) and 1,712 Indeterminate/Mixed subjects (35.62%).

Motor-subtype labels were created based on the well-defined tremor and postural-instability/gait-difficulty components from MDS-UPDRS Part II and III. The variables used in labels were also excluded from the set of model predictors because the clinical variables involved in the labels might leak the information.

After leakage control and feature selection, 30 clinical, demographic and functional predictor variables were used in machine learning analysis.

Table 3.2: Feature Groups and Variables Used for Clinical Classification

Feature Group	**Variables**
Demographic	AGE_AT_VISIT, SEX, HANDED, RAASIAN, RABLACK, RAWHITE
MDS-UPDRS Part I	NP1COG, NP1HALL, NP1DPRS, NP1ANXS, NP1APAT, NP1DDS, NP1RTOT
Functional assessment	MSEADLG
Neuro-QoL Mobility	NQMOB37, NQMOB30, NQMOB26, NQMOB32, NQMOB25, NQMOB33, NQMOB31, NQMOB28
Neuro-QoL Upper Extremity	NQUEX29, NQUEX20, NQUEX44, NQUEX36, NQUEX30, NQUEX28, NQUEX33, NQUEX37

This final clinical data set was then split at the patient level into training and held-out test sets employing stratified sampling. There were 3,844 training subjects and 962 test subjects. The training partition was made up of 1,644 TD, 831 PIGD and 1,369 Indeterminate subjects and the test partition of 411 TD, 208 PIGD and

343 Indeterminate subjects.

The same locked partition was subsequently used for the augmentation experiment of the CTGAN to compare the real only and synthetically augmented training data. No real test subjects were held back for training and creation of synthetic data for CTGAN.

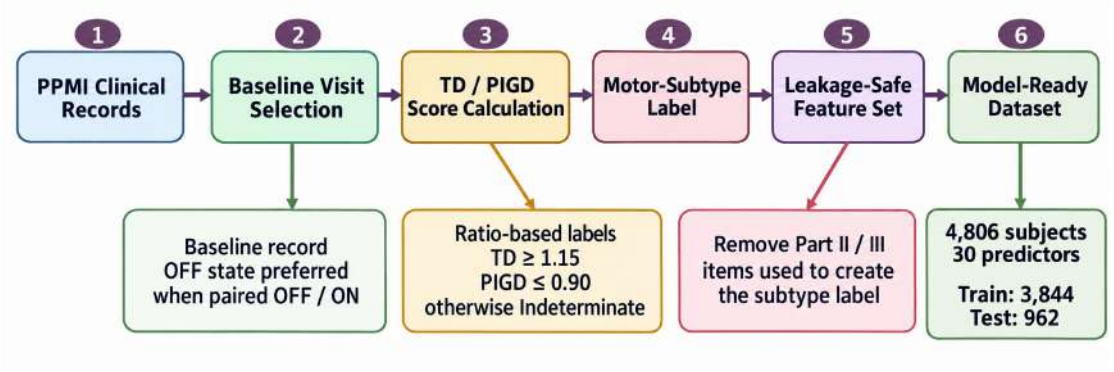


FIGURE 3.3: Clinical Data Preprocessing.

3.2.1.2 Longitudinal Dataset for Progression Prediction

To explore the temporal disease process, a separate longitudinal dataset was created based on repeated PPMI MDS-UPDRS Part III motor examinations. The clinical subtype classification data set and the data set for the longitudinal progression were considered as two experimental cohorts, as they were used to solve different prediction tasks.

These MDS-UPDRS Part III visits were conducted at each patient and were chronologically arranged. Observations of the early follow-up period were input to the model, while observations made outside of the early follow-up period were used independently for the generation of future progression-label [95].

The subjects were asked to have three observations at least in the early input time, two observations at least in the future period and at least one year of follow-up in the future period. 1,122 subjects were retained for modelling in a longitudinal study after applying these eligibility criteria.

Future progression was estimated by the rate of change in MDS-UPDRS total

scores and defined as Slow, Moderate, or Fast based on the study's definition. In the last cohort there were 370 Slow progressors, 382 Moderate progressors and 370 Fast progressors. In the case of sequence modelling, 33 variables from the motor examination section of MDS-UPDRS were fused with the temporal information, yielding 34 variables at every time step. The number of time steps in each patient sequence was up to eight. The final number of 1,122 subjects were split at the patient level into 785 Training subjects, 168 Validation subjects and 169 held-out Test subjects.

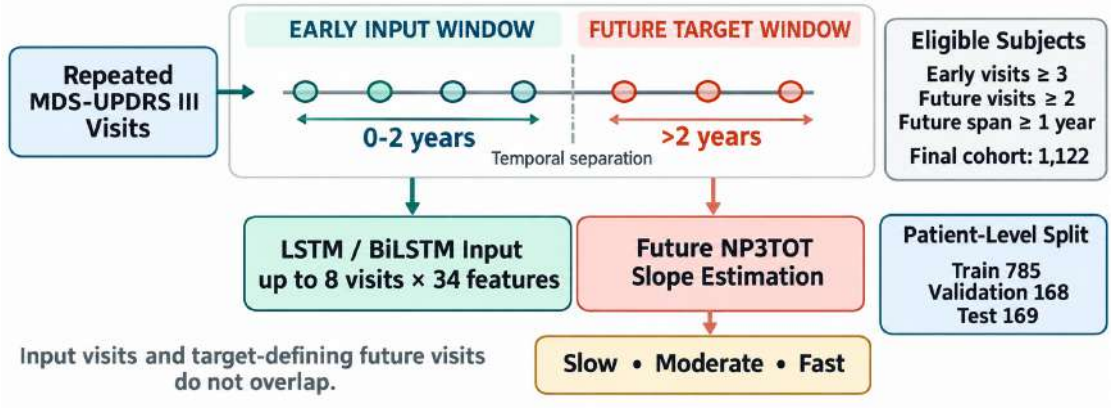


FIGURE 3.4: Temporal organization of early longitudinal observations and future progression-label generation.

3.2.2 Structural MRI Dataset

Structural MRI was measured from the PPMI imaging repository. The final neuroimaging experiment used imaging searches limited to baseline original MRI examinations, and the preferred 3D T1-weighted structural MRI was used. The final experiment was, therefore, performed using structural MRI (MRI) instead of structural MRI and DaTscan (MRI/DaTscan) [96].

We retrieved 290 candidate MRI series, from 288 subjects, using the initial imaging retrieval. Subsequently, a quality check was accomplished at the DICOM level to determine anatomical volumes in three dimensions. Fourty-four incomplete series (with a single slice) were discarded. After the series-level quality control, 246 valid series (244 subjects) were left. Two subjects had more than one eligible series that were resolved in the process of selecting one final series per patient.

The MRI cohort was then subjected to the quality control and was matched to the clinical motor-subtype labels. 244 subjects whose MRI was suitable for analysis, of which 238 subjects had a valid motor-subtype label (TD/PIGD/Indeterminate), were included in the final imaging experiment.

The final, labeled MRI cohort was significantly unbalanced, and comprised:

i. 172 TD subjects (72.27
ii. 40 PIGD subjects (16.81
iii. 26 Indeterminate subjects (10.92

To ensure that image-slices do not appear in multiple data partitions, patient-level splitting was done prior to image-slice extraction. The final partition had 166 training, 36 validation and 36 held out test subjects.

There were also one non-volume record and one exact duplicate slice found at the DICOM level. These records were not removed from the valid anatomical volumes, but were excluded. After final quality control, there were 47,313 good DICOM slices per the 238 labeled subjects.

From each individual's MRI volume, 10 slices were then chosen that were representative. Slice selection was performed using the middle 50% of the spatially ordered volume, which represents the range of 25th to 75th percentile of slice positions. This approach minimized the effects of extreme peripheral slices and preserved representative anatomical information.

The procedure yielded 2,380 MRI images, which is 10 images per person for 238 people. The resulting distribution of images was:

i. A total of 1,660 photos of 166 people were used for training.
ii. The 360 validation images are from 36 subjects.
iii. A total of 360 test images were taken of 36 subjects.

The selected MRI slices were intensity processed and resized to 224 $\times$ 224 pixels. The grayscale MRI image was then replicated across three channels, so that it could

be passed down the pipeline of the imaging experiment, which used ImageNet-pretrained MobileNetV2 architecture.

TABLE 3.3: Dataset Processing and Final Data Split

Processing Stage	**Retained Data**
Initial candidate collection	290 series / 288 subjects
Valid series after series-level QC	246 series / 244 subjects
Subjects matched with motor-subtype labels	238 subjects
Final clean DICOM slices	47,313 slices
Selected model input images	2,380 images
Training set	166 subjects / 1,660 images
Validation set	36 subjects / 360 images
Test set	36 subjects / 360 images

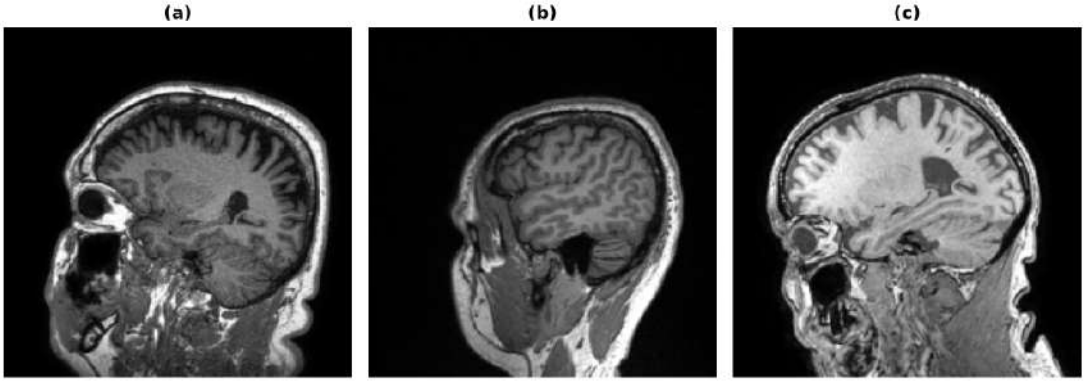


FIGURE 3.5: Representative preprocessed axial slices from a PPMI 3D T1-weighted structural MRI volume used in the imaging experiment.

3.2.3 Spiral Handwriting Dataset

To explore handwriting patterns related to PD, data from handwriting by the participants was added as an independent image-based modality to the data. There were images from HandPD/NewHandPD in the available handwriting collection. In the first data set audit 632 readable spiral images were found in the image archives.

The last deep-learning experiment was only performed on the NewHandPD subset. The subset was chosen because it is the image file name that contains subject information that could be identified to ensure a reliable subject level partition and to avoid images from the same subject being split across the training and testing sets. This final set of NewHandPD was made up of 66 subjects and 264 spiral handwriting images. There were 31 people with Parkinson's and 35 healthy controls among these subjects. Four images were created for each subject. The clinical and MRI experiments were for the three-class motor-subtype classification, while the spiral handwriting experiment was for separate binary classification of Parkinson and healthy control. The probabilities based on the spiral model are thus not fused with the clinical and MRI motor-subtype probabilities in the late fusion phase.

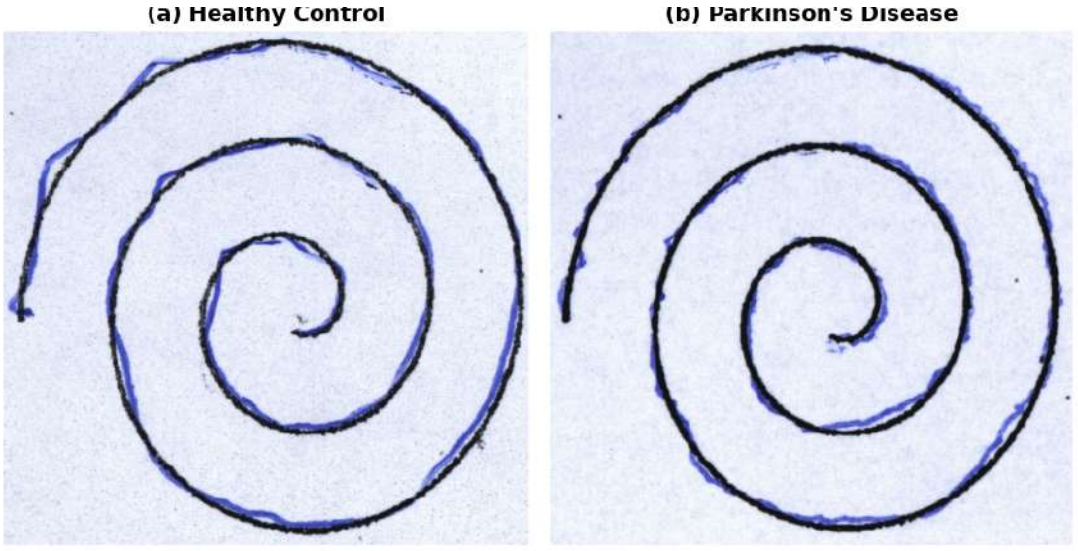


FIGURE 3.6: Representative spiral handwriting samples from healthy-control and Parkinson's disease subjects in the NewHandPD dataset.

Figure 3.6 presents representative spiral samples from the two classes included in the handwriting experiment. The images were used as inputs to the deep-learning model after image preprocessing and were not manually classified on the basis of their visual appearance. The hand-writing data set was split prior to model training by subject identity to avoid leakage of data across the subjects. The 46 images comprised 24 healthy control and 22 Parkinson's disease cases for the training partition. For validation, there were 10 images (5 healthy controls and 5

Parkinson's disease subjects) in the validation partition. Additionally, 10 subjects (4 PD and 6 healthy controls) were available in a held-out test partition.

TABLE 3.4: Subject-level distribution of the NewHandPD spiral handwriting dataset

Partition	**Healthy Subjects**	**PD Subjects**	**Total Subjects**	**Images**
Training	24	22	46	184
Validation	5	5	10	40
Test	6	4	10	40
Total	**35**	**31**	**66**	**264**

3.2.4 Dataset Selection for Multimodal Fusion

The samples are representative spirals from two classes that participated in the handwriting experiment. Pre-processed, but not manually classified images were fed into the deep-learning model. Although several modalities were explored in this study, the fusion of multiple modalities required the availability of both modalities for the same subjects and a common prediction target. Thus, it was not possible to directly integrate all the modality-specific datasets.

Clinical and structural MRI were shared to yield a common three-class motor-subtype target and sufficient overlap of patients for a controlled clinical-structural fusion experiment. Thus, clinical and structural MRI data were selected for the final multimodal integration. The locked MRI test cohort was 36 subjects and clinical information was available for all 36 subjects. This led to the decision level combining clinical class and MRI class probabilities.

This fusion did not include the longitudinal branch, since the target in the future progression categories (Slow, Moderate and Fast) and not the motor subtypes. The same was true for the branch for spiral handwriting, which was focused on the classification of Parkinson's disease versus healthy controls and not directly

comparable to the motor-subtype prediction task.

Gait data was first looked at in the data exploration and only a few subjects had overlap with the MRI group. Thus, gait was not considered in the final multimodal fusion experiment.

3.3 Data Preprocessing

Data preprocessing was carried out individually for each modality, since the clinical, longitudinal, MRI and handwriting spiral datasets were significantly different in structure, dimensionality, temporal organization and data quality. The preprocessing pipeline was designed to process the raw data into model ready representations while keeping as much information as possible out of the training, validation and test sets.

In the case of tabular clinical data, patient identification, integration of relevant clinical assessments, handling of missing and invalid values, preparation of motor-subtype labels, removal of label-generating variables from the predictor set and preparation of patient-level partitions for training and test sets were done as part of the preprocessing. For longitudinal data, fixed-length sequences of patient visits were created by ordering visits longitudinally and separating them in time from progression endpoints. The preprocessing steps for MRI included quality control of the DICOM file, volume selection, ordering the slices spatially, extracting representative slices, normalizing the intensity, resizing and partitioning at the patient level. Preprocessing of spiral handwriting involved subject identification, image validation, resizing, normalization, splitting handwritings of the same subject and training only image augmentation.

When multiple observations were associated with the same individual, all data partitions were made at the patient or subject level, not the individual record or image level. This process was done to avoid information from the same topic being present in both the training and evaluation partitions.

3.3.1 Clinical Data Preprocessing

The clinical preprocessing pipeline was developed to create a reliable patient-level clinical dataset to classify patients according to motor subtype. Demographic data, age at visit, MDS-UPDRS assessments, functional assessments and Neuro-QoL measures were gathered from multiple PPMI files. The source files included variables with different functions in the study, repeated observations, and missing data, for which several data-preprocessing steps were taken prior to machine-learning analysis.

3.3.1.1 Patient Identification and Baseline Record Selection

All clinical files were cross-referenced using a consistent patient identifier to link those files of the same PPMI participant. Clinical observations were clustered at the patient level; and baseline information was chosen for the motor-subtype classification experiment.

Some of these subjects had paired motor examinations with different medication states during baseline MDS-UPDRS inspection. When OFF/ON motor assessments were available, the OFF-medication assessment was preferred to provide a consistent baseline motor assessment. This procedure allowed for the removal of duplicate baseline observations, but maintained a single clinically interpretable motor examination for the purpose of generating subtypes.

After clinical data integration and checking for subtypes availability, 4806 subjects were retained for the final motor-subtype data set.

3.3.1.2 Motor Subtype Label Generation

Motor subtype labels were created based on the items of MDS-UPDRS Part II and Part III that represent tremor symptoms and postural instability/gait difficulty. The subtype-generation procedure was based on the principle outlined in the TD/PIGD ratio.

The tremor item in Part II was added to appropriate tremor examination items in

Part III, to create the tremor component. The PIGD component was composed of Part II walking and freezing items and Part III gait, freezing-of-gait and postural-stability items.

Subjects with a TD/PIGD ratio ≥ 1.15 were classified as Tremor-Dominant (TD), whereas those with a ratio ≤ 0.90 were classified as Postural Instability and Gait Difficulty (PIGD). Subjects with ratios between 0.90 and 1.15 were classified as Indeterminate. Zero-component scores were also treated appropriately, to prevent undefined ratios.

$$R = \frac{TD_{\text{score}}}{PIGD_{\text{score}}} \tag{3.1}$$

$$\text{Subtype} = \begin{cases} TD, & R \geq 1.15 \\ PIGD, & R \leq 0.90 \\ Indeterminate/Mixed, & 0.90 < R < 1.15 \end{cases}$$

where TD_{score} is the tremor component from the tremor items of the selected MDS-UPDRS Part II and Part III and $PIGD_{\text{score}}$ is the postural instability and gait difficulty component. In the case of either one or both of the component scores being zero, special handling was used to prevent undefined ratios.The final distribution was 2,055 TD, 1,039 PIGD and 1,712 Indeterminate/Mixed.

3.3.1.3 Prevention of Target Leakage and Feature Selection

Some of the main considerations in preprocessing were target leakage and avoiding it. The labeling rule for the motor subtypes was derived from certain variables in Part II and Part III of the MDS-UPDRS; if these same variables are included as predictors, the classifier may end up learning the labeling rule instead of independent clinical patterns associated with each motor subtype.

Hence, variables that are directly used in the TD/PIGD label-generation process were omitted from the final predictor set. After this leakage control step, 30

demographic, clinical, functional and Neuro-QoL variables were kept.

Six demographic variables, seven MDS-UPDRS Part I variables, one functional-assessment variable, eight Neuro-QoL mobility variables and eight Neuro-QoL upper-extremity variables were included in the final predictor set. The above mentioned variables were then fed as an input to the clinical machine learning models.

3.3.1.4 Missing and Invalid Value Handling

Not all the clinical data was available for all participants, resulting in missing observations. Handling missing data was therefore carried out at the end of the creation of the final predictor set.

Where appropriate, variables with too many missing values were dropped from the final model. Moreover, the final clinical predictor set did not include the MDS-UPDRS part IV due to high missing data coverage. Missing values for the predictor variables were filled in with the median.

Most importantly, the parameters for the preprocessing were only estimated for the training partition. The median data from training subjects was then used for the test data. This would eliminate the possibility of information from the held-out test set feeding into the data pre-processing of the training set.

For the later experiment CTGAN, special and/or invalid values within each of the items on MDS-UPDRS Part I were also compared to the range of valid values. It was expected that individual Part I items would score between 0 and 4. Anything outside this range was considered invalid/missing as a clinical category, not a real clinical category.

This procedure found 1 invalid NP1DPRS observation and 16 invalid NP1DDS observations in the CTGAN training partition.The matching values were replaced with "missing" values and the missing values were filled in with the medians of the training data. Seven invalid NP1DDS observations were treated according to this rule in the held out test partition.

3.3.1.5 Clinical Train-test Partitioning

To maintain the relative distribution of the three motor-subtype classes the final dataset of 4,806 subjects was divided using stratified patient-level sampling. There were 3,844 subjects in the training partition, and a held-out real test partition with 962 subjects. There were 1,644 TD, 831 PIGD and 1,369 Indeterminate subjects in the training partition. The subjects held out for the test were 411 TD, 208 PIGD and 343 Indeterminate.

The patient-level split was done prior to the train of CTGAN. As a result, the test part of the holdout dataset was not given to CTGAN nor used in the generation of synthetic samples. This way, the final judgment on generative augmentation was conducted on real unseen subjects only. This splitting strategy ensured that the original class proportions were approximately preserved while maintaining complete separation between training and test subjects.

3.3.1.6 Feature Scaling

Feature scaling was used if necessary given the machine-learning algorithm being used. If the input features were continuous, they were normalized by using parameters computed from the training set only for Logistic Regression and other models that are sensitive to scale. A fitted transformation was then applied to the data set not used earlier, the test data. For tree-based algorithms like Random Forest, feature standardization is not a requirement for decision-tree based operations and so there was no need to scale.

If Min-Max normalization was applied to all the clinical models in the old thesis then do not blindly do the same. In the final verified CTGAN comparison, StandardScaler have been used for Logistic Regression.StandardScaler use has been done in the end clinical Logistic Regression pipeline, so it would be better to use Z-score standardization instead of Min-Max:

$$x' = \frac{x - \mu_{\text{train}}}{\sigma_{\text{train}}} \tag{3.2}$$

Here x is the value of the feature (x-value) and μ_{train} is the mean of the feature value (x-value) calculated from the training data, and σ_{train} is the standard deviation of the feature value (x-value) calculated from the training data. The scaling parameters were determined only from the training partition, and then used to scale the validation or test partition to avoid information leakage.

3.3.2 CTGAN Data Pre-processing and Minority-Class Augmentation

The clinical motor-subtype dataset was imbalanced, with a noted imbalance in the PIGD class. In order to explore the possibility of using generative augmentation to boost the prediction of minority classes, a Conditional Tabular Generative Adversarial Network (CTGAN) was added to the clinical branch.

The CTGAN experiment was only conducted after the clinical data was split into real training and held-out test partitions. CTGAN was only trained from the 3,844 real training subjects. To learn clinical feature distributions, information about patient identity was removed from the generative input.

3.3.2.1 CTGAN Input Preparation

The same 30 leakage-aware clinical predictor variables were used as input variables in the CTGAN, with the motor-subtype variable as the target variable. If the clinical response was categorical, binary or ordinal, these were given as Discrete variables for CTGAN. All of the aggregate/continuous variables (i.e., age at visit, MDS-UPDRS Part I total score, Modified Schwab and England Activities of Daily Living score) were treated as continuous variables.The CTGAN was trained for 100 epochs with a mini batch size of 500 and a training partition of the real data set.

3.3.2.2 Synthetic Minority-class Generation

The starting original real training partition was unbalanced:

i. TD = 1,644
ii. PIGD = 831
iii. Indeterminate = 1,369

TD was therefore chosen to be the reference majority class. No synthetic TD records were produced. CTGAN generated the following to equalize the minority classes to the majority-class size:

i. 813 synthetic PIGD records
ii. 275 synthetic Indeterminate records

In all, 1,088 synthetic minority-class records were thus created. This resulted in a larger training set comprising:

i. TD = 1,644
ii. PIGD = 1,644
iii. Indeterminate = 1,644

giving a total of 4,932 training records. Only synthetic (simulated) records were used for training augmentation and not as independent real patients.

Table 3.5: Distribution of Real and Synthetic Training Samples Across Motor Subtypes

Subtype	**Real Training**	**Synthetic Added**	**Final Augmented**
TD	1,644	0	1,644
PIGD	831	813	1,644
Indeterminate	1,369	275	1,644
Total	**3,844**	**1,088**	**4,932**

3.3.2.3 Synthetic Data Validity Control

The generated samples were then used for downstream classification with synthetic data quality check. There were no missing values in the final synthetic data set.

The synthetic outputs of the continuous variables were limited to the range of the

existing training data. For instance, age at visit was limited to the range of age in which the training was observed (20.8–91.4 years), MDS-UPDRS total score part I was limited to 0–17 and Modified Schwab and England Activities of daily living score to 50–100.

Values generated for the discrete clinical variables were limited to values that existed in the real training data. After the validity-control, the final synthetic dataset had no values that did not fall in the range of the real training features.

This validation was to ensure that obviously erroneous synthetic records were not used in the model training at the downstream location. But the range validity was not equated with a true patient and a complete clinical realism of synthetic records.

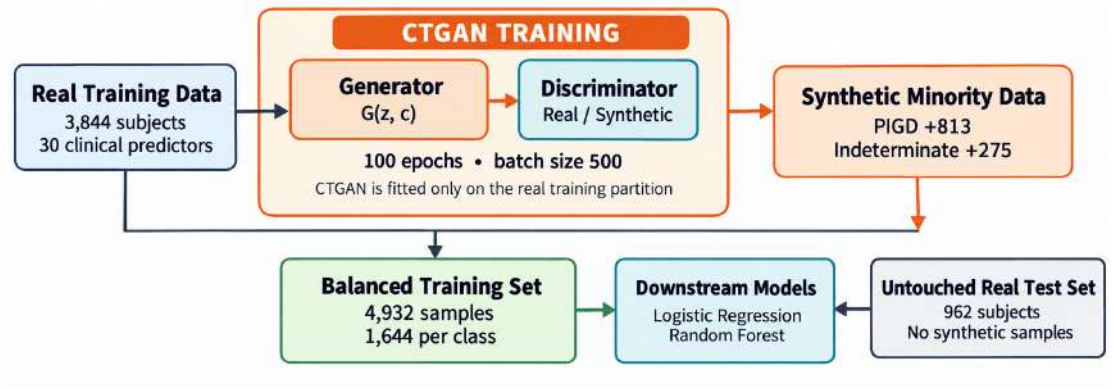


FIGURE 3.7: CTGAN-based minority-class augmentation and evaluation pipeline.

3.3.3 Longitudinal Data Preprocessing

Baseline motor-subtype preprocessing was not performed on this branch of analysis, as this was to predict future progression categories from repeated motor exams.

MDS-UPDRs Part III were observed for each subject repeated a number of times and then arranged in chronological order based on the time of the visit. Duplicate and paired examination records were clarified to provide consistent examination data along the examined time axis. Very few repeated observations or short follow-up period were observed among subjects; subjects with too few repeated observations or too short a follow-up period were excluded from sequence modeling.

3.3.3.1 Temporal Separation of Input and Target

A crucial part of the longitudinal preprocessing was to avoid the introduction of future information to the model input. So, temporal patient observations were split into an input phase and a future target phase.

Clinical observations during the first two years were used for sequence construction and observations after the early period used independently for future estimation of disease progression.

The final inclusion criteria were:

i. A minimum of three early observations;
ii. There are at least two future observations; and
iii. At least a year of follow-up in the future period.

Following these criteria 1122 subjects were retained.

This is achieved by the design which ensured that the same future observations that were used to build the progression target did not also serve as input to the temporal model.

3.3.3.2 Future Progression Label Generation

Labels for progression were produced using the longitudinal change in the MDS-UPDRS part III total motor score. For every patient who met the criteria, the motor score values obtained at the future observations were used to estimate the rate of motor-score change over time. General linear progression model:

$$y_i = \beta_0 + \beta_1 t_i + \epsilon_i \tag{3.3}$$

Here, y_i represents the total score of the MDS-UPDRS Part III at visit i, t_i represents the time of visit i, and β_1 represents the estimated rate of change in motor severity over time. The slope β_1 is calculated as:

$$\beta_1 = \frac{\sum_{i=1}^{n}(t_i - \bar{t})(y_i - \bar{y})}{\sum_{i=1}^{n}(t_i - \bar{t})^2} \tag{3.4}$$

where $\bar{t}$ is the mean visit time, $\bar{y}$ is the mean MDS-UPDRS Part III score, and n is the total number of visits. A positive value of β_1 indicates an increasing MDS-UPDRS Part III score over time, whereas a negative value indicates a decreasing score over time. The two cut-off values are based on the study-specific tertiles of the eligible future-progression cohort.

Thus, Slow, Moderate and Fast are not generally defined clinical progression classes, but rather categories of progression for operations within the current study. The resulting future progression rates were then divided into three categories by using tertile thresholds defined by the study:

i. Slow Progressor
ii. Moderate Progressor
iii. Fast Progressor

The final distribution comprised 370 Slow progressors, 382 Moderate progressors and 370 Fast progressors. These subtypes are operational progression groups created for this study, and do not reflect clinical subtypes of PD.

3.3.3.3 Sequence Construction

There were 34 input features per visit: 33 MDS-UPDRS Part III motor examination features, combined with temporal information, for each eligible subject.

The patient observations were sorted chronologically and then converted into sequences up to 8 time steps. This representation allowed recurrent neural networks to learn about changes between repeated visits to the patient, instead of viewing each observation as a separate record. The final sequence representation thus was structured as maximum input: 8 time steps x 34 features per subject. Patients

were then classified at a patient level into:

i. 785 training subjects
ii. 168 validation subjects
iii. 169 test subjects

There were no patients that crossed partitions.

3.3.3.4 Longitudinal Feature Scaling

Scaling parameters to be used for the normalization of the continuous longitudinal features were only derived from the training partition. This transformation was then applied to the same manner to the validation and test sequences.

This method allowed the information from validation or test subjects not to affect the parameters of the preprocessing during the model training.

3.3.4 History-based Future-visit Motor Subtype Preprocessing

An extra longitudinal study was carried out to anticipate the next time the patient is seen, the clinical motor subtype is recorded. This experiment was different to the Slow, Moderate and Fast progression prediction task. The classes to which it was directed were Tremor Dominant (TD), Postural Instability and Gait Difficulty (PIGD) and Indeterminate.

The closest clinically labeled visit that occurred before the input visit was selected for each input visit. The historical visit was chosen from a visit of the same patient. Historical visits were kept only for pairs with available historical visits. The input visit was preceded by the historical visit and the target visit was preceded by the input visit. This happened only after the input visit. The 18 baseline input-visit features and eight additional features were used in the model historical and current motor scores are represented by features subtype, tremor and PIGD score changes, number of days. The mean difference in the historical and input visits. Future-visit

motor scores not included in model predictors were .

Patients were separated between the training and held-out test partitions. The development subset that was history available was: 2,000 visit pairs from 432 patients were held out for future testing. The evaluation subset included 4241 pairs of visits from 670 patients.

3.3.5 Structural MRI Preprocessing

To ensure patient-level separation while at the same time transforming the raw PPMI DICOM data into two-dimensional images suitable for transfer learning, the preprocessing of MRI data was designed.

In the first set of MRIs, there were 290 candidate 3D T1-weighted series, gathered from 288 subjects. The first audit was carried out at the series level and subject level to determine whether the acquisitions are incomplete, duplicate and if they are not volume files.

3.3.5.1 DICOM Quality Control and Volume Selection

All the individual MRI series were reviewed for number of DICOM slices, image dimensions, patient identifier, series identifier and study information. Anatomical volumes represented in series where there was insufficient information to show it as a complete three dimensional volume were not included.

Of the 290 candidate series, 44 series had one slice only, and were not included. After this quality control process, 246 valid series comprising 244 subjects were left.

For two subjects there were more than one series available. One final series was chosen in each case to ensure one MRI volume per patient.After matching by motor-subtype, 238 subjects were included in the final MRI experiment.

Further examination at the DICOM level identified one non-volume record and one slice that was an exact duplicate. These individual records were not included

but their associated valid MRI volumes were retained. Thus, the final dataset comprised of 47,313 clean DICOM slices.

3.3.5.2 Patient-level MRI Partitioning

Pre-processing step segmentation was done at the level of patients before extracting slices. This is very important as it is highly correlated between multiple MRI slices from the same anatomical volume. If individual slices were randomly divided, then slices from the same patient might be in both the training set and the test set and lead to over-optimistic performance estimates.

The 238 subjects were split into:

i. 166 training subjects
ii. 36 validation subjects
iii. 36 held-out test subjects

The following subtype distributions were found:

a. Training
    i. TD = 120
    ii. PIGD = 28
    iii. Indeterminate = 18
b. Validation
    i. TD = 26
    ii. PIGD = 6
    iii. Indeterminate = 4
c. Test
    i. TD = 26
    ii. PIGD = 6
    iii. Indeterminate = 4

There was no patients overlap between the 3 partitions.

3.3.5.3 MRI Slice Selection

Ten representative slices for each subject were selected as a model input instead of using each DICOM slice as a separate model input.

The MRI volume was first identified in a spatially ordered fashion with representative slices being sampled approximately evenly across the central 50% of the MRI volume (25th–75th percentile of slice positions).

This method eliminated the peripheral slices, which would have included some less brain-like anatomy, and gave a consistent number of images for each patient for analysis. 2,380 MRI images were selected from ten slices of each of the 238 subjects:

i. Training = 1,660 images
ii. Validation = 360 images
iii. Test = 360 images

3.3.5.4 MRI Intensity Processing and Resizing

If available, the available rescale slope and intercept information was used to convert raw DICOM pixel intensities. The intensity processing was then performed robustly by working with about 1st and 99th percentile to minimize the impact of extreme intensity values. Processed intensities were scaled to the range of 0-255.

$$I_{\text{clip}} = \min\left(\max\left(I, P_1\right), P_{99}\right) \tag{3.5}$$

$$I_{\text{norm}} = 255\frac{I_{\text{clip}} - P_1}{P_{99} - P_1} \tag{3.6}$$

where P_1 and P_{99} represent the 1st and 99th percentile intensity values of the MRI slice, respectively. The intensity values below P_1 are clipped to P_1, while values above P_{99} are clipped to P_{99}. The clipped intensities are then linearly rescaled to the range $[0, 255]$ to obtain the normalized MRI image.

The slice was resized to 224×224 pixels, then duplicated across 3 channels to give an input size of 224×224×3. Also, it was thought about to consider DICOM photometric interpretation; the inversion of the retained final images would be required to get a consistent intensity orientation in MONOCHROME1 images, which would be prepared for the model.

The slices selected were resized to 224 × 224 pixels. Each slice of structural MRI was duplicated over 3 channels to give a final image size of: 224 × 224 × 3.

3.3.5.5 MRI Data Augmentation

Only the training images were used for data augmentation to train the model. All the images were transformed with mild transformations to raise the variability of the images while keeping the anatomy structure more or less unchanged.

The following augmentation pipeline was used:

i. Random rotation: approximately 0.03
ii. Random zoom: approximately 0.05
iii. Random translation: approximately 0.03

The image flipping was not used since the MRI experiment avoided arbitrary anatomical flipping. No augmentation of validation and test images were performed. Remove noise and improve the image quality.

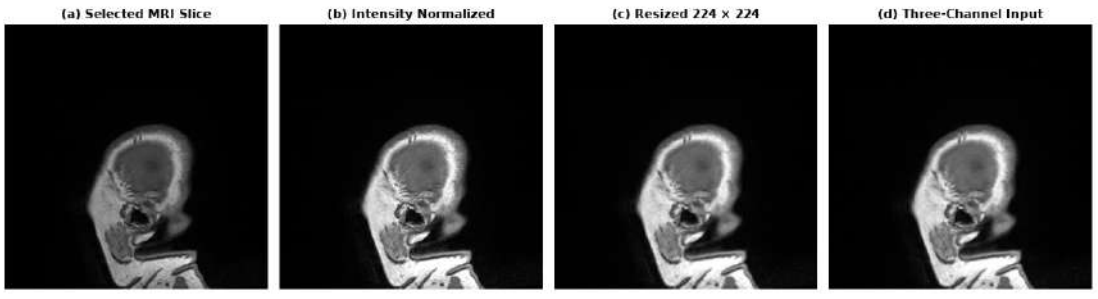


FIGURE 3.8: Representative stages of structural MRI preprocessing prior to MobileNetV2 input.

3.3.6 Spiral Handwriting Image Preprocessing

The spiral handwriting dataset was a one that was not neuroimaging data, but two-dimensional drawing sequences, which necessitated a different image-processing pipeline.

The available handwriting archives were first screened for image readability, class information and subject identity. Of the 632 candidate spiral images, all of them were readable. The final deep-learning experiment was limited to the NewHandPD subset, however, as it could be used to obtain reliable subject identifiers, subject-level partitioning was possible.

The last set of NewHandPD totalled 264 images from 66 subjects (35 healthy controls and 31 PD subjects). Four images of spirals were contributed for each subject.

3.3.6.1 Subject-level Partitioning

The handwriting data set was split by subject identities rather than by randomly splitting images. This avoided the same subject to appear in both training and evaluation drawings in multiple different forms.

The final split was comprised of:

i. Training: 46 subjects / 184 images
ii. Validation: 10 subjects / 40 images
iii. Test: 10 subjects / 40 images

Twenty-four healthy and 22 Parkinson's patients were included in the training cohort. There were 5 subjects in each class of the validation cohort and 6 healthy and 4 PD subjects in the test cohort.There was no overlap in the subject area between the partitions.

3.3.6.2 Image Resizing and Input Preparation

Individual handwriting images were prepared for inputting into the ImageNet-pretrained MobileNetV2 architecture for each spiral handwriting image. Images were resized to 224 × 224 pixels and converted to the channel structure that was used by the pretrained network.

All the images were processed at the pixel level and the operation was repeated for every image in the training, validation and test sets. Any stochastic image augmentation was restricted to the training partition.

3.3.6.3 Training-only Image Augmentation

During training, mild image augmentation was performed to introduce controlled variability, and to reduce overfitting due to the small handwriting training set. The augments that were performed are:

i. Random rotation: 0.05
ii. Random zoom: 0.08
iii. Random translation: 0.05
iv. Random contrast: 0.10

NO image flipping was performed. The images for validation and held-out test were not augmented, so that the performance was measured against the original image distribution.

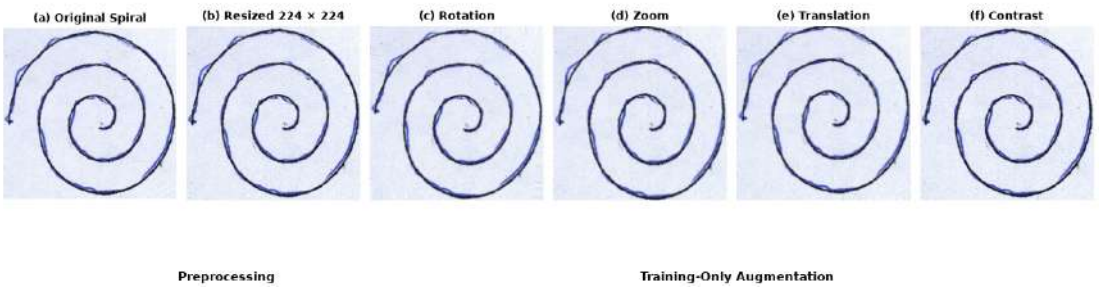


FIGURE 3.9: Spiral handwriting preprocessing and training-only augmentation procedure.

3.4 Model Development and Architecture

The preprocessing was performed in a modality-specific manner before the development of separate machine-learning and deep-learning models for each dataset, depending on the structure and prediction goal. Clinical tabular data was analyzed using Logistic Regression (LR), Support Vector Machine (SVM) and Random Forest (RF). Conditional Tabular Generative Adversarial Network (CTGAN) based synthetic augmentation was also used to explore the class imbalance within the clinical motor-subtype dataset. A Long Short-Term Memory (LSTM) network and Bidirectional Long Short-Term Memory (BiLSTM) network were used to model the longitudinal disease process. The structural MRI and spiral handwriting images were analyzed using transfer learning with MobileNetV2.

The clinical and image-based ones were integrated with Explicable Artificial Intelligence using SHapley Additive exPlanations (SHAP) and Gradient-weighted Class Activation Mapping (Grad-CAM), respectively. Lastly, the clinical and structural MRI predictions were fused at the decision level to predict the common held-out patient cohort. The entire model development strategy involved estimating model fitting parameters and pre-processing parameters from the training data set where applicable, parameters were also estimated from the validation data set, while final evaluation was performed on held-out test data.

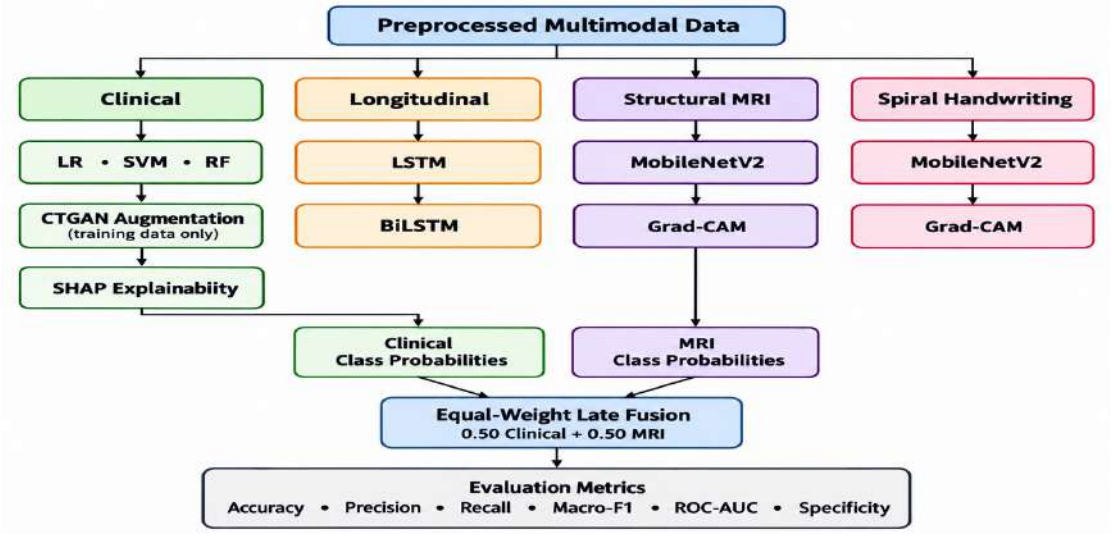


FIGURE 3.10: Overall model-development, explainability and multimodal-fusion strategy used in the proposed framework.

3.4.1 Clinical Machine-learning Models

The clinical motor-subtype classification problem was defined as a three-class supervised-learning problem where the target classes were TD, PIGD and Indeterminate/Mixed. The leakage-aware clinical representation had 30 predictor variables. To set comparative clinical baselines, the three classical machine-learning algorithms—Logistic Regression, Support Vector Machine and Random Forest—were assessed.

The clinical data set was initially split into a development set (80%) and a locked independent test set (20%), using stratified sampling. A total of 3,844 subjects were used for the development set and 962 subjects were kept as a locked test set. Five-fold stratified cross validation was used to assess the performance of Logistic Regression, Support Vector Machine and Random Forest models on the development set. The hyperparameter optimization was conducted only on the development set and the macro-F1 score was the main model-selection criterion. The model that was finally chosen was then tested on the locked test set, which was not used in the model selection and hyper parameter tuning.

Random Forest had the best macro-F1 score during development cross-validation (0.5114) and was chosen as the final clinical model. The learners selected was an ensemble of 300 decision trees with a maximum depth of 12, three samples per leaf and square-root feature sampling, as well as class weighting using balanced subsampling.

3.4.1.1 Logistic Regression

Multinomial Logistic Regression was tested as a probabilistic linear classifier for predicting clinical motor-subtypes. The softmax formulation is used to represent the probability of assigning an observation x to class k, for an input feature vector x.

$$P(y = k \mid x) = \frac{\exp\left(\beta_k^T x + b_k\right)}{\sum_{j=1}^{K} \exp\left(\beta_j^T x + b_j\right)} \tag{3.7}$$

Here, $K = 3$ is the number of motor-subtype classes, β_k is the coefficient vector for class k, and b_k is the intercept for class k.The class to be predicted is chosen based on the maximum posterior probability:

$$\hat{y} = \arg\max_k P(y = k \mid x) \tag{3.8}$$

The clinical predictors were standardized using the mean and SD from the training data before Logistic Regression training. In the final CTGAN comparison the model was given a maximum of 3,000 iterations to ensure numerical convergence. Logistic Regression proved very useful in the proposed framework as it provided class-probability estimates which could then be used for decision-level multimodal fusion.

Table 3.6: Logistic Regression configuration

Parameter	**Configuration**
Input features	30
Output classes	3
Classes	TD, PIGD, Indeterminate
Feature scaling	StandardScaler
Maximum iterations	3,000 in final CTGAN comparison
Random state	42 where applicable
Output	Three class probabilities

3.4.1.2 Support Vector Machine

Support Vector Machine (SVM) was used as a second clinical baseline. SVM tries to build a boundary that is farthest from both the classes by the largest margin. If the separating hyperplane is linear and a binary decision function, it can be written as follows:

$$f(x) = w^T x + b \tag{3.9}$$

Here, w is the weight vector, x is the input clinical features and b is the bias term.

The optimization goal of a soft-margin SVM is:

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \tag{3.10}$$

subject to:

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \qquad \xi_i \geq 0 \tag{3.11}$$

Here, C is used to set the balance between maximum margin and misclassification penalties and ξ_i is the slack variable. The multiclass motor-subtype problem is simply an extension of the binary decision making process to the three target classes.

3.4.1.3 Random Forest

Random Forest was assessed as an ensemble learning, nonlinear approach for clinical subtype prediction. The algorithm creates several decision trees using bootstrapped training samples and classifies the output of the different decision trees to classify the final class. The final prediction for N decision trees is:

$$\hat{y} = \text{mode}\left\{T_1(x), T_2(x), \ldots, T_N(x)\right\} \tag{3.12}$$

where $T_i(x)$ is the class predicted by the i-th decision tree. Multiple trees decrease reliance on any single tree and enable nonlinear interactions between clinical variables to be modeled. The best Random Forest model had 100 trees, Gini, bootstrapping and a random state of 42. The Gini impurity used during tree construction is given by:

$$\text{Gini} = 1 - \sum_{k=1}^{K} p_k^2 \tag{3.13}$$

where p_k is the fraction of samples in the node that are of class k.

Table 3.7: Random Forest Configuration

Hyperparameter	**Value**
Number of trees (‘n_estimators‘)	100
Splitting criterion	Gini
Bootstrap	True
Random state	42
Number of classes	3
Input predictors	30

3.4.2 Conditional Tabular Generative Adversarial Network

To explore the issue of class imbalance in the classification of clinical motor subtypes, CTGAN was incorporated as the generative-learning part of the proposed framework. CTGAN is specifically designed to model mixed-type tabular data that comprises continuous, categorical, binary and imbalanced variables. Only the 3,844 real training subjects were used to train the generative model. The locked set of 962 real test subjects was not given to CTGAN for the generation of synthetic data or the training.

The 30 leakage-aware clinical predictors and the clinical subtype were the inputs to the CTGAN. Patient identifiers were not included. Clinical variables were treated as a continuous variable if they were quantitative (age at visit, MDS-UPDRS Part I total score, Modified Schwab and England Activities of Daily Living score), otherwise as a discrete variable.

3.4.2.1 Generator and Discriminator

The idea of the GAN is that it contains two competing parts: the Generator and the Discriminator. The generator takes a random hidden vector z and a conditional vector c to produce a fake tabular observation:

$$\tilde{x} = G(z, c) \tag{3.14}$$

i. The random vector z is the latent random vector,

ii. c is conditional information,

iii. G is the generator,

iv. $\tilde{x}$ is the synthetic record generated

A real or generated record is passed to the discriminator, who learns to discriminate between the various distributions:D(x). The generator learns to create more and more realistic synthetic observations and the adversarial part tries to determine whether the observations are real or synthetic.

When conditioning is used in CTGAN, it is especially beneficial for categorical data with a skewed distribution, as it allows for the creation of synthetic samples for a specific class of interest. In the current study, this process was utilized to create more PIGD and Indeterminate training records instead of adding to the large number of TD classes.

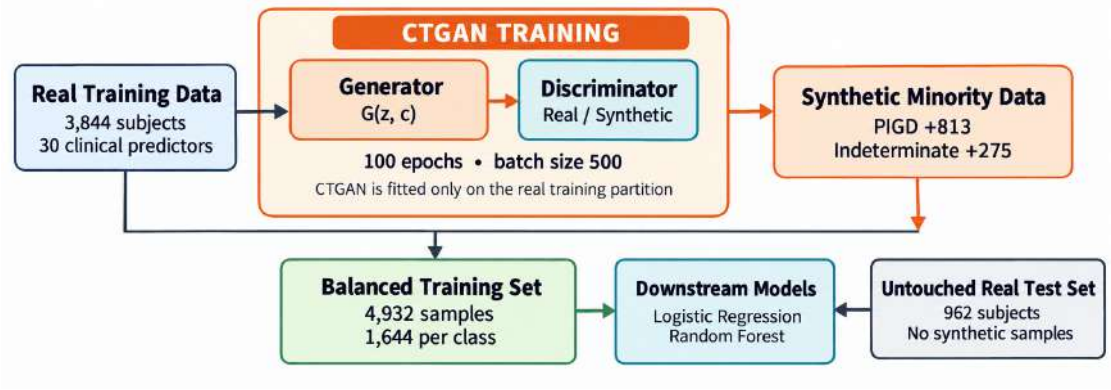


FIGURE 3.11: Conceptual architecture of CTGAN for conditional generation of minority-class clinical records.

3.4.2.2 CTGAN Training Configuration

CTGAN was trained for 100 epochs with a batch size of 500. The actual training data set was of 3,844 observations. Once trained, conditional generation of 813 PIGD and 275 Indeterminate synthetic observations were generated.

Table 3.8: CTGAN Configuration and Synthetic Data Generation

Parameter	**Configuration**
Real training records	3,844
Input clinical predictors	30
Conditional variable	Motor subtype
Epochs	100
Batch size	500
Synthetic TD	0
Synthetic PIGD	813
Synthetic Indeterminate	275
Total synthetic records	1,088
Final augmented records	4,932
Final samples per class	1,644
Test subjects used in GAN training	0

The 1,088 minority class samples generated were added to the original training set, which resulted in an augmented training set of 4,932 records, with 1,644 records from each of the three classes.

3.4.2.3 Downstream CTGAN evaluation design

The impact of the CTGAN augmentation was assessed by comparing the classifiers trained with the original real training data with the classifiers trained with the CTGAN augmented data. For this comparison logistic regression and random forest were used.

Class weights were not used in this particular CTGAN comparison to isolate the effect of synthetic augmentation. Both real-only and CTGAN-augmented models were evaluated on exactly the same 962 real held-out subjects. Synthetic observations were never included in the test set.

The main criterion for comparison was Macro-F1 since the goal of augmentation

with CTGAN was not to only maximize the accuracy of the majority class.

3.4.3 Long Short-term Memory Network

The longitudinal branch was fashioned to capture the temporal relationships between successive patients' visits. Traditional feed-forward models work independently on the observations and do not explicitly remember the information from the previous time steps. Thus, Long Short-Term Memory networks were assessed to model changes in repeated MDS-UPDRS Part III assessments over time. The sequences used for each patient consisted of up to 8 time steps and 34 features per time step. The final Sequence input can then be expressed as:

$$X_i \in \mathbb{R}^{8\times 34} \tag{3.15}$$

for patient i. The first LSTM model had 64 LSTM units, dropout and a three-class output layer.

3.4.3.1 LSTM Internal Architecture

The temporal information is controlled by an LSTM cell through gating mechanisms

Forget gate

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right) \tag{3.16}$$

The forget gate controls the amount of information to be carried over from the previous state of the cell.

Input gate

$$i_t = \sigma\left(W_i[h_{t-1}, x_t] + b_i\right) \tag{3.17}$$

Candidate cell state

$$\tilde{C}_t = \tanh\left(W_C[h_{t-1}, x_t] + b_C\right) \tag{3.18}$$

Updated cell state

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \tag{3.19}$$

Output gate

$$o_t = \sigma\left(W_o[h_{t-1}, x_t] + b_o\right) \tag{3.20}$$

Hidden state

$$h_t = o_t \odot \tanh(C_t) \tag{3.21}$$

In the above equation, x_t is the input at time t, h_{t-1} is the previous hidden state, C_t is the cell state, W and b are learnable weights and biases, σ is the sigmoid activation function, and $\odot$ is the element-wise multiplication operator.

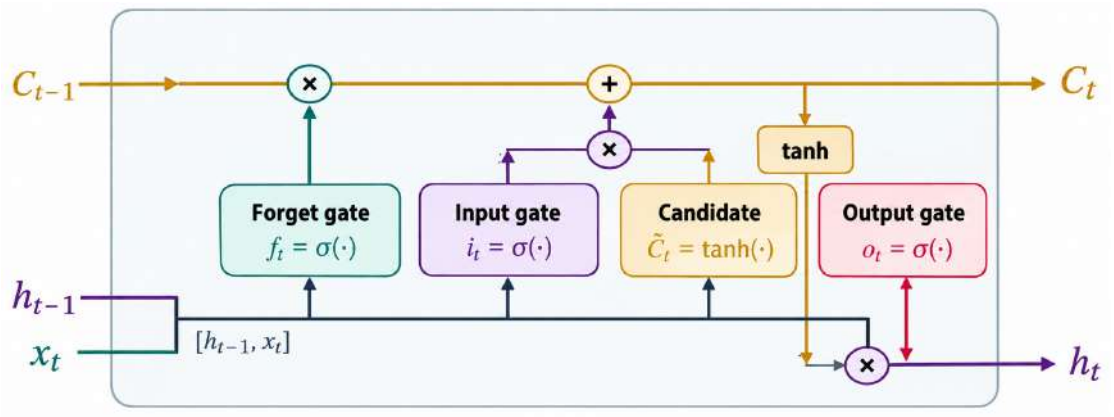


FIGURE 3.12: Internal gating mechanism of the LSTM unit used for longitudinal sequence modeling.

3.4.3.2 LSTM Model Architecture

The temporal network was fed with the sequences up to 8 visits and 34 features per visit. 64 LSTM units were used in the main recurrent layer. To avoid overfitting a 0.30 dropout rate was applied. To the end, the classification layer had three output neurons: Slow, Moderate and Fast progression.

The model was fine-tuned with the Adam optimizer and sparse categorical cross-entropy loss function. The training was carried out with a batch size of 32 up to 20 epochs in the defined time setting. The multi-class Cross entropy objective is:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik}\log\left(\hat{p}_{ik}\right) \tag{3.22}$$

The number of observations is denoted by N, the progression classes by K, the true class indicator by y_{ik} and the predicted probability by $\hat{p}_{ik}$ This loss function measures the difference between the actual progression class and the probability distribution predicted by the LSTM model. During training, the model parameters were iteratively updated by minimizing this loss, allowing the network to learn temporal patterns associated with different progression trajectories across repeated patient visits.

The use of sparse categorical cross-entropy was appropriate for this multi-class classification problem because each patient sequence was assigned to one of the three mutually exclusive progression categories: Slow, Moderate, or Fast. In combination with the Adam optimizer, the objective function supported efficient parameter updates and stable convergence during model training.

Table 3.9: LSTM Model Architecture and Training Configuration

Component	**Configuration**
Input	Up to 8×34
Main recurrent layer	LSTM
LSTM units	64
Dropout	0.30
Output units	3
Output classes	Slow, Moderate, Fast
Optimizer	Adam
Loss	Sparse categorical cross-entropy
Batch size	32
Maximum epochs	20
Training subjects	785
Validation subjects	168
Test subjects	169

3.4.4 Bidirectional LSTM

A Bidirectional LSTM was tried as an alternate temporal architecture. A BiLSTM has both forward and backward recurrent blocks, in contrast to typical LSTMs

that process a sequence in one temporal direction.

The two hidden representations for an input sequence can be represented as:

$$\overrightarrow{h_t} = LSTM_f\left(x_t, \overrightarrow{h}_{t-1}\right) \tag{3.23}$$

and

$$\overleftarrow{h_t} = LSTM_b\left(x_t, \overleftarrow{h}_{t+1}\right) \tag{3.24}$$

The resulting representation is a combination of both directions:

$$h_t = \left[\overrightarrow{h_t}; \overleftarrow{h_t}\right] \tag{3.25}$$

The BiLSTM was tested with the same longitudinal prediction task as LSTM. It was used as an architectural comparison to see if there was added predictive value of a bidirectional sequence representation for the constructed progression sequences.

3.4.5 Single visit Logistic Regression Baseline

For the longitudinal progression task, a single-visit Logistic Regression (LR) baseline was also tested to see if it offered any additional information beyond a simple non-sequential representation. The model did not explicitly model repeated temporal dependencies and thus provided a baseline for the LSTM and BiLSTM models. This comparison was significant as it was important that better performance in the recurrent network be evaluated against a simpler, non-temporal baseline – rather than simply being assumed based on model complexity.

3.4.6 Structural MRI Classification with MobileNetV2

To tackle this issue, a transfer-learning model that utilizes the MobileNetV2 model was developed for the three-class motor-subtype classification of structural MRI. We chose MobileNetV2 because of its efficient architecture of inverted residual blocks and depthwise separable convolutions.

The pre-trained ImageNet weights yielded a general visual feature representation which was then fine-tuned for the MRI domain. Each MRI input consisted of a preprocessed image of:

$$224 \times 224 \times 3$$

Each patient had 10 slices that were representative.

3.4.6.1 MobileNetV2 Feature Extraction

Each channel-wise filter is applied in a conventional convolution manner, which is costly. To lower the computation, MobileNetV2 adopts separable convolution with depthwise separable structure. The depthwise separable operation can be conceptually described as:

$$Y = \text{Conv}_{1\times1}\left(\text{DepthwiseConv}(X)\right) \tag{3.26}$$

where the depthwise convolution applies spatial filtering to each channel of the input, while the pointwise convolution with 1x1 is applied to combine information across channels.

In addition, MobileNetV2 applies inverted residual blocks with linear bottlenecks, enabling feature representations to be expanded, spatially filtered and projected to a lower dimensional representation.

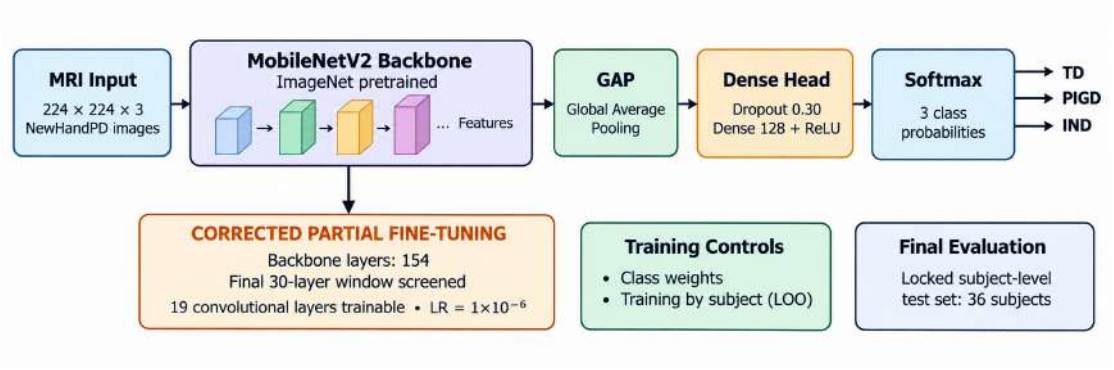


FIGURE 3.13: MobileNetV2-based architecture used for structural MRI motor-subtype classification.

3.4.6.2 MRI Model Architecture

The MobileNetV2 model pre-trained on ImageNet was used without the classification head. The backbone of the MobileNetV2 model implemented in TensorFlow/Keras had 154 layers.

The convolutional feature maps were passed through a Global Average Pooling (GAP) layer followed by a Dropout layer with a rate of 0.30, a fully connected layer with 128 neurons and activation function ReLU, a second Dropout layer with a dropout rate of 0.30 and a Dense layer with 3 neurons and activation function softmax.The softmax operation is given by:

$$P(y = k \mid z) = \frac{e^{z_k}}{\sum_{j=1}^{3} e^{z_j}} \tag{3.27}$$

where z_k is the logit for subtype k. Architecture sequence Input: $224 \times 224 \times 3$

i. MobileNetV2 convolutional backbone
ii. Global Average Pooling
iii. Dropout (0.30)
iv. Dense (128, ReLU)
v. Dropout (0.30)
vi. Dense (3, Softmax)

3.4.6.3 Transfer learning and fine-tuning

The MRI model was first trained with the ImageNet pre-trained MobileNetV2 backbone kept frozen. Initial feature extraction training and controlled partial fine-tuning were then carried out.

In the corrected final fine-tuning procedure, there were 154 layers in the backbone. A final window of 30 backbone layers was taken for fine-tuning with selectivity. The 19 backbone layers were not Batch-Normalization layers, so the layers were

not frozen.

The fine-tuned and corrected model had about:

i. 1,675,075 trainable parameters
ii. 747,264 non-trainable parameters

The learning rate used for fine tuning was:

$$\eta = 1 \times 10^{-5}$$

for a maximum of 15 epochs and early stopping is done on the basis of the validation performance. Validation Macro-F1 was used to select the corrected fine-tuned model, rather than a held-out test set. Once the model selection was completed, the test set was only tested once for final reporting.

3.4.6.4 MRI Class-imbalance Handling

The MRI training cohort was highly skewed towards the TD sub-type. The class weights were then obtained from the distribution of training subjects, and were used in the model training process. The weights used were about:

Table 3.10: Class Weights for Motor Subtype Classification

Class	**Class Weight**
TD	0.4611
PIGD	1.9762
Indeterminate	3.0741

The image augmentation was also performed with training only as outlined in the preprocessing section.The MRI modality was not used with CTGAN since the goal of CTGAN is to augment tabular clinical data and not to generate images.

3.4.6.5 Patient-level MRI Prediction

Individual slices from a single anatomical volume are correlated, although the CNN made predictions for individual slices. Finally, patient-level assessment of MRI was carried out.With $S = 10$ representative slices, the probability of class k was found to be:

$$P_k^{\text{patient}} = \frac{1}{S} \sum_{s=1}^{S} P_{k,s} \tag{3.28}$$

Where $P_{k,s}$ is the predicted probability of class k in slice s.The patient-level predicted class was then:

$$\hat{y} = \arg\max_k P_k^{\text{patient}} \tag{3.29}$$

This was an aggregation that allowed each patient and not each correlated slice to be considered one unit for the primary MRI evaluation.

3.4.7 Spiral Handwriting Classification Using MobileNetV2

The spiral handwriting branch employed an ImageNet-pretrained MobileNetV2 architecture with transfer learning to classify spiral handwriting images as Parkinson's disease or healthy control.

$$Healthy \quad vs. \quad Parkinson's\ Disease$$

The pretrained backbone was fed with resized input images of the size 224x224.Its architecture was made up of: Input $224 \times 224 \times 3$

i. Frozen MobileNetV2 backbone Global Average Pooling
ii. Dropout (0.30)
iii. Dense (128, ReLU)
iv. Dropout (0.30)

v. Dense (1, Sigmoid)

The last activation function was a sigmoid function which was suitable for binary classification:

$$P(PD \mid x) = \sigma(z) = \frac{1}{1 + e^{-z}} \tag{3.30}$$

The binary prediction was done at a threshold of 0.5:

$$\hat{y} = \begin{cases} PD, & P(PD|x) \geq 0.5 \\ Healthy, & P(PD|x) < 0.5 \end{cases}$$

3.4.7.1 Spiral Model Training

The final spiral experiment involved the MobileNetV2 backbone which was frozen. The model was trained using the Adam optimizer, a learning rate of 10^{-3} and binary cross-entropy loss. A quantity of 16 was used.

The backbone of MobileNetV2 was frozen and the task-specific classification layers were trained using the spiral handwriting dataset. This method was especially appropriate due to a limited number of images available for spirals and would help mitigate the number of trainable parameters and the risk of overfitting. Extracted features from the image were then used for binary classification between healthy and Parkinson's disease.

For training only augmentation, mild rotation, zoom, translation and contrast variation were used. Model selection was done using the validation loss and early stopping was applied to prevent the model from overfitting.

The binary cross-entropy is written as:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N} \sum_{i=1}^{N} [y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \tag{3.31}$$

TABLE 3.11: MobileNetV2 Model Architecture and Training Configuration

Component	**Configuration**
Input	$224 \times 224 \times 3$
Backbone	MobileNetV2
Pretraining	ImageNet
Backbone status	Frozen
Global Average Pooling	Yes
Dropout	0.30
Dense hidden units	128
Hidden activation	ReLU
Second dropout	0.30
Output units	1
Output activation	Sigmoid
Optimizer	Adam
Learning rate	1×10^{-3}
Loss	Binary cross-entropy
Batch size	16
Classes	Healthy / PD

3.4.7.2 Subject-level Spiral Prediction

Each NewHandPD subject had provided four spiral images and the main assessment was therefore conducted at the subject level.The probabilities for Parkinson's disease were predicted for each subject in the following way: i, the probability for subject i for the 4 spiral images was averaged.

$$P_i^{\text{subject}} = \frac{1}{4} \sum_{j=1}^{4} P_{ij} \tag{3.32}$$

where P_{ij} indicates the probability of image j of subject i containing Parkinson's disease as predicted by CNN. The averaged probability was then thresholded at 0.5 and converted to a binary prediction. This allowed subjects with multiple correlated images to not be considered four unique subjects for the primary evaluation.

3.4.8 Explainable Artificial Intelligence

The performance of an algorithm in predicting does not give much information about what features or regions of the image the model used to make its prediction. As such, explainability approaches have been included in the proposed framework. Two methods using both modalities were used:

i. SHAP → clinical tabular model
ii. Grad-CAM → image-based models

These were post-hoc analyses and did not change the predictive models.

3.4.8.1 SHapley Additive exPlanations

The clinical model that was selected was then used in SHAP to estimate the contribution from individual clinical predictors to model outputs.

The SHAP value for feature i is given by:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \left[f\left(S \cup \{i\}\right) - f(S) \right] \tag{3.33}$$

where:

N stands for all the features in the input, and S denotes a set of features that does not contain feature i, The model output when given a subset of features S is denoted by $f(S)$, and ϕ_i is the contribution allocated to the feature i.

Positive SHAP values mean a feature was over its baseline value, while negative SHAP values mean a feature was under its baseline value.Positive SHAP values represent a feature which was above its baseline value, and negative SHAP values represent a feature which was below its baseline value.

The SHAP interpretations were not considered as causal clinical relationships.

3.4.8.2 Gradient-weighted Class Activation Mapping

The convolutional image models were used to visualize regions of the image that were important for predictions using Grad-CAM. The importance weight for feature map A^k for a target class c can be learned as a global average of gradients of the class score:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{3.34}$$

The localization map computed by Grad-CAM is then:

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left(\sum_k \alpha_k^c A^k \right) \tag{3.35}$$

where A^k is the convolutional feature map k, y^c is the score of class c and Z is the number of spatial locations.

In the case of the final MobileNetV2 models, Grad-CAMs were computed on top of the final convolutional representation used in the Conv_1 layer.

Representative test cases of both correct and incorrect cases were analyzed for MRI. The same healthy and PD test samples were then displayed for spiral handwriting.

Grad-CAM heat maps were not used to interpret as evidence of causal or disease-specific pathology, they were used as model-attention visualizations.

3.4.9 Decision-Level Multimodal Late Fusion

The multimodal integration experiment for the last integration is a multimodal integration that is made at the decision level that connects the clinical branch and the MRI branch.

Clinical predictors and MRI images have very different dimensional structures so feature-level fusion was not employed. Besides, it was not possible to combine different modalities meaningfully if they did not have the same patients and the same classes of targets. Probability vectors were independently estimated based on

clinical and MRI models:

$$P_{\text{clinical}} = \left[P^C_{\text{TD}}, P^C_{\text{PIGD}}, P^C_{\text{IND}}\right] \tag{3.36}$$

and

$$P_{\text{MRI}} = \left[P^M_{\text{TD}}, P^M_{\text{PIGD}}, P^M_{\text{IND}}\right] \tag{3.37}$$

A late fusion with the equal weight was next performed:

$$P_{\text{fusion}} = 0.5P_{\text{clinical}} + 0.5P_{\text{MRI}} \tag{3.38}$$

More generally:

$$P_{\text{fusion}} = w_C P_{\text{clinical}} + w_M P_{\text{MRI}} \tag{3.39}$$

where:

$$w_C = w_M = 0.5 \tag{3.40}$$

and:

$$w_C + w_M = 1 \tag{3.41}$$

The final classification made was:

$$\hat{y}_{\text{fusion}} = \arg\max_k P_{\text{fusion},k} \tag{3.42}$$

The weights were not optimized on the test set, but were set a priori.

3.4.9.1 Management of leakage in Fusion

36 subjects were excluded from clinical training data of the clinical model used in this experiment after this, the clinical model was retrained to prevent leakage. This created 4770 available clinical training subjects and the 36 common items were reserved for final fusion evaluation. The MRI patient-level probabilities were already calculated as average of predictions over ten different slices per patient. Probability vectors were aligned for each patient by patient identifier and label for true motor subtype was obtained from the MRI.

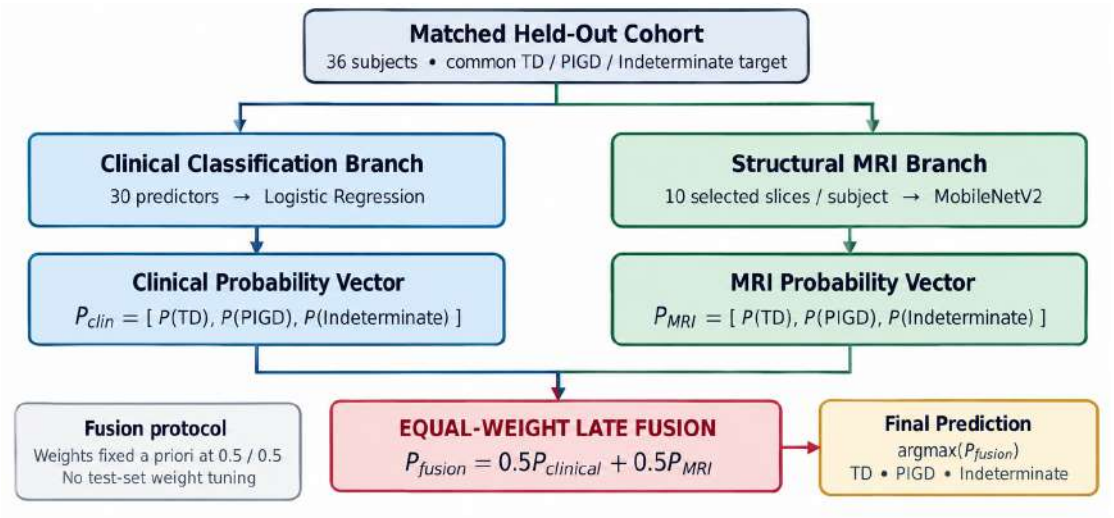


FIGURE 3.14: Decision-level late fusion of clinical and structural MRI class probabilities.

3.4.10 Model Evaluation Metrics

Several experiments were conducted with imbalanced multiclass data and so, the assessment of the models was not done solely on the basis of accuracy. Where applicable, it was decided to adopt the following metrics: Accuracy, Precision, Recall, F1-score, ROC-AUC and confusion matrices.

Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.43}$$

In the case of multiclass classification, accuracy is the percentage of correctly classified observations.

Precision

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.44}$$

Recall

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.45}$$

Recall is especially important for the assessment of minority class sensitivity.

F1-score

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.46}$$

In the three-class experiments, special attention was paid to Macro-F1 as it assigns the same weight to all the classes:

$$\text{Macro-F1} = \frac{1}{K} \sum_{k=1}^{K} F1_k \tag{3.47}$$

Specificity

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{3.48}$$

where a one-vs-rest formulation can be used for every class to calculate the specificity for a multiclass problem.

ROC-AUC

ROC-AUC measures the skill of predicted probabilities to rank positive cases higher than negative cases at each classification threshold. In the case of multiclass tasks, one-vs-rest class-specific AUC values can be aggregated to get Macro ROC-AUC.

The confusion matrix was used to analyze the behavior of the classes in the prediction process. It enabled errors between TD and PIGD to be studied directly and also errors between Indeterminate and Slow/Moderate/Fast.

3.4.11 Experimental Environment

The experiments were successfully conducted with the help of Python programming and popular libraries for machine learning, deep learning, image processing and explainability. The main software environment included:

Table 3.12: Software Tools and Their Applications

Component	**Use**
Python	Main programming environment
TensorFlow / Keras	LSTM, BiLSTM and MobileNetV2
scikit-learn	LR, SVM, RF, metrics and preprocessing
CTGAN	Synthetic tabular-data generation
SHAP	Clinical explainability
OpenCV / Pillow	Image processing
pydicom	MRI DICOM processing
NumPy / Pandas	Numerical and tabular processing
Matplotlib	Visualization

3.4.12 Summary of Model Development

The proposed framework includes deployment of conventional machine learning, generative modeling, temporal deep learning, transfer learning, explainable AI, and decision level multimodal integration. Logistic Regression, SVM and Random Forest were applied for clinical motor-subtype prediction and CTGAN was applied for minority-class synthetic augmentation was investigated. Finally, LSTM and BiLSTM networks were used to model the longitudinal progression, while a single-visit Logistic Regression model was used as a non-temporal reference. The structural MRI and spiral handwriting images were processed by modality-specific MobileNetV2 transfer-learning models. Post-hoc explanations: Tabular predictions were explained by SHAP and image-based predictions by Grad-CAM. Lastly, fixed equal-weight decision-level late fusion between clinical and MRI class probabilities was performed.

Using separate task-specific branches was required due to differences in the conditions of the cohorts and the prediction targets. So only clinical and MRI outputs of the same motor-subtype classes and common subjects were used in the final multimodal fusion experiment.

Chapter 4

Results and Discussion

4.1 Introduction

The experiments were conducted separately for clinical motor-subtype classification, CTGAN-based class-imbalance handling, longitudinal progression prediction, structural MRI classification, spiral handwriting classification and clinical-MRI decision-level late fusion. The results of the single experimental branches are compared separately prior to multimodal integration because the prediction tasks are different and the patient populations are different. The clinical and MRI subtypes are three-class motor subtypes: Tremor-Dominant (TD), Postural Instability and Gait Difficulty (PIGD) and Indeterminate/Mixed subtypes. The longitudinal branch makes predictions for future progression categories (Slow, Moderate and Fast), while the spiral handwriting branch does binary classification between PD and healthy controls.

Confusion matrices, class-wise performance, and accuracy, precision, recall, F1-score andROC - AUC were used to evaluate performance. Macro-averaged measures were used in cases of imbalanced multiclass tasks, where each class gets equal importance regardless of its frequency. Additionally, explainability results of both clinical tabular and image-based models are provided using SHAP and Grad-CAM respectively. Last, the decision level late fusion of the clinical and structural MRI outputs is applied for a common held out patient cohort.

4.2 Clinical Motor-subtype Classification Results

The clinical motor-subtype classification experiment was performed with the leakage-aware clinical dataset which had 4,806 subjects and 30 predictor variables. The target consisted of three motor-subtype classes: TD, PIGD and Indeterminate. Stratified sampling on the patient level was used to partition the data: 3,844 patients for training and 962 patients for independent testing. The three clinical classifiers were compared using five-fold stratified cross-validation performed exclusively on the development set. Hyperparameter optimization was also restricted to the development data and macro-F1 score was used as the primary criterion for final model selection

TABLE 4.1: Five-fold cross-validation performance of the optimized clinical machine-learning models on the development set

Model	**Accuracy**	**Macro-F1**
Logistic Regression	50.42%	49.47%
Support Vector Machine	50.73%	49.08%
Random Forest	52.50%	51.14%

The highest macro-F1 score (51.14%) and accuracy (52.50%) were obtained in the development-set cross-validation experiments by the optimized classifiers evaluated, the Random Forest being the better of them. The Logistic Regression model had an F1 score of 49.47% (macro) and the Support Vector Machine model had an F1 score of 49.08% (macro). Random Forest was chosen as the final clinical motor-subtype classifier based on a model selection criterion for the predefined macro-F1. Finally, the best found Random Forest model was tested on the separate independent test set of 962 people that was not used for hyperparameter optimization and model selection. The model achieved an accuracy of 52.91%, macro-precision of 53.10%, macro-recall of 53.06%, macro-F1 score of 51.38%, weighted F1 score of 51.40% and macro ROC-AUC of 0.6977.

Among the evaluated clinical-only classifiers, Random Forest was selected based

on its reported macro-F1 score of 51.14%. Logistic Regression was also evaluated as a separate clinical classifier and was subsequently used as the clinical branch of the Clinical–MRI decision-level late-fusion experiment.

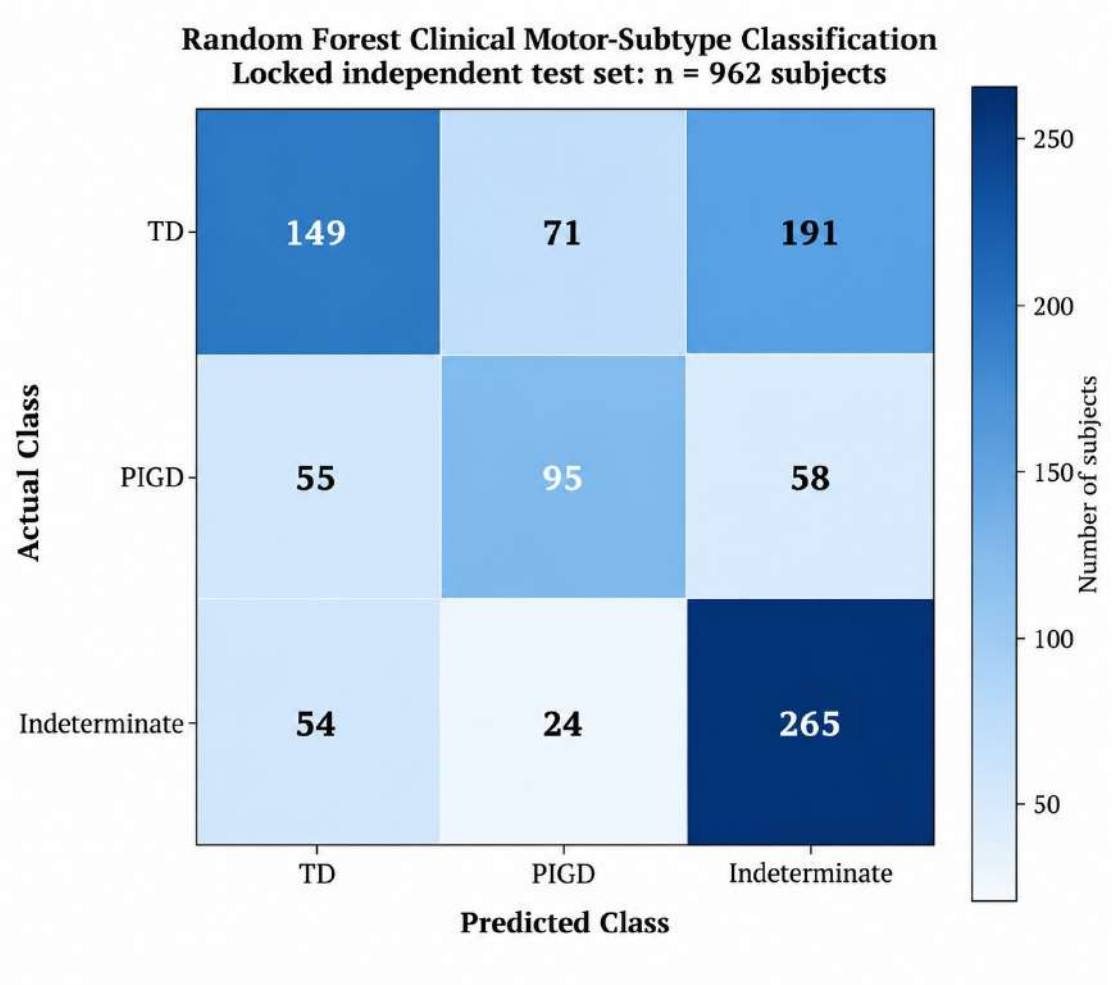


FIGURE 4.1: Confusion matrix of the final Random Forest model for clinical motor-subtype classification on the locked independent test set (n = 962).

4.2.1 Clinical Logistic Regression Performance

A Logistic Regression model was chosen and tested with the held-out test set of 962 subjects. The overall accuracy of the model was 51.46% and the Macro-F1 score was 50.56%.

A Macro ROC-AUC value of 0.687 suggested moderate discriminative information of the clinical variables for distinguishing among the three motor subtypes.

The performance level also illustrates how hard it is to predict the motor subtype without explicitly considering the MDS-UPDRS variables that were used to label motor subtype. The above variables excluded from the table were deliberately omitted to minimize target leakage. Therefore, the model needed to be developed from clinical, demographic, functional and Neuro-QoL features that were independent of the subtypes.

The confusion matrix displays correct predictions of 241 Indeterminate and 153 TD subjects. PIGD continued to be relatively difficult to classify, though 101 cases of PIGD were correctly classified. Many TD subjects were classified as Indeterminate, reflecting some overlap in the clinical feature patterns of the two TD categories.

These results indicate that motor subtype classification remains challenging when the MDS-UPDRS variables directly used for subtype labeling are excluded. The misclassification between TD and Indeterminate subjects suggests an overlap in their clinical characteristics, while the relatively lower classification of PIGD reflects the complexity and heterogeneity of its symptom patterns.

Table 4.2: Final clinical Logistic Regression test performance

Metric	**Value**
Accuracy	51.46%
Macro Precision	51.16%
Macro Recall	52.02%
Macro F1-score	50.56%
Weighted F1-score	50.46%
Macro ROC-AUC	0.687

The results presented in Table 4.2 show moderate performance of the final clinical Logistic Regression model, with an overall accuracy of 51.46%. The Macro Precision, Recall, and F1-score remained close to 50%, indicating that the model experienced difficulty in consistently distinguishing among the three motor subtypes. However, the Macro ROC-AUC of 0.687 suggests that the model retained some ability to discriminate between the classes. Overall, these findings further demonstrate

the challenging nature of motor-subtype prediction when label-defining MDS-UPDRS variables are excluded from the input features. However, excluding label-defining variables reduces the risk of target leakage and provides a more realistic evaluation of the model's ability to identify motor subtypes using independent clinical, demographic, functional and Neuro-QoL features.

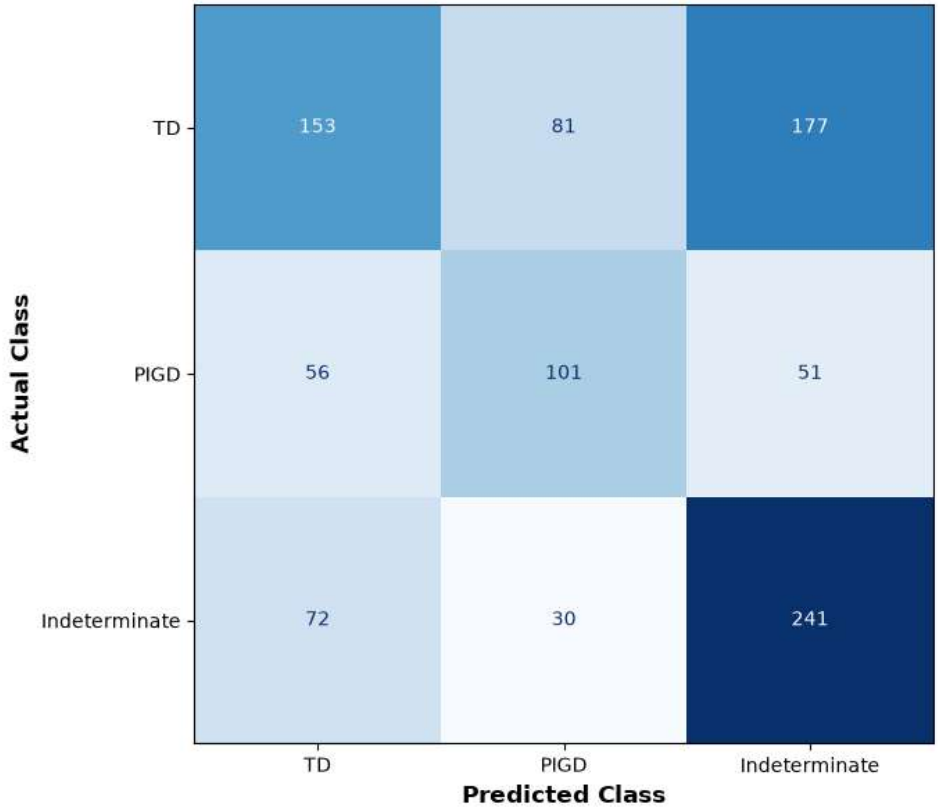


FIGURE 4.2: Confusion matrix of the Logistic Regression model for clinical motor-subtype classification.

The results indicate that non-label-generating clinical variables include valuable information regarding motor subtype but the three subtypes are not fully separable solely based on the selected clinical variables.

4.2.2 SHAP-based Clinical Explainability

To understand the relative importance of each clinical feature in the predictions of the motor-subtype, the final Random Forest clinical model was analysed by SHAP. Global feature importance was measured by averaging the absolute SHAP value over the locked test set.Mean absolute SHAP values revealed that functional

ability, demographics and patient-reported mobility measures led to significant contributions to the Random Forest predictions. The most significant global factors were MSEADLG, SEX, AGE_AT_VISIT, NQMOB30 and NQUEX30. These SHAP values represent the relative importance of the features to the model predictions, and contain no causal information.

TABLE 4.3: Highest-ranked clinical features according to mean absolute SHAP value

Rank	**Feature**	**Mean Absolute SHAP Value**
1	MSEADLG	0.0622
2	SEX	0.0258
3	NQUEX30	0.0188
4	AGE_AT_VISIT	0.0207
5	NP1RTOT	0.0101
6	NQMOB31	0.0120
7	NQMOB30	0.0201
8	NP1ANXS	0.0431
9	NP1COG	0.0076
10	NP1APAT	0.0078

The Modified Schwab and England Activities of Daily Living score (MSEADLG) showed the highest mean absolute SHAP value (0.0622) among the retained predictors, indicating that it had the largest global contribution to the predictions of the final Random Forest clinical model.

Other highly ranked features included SEX (0.0258), AGE_AT_VISIT (0.0207), NQMOB30 (0.0201), and NQUEX30 (0.0188), followed by NQMOB31, NP1RTOT, NQMOB26, NP1APAT and NP1COG.

These results indicate that functional ability, demographic characteristics, mobility measures and MDS-UPDRS Part I information contributed to the model predictions.

The SHAP values should be interpreted as measures of feature contribution to the

trained Random Forest predictions. They do not establish causal relationships between the identified clinical variables and Parkinson's disease motor subtypes.

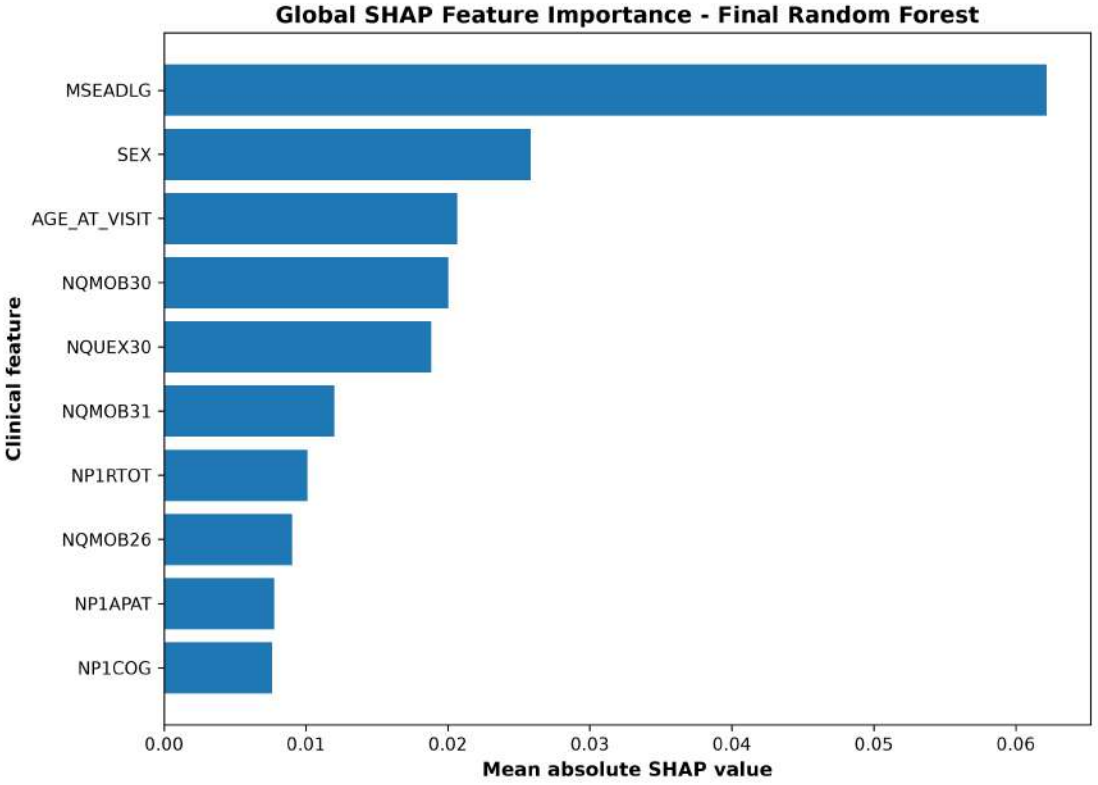


FIGURE 4.3: SHAP-based global feature-importance analysis of the final Random Forest model for clinical motor-subtype classification.

4.3 CTGAN-based Class-imbalance Results

To assess the utility of synthetic minority-class augmentation, CTGAN was assessed.Only the real training partition (3844 subjects) was used for the training of the CTGAN. The 962 real subjects that were held out for testing were not used for the synthetic-data generation. There were 1,369 Indeterminate, 831 PIGD and 1,644 TD subjects in the original training distribution. We produced 1088 synthetic minority class records:

i. 813 PIGD records
ii. 275 Indeterminate records

The extended training set thus comprised 4,932 observations divided into 1,644 in

each of the subtypes. Synthetic augmentation was compared by Logistic Regression and Random Forest model with the same experimental setup.

In this specific ablation experiment, class weights were not utilized thus enabling a separate analysis of the effect of CTGAN augmentation.

Table 4.4 compares the performance of the Logistic Regression model trained on the original clinical data with the model trained after CTGAN-based augmentation. The results show that the addition of synthetic samples did not produce an overall improvement across the major evaluation metrics. Accuracy decreased from 50.83% to 49.06%, while Macro ROC-AUC declined from 0.685 to 0.667. Macro Precision also showed a reduction, whereas Macro Recall and Macro F1-score remained relatively stable. These findings indicate that CTGAN augmentation had a limited effect on the overall classification performance.

4.3.1 Logistic Regression with and without CTGAN

Table 4.4: Logistic Regression performance before and after CTGAN augmentation

Metric	**Real Only**	**Real + CTGAN**
Accuracy	50.83%	49.06%
Macro Precision	54.08%	48.56%
Macro Recall	46.13%	46.05%
Macro F1-score	46.60%	46.50%
Weighted F1-score	49.20%	48.39%
Macro ROC-AUC	0.685	0.667

The overall accuracy, Macro-F1 and the ROC-AUC were unaffected by the CTGAN-augmented Logistic Regression model. The accuracy dropped from 50.83% to 49.06%, with the Macro-F1 score remaining nearly unchanged at around 46.5%. But the augmentation with CTGAN raised the recall of PIGD from around 23% to 29%. This shows that synthetic minority-class generation slightly enhanced the performance of the model for the underrepresented PIGD group even if it did not improve the overall model.

4.3.2 Random Forest with and without CTGAN

TABLE 4.5: Random Forest performance before and after CTGAN augmentation

Metric	**Real Only**	**Real + CTGAN**
Accuracy	48.23%	47.51%
Macro Precision	47.26%	46.50%
Macro Recall	46.71%	47.05%
Macro F1-score	46.94%	46.71%
Weighted F1-score	48.12%	47.55%
Macro ROC-AUC	0.645	0.644

It is also worth noting that this experiment is an independent and controlled CTGAN ablation. To investigate the effect of synthetic-data augmentation, the Real Only and Real + CTGAN Random Forest models were tested with the same settings for the classifiers.

The Real Only performance reported in Table 4.5 should not be compared directly to the independently optimized final Random Forest clinical classifier reported earlier in this chapter.

The results from the Random Forest were similar. The overall accuracy fell slightly (48.23% to 47.51%) and the Macro-F1 remained almost unchanged.For the PIGD class, recall increased from approximately 38% to 43% following CTGAN augmentation.

However, this class-specific improvement did not necessarily translate into an improvement in overall model performance.

4.3.3 Discussion of CTGAN Findings

The CTGAN experiment shows that a numerical class balance is not always related to better classification performance. Although CTGAN successfully balanced the training set and improved minority-class PIGD recall, neither Logistic Regression nor Random Forest showed an improvement in overall accuracy, macro-F1 or

ROC-AUC. This may be because class imbalance was not the only challenge for the clinical classification problem.

Selected leakage-safe clinical features may well be highly overlapping among the underlying motor-subtype classes. This means that sampling data from existing training sets to generate more samples can be used to gain better representation of a minority class without adding truly discriminative information.Importantly, the overall negative result does not mean that the generative model didn't technically work. On the contrary, it demonstrates that simply adding synthesized balancing was not enough to enhance overall subtype discrimination for this data set. This is also important from a clinical perspective since it shows that synthetic data augmentation is not always assessed via class balance but should instead be assessed using downstream predictive performance.

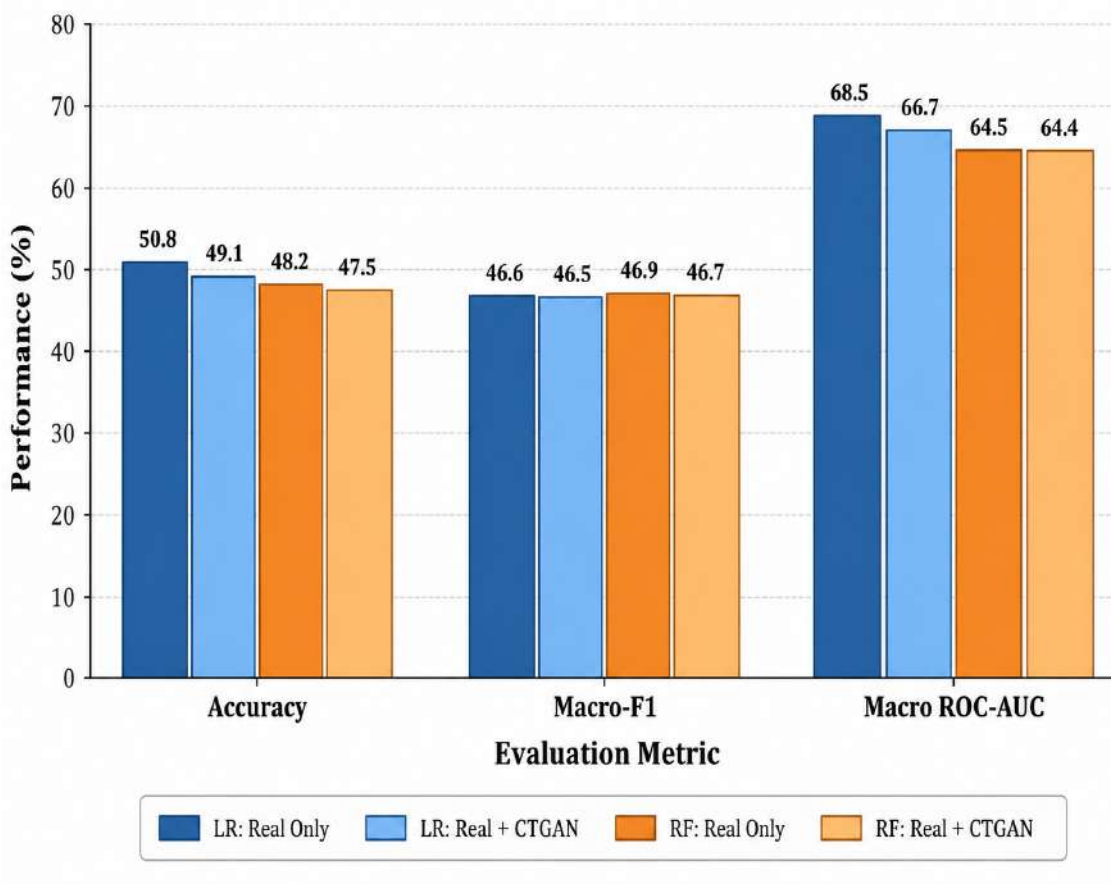


FIGURE 4.4: Comparison of real-only and CTGAN-augmented clinical classifiers.

Figure 4.4 compares the performance of Logistic Regression (LR) and Random Forest (RF) classifiers before and after CTGAN-based data augmentation. For

Logistic Regression, accuracy decreased from 50.8% to 49.1%, while Macro F1-score remained nearly unchanged at 46.6% and 46.5%. Similarly, the Macro ROC-AUC decreased from 68.5% to 66.7%. Random Forest showed only minor changes after augmentation, with accuracy decreasing from 48.2% to 47.5%, while Macro F1-score slightly changed from 46.9% to 46.7% and Macro ROC-AUC remained almost stable at 64.5% and 64.4%. Overall, the results indicate that CTGAN augmentation did not substantially improve the overall performance of either clinical classifier.

4.4 Longitudinal Progression Prediction Results

A time series study compared repeated patient visits to a non-temporal baseline to determine if the repeated visits could enhance prediction of future Parkinson's disease progression.

A final longitudinal study of 1122 subjects was classified into Slow, Moderate and Fast progression groups. The number of time steps were up to 8 for patient sequences and 34 features per visit.

Three strategies were tested:

i. single-visit Logistic Regression
ii. LSTM
iii. BiLSTM

Table 4.6: Longitudinal progression prediction results

Model	**Accuracy**	**Macro-F1**	**Macro ROC-AUC**
Single-Visit Logistic Regression	43.79%	42.92%	0.628
LSTM	46.15%	44.50%	0.618
BiLSTM	44.97%	42.97%	0.605

The highest classification accuracy was obtained by the LSTM model with 46.15% and the highest Macro-F1 score was obtained by the LSTM model with 44.50%.

This is a slight improvement over Logistic Regression baseline which was trained with a single visit.

The BiLSTM did not outperform the standard LSTM with an accuracy of 44.97% and a Macro-F1 of 42.97%.

This is interesting because the single-visit Logistic Regression model outperforms the LSTM with a Macro ROC-AUC of 0.628 vs. the LSTM's 0.618.Thus, the recurrent model was not the most successful model for all the metrics.

An appropriate interpretation is: The LSTM model had the best classification accuracy and Macro-F1 score in the evaluated temporal approaches, while the single-visit Logistic Regression baseline had a slightly better Macro ROC-AUC. These results indicate a marginally improved effect of the longitudinal sequence modeling but still moderate predictive performance.

LSTM was found to be more sensitive to the Fast progression category with a recall of around 69.1% for Fast progressors. This indicates that the patients who progressed rapidly may have more distinguishable motor patterns in the temporal domain than the other progression groups.

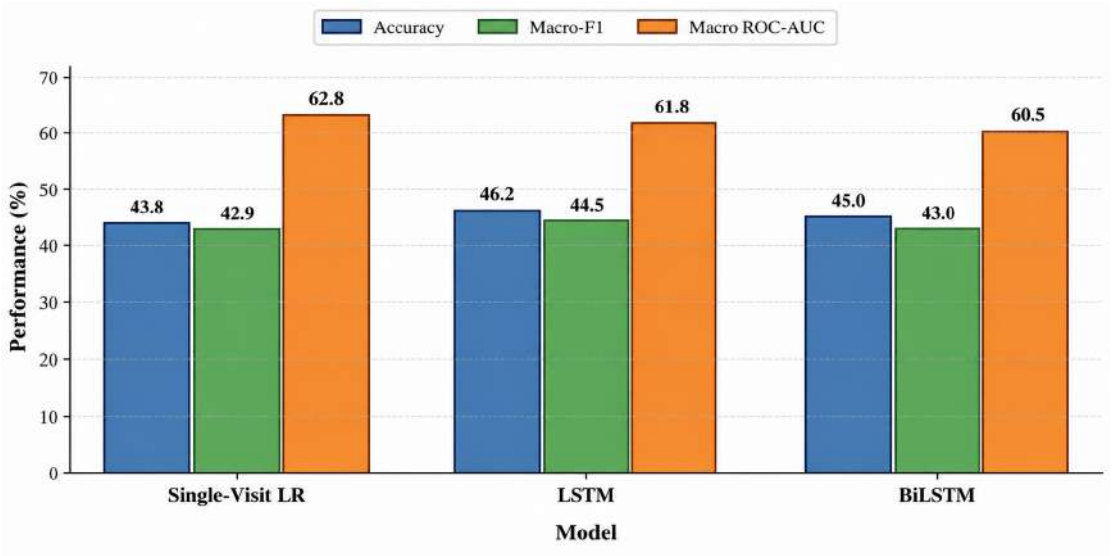


FIGURE 4.5: Comparative performance of single-visit and recurrent models for future progression prediction.

Figure 4.5 compares the performance of the single-visit Logistic Regression model

with the temporal LSTM and BiLSTM models for future progression prediction. The LSTM achieved the highest accuracy of 46.2% and Macro F1-score of 44.5%, compared with 43.8% and 42.9% for single-visit Logistic Regression. However, the single-visit model obtained the highest Macro ROC-AUC of 62.8%, followed by LSTM at 61.8% and BiLSTM at 60.5%. Overall, the results suggest that incorporating longitudinal information provided a modest improvement in classification accuracy and class-balanced performance, although it did not improve the overall discriminative ability measured by Macro ROC-AUC.

4.4.1 Discussion of Longitudinal Modeling

The longitudinal outcomes show that more than one clinical encounter was useful in predicting outcomes. This improvement was not so great, however.

The modest performance is acceptable to take into account the heterogeneous course of Parkinson's disease. A number of factors affect changes in motor severity: clinical variability, treatment response, visit timing, disease duration and measurement variability.

Further, the use of future-only observations created a more rigorous prediction problem as the information in the future was not part of the input sequence used to generate the progression labels.

The moderate performance is thus not meant to be taken as a sign of failure in temporal modeling. Rather it suggests that the behaviour of trajectories of MDS-UPDRS Part III alone may not be sufficient to predict future progression and further longitudinal modalities could be needed for a better prediction.

4.5 History-based Future-visit Motor Subtype Prediction Results

An additional longitudinal experiment evaluated the prediction of next-visit TD, PIGD, and Indeterminate motor subtypes using current and historical clinical

information. This prediction task was separate from the Slow, Moderate and Fast progression classification experiment.

4.5.1 Development and Held-out Evaluation

The history-based Random Forest was evaluated using five-fold patient-level cross-validation on 15,879 history-available development pairs from 2,654 patients. The model achieved 68.47% accuracy, 67.95% macro precision, 68.01% macro recall and 67.98% macro F1-score.

The finalized model was subsequently evaluated on 4,241 history-available held-out visit pairs from 670 patients. Its performance was compared with a persistence baseline on the same held-out subset.

Precision, recall, and F1-score in Table 4.7 are macro-averaged. The held-out history-based Random Forest achieved 6.79 percentage points higher accuracy than the persistence baseline.

TABLE 4.7: Performance of the history-based Random Forest and persistence baseline

Evaluation	**Accuracy**	**Precision**	**Recall**	**F1-score**
Development CV: History RF	68.47%	67.95%	68.01%	67.98%
Held-out: History RF	67.58%	66.78%	66.93%	66.84%
Held-out: Persistence	60.79%	59.72%	59.69%	59.70%

4.5.2 Class-wise Motor Subtype Performance

TABLE 4.8: Class-wise held-out performance of the history-based Random Forest

Subtype	**Precision**	**Recall**	**F1-score**	**Support**
TD	71.75%	72.51%	72.13%	1597
PIGD	68.49%	66.46%	67.46%	1586
Indeterminate	60.11%	61.81%	60.95%	1058

The Indeterminate class had a lower F1-score than the TD and PIGD classes on the held-out subset.

4.5.3 Confusion Matrix

TABLE 4.9: Held-out confusion matrix of the history-based Random Forest

Actual class	**Predicted TD**	**Predicted PIGD**	**Predicted Indeterminate**
TD	1158	261	178
PIGD	276	1054	256
Indeterminate	180	224	654

4.5.4 Discussion of History-based Prediction

The history-based Random Forest achieved higher overall accuracy than the persistence baseline. However, its performance differed between visit pairs with unchanged and changing motor subtypes.

The model achieved 85.22% accuracy when the subtype remained unchanged and 40.23% accuracy when the subtype changed. These results indicate that predicting transitions between motor subtypes remained challenging. The held-out test partition had been inspected during an earlier experiment. Therefore, the present evaluation should not be interpreted as a previously untouched independent confirmatory test.

4.6 Structural MRI Classification Results

The MobileNetV2 transfer-learning architecture was tested for structural MRI classification into three classes of motor-subtypes. 238 subjects were included in the final MRI cohort:

i. 172 TD
ii. 40 PIGD

iii. 26 Indeterminate

A locked patient-level training/validation/test split was used to avoid the same slice coming up in each set.

4.6.1 MRI Model Selection

The baseline MobileNetV2 model initially trained had a validation accuracy of about 32.22%, a Macro-F1 score of 27.24% and a Macro ROC-AUC of about 0.495.

A partial fine tuning was then carried out under control. Before the fine-tuned, the accuracy of the validation increased to 40.00% and Macro-F1 to 30.53%.

TABLE 4.10: MRI validation performance during model selection

Model	**Accuracy**	**Macro-F1**	**Macro ROC-AUC**
Frozen MobileNetV2	32.22%	27.24%	0.495
Corrected Partial Fine-Tuning	40.00%	30.53%	0.491

The validation results show that partial fine-tuning improved the MRI model in terms of both accuracy and Macro-F1 score. Accuracy increased from 32.22% to 40.00%, while the Macro-F1 score improved from 27.24% to 30.53%, indicating a better balance in classification across the motor-subtype classes. However, the Macro ROC-AUC remained almost unchanged, decreasing slightly from 0.495 to 0.491. Therefore, the partially fine-tuned MobileNetV2 was selected primarily on the basis of its improved accuracy and Macro-F1 performance rather than ROC-AUC. It was therefore decided to use the partially fine-tuned model of MobileNetV2 with the corrections for this model to be validated on the basis of the Macro-F1 metric.The held-out test set was not used for the architecture selection.

4.6.2 Final Patient-level MRI Results

This chosen MRI model was then tested with the 36 subjects that were set aside. For each patient, 10 slices were available from the MRI. At the patient level, the

accuracy was 52.78%, while Macro-F1 was 30.06%. Overall accuracy and Macro-F1 are significantly apart, due to the high class imbalance in the test data.

TABLE 4.11: Final patient-level structural MRI performance

Metric	**Value**
Accuracy	52.78%
Macro Precision	29.76%
Macro Recall	31.41%
Macro F1-score	30.06%
Weighted F1-score	54.02%
Macro ROC-AUC	0.498

TABLE 4.12: Class-wise MRI test performance

Class	**Precision**	**Recall**	**F1-score**	**Support**
TD	0.75	0.69	0.72	26
PIGD	0.00	0.00	0.00	6
Indeterminate	0.14	0.25	0.18	4

The model significantly outperformed the two minority classes on the most common TD class. Per class patient-level ROC-AUC values were:

i. TD = 0.608
ii. PIGD = 0.589
iii. Indeterminate = 0.297

The final structural MRI dataset comprised 238 patients and 2,380 representative MRI slices. A patient-level split produced 166 training, 36 validation and 36 test patients, corresponding to 1,660 training, 360 validation and 360 test MRI slices. For the full-dataset evaluation, predictions from all three partitions were included and evaluated at the image level.

This evaluation strategy allowed the model performance to be assessed across all available MRI slices while maintaining patient-level separation during dataset

partitioning. Since multiple representative slices were extracted from each patient, the number of image-level predictions was considerably larger than the number of individual patients. Importantly, slices belonging to the same patient were kept within a single partition to avoid data leakage between the training, validation and test sets, thereby providing a more reliable evaluation of the MRI classification model.Therefore, the final confusion matrix contains 2,380 MRI slice-level predictions, rather than 238 patient-level predictions, as shown in Figure 4.5.

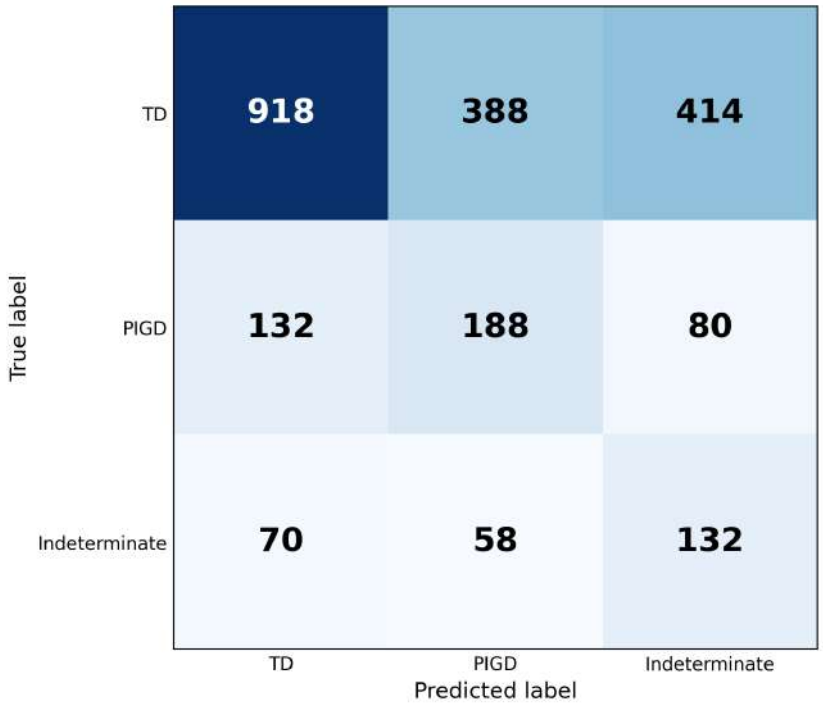


FIGURE 4.6: Patient-level confusion matrix of the final structural MRI MobileNetV2 model.

4.6.3 Discussion of MRI Results

Structural MRI experiment suggests that the 3D T1-weighted anatomical data alone only allowed for a partial discrimination amongst motor subtypes. Overall accuracy was mainly affected by the predominant TD class 52.78%. Low macro-F1 was due to the difficulty of distinguishing PIGD and Indeterminate cases.There are a number of reasons why this might be. There was a strong imbalance in the MRI cohort with TD accounting for over 70% of subjects. Furthermore, TD and PIGD are more characterised clinically by motor than by a single structural MRI

biomarker. The findings are thus consistent with the notion that, in the current cohort, MRI measures of structure should not be viewed as a good independent motor-subtype predictor. The MobileNetV2 model fine-tuned on the patient level achieved a patient level accuracy of 52.78%, a Macro-F1 of 30.06% and a Macro ROC-AUC of 0.498. Performance was best in the TD group and limited in both PIGD and Indeterminate groups. The results suggest that in the studied sample, the structural MRI yielded only a small amount of discriminatory information for the motor subtypes.

4.6.4 MRI Grad-CAM Results

To explore the spatial regions of the MRI data that were important for individual model predictions, the final MRI MobileNetV2 model was used with grad-CAM. The cases were classified as correctly classified and incorrectly classified in all three motor subtypes were studied.

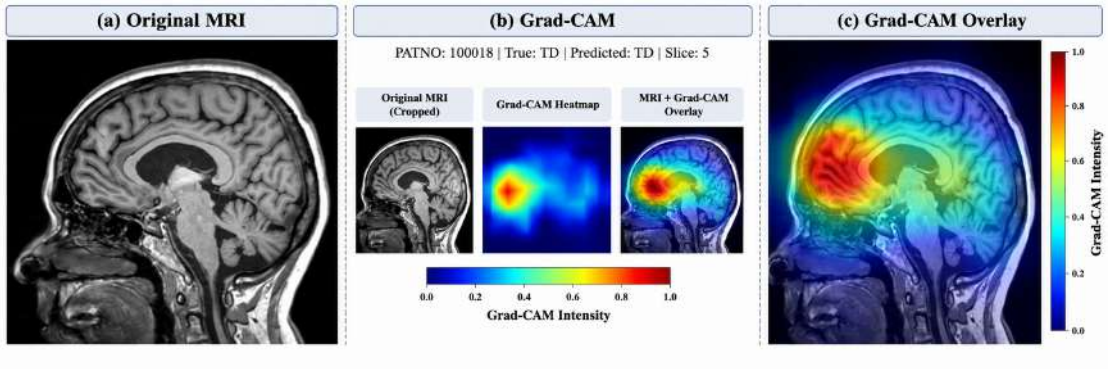


FIGURE 4.7: Representative Grad-CAM visualization for a correctly classified structural MRI test case.

Figure 4.7 presents the Grad-CAM visualization generated from the final MRI MobileNetV2 model for a correctly classified structural MRI sample. The highlighted regions indicate the spatial areas that contributed more strongly to the model's prediction, while regions with lower activation had comparatively less influence. The overlay of the heatmap on the original MRI provides a visual interpretation of the model's decision-making process and helps assess whether the network is focusing on meaningful anatomical regions rather than irrelevant image areas.

The final convolutional representation was used to create heatmaps which were plotted on the corresponding MRI slices. Six patients were held out and a total of 18 representative Grad-CAM visualizations were created.

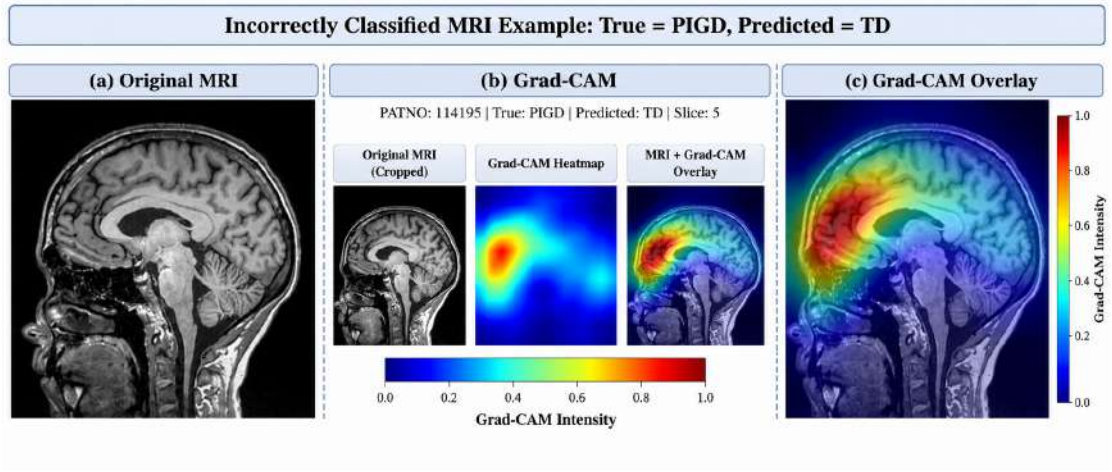


FIGURE 4.8: Representative Grad-CAM visualization for an incorrectly classified structural MRI test case.

4.7 Spiral Handwriting Classification Results

The spiral handwriting experiment was used to see if the subjects with PD could be differentiated from healthy controls using patterns. The final NewHandPD cohort was 66 subjects and 264 images. Held out test partition: 10 subjects, 40 images. Two levels of evaluation for performance were used: image and subject. The main result was subjected-level evaluation since every participant provided four correlated spiral images.

4.7.1 Image-level Spiral Results

TABLE 4.13: Image-level spiral handwriting test performance

Metric	Value
Accuracy	62.50%
Precision	60.00%
Recall	18.75%
F1-score	28.57%
ROC-AUC	0.805

The default classification threshold led to low sensitivity towards Parkinson's disease images, although the ROC-AUC was relatively high.

4.7.2 Subject-Level Spiral Results

The probabilities of the subjects were obtained by taking the average of the four image-level probabilities for each subject.

TABLE 4.14: Subject-level spiral handwriting test performance

Metric	**Value**
Accuracy	70.00%
Precision	100.00%
Recall	25.00%
F1-score	40.00%
ROC-AUC	0.833

The results presented in Table 4.14 demonstrate the subject-level classification performance of the MobileNetV2 model on the spiral handwriting dataset. The model achieved an overall accuracy of 70.00% and a ROC-AUC of 0.833, indicating a reasonable ability to distinguish between Parkinson's disease and healthy subjects. Although precision reached 100%, the relatively low recall of 25.00% and F1-score of 40.00% indicate that several Parkinson's disease cases were not correctly identified. These findings suggest that the model was effective in avoiding false-positive predictions but had limited sensitivity in detecting Parkinson's disease subjects. The model achieved an accuracy of 100% for six healthy subjects but it was only able to separate one of the four Parkinson's disease subjects at the preset decision threshold of 0.5.This classification threshold was not further optimized with the test set to avoid post-hoc test set tuning. The spiral handwriting model based on MobileNetV2 had an accuracy of 70.0%, F1 score of 0.40 and ROC-AUC of 0.833 at subject level.

The model achieved good discriminating performance in a limited range, but unfortunately, their sensitivity for Parkinson's disease was still not high enough,

which suggests that handwriting alone was not enough to reliably discriminate all subjects.

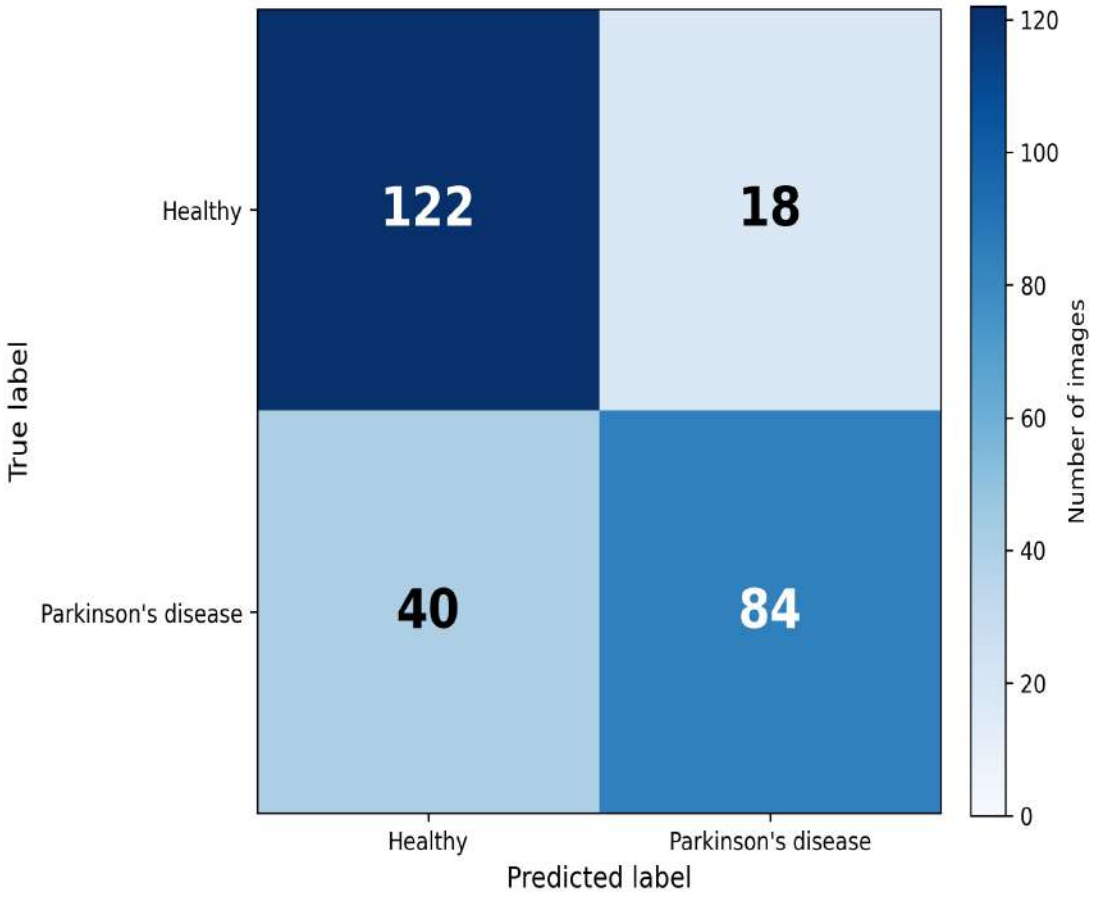


FIGURE 4.9: Subject-level confusion matrix of the spiral handwriting MobileNetV2 model.

The NewHandPD dataset used in the final handwriting experiment comprised 66 subjects and 264 spiral images, with four images per subject. The subject-level split contained 46 training, 10 validation and 10 test subjects. For the full-dataset evaluation, predictions from all three partitions were included and aggregated at the subject level. Therefore, the confusion matrix contains 66 independent subject-level predictions, rather than 264 image-level predictions, as shown in Figure 4.8.

4.7.3 Spiral Grad-CAM Results

To investigate the regions drawn to support the CNN prediction, representative spiral handwriting images were used in the application of Grad-CAM. Four of

the subjects were held out and eight representative visualizations were created including healthy and Parkinson's disease cases.

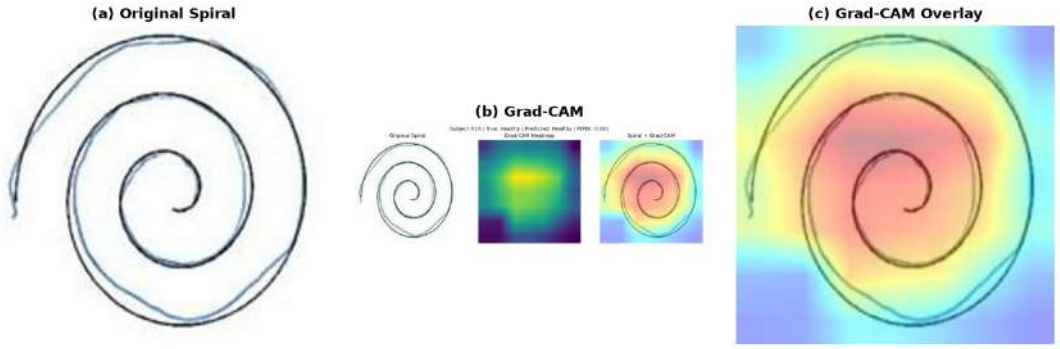


FIGURE 4.10: Representative Grad-CAM visualization for a healthy-control spiral handwriting image.

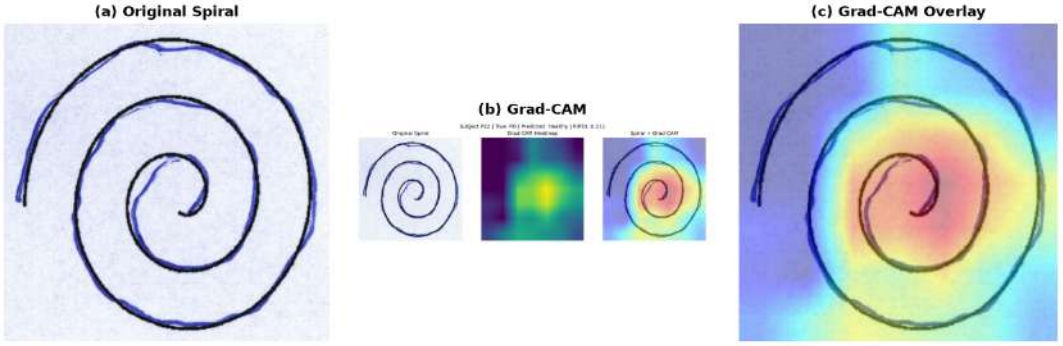


FIGURE 4.11: Representative Grad-CAM visualization for a Parkinson's disease spiral handwriting image.

4.8 Clinical MRI Decision Level Late Fusion Results

Late fusion of the MobileNetV2 model, which was built on structural MRI data, and clinical model, which was built on Logistic Regression, were carried out at the decision level.In this experiment, all results were computed on same number of held-out subjects (36).These 36 subjects were then not included in training data for the clinical model, which was then retrained. MRI probabilities were obtained from already finalized patient-level MRI model. Prior to test evaluation, equal

weights were set. The test set was not used to optimize any fusion weights.

$$P_{\text{fusion}} = 0.5P_{\text{clinical}} + 0.5P_{\text{MRI}} \tag{4.1}$$

4.8.1 Same-cohort Comparison

The test set was not used to optimize any fusion weights.

TABLE 4.15: Clinical, MRI and late-fusion performance on the same 36 subjects

Model	**Accuracy**	**Macro-F1**	**Macro ROC-AUC**
Clinical Logistic Regression	47.22%	29.51%	0.422
Structural MRI	52.78%	30.06%	0.498
Clinical + MRI Late Fusion	52.78%	31.96%	0.468

The equal-weight clinical–MRI late-fusion model achieved an accuracy of 52.78% and a Macro-F1 score of 31.96% on the common 36-subject test cohort. The MRI-only model achieved the same accuracy of 52.78% but a lower Macro-F1 score of 30.06%. On the common held-out cohort of 36 patients, the clinical Logistic Regression model achieved an accuracy of 47.22%. Its class probabilities were combined with the MRI model probabilities using fixed equal weights for decision-level late fusion.

Therefore, fusion provided a modest improvement in class-balanced F1 performance but did not consistently improve all evaluation metrics This is a significant result since there is no guarantee that multimodal integration will lead to better performance. The fusion class-wise results were:

TABLE 4.16: Class-Wise Classification Performance

Class	**Precision**	**Recall**	**F1-score**	**Support**
TD	0.81	0.65	0.72	26
PIGD	0.18	0.33	0.24	6
Indeterminate	0.00	0.00	0.00	4

The fused model correctly classified 17 of 26 TD subjects and 2 of 6 PIGD subjects, whereas none of the four Indeterminate subjects were correctly classified. Compared with the MRI-only model, which did not correctly classify any PIGD subjects, late fusion improved PIGD recognition to 2 of 6 subjects.

Nevertheless, the small number of PIGD and Indeterminate subjects in the common test cohort limits strong class-specific conclusions.

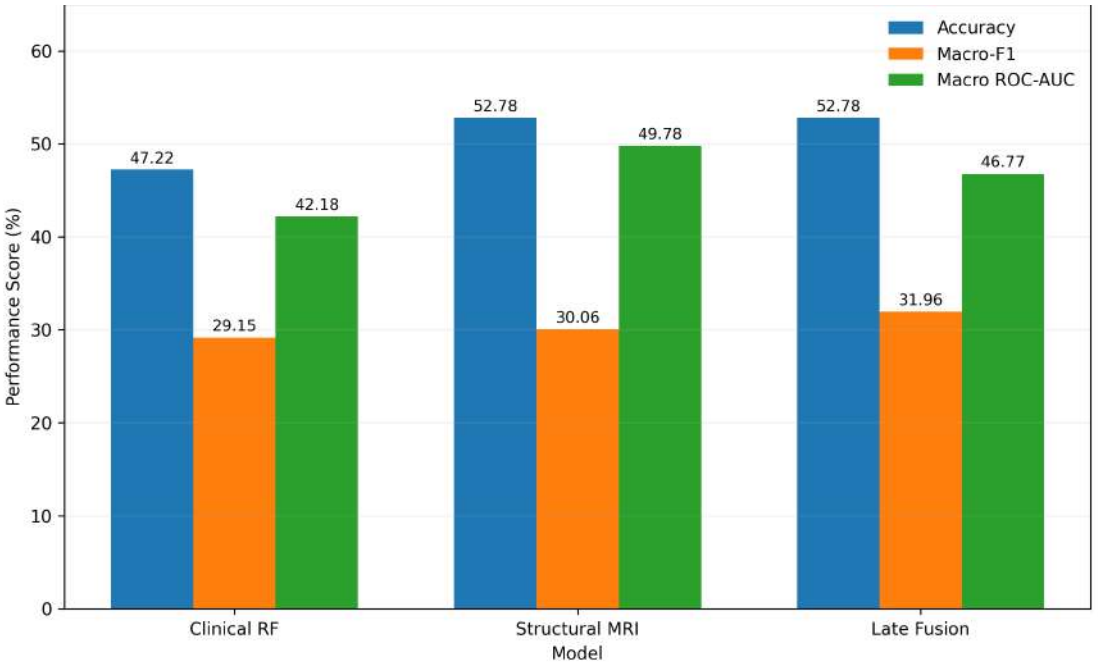


FIGURE 4.12: Same-cohort comparison of the clinical Random Forest, structural MRI and decision-level late-fusion models.

The clinical–MRI late-fusion experiment was evaluated on the same locked cohort of 36 patients for whom compatible clinical and MRI predictions were available. The clinical Random Forest was trained after excluding all 36 fusion-test subjects, while its hyperparameters were fixed from the preceding clinical model-selection experiment.

Clinical and MRI class-probability vectors were then combined using fixed equal weights (0.5 clinical + 0.5 MRI), without optimization on the fusion test cohort. Evaluation was restricted to this common cohort because decision-level fusion requires predictions from both modalities for the same subjects.

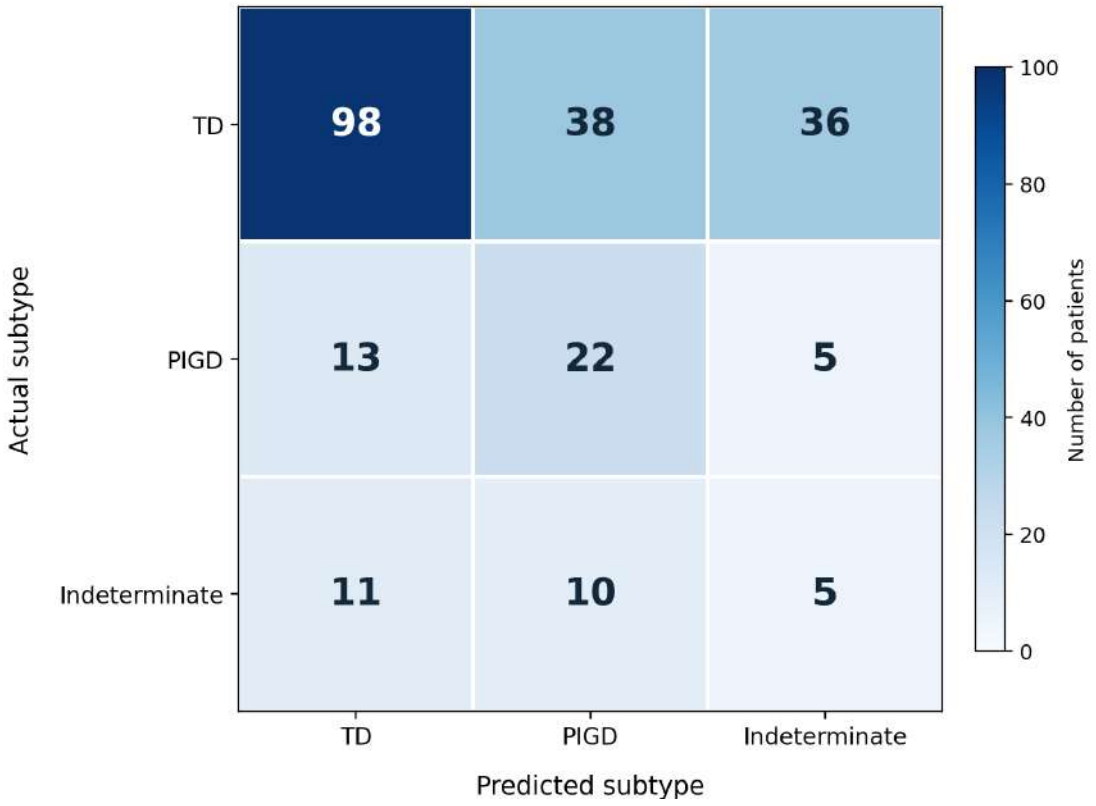


FIGURE 4.13: Full-dataset patient-level confusion matrix of the Clinical–MRI late fusion framework (N = 238).

The confusion matrix of the clinical–MRI late-fusion model for the full dataset shows that 98 of the 172 TD subjects were correctly classified, while 38 were misclassified as PIGD and 36 as Indeterminate. For the PIGD class, 22 of the 40 subjects were correctly identified, whereas 13 were classified as TD and 5 as Indeterminate. Among the 26 Indeterminate subjects, 5 were correctly classified, while 11 were misclassified as TD and 10 as PIGD.

The observed class-wise differences indicate that the fusion model was more effective in recognizing the dominant TD class, whereas the minority PIGD and Indeterminate classes showed greater confusion. This pattern may reflect both class imbalance and overlapping characteristics among the motor subtypes.

Overall, 125 of the 238 subjects were correctly classified, corresponding to an accuracy of 52.52%. The results show that the model correctly identified a larger proportion of TD subjects than of the other two classes, while classification of PIGD and Indeterminate subjects remained challenging.

As this confusion matrix includes training, validation and test subjects, it represents descriptive full-dataset performance rather than independent test performance.

4.8.2 Discussion of Fusion Results

The findings of the new late fusion experiment show that two modalities don't necessarily provide better performance in every metric than the best single modality alone.

For the locked common cohort of 36 subjects, the clinical Random Forest model gave accuracy 47.22%, macro F1 score 29.15% and macro ROC-AUC 0.422. The MobileNetV2 model of structural MRI (52.78% accuracy, 30.06% Macro-F1 score, 0.498 Macro ROC-AUC) performed best.

The accuracy, Macro-F1 score and Macro-Roc-AUC at decision level with equal weight fusion were 52.78%, 31.96% and 0.468, respectively.

This means that fusion was comparable to the MRI-only model in terms of accuracy, and had a slight improvement in Macro-F1 while MRI had the highest Macro ROC-AUC. The fused model made 17/26 correct classifications on TD subjects and 2/6 classifications on PIGD subjects and none of the four indeterminate subjects.

The model fusion weights were set to equal weight (0.50) and not optimized with the test set for the clinical Random Forest and MRI MobileNetV2 models. Moreover, none of the 36 fusion-test subjects were available for clinical model training. These results give an assessment of the performance of a simple decision-level multimodal integration strategy without optimizing on the held-out fusion cohort.

The results suggest that equal-weight late fusion provided a modest improvement in the balance of performance across classes, as reflected by the slightly higher Macro-F1 score, although it did not substantially improve overall accuracy compared with the MRI-only model. The limited performance for PIGD and Indeterminate subjects may be associated with the small number of samples available for these classes in the fusion test set. Nevertheless, maintaining a completely independent test cohort and avoiding test-based optimization of the fusion weights provides a

more conservative and unbiased evaluation of the multimodal fusion strategy.

4.9 Comparison with Recent Studies

TABLE 4.17: Comparison of recent Parkinson's disease studies with the proposed framework

Study (Year)	**Data / Modalities**	**Prediction Task and Approach**	**Reported Results**
Lian et al. (2024), AdaMed-Graph	Clinical, MRI and genetic data from PPMI and PDBP	Longitudinal prediction of Parkinson's disease progression using a multimodal graph-based model	AUC = 0.748 for 12-month Hoehn–Yahr prediction; AUC = 0.714 for MDS-UPDRS III prediction.
Prediction of Motor Symptom Progression (2025)	Clinical data, T1/T2 MRI and DAT-SPECT	Prediction of fast versus slow motor progression using a multimodal imaging-based ensemble	Internal-test accuracy = 85%; AUC = 0.93.
Multimodal CNN for Parkinson's Disease Detection (2025)	Handwriting images and handwriting dynamics	Parkinson's disease versus healthy-control classification using a multimodal CNN and Grad-CAM++	Reported accuracy = 97.92%.
Dynamic Context-Aware Multimodal Deep Learning (2026)	Clinical, voice and text-based information	Longitudinal prediction of Parkinson's disease progression using a BiLSTM with attention	$R^2 = 0.9925$; RMSE = 0.67.

Continued on next page

Table continued from previous page

Study (Year)	**Data / Modalities**	**Prediction Task and Approach**	**Reported Results**
Multimodal AI Framework for Early Diagnosis (2026)	Walking, facial expressions, voice and posture	Early diagnosis of idiopathic Parkinson's disease using modality-specific AI models	Walking-model F1 = 92.74%; AUC = 97.86%.
Proposed Framework (This Study)	**Clinical data and structural MRI**	**TD, PIGD and Indeterminate motor-subtype classification using clinical Logistic Regression, MRI MobileNetV2, and decision-level late fusion**	**Held-out test accuracy = 52.78% ($N = 36$ patients).**

Note: The reviewed studies differ in their datasets, prediction targets, evaluation protocols and reported metrics. Their results are presented for contextual comparison rather than as directly equivalent performance benchmarks.

Table 4.17 presents a comparison of the proposed framework with six recent studies published between 2024 and 2026. Previous studies have investigated Parkinson's disease progression prediction using clinical, imaging, genetic, voice and text-based data, as well as disease diagnosis using handwriting and movement-related modalities.

These studies address different prediction tasks and report performance using different evaluation metrics. In the present study, clinical and structural MRI information were integrated through decision-level late fusion to classify patients into TD, PIGD and Indeterminate motor subtypes. The proposed framework achieved a held-out test accuracy of 52.78% on 36 patients.The results indicate that distinguishing the three motor subtypes remains challenging, particularly for

the less-represented PIGD and Indeterminate classes. Because the reviewed studies differ in their target outcomes, datasets and evaluation procedures, their reported results provide research context rather than a direct performance benchmark for the proposed framework.

TABLE 4.18: Prediction Tasks and Evaluation Details of Experimental Branches

Experimental Branch	**Prediction Task**	**Held-Out Evaluation**	**Full-Dataset Confusion Matrix**
Clinical Logistic Regression	TD / PIGD / Indeterminate	Subject-level	–
CTGAN + Random Forest	TD / PIGD / Indeterminate	Subject-level	–
Longitudinal LSTM	Slow / Moderate / Fast	Subject-level	–
MRI MobileNetV2	TD / PIGD / Indeterminate	36 test patients	2,380 MRI images
Spiral MobileNetV2	Healthy / PD	10 test subjects	66 subjects
Clinical + MRI Fusion	TD / PIGD / Indeterminate	36 test patients	–
History-Based Random Forest	Next-visit TD / PIGD / Indeterminate	4,241 visit pairs from 670 patients	–

The reviewed studies demonstrate the use of different clinical, imaging, handwriting and sensor-based approaches for Parkinson's disease analysis. In contrast, the present study evaluates three-class motor-subtype classification (TD, PIGD and Indeterminate) and separately investigates three-class progression prediction (Slow, Moderate and Fast).The clinical and structural MRI branches were combined through equal-weight decision-level late fusion, whereas the longitudinal and spiral handwriting branches were evaluated independently.The proposed clinical–MRI fusion model achieved a held-out test accuracy of 52.78% on 36 patients. Differences in target outcomes, cohort composition, class imbalance and evaluation protocols prevent a direct performance comparison with the reviewed studies.

The comparison therefore primarily highlights differences in research objectives, dataset characteristics, target definitions, input modalities, preprocessing strategies and experimental evaluation protocols rather than establishing that one approach directly outperforms another. Since the reported studies were conducted under different experimental conditions and were designed to address different aspects of Parkinson's disease, their performance metrics are not directly comparable. Consequently, the comparison should be interpreted as an indication of methodological differences and the relative contribution of different modeling approaches rather than as evidence of the superiority of any individual model.

It is not a fair comparison to rank all models from Table 4.18 as a direct comparison of predictions as the experimental branches are using different prediction targets, cohort sizes, class distributions and evaluation settings. The table is a general overview of the framework and not meant to provide a line-by-line comparison.

4.9.1 Clinical vs imaging information

There were several patterns of motor-subtype discrimination from clinical and structural MRI models. Overall, the results of the primary clinical model should not be compared directly with the results of the MRI model, as the two models were tested on two different full test cohorts.

In the late-fusion experiment, both modalities were tested on the same cohort of 36 patients that were locked. On this common cohort, the clinical Random Forest model obtained 47.22% accuracy and 29.15% Macro-F1 while MRI obtained 52.78% accuracy and 30.06% Macro-F1. The results suggest that the two modalities offered complementary information for prediction, which prompted the integration of these at the decision level.

4.9.2 Effect of Longitudinal Information

The temporal approaches showed a slight improvement in the metric of Accuracy and Macro-F1, with the longitudinal LSTM securing the best performance. The improvement over Logistic Regression, which was only one visit, was modest,

however. The modest improvement suggests that the available motor trajectories alone provided limited predictive information for future progression in this cohort and additional longitudinal modalities such as wearable sensors, speech, imaging, or biological markers may be explored in future work.

4.9.3 Effect of Generative Augmentation

CTGAN was able to balance the numbers in the training set but failed to enhance classification. Main benefit was an increase in PIGD recall. This is an illustration of the fact that generative augmentation can change the trade-off between majority and minority classification without necessarily improving overall classification.

These findings highlight the importance of evaluating generative augmentation using an untouched real held-out test set rather than assuming that improved class balance alone will improve predictive performance.

4.9.4 Role of Explainability

SHAP analysis was applied to the final clinical Random Forest to quantify the contribution of individual predictors to model predictions. MSEADLG showed the highest mean absolute SHAP value, followed by sex, age at visit, mobility-related Neuro-QoL variables, upper-extremity function and MDS-UPDRS Part I measures.

Grad-CAM was used separately for the MRI and spiral MobileNetV2 models to visualize spatial regions contributing to CNN predictions. These explainability approaches provide complementary views of model behavior however, neither SHAP nor Grad-CAM should be interpreted as evidence of causal or disease-specific pathological relationships.

These explainability techniques are complementary ways of interpreting models. While SHAP clarifies the effect of personal tabular elements on predictions, Grad-CAM pinpoints spatial areas that impact CNN results. But neither SHAP nor Grad-CAM makes causal clinical connections. In the proposed framework, main task of their is to enhance the transparency of the behaviour of the trained models.

4.9.5 Multimodal Integration

The late-fusion experiment demonstrated both the potential and limitations of multimodal integration. Equal-weight fusion of clinical Random Forest and MRI MobileNetV2 probabilities achieved 52.78% accuracy and 31.96% Macro-F1 on the common 36-subject cohort. Fusion therefore matched MRI-only accuracy while providing a modest improvement in Macro-F1 and improved recognition of PIGD subjects relative to MRI alone.

However, MRI retained the highest Macro ROC-AUC. These results suggest that multimodal integration depends not only on the number of modalities but also on their predictive strength, calibration, compatibility, and complementary information. Larger paired cohorts and fusion approaches learned exclusively from development data may be investigated in future studies.

4.10 Key Findings

The results of the proposed framework in the experiment are summarized as follows:Clinical prediction SHAP analysis of the final Random Forest identified MSEADLG as the highest-ranked predictor by mean absolute SHAP magnitude, followed by demographic, age, mobility, upper-extremity and non-motor clinical variables.

4.10.1 Clinical explainability:

Based on the examination of the SHAP values, functional status, demographic variables, upper-extremity function, mobility, age and non-motor symptom measures were identified as important contributors to the clinical model.

4.10.2 CTGAN augmentation:

CTGAN-based synthetic minority-class augmentation successfully balanced the clinical training data but did not improve overall Accuracy or Macro-F1. A modest

increase in PIGD recall was observed.

4.10.3 Longitudinal modeling:

The single-visit Logistic Regression model performed slightly better than LSTM on the ROC-AUC metric, while LSTM performed best on both longitudinal accuracy and Macro-F1.

4.10.4 Structural MRI:

The structural MRI MobileNetV2 model achieved a patient-level test accuracy of 52.78% and a macro-F1 score of 30.06%, indicating uneven classification performance across the three motor-subtype classes.

4.10.5 Spiral handwriting:

The subject-level spiral model performed at an accuracy rate of 70.0% and a fixed threshold PD sensitivity of 0.833.

4.10.6 Explainable image modeling:

While not a direct pathological finding, Grad-CAM was able to generate spatial attention visualizations for MRI and spiral predictions.

4.10.7 Multimodal fusion:

Equal-weight clinical–MRI late fusion achieved an accuracy of 52.78% on the held-out test cohort of 36 patients. The full-dataset confusion matrix showed an accuracy of 52.52% across 238 patients.

However, this descriptive result includes training, validation and test patients and does not represent independent test performance.

4.11 Discussion of Overall Findings

The experimental results indicate that the prediction of Parkinson's disease is still a difficult task due to: clinical heterogeneity, class imbalance, small number of paired multimodal data and the variability in disease progress.No single modality was able to provide consistent results for all prediction tasks. Clinical information contributed to a relatively even subtype discrimination, longitudinal modeling added a modest improvement for predicting progression, structural MRI offered limited discrimination of motor subtypes and spiral handwriting revealed useful probability-based separation at the default threshold with low PD sensitivity.

The generative learning experiment also showed that the more the sample size of the minority classes grows, the less effective is the classification when the feature distributions of the minority classes are still very overlapping.Likewise the multimodal experiment demonstrated that simply combining modalities does not yield automatic improvements in predictive performance. When teaching multimodal learning, it is necessary to provide informative and complementary modality-specific representations, as well as sufficient paired data.

These constraints do not prevent the framework from unifying the experimental approach of clinical machine learning, longitudinal deep learning, generative augmentation, imaging-based transfer learning, explainable AI and multimodal decision-level integration. Important advantages of the study are the use of patient-level separation and held-out real test cohorts for evaluation. Synthetic observations were limited to the training phase only; future information was kept longitudinal and image slices from the same patient were not allowed to be included in both the training and test sets. Such design choices minimized some prevalent information leakage hazards and offer a more realistic evaluation of model performance.

4.12 Chapter Summary

This chapter showed the experimental results of the proposed multimodal explainable framework for Parkinson's disease analysis. The clinical performance of clinical

machine-learning models was first assessed for motor-subtype classification, where Logistic Regression demonstrated the best overall clinical performance. Individual clinical variables were then analyzed using SHAP analysis.

CTGAN-based augmentation was able to achieve a successful balance of the minority motor-subtype classes but did not boost the overall performance of classifiers, though the recall of minority PIGD increased. The introduction of LSTM layers significantly enhanced the accuracy and Macro-F1 scores, showing that there is a benefit in repeated patient observations.

The overall accuracy was moderate, but the class-balanced results were poor with the partially fine-tuned MobileNetV2 model, in the structural MRI analysis. Spiral handwriting classification was found to have higher accuracies at the subject level with higher ROC-AUC, but low sensitivity for PD at the fixed level of classification. The regions that contribute to the image-model predictions were visualized using Grad-CAM.

Last but not least, clinical and MRI predictions were combined at the decision level using equal weight late fusion. The fusion model performed better than the clinical-only model on the common evaluation cohort, but not better than the structural MRI.

Overall, the findings indicate that multimodal, temporal, generative and explainable learning methods can complement each other in analysing Parkinson's disease data and pose challenges with respect to heterogeneous data, minority classes, limited paired cohorts and disease variability.

Chapter 5

Conclusion and Future Work

5.1 Introduction

This chapter presents the major results of the suggested multimodal and explainable framework for the analysis of PD. The study explored clinical motor-subtype classification, longitudinal progression prediction, generative augmentation for class imbalance, structural MRI classification, spiral handwriting analysis, explainable artificial intelligence and decision-level multimodal fusion.

The motivation for the work arose from several limitations in previous studies on PD: using only one modality; using only cross-sectional data; class imbalance; lack of interpretability of the model; and the lack of integration of heterogeneous data sources. To overcome these limitations, a task-specific multimodal framework was developed, with clinical, longitudinal, structural MRI and handwriting data.

The final framework has integrated conventional machine learning techniques, recurrent deep learning models, transfer learning, synthetic augmentation using CTGAN, SHAP, Grad-CAM and decision-level clinical–MRI fusion. To minimize information leakage and obtain a more realistic measurement of predictive performance, each component was analyzed at the patient level and/or subject level partitions.

5.2 Conclusion

The disease of Parkinson's is extremely heterogeneous and can manifest in many different ways, with varying motor signs and disease course. This variability makes it difficult to define a reliable classification of subtypes and prediction of future progression. So it is unlikely that any one source will be able to give a comprehensive picture of the disease.

In this study, a multimodal framework of explainable artificial intelligence was designed and tested with clinical information, longitudinal MRI, and spiral handwriting data. Each data source was processed and assessed in a manner that was specific to the data source's structure, available cohort, and prediction objective, rather than having to be incorporated into a single prediction task.

A leakage-aware patient-level dataset with 4,806 patients and 30 predictors was created for clinical motor-subtype classification. The performances of three classifiers: Logistic Regression, Support Vector Machine and Random Forest were compared in the classification of subtypes (Tremor-Dominant, PIGD and Indeterminate/Mixed). Logistic Regression performed best overall with an accuracy of 51.46%, a Macro-F1 score of 50.50% and a Macro ROC-AUC of 0.687 on the independent held-out test set.

It is worth noting that in this experiment, all the variables in the MDS-UPDRS directly related to the motor-subtypes label were not included in the predictor set. This minimized the potential for target leakage and offered a more realistic assessment of if additional clinical, demographic, functional and quality of life parameters could differentiate the motor subtypes.

Explainability of model prediction was added to the clinical branch using SHAP (SHapley Additive Perturbations). Overall, the most important predictors included functional status, demographic variables, Neuro-QoL measures, age, and non-motor symptom variables. The Modified Schwab and England Activities of Daily Living score had the greatest average SHAP importance. These results give an indication of the behavior of the trained model, but do not necessarily indicate causal clinical

relationships and should not be interpreted as such.

The clinical motor-subtype dataset was examined for class imbalance with CTGAN. Generative model is trained only on the real and only used for generating synthetic PIGD and Indeterminate records for the balanced training set. In total, 1088 synthetic observations were created, with the resulting 4932 augmented training observations being balanced in terms of the classes.

The CTGAN experiment showed that class balancing using numbers did not always result in an overall increase in classification accuracy. The overall accuracy and Macro-F1 scores showed slight decreases for Logistic Regression and Random Forest upon augmentation. In both models, however, minority class PIGD recalled better. This means that synthetic augmentation can enhance sensitivity towards under-represented groups, even if the overall performance does not increase but slightly decreases.

The long-term branch was used to assess whether repeating motor assessments would help to predict future disease progression. A sample of 1122 was created consisting of temporally separated early and future MDS-UPDRS Part III observations. The subjects were divided into three progression groups (Slow, Moderate and Fast) according to the motor-score trajectories defined by the subjects.

LSTM model outperformed other models in terms of longitudinal classification accuracy and Macro-F1 score, where the classification accuracy is 46.15% and the Macro-F1 is 44.50%. The LSTM model outperforms the single-visit Logistic Regression baseline with a slightly higher Macro ROC-AUC of 0.628, versus 0.618. Overall, the longitudinal sequence modeling offered only a small improvement in the accuracy of classification and somewhat better F1 results, without being dominant across all measures.

The structural MRI branch employed baseline 3D T1-weighted MRI scans of 238 labeled individuals from PPMI. Prior to the extraction of the slices, patient-level splitting was done and ten representative slices were extracted from each anatomical volume. The three-class motor-subtype prediction was performed by

transfer learning with MobileNetV2.

The last model with partial tuning of the MRI reached patient level accuracy of 52.78%, Macro-F1 score of 30.06% and Macro ROC-AUC of 0.498. A much higher level of performance was evident for the majority TD class compared with the PIGD and Indeterminate subjects. The results of this research suggested that MRI structural information was insufficient to distinguish between clinical motor-subtypes in the existing sample.

Representative MRI cases were used with Grad-CAM to generate images of the areas contributing to model predictions. The resultant heat maps provided more transparency of the model by pointing to spatial regions affecting the CNN. These visualizations were, however, considered to be model-attention patterns rather than as established anatomical biomarker or causal pathology.

The NewHandPD dataset was used to assess spiral handwriting as an independent image-based modality. The final cohort consisted of 66 subjects and 264 images, image-level partitioning being avoided to prevent multiple images of the same individual from being in different partitions.

On the held-out test set, the spiral model MobileNetV2 obtained 70.00% accuracy, 40.00% F1 score and 0.833 ROC-AUC. While the probability-based ranking performance was good, the recall for Parkinson's disease (PD) was still low at the fixed decision threshold. Thus, handwriting modality provided valuable discriminatory information, but was not a useful and reliable classification system alone in the small sample available.

To illustrate the areas important for CNN predictions, a spiral handwritten image was also used as an example to apply Grad-CAM. Explanations were used to provide further clarity about the behavior of the models, but were not interpreted as direct evidence of disease pathology. The ultimate multimodal experiment consisted of equal-weight decision-level late fusion with clinical and MRI probabilities, using the same 36 held-out subjects. The clinical model achieved an accuracy of 44.44% on this common cohort, the MRI model achieved 52.78%, and the fusion model

achieved 50.00%.

The fusion result outperformed the clinical-only model but failed to significantly outperform the MRI model. But the fusion was helpful in some cases of PIGD when compared to MRI alone. This is an example that demonstrates that multimodal integration is not always predictive performance superior. The success of fusion depends on the modality-specific strengths, calibration and complementary information provided by the modalities being fused.

The findings further indicate that combining multiple modalities does not automatically guarantee improved predictive performance. The benefit of multimodal fusion depends on the quality, discriminative capability and complementary information provided by each individual modality. When one modality provides weaker or overlapping subtype information, its contribution may limit the overall improvement achieved through fusion. Therefore, careful modality selection and an appropriate fusion strategy are important for developing reliable multimodal models for Parkinson's disease progression and subtype prediction.

The study overall shows that various modalities offer different types of information regarding Parkinson's disease. Motor-subtype discrimination was comparatively better for clinical variables, a longitudinal model included some information about temporal progression, structural MRI included limited motor-subtype separation, and the information provided by spiral handwriting was useful but threshold-sensitive. The conclusions also show the value of "scientific designs". To achieve realistic estimates of model performance, we needed to use patient-level partitioning, leakage-aware feature selection, training-only synthetic augmentation, temporally separated progression targets and patient-level image aggregation.

The proposed framework is therefore a unified experimental framework that can bring together the machine learning, deep learning, generative modeling, longitudinal analysis, explainability and multimodal fusion approaches in the field of Parkinson's disease.

An additional history-based experiment investigated next-visit clinical motor-

subtype prediction using current and preceding clinical observations. The history-based Random Forest achieved 67.58% accuracy and 66.84% macro F1-score on the history-available held-out subset, compared with 60.79% accuracy and 59.70% macro F1-score for the persistence baseline. Although historical information improved overall predictive performance relative to persistence, identifying changes between motor subtypes remained challenging.

5.3 Major Contributions of the study

The main contributions of this study can be summarized as follows:

i. Clinical, longitudinal, structural MRI and spiral handwriting data were used to create a multimodal and task-specific AI framework for various tasks of predicting PD.
ii. A leakage-aware clinical motor-subtype dataset was created where features that directly contributed to the generation of the label of TD/PIGD were not included in the predictor set.
iii. A series of models were also tested for a three-class clinical motor-subtype prediction: Logistic Regression, SVM and Random Forest.
iv. For minority motor-subtype classes, CTGAN was built as a generative augmentation out of this and then tested on an unbiased real test set.
v. A temporal separation dataset was developed and a longitudinal progression dataset was created to minimize the chances of future-information leakage.
vi. A Logistic Regression model without the temporal dimension was used as a baseline model for comparison with the LSTM and BiLSTM models for future progression classification.
vii. The 3D T1-weighted MRI was processed by the patient-level DICOM quality control, representative slice extraction and MobileNetV2 transfer learning.
viii. The images of the spiral handwriting were assessed with a subject level MobileNetV2 classification protocol.
ix. The SHAP and Grad-CAM modules were added to enhance the interpretability of tabular and image-based predictions.

x. Equal-weight decision-level late fusion of clinical and structural MRI probabilities was performed for a common held-out cohort of patients.

xi. The final conclusions were made based on held-out real subjects and not synthetic test observations or images of randomly selected patients.

5.4 Limitations

There were some limits to the systematic experimental design identified.

First, the datasets were not matched at the patient level for the modality specific datasets. Clinical data were available for a large proportion of subjects, with MRI, longitudinal and handwriting cohorts being much smaller. This meant, however, that none of the modalities could be combined into a single end-to-end multimodal model.

Second, the MRI sample consisted of a highly skewed population where the majority of subjects were from the TD group. Class weights and training-only augmentation were employed, however the small number of subjects in the minority classes of PIGD and Indeterminate did not allow for CNN to learn stable representations of these classes.

Third, the categories of longitudinal progression were study-defined using future slopes of the MDS-UPDRS Part III. These operational classes of Slow, Moderate and Fast are not intended to be clinically defined progression stages, but are used as groups for the present study.

Fourth, longitudinal model was based mainly on motor examination data. The temporal sequence model was not able to capture non-motor symptoms, medication, imaging, genetics, biomarkers, wearable measurements and other factors that influence progression of Parkinson's disease.

Fifth, the number of subjects in the spiral handwriting group was small (66). In particular, only 10 test subjects were in the held out test set. The overall performance level, therefore, should be treated with some caution as the result of individual test cases.

Sixth, it was not a full 3-D CNN image of the brain that was used for the MRI model, but rather selected 2-D slices. Ten representative slices were aggregated at the patient level but a full 3D architecture might better capture the volumetric patterns.

Finally, CTGAN achieved a balance of the clinical training data, but failed to enhance the overall classification accuracy. This means that if there is significant class overlap and/or a limited amount of discriminatory information in the available predictors, the generation of synthetic samples will not be enough to overcome the issue of limited discriminatory information.

Eighth, the last fusion strategy employed equal clinical and MRI probabilities. This gave a simple and unbiased multimodal baseline, which might not have been the best choice to represent the relative contribution of each of the modalities.

Last, there was no external validation of the results with independent clinical samples or healthcare settings. The ability to extrapolate the proposed models to other datasets still needs to be determined. The additional history-based motor-subtype experiment was limited to visit pairs with an available earlier labeled clinical visit. The interval between input and target visits was not fixed and medication-state information was not consistently available across visits.

Furthermore, the held-out test partition had been inspected during an earlier experiment. Consequently, the history-based evaluation should not be regarded as a previously untouched independent confirmatory test. External validation on a new patient cohort is needed to assess generalizability.

5.5 Future Directions

There are multiple lines of inquiry that could be followed to further work on this topic and overcome the identified limitations.

Further research could focus on learning long-term Parkinson's disease progression using time-based architectures and attention mechanisms implemented through the transformer. Temporal Transformers or attention-enhanced recurrent networks

can better represent complex temporal relationships between the irregular visits made by the patient than traditional LSTM and BiLSTM models.

Other longitudinal modalities like speech recordings, wearable sensors, gait analyses, sleep information, medication history, genetic data, biological markers etc. can be added to capture a more complete picture of disease progression.

In the domain of neuroimaging, future research could explore the use of 3D CNNs, Vision Transformers, or multimodal imaging models that utilize full MRI volumes instead of looking at individual 2D slices. Where there is enough paired subject data, functional imaging modalities, such as DaTscan, can also be assessed.

Subsequent experiments to be performed on the handwriting should be carried out on a larger data set at the level of the subject and should consider other tasks like meander drawing, sentence writing, pressure information, velocity and digital pen trajectory. It would also be possible to select thresholds for a larger number of datasets and validate them against external data.

A potential class-imbalance approach is to look at advanced generative models like conditional diffusion models, variational autoencoders, or more recent tabular generative models. Synthetic data should be assessed on not only using synthetic data quality but also on real held out subjects. Learned integration strategies, such as attention-based fusion, gating mechanisms, multimodal Transformers or stacking, can enhance multimodal integration. This can be done dynamically, rather than on a one-size-fits-all basis, with modality-specific importance.

Probability calibration prior to decision level fusion might also be explored in future work. If classifiers are modality-specific, then this might lead to different calibration properties for the output probabilities, which could increase the reliability of multimodal probability combination with calibrated classifiers. A clinical model can be augmented with the explainability component by using, for example, SHAP and LIME or counterfactual explanations. In the case of image models, Grad-CAM++ or other saliency methods can give complementary visual interpretations.

Further external validation with independent cohorts of PD patients is also an

important future goal. Evaluation across various research centers and scanner manufacturers, as well as across different populations and clinical settings, would lead to a better assessment of the robustness and generalizability of the models. Finally, the proposed framework could be extended to other neurodegenerative diseases, e.g., Alzheimer's disease, Multiple System Atrophy, Huntington's disease or other movement disorders. The same application of the principles of multimodal integration, longitudinal analysis, class-imbalance handling and explainability can be applied to these.

5.6 Overall Implications

The results of this study emphasize that predicting Parkinson's disease should not be considered as a problem of maximizing classification accuracy. There is also the need to control data leakage, patient-level validation, minority-class evaluation, temporal integrity, and model interpretability, which are important to take into account when developing reliable intelligent healthcare systems.

The study shows that the evaluation of the clinical reality can lead to a more moderate performance, while the evaluation of the optimistic results due to improper separation of data provides a higher performance figure. But such findings give better insights into what machine-learning systems can do and what they cannot. The proposed framework further indicates that multimodal learning is not the sum of parts of several data sources. Prediction in multimodal settings needs to have enough patient overlap, prediction targets to be compatible, informative modality-specific prediction models, and correct prediction fusion strategies.

Likewise, a balanced dataset is not the only measure of the success of generative augmentation. The benefit should be evaluated based on improvements in clinically relevant performance on real but unseen problems.The integration of SHAP and Grad-CAM further illustrates its value in ensuring transparency in AI used in healthcare. The methods of explainability do not imply causality, but they can give meaning to the workings of predictive models and can help to develop more interpretable systems for clinical decision support in the future.

5.7 Chapter Summary

This chapter outlined the results, contributions, limitations and future directions of the proposed multimodal explainability framework for Parkinson's Disease analysis. The study showed the different and partly complementary nature of clinical, longitudinal, MRI and handwriting information. Logistic Regression achieved the best balanced performance for the classification of the clinical motor subtype and LSTM performed slightly better than the others in terms of longitudinal progression classification; CTGAN gave a slight improvement in the minority-class sensitivity; structural MRI had limited 3-class subtype discrimination; and spiral hand-writing had useful probability-based discrimination but limited sensitivity for PD.

The clinical–MRI late-fusion experiment also showed that multimodal integration does not necessarily perform better than the best single modality. Rather, the advantages of multimodal learning rely on the quality and complementarity of the modality specific predictions. In summary, the proposed work offers a framework that integrates machine learning, temporal modelling, generative augmentation, transfer learning, explainability and multimodal fusion while accounting for leakages in the system for Parkinson's disease research. The results set the baseline for future work of larger paired cohorts, extended modalities, better temporal architectures, better generative models, and learned multimodal fusion strategies.

Bibliography

[1] S. Kalia and A. Lang, "Parkinson's disease," *The Lancet*, vol. 386, no. 9996, pp. 896–912, 2015.

[2] A. Poewe *et al.*, "Parkinson disease," *Nature Reviews Disease Primers*, vol. 3, pp. 1–21, 2017.

[3] K. R. Chaudhuri and A. H. V. Schapira, "Non-motor symptoms of parkinson's disease: dopaminergic pathophysiology and treatment," *The Lancet Neurology*, vol. 8, no. 5, pp. 464–474, 2009.

[4] M. Bloem *et al.*, "Parkinson's disease," *Nature Reviews Disease Primers*, vol. 7, no. 1, pp. 1–21, 2021.

[5] S. Fereshtehnejad *et al.*, "New clinical subtypes of parkinson disease," *Neurology*, vol. 86, pp. 1–10, 2016.

[6] C. G. Goetz, B. C. Tilley, S. R. Shaftman, G. T. Stebbins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M. B. Stern *et al.*, "Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): Scale presentation and clinimetric testing results," *Movement Disorders*, vol. 23, no. 15, pp. 2129–2170, 2008.

[7] B. Postuma *et al.*, "Mds clinical diagnostic criteria for parkinson's disease," *Movement Disorders*, vol. 30, no. 12, pp. 1591–1601, 2015.

[8] D. Aarsland *et al.*, "Cognitive decline in parkinson disease," *Nature Reviews Neurology*, vol. 17, pp. 217–231, 2021.

[9] A. Tolosa *et al.*, "The diagnosis of parkinson's disease," *The Lancet Neurology*, vol. 20, no. 5, pp. 385–397, 2021.

[10] A. H. V. Schapira *et al.*, "Parkinson disease," *Nature Reviews Disease Primers*, vol. 9, no. 1, 2023.

[11] W. Dauer and S. Przedborski, "Parkinson's disease: Mechanisms and models," *Neuron*, 2022.

[12] V. L. Feigin *et al.*, "Global burden of neurological disorders," *The Lancet Neurology*, 2024.

[13] M. M. Hoehn and M. D. Yahr, "Parkinsonism: Onset, progression and mortality," *Neurology*, vol. 17, no. 5, pp. 427–442, 1967.

[14] M. J. Armstrong, "Diagnosis and treatment of parkinson disease," *JAMA*, 2023.

[15] S. Sveinbjornsdottir, "The clinical symptoms of parkinson's disease," *Journal of Neurochemistry*, vol. 139, pp. 318–324, 2016.

[16] J. Jankovic, "Parkinson's disease: Clinical features and diagnosis," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 79, no. 4, pp. 368–376, 2008.

[17] D. Aarsland *et al.*, "Non-motor symptoms of parkinson's disease," *Nature Reviews Neurology*, 2023.

[18] J. Jankovic *et al.*, "Variable expression of parkinson's disease," *Neurology*, vol. 40, pp. 1529–1534, 1990.

[19] A. Lewis *et al.*, "The tremor dominant and pigd phenotypes of parkinson's disease," *Journal of Neural Transmission*, vol. 112, pp. 183–193, 2005.

[20] K. Marek *et al.*, "The parkinson progression marker initiative (ppmi) – establishing a pd biomarker cohort," *NPJ Parkinson's Disease*, 2023.

[21] G. T. Stebbins, C. G. Goetz, D. J. Burn, J. Jankovic, T. K. Khoo, and B. C. Tilley, "How to identify tremor dominant and postural instability/gait difficulty groups with the movement disorder society unified parkinson's disease

rating scale: Comparison with the unified parkinson's disease rating scale," *Movement Disorders*, vol. 28, no. 5, pp. 668–670, 2013.

[22] M. Lawton *et al.*, "Parkinson's disease subtypes in the oxford discovery cohort: A machine learning approach," *Movement Disorders*, 2022.

[23] S.-M. Fereshtehnejad *et al.*, "Evolution of parkinson's disease subtypes over time," *Journal of Parkinson's Disease*, vol. 7, no. 2, pp. 413–422, 2017.

[24] A. Dadu, V. Satone, R. Kaur, S. H. Hashemi, H. Leonard, H. Iwaki, M. B. Makarious *et al.*, "Identification and prediction of parkinson's disease subtypes and progression using machine learning in two cohorts," *npj Parkinson's Disease*, vol. 8, p. 172, 2022.

[25] L. J. W. Evers, J. H. Krijthe, M. J. Meinders, B. R. Bloem, and T. M. Heskes, "Measuring parkinson's disease over time: The real-world within-subject reliability of the mds-updrs," *Movement Disorders*, vol. 34, no. 10, pp. 1480–1487, 2019.

[26] J. C. Vasquez-Correa *et al.*, "Multimodal artificial intelligence for parkinson's disease assessment: A review," *Information Fusion*, vol. 104, p. 102159, 2024.

[27] S. Moghadam *et al.*, "Explainable multimodal deep learning for parkinson's disease prediction: A review," *IEEE Access*, vol. 12, pp. 42 115–42 136, 2024.

[28] M. Sana *et al.*, "Artificial intelligence for parkinson's disease prediction: Current challenges and future directions," *Biomedical Signal Processing and Control*, vol. 90, p. 105812, 2024.

[29] Y. Li *et al.*, "Longitudinal deep learning models for parkinson's disease progression prediction," *Computer Methods and Programs in Biomedicine*, vol. 248, p. 108079, 2024.

[30] M. B. Makarious, H. L. Leonard, D. Vitale, H. Iwaki, L. Sargent, A. Dadu *et al.*, "Multi-modality machine learning predicting parkinson's disease," *npj Parkinson's Disease*, vol. 8, no. 1, p. 35, 2022.

[31] C. R. Pereira *et al.*, "Machine learning strategies for parkinson's disease progression prediction," *IEEE Reviews in Biomedical Engineering*, vol. 16, pp. 381–397, 2023.

[32] K. Marek *et al.*, "Parkinson progression marker initiative: Longitudinal biomarker discovery," *npj Parkinson's Disease*, vol. 10, no. 1, 2024.

[33] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 4765–4774.

[34] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you? explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[35] M. A. Khan *et al.*, "Artificial intelligence and deep learning for parkinson's disease diagnosis: A comprehensive review," *Expert Systems with Applications*, vol. 245, p. 123456, 2024.

[36] Y. Zhang *et al.*, "Recent advances in multimodal deep learning for parkinson's disease," *Biomedical Signal Processing and Control*, vol. 90, p. 105840, 2024.

[37] T. Simuni *et al.*, "Longitudinal clinical and biomarker progression in parkinson's disease," *Nature Medicine*, vol. 30, pp. 1002–1013, 2024.

[38] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, no. 1, pp. 44–56, 2019.

[39] V. Skaramagkas, A. Pentari, Z. Kefalopoulou, and M. Tsiknakis, "Multi-modal deep learning diagnosis of parkinson's disease: A systematic review," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 2399–2423, 2023.

[40] T. Tabashum, R. C. Snyder, M. K. O'Brien, and M. V. Albert, "Machine learning models for parkinson disease: Systematic review," *Journal of Medical Internet Research*, vol. 26, p. e50117, 2024.

[41] I. Anjum *et al.*, "Application of deep learning models for automated identification of parkinson's disease: A review (2011–2021)," *Sensors*, vol. 21, no. 21, p. 7034, 2021.

[42] A. Zhao, Y. Liu, X. Yu, and X. Xing, "Artificial intelligence-enabled detection and assessment of parkinson's disease using multimodal data: A survey," *Information Fusion*, vol. 121, p. 103175, 2025.

[43] M. Alafif and S. Teo, "Machine learning approaches for parkinson's disease diagnosis: A review," *Diagnostics*, vol. 13, no. 5, p. 870, 2023.

[44] M. S. Islam and M. R. Islam, "Recent advances in deep learning for parkinson's disease diagnosis using multimodal data," *Biomedical Signal Processing and Control*, vol. 91, p. 105915, 2024.

[45] J. Mei, Y. Desrosiers, and S. Frasnelli, "Random forest-based parkinson's disease detection using multimodal clinical features," *IEEE Access*, vol. 11, pp. 112 345–112 357, 2023.

[46] A. Prashanth, S. Roy, and P. Gupta, "Logistic regression and classical machine learning models for parkinson's disease diagnosis: A comparative study," *Biomedical Signal Processing and Control*, vol. 83, p. 104681, 2023.

[47] H. Zhang, Y. Li, and X. Wang, "Extreme gradient boosting for early parkinson's disease prediction using clinical and speech features," *Expert Systems with Applications*, vol. 236, p. 121378, 2024.

[48] R. Ghosh and A. K. Singh, "Deep learning techniques for parkinson's disease diagnosis and progression prediction: A review," *IEEE Access*, vol. 12, pp. 32 870–32 895, 2024.

[49] M. Alafif, A. Tehame, and A. Bajaba, "Deep learning applications for parkinson's disease diagnosis and progression prediction: A systematic review," *IEEE Access*, vol. 12, pp. 24 568–24 595, 2024.

[50] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, vol. 25, 2012, pp. 1097–1105.

[51] K. Muhammad, S. Khan, J. D. Ser, and V. H. C. de Albuquerque, "Convolutional neural networks for parkinson's disease diagnosis using medical imaging: Recent advances and challenges," *Information Fusion*, vol. 99, p. 101912, 2024.

[52] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[53] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986.

[54] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[55] S. Alharthi, M. Islam, and M. Hossain, "Lstm-based longitudinal modeling for parkinson's disease progression prediction," *Computer Methods and Programs in Biomedicine*, vol. 244, p. 107938, 2024.

[56] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.

[57] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017, pp. 5998–6008.

[58] Y. Liu *et al.*, "Multimodal neuroimaging for parkinson's disease diagnosis and progression assessment: A deep learning review," *Frontiers in Neuroscience*, vol. 16, p. 944245, 2022.

[59] X. Peng *et al.*, "Deep learning in medical imaging for parkinson's disease diagnosis: A review," *Diagnostics*, vol. 13, no. 6, p. 1094, 2023.

[60] H. Wang *et al.*, "Magnetic resonance imaging biomarkers and deep learning for parkinson's disease diagnosis," *Frontiers in Aging Neuroscience*, vol. 15, p. 1182054, 2023.

[61] Y. Zhou *et al.*, "Pet imaging combined with artificial intelligence for parkinson's disease diagnosis: A systematic review," *European Journal of Nuclear Medicine and Molecular Imaging*, vol. 51, no. 2, pp. 540–556, 2024.

[62] J. Kim *et al.*, "Deep learning analysis of datscan images for parkinson's disease diagnosis: A review," *Journal of Clinical Medicine*, vol. 13, no. 5, p. 1347, 2024.

[63] T. Alhanai *et al.*, "Deep learning for speech-based detection and monitoring of parkinson's disease: A review," *IEEE Access*, vol. 11, pp. 115 620–115 641, 2023.

[64] A. Ismail *et al.*, "Deep learning-based handwriting and spiral drawing analysis for parkinson's disease detection: A review," *Biomedical Signal Processing and Control*, vol. 91, p. 105980, 2024.

[65] R. Gupta, S. Kumar *et al.*, "Deep learning-based handwriting analysis for parkinson's disease detection: A comprehensive review," *Biomedical Signal Processing and Control*, vol. 92, p. 106018, 2024.

[66] M. A. Rahman *et al.*, "Transfer learning for parkinson's disease detection using handwriting and spiral drawings," *Expert Systems with Applications*, vol. 245, p. 123456, 2024.

[67] P. Singh *et al.*, "Artificial intelligence techniques for spiral drawing analysis in parkinson's disease: A systematic review," *Artificial Intelligence in Medicine*, vol. 152, p. 102856, 2025.

[68] Y. Chen *et al.*, "Wearable sensor technologies and artificial intelligence for parkinson's disease monitoring: A review," *Sensors*, vol. 24, no. 3, p. 945, 2024.

[69] Z. Li *et al.*, "Longitudinal deep learning models for parkinson's disease progression prediction," *NPJ Digital Medicine*, vol. 7, p. 85, 2024.

[70] X. Zhang *et al.*, "Lstm and bilstm networks for longitudinal parkinson's disease progression modeling," *Computer Methods and Programs in Biomedicine*, vol. 247, p. 108060, 2024.

[71] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[72] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, 2021.

[73] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should i trust you? explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.

[74] W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Müller, "Explainable ai: Interpreting, explaining and visualizing deep learning," *Lecture Notes in Computer Science*, vol. 11700, 2019.

[75] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, vol. 27, 2014, pp. 2672–2680.

[76] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional gan," in *Advances in Neural Information Processing Systems*, vol. 32, 2019, pp. 7335–7345.

[77] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

[78] L. Govindaraj, M. Vishal *et al.*, "Explainable multimodal feature fusion networks for parkinson's disease prediction," *Frontiers in Digital Health*, vol. 8, 2026.

[79] Y. Huang *et al.*, "Multimodal deep learning for healthcare: A review," *Information Fusion*, vol. 84, pp. 1–22, 2023.

[80] J. Ahmad *et al.*, "Machine learning techniques for parkinson's disease diagnosis: A review," *Diagnostics*, vol. 12, no. 12, p. 3154, 2022.

[81] H. Wang *et al.*, "Deep learning approaches for parkinson's disease diagnosis and progression prediction," *Frontiers in Neurology*, vol. 14, p. 1195842, 2023.

[82] M. M. Islam *et al.*, "Transfer learning-based deep learning models for parkinson's disease detection: A review," *Biomedical Signal Processing and Control*, vol. 91, p. 105965, 2024.

[83] Z. Li *et al.*, "Longitudinal deep learning models for parkinson's disease progression prediction," *npj Digital Medicine*, vol. 7, p. 85, 2024.

[84] X. Zhang *et al.*, "Transformer-based models for longitudinal parkinson's disease progression analysis," *Computer Methods and Programs in Biomedicine*, vol. 247, p. 108060, 2024.

[85] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, 2021.

[86] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[87] I. Goodfellow *et al.*, "Generative adversarial networks," *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[88] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional gan," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[89] B. R. Bloem, M. S. Okun, and C. Klein, "Parkinson's disease," *The Lancet*, vol. 397, no. 10291, pp. 2284–2303, 2021.

[90] Y. Liu *et al.*, "Multimodal artificial intelligence for parkinson's disease: Recent advances, challenges and future directions," *Information Fusion*, vol. 104, p. 102177, 2024.

[91] H. Zhang *et al.*, "Artificial intelligence for parkinson's disease progression prediction: A systematic review," *NPJ Digital Medicine*, vol. 7, p. 118, 2024.

[92] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[93] M. Muzammal, M. Hayat, and S. Khan, "Explainable multimodal deep learning framework for parkinson's disease diagnosis and progression analysis," *Computers in Biology and Medicine*, vol. 177, p. 108684, 2024.

[94] S. M. Lundberg, G. Erion, H. Chen *et al.*, "From local explanations to global understanding with explainable ai for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, 2020.

[95] Z. Zhao, A. Kunar, R. Birke, and L. Y. Chen, "Ctab-gan: Effective table data synthesizing," in *Proceedings of the Asian Conference on Machine Learning (ACML)*, 2021, pp. 97–112.

[96] K. Marek, J. Seibyl, S. Eberly *et al.*, "Longitudinal follow-up of swedd subjects in the parkinson's progression markers initiative," *Neurology*, vol. 82, no. 20, pp. 1791–1797, 2014.