final_merged.pdf

Shri Vile Parle Kelavani Mandal

Document Details

Submission ID
trn:oid:::9832:140908329

Submission Date
May 29, 2026, 11:01 AM GMT+5:30

Download Date
May 29, 2026, 11:18 AM GMT+5:30

File Name
final_merged.pdf

File Size
526.1 KB

51 Pages

13,141 Words

78,256 Characters

13% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 8 words)
- Internet sources

Match Groups

129 Not Cited or Quoted 13%
Matches with neither in-text citation nor quotation marks

0 Missing Quotations 0%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

0% Internet sources
3% Publications
12% Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- **129** Not Cited or Quoted 13%
  Matches with neither in-text citation nor quotation marks
- **0** Missing Quotations 0%
  Matches that are still very similar to source material
- **0** Missing Citation 0%
  Matches that have quotation marks, but no in-text citation
- **0** Cited and Quoted 0%
  Matches with in-text citation present, but no quotation marks

Top Sources

0% Internet sources
3% Publications
12% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Submitted works	Delhi Technological University on 2026-05-28 <1%
2	Submitted works	Delhi Technological University on 2026-05-29 <1%
3	Submitted works	Delhi Technological University on 2026-05-07 <1%
4	Submitted works	Delhi Technological University on 2026-05-26 <1%
5	Submitted works	Strathmore University (Main Account) on 2026-03-26 <1%
6	Submitted works	Tshwane University of Technology on 2019-04-16 <1%
7	Submitted works	Charles University in Prague on 2026-04-30 <1%
8	Submitted works	The University of the West of Scotland on 2026-04-23 <1%
9	Submitted works	University College London on 2026-04-26 <1%
10	Submitted works	University of Sheffield on 2025-05-23 <1%

11 Submitted works
Heriot-Watt University on 2025-11-20 <1%

12 Publication
Yuandong Liao, Wenyong Li, Guan Lian, Junzhuo Li. "A dual-path GNN with two-st... <1%

13 Submitted works
University of East London on 2026-05-08 <1%

14 Submitted works
Colorado Technical University Online on 2026-05-24 <1%

15 Submitted works
University of Ulster on 2026-04-26 <1%

16 Submitted works
King's College on 2026-04-22 <1%

17 Submitted works
Delhi Technological University on 2026-05-27 <1%

18 Submitted works
Tilburg University on 2026-05-25 <1%

19 Submitted works
ICTS on 2025-07-15 <1%

20 Submitted works
University of Melbourne on 2023-11-14 <1%

21 Submitted works
Delhi Technological University on 2026-05-21 <1%

22 Submitted works
Delhi Technological University on 2026-05-25 <1%

23 Submitted works
Queen Mary and Westfield College on 2024-08-22 <1%

24 Publication
Srinjay Dasgupta, Debopriya Dey, Debrupa Pal, Samadrita Karmakar, Prabuddha ... <1%

25 Submitted works
University of Bristol on 2026-05-05 <1%

26 Submitted works
African Leadership University on 2026-05-24 <1%

27 Publication
Mahmoud Essam, Mazen Balat, Ahmed B. Zaky, Mervat Samy, Ahmed M. Anter. "A... <1%

28 Submitted works
Sabanci Universitesi on 2026-02-17 <1%

29 Submitted works
Glyndwr University on 2025-08-08 <1%

30 Submitted works
Johns Hopkins Unversity on 2025-03-03 <1%

31 Submitted works
Tilburg University on 2026-05-23 <1%

32 Submitted works
University of Leeds on 2026-05-11 <1%

33 Publication
Erik Štrumbelj, Igor Kononenko. "Explaining prediction models and individual pre... <1%

34 Publication
Ji Zhang, Zhixuan Wang, Zhuo Wang, Haibo Yuan, Liansheng Cheng, Zhenhua Bai.... <1%

35 Submitted works
Tilburg University on 2026-05-24 <1%

36 Submitted works
Uttar Pradesh Technical University on 2019-05-13 <1%

37 Submitted works
Tilburg University on 2026-05-25 <1%

38 Submitted works
Tilburg University on 2026-05-28 <1%

39 Submitted works
Islington College,Nepal on 2026-05-12 <1%

40 Submitted works
Lead College Pty Ltd on 2024-03-23 <1%

41 Submitted works
Sunway Education Group on 2026-04-19 <1%

42 Submitted works
Bocconi University on 2026-03-12 <1%

43 Submitted works
University of Cape Town on 2025-11-01 <1%

44 Submitted works
University of Exeter on 2024-12-08 <1%

45 Submitted works
Delhi Technological University on 2026-04-30 <1%

46 Submitted works
Delhi Technological University on 2026-05-28 <1%

47 Publication
Qingchuan Zhang, Zhenqiao Liu, Zexi Song, Shaoyi Song, Xuan Li, Zihan Li, Min Zu... <1%

48 Submitted works
Tilburg University on 2026-05-19 <1%

49 Submitted works
Copenhagen Business School on 2026-02-09 <1%

50 Submitted works
JIS University on 2025-11-04 <1%

51 Submitted works
Malta College of Arts,Science and Technology on 2026-05-23 <1%

52 Submitted works
UCL on 2025-09-08 <1%

53 Submitted works
University of Central Florida on 2018-11-14 <1%

54 Submitted works
University of East London on 2025-09-06 <1%

55 Submitted works
University of Edinburgh on 2025-04-18 <1%

56 Submitted works
University of Surrey on 2023-01-30 <1%

57 Submitted works
University of Sydney on 2025-10-26 <1%

58 Submitted works
City University of Hong Kong on 2025-04-11 <1%

59 Submitted works
Strathmore University (Main Account) on 2026-03-17 <1%

60 Submitted works
The University of Manchester on 2024-04-15 <1%

61 Submitted works
University of Central Lancashire on 2026-05-03 <1%

62 Submitted works
University of East London on 2026-05-08 <1%

63 Submitted works
University of Exeter on 2024-05-01 <1%

64 Submitted works
African Leadership University on 2025-11-17 <1%

65 Submitted works
Australian National University on 2024-03-26 <1%

66 Submitted works
Dublin Business School on 2026-05-11 <1%

67 Publication

Fatima Zahra Salmam, Abdellah Madani, Mohamed Kissi. "Emotion Recognition fr... <1%

68 Submitted works

Middle East Technical University on 2016-05-23 <1%

69 Submitted works

Rijksuniversiteit Groningen - Tii on 2026-04-16 <1%

70 Submitted works

University of Oulu on 2026-05-28 <1%

71 Submitted works

University of Surrey on 2023-05-17 <1%

72 Submitted works

Ajman University on 2026-03-11 <1%

73 Submitted works

Indian Institute of Science Education and Research on 2025-05-20 <1%

74 Publication

K. Raajasree, R. Jaichandran. "Enhanced EfficientNet-Extended Multimodal Parkin... <1%

75 Submitted works

Tilburg University on 2026-05-21 <1%

76 Submitted works

Universidad Nacional de Colombia on 2025-05-16 <1%

77 Submitted works

University of Greenwich on 2024-11-29 <1%

78 Submitted works

Associatie K.U.Leuven on 2019-08-19 <1%

79 Publication

David Cortés-Polo, Jesús Calle-Cancho, Mercedes E. Paoletti, Juan M. Haut. "Transf... <1%

80 Submitted works

FH Campus Wien on 2026-05-28 <1%

81 Submitted works
Hamdan Bin Mohammed Smart University on 2026-04-25 <1%

82 Submitted works
Indian Institute of Technology Mandi on 2026-05-11 <1%

83 Submitted works
Islington College,Nepal on 2026-05-10 <1%

84 Submitted works
Liverpool John Moores University on 2023-12-17 <1%

85 Submitted works
London School of Economics and Political Science on 2024-08-07 <1%

86 Publication
Melanie Maliti, Aaron Zimba. "A White-box Approach to Forecasting Petrol Prices i... <1%

87 Submitted works
National Institute of Technology, Warangal, Telangana, India on 2026-04-24 <1%

88 Submitted works
Selçuk Üniversitesi on 2017-01-17 <1%

89 Submitted works
Technical University of Košice on 2026-05-26 <1%

90 Submitted works
Tilburg University on 2026-05-25 <1%

91 Submitted works
Universiti Malaysia Kelantan on 2026-05-08 <1%

92 Submitted works
Universiti Teknologi Malaysia on 2016-04-21 <1%

93 Submitted works
University of Birmingham on 2025-08-31 <1%

94 Submitted works
University of Central England in Birmingham on 2022-12-16 <1%

95 Submitted works

University of East London on 2026-05-08 <1%

96 Submitted works

University of Essex on 2024-08-27 <1%

97 Submitted works

University of Nottingham on 2025-09-23 <1%

98 Submitted works

University of Technology, Sydney on 2024-11-05 <1%

99 Submitted works

University of Westminster on 2026-03-30 <1%

EXPLAINABLE TEMPORAL TRANSFORMER FOR DISEASE PROGRESSION PREDICTION USING ATTENTION AND SHAP ANALYSIS

A Project Report Submitted in Partial Fulfillment of the

Requirements for the Degree of

MASTER OF TECHNOLOGY

in

COMPUTER SCIENCE & ENGINEERING

by

Ajitesh Chhibber

(24/CSE/31)

Under the Supervision of

Prof. Vinod Kumar

Assistant Professor, Department of CSE

Delhi Technological University

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042
May 2026

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042. India

CANDIDATE'S DECLARATION

I Aman Kumar Gaur hereby certify that the work which is being presented in the thesis entitled **"Explainable Temporal Transformer for Disease Progression Prediction using Attention and SHAP Analysis"** in partial fulfillment of the requirements for the award of the Master of Technology Degree, submitted in the Department of Computer Science and Engineering, Delhi Technological University is an authentic record of my own work carried out during the period from [Start Date, e.g., August 2025] to May 2026 under the supervision of **Prof. Vinod Kumar**

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

_____

Candidate's Signature

This is to certify that the student has incorporated all the corrections suggested by the examiners in the thesis and the statement made by the candidate is correct to the best of our knowledge.

Signature of Supervisor(s) **Signature of External Examiner**

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Bawana Road, Delhi-110042. India

CERTIFICATE

I hereby Certified that **Ajitesh Chhibber (24/CSE/31)** has carried out their research work presented in this thesis entitled **"Explainable Temporal Transformer for Disease Progression Prediction using Attention and SHAP Analysis"** for the award of Master of Technology from Department of Computer Science and Engineering, Delhi Technological University, Delhi, under my supervision.

The thesis embodies results of original work, and studies are carried out by the student himself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Place: Delhi

Date:

Prof. Vinod Kumar
Professor
Delhi Technological University

ABSTRACT

Parkinson's disease is a progressive neurodegenerative disorder characterised by the gradual deterioration of motor function over time. Accurately forecasting this progression is of considerable clinical relevance, as it can guide treatment decisions, support early intervention strategies, and improve patient outcomes. The Parkinson Telemonitoring dataset, sourced from the UCI Machine Learning Repository, provides a longitudinal record of 5,875 voice measurements collected from 42 patients, making it a natural substrate for temporal modelling approaches. Yet despite the sequential nature of this data, the dominant paradigm in the literature has relied on traditional machine learning models that process each observation in isolation, ignoring the temporal context in which successive measurements are embedded.

This thesis investigates whether Transformer-based temporal modelling can improve the prediction of motor_UPDRS scores relative to classical and recurrent baselines, and whether the resulting models can be made interpretable enough to be clinically useful. Three model families are developed and systematically evaluated: a Random Forest baseline representing the conventional supervised learning paradigm, a Long Short-Term Memory network capturing sequential dependencies through gated recurrence, and a Transformer model leveraging multi-head self-attention to model dependencies across arbitrary time-step distances without the vanishing-gradient limitations of recurrent architectures.

Explainability is treated not as an afterthought but as a first-class design requirement. Two complementary interpretation frameworks are integrated: attention weight visualisation, which reveals which temporal observations within an input sequence the model considers most informative, and SHAP (SHapley Additive exPlanations) analysis [2], which decomposes the contribution of individual input features to each prediction in a theoretically grounded manner rooted in cooperative game theory. Together, these mechanisms transform the Transformer from a black-box regression system into a clinically interrogable diagnostic aid.

Experiments conducted on the full 5,875-sample dataset demonstrate that the Transformer model achieves the best predictive accuracy among all three approaches, with performance improving substantially when longer temporal sequence lengths are used—a finding that confirms the clinical hypothesis that disease progression signals accumulate over extended observation windows. The SHAP analysis identifies the most influential vocal biomarkers driving the predictions, providing actionable insights that align with the established neurological understanding of Parkinson's disease symptom progression.

Keywords: Parkinson's disease, disease progression, temporal modelling, Transformer,

LSTM, SHAP analysis, attention mechanism, explainable AI, motor UPDRS, telemonitoring.

ACKNOWLEDGEMENTS

I am grateful to the faculty members of the Department of Computer Science and Engineering, Delhi Technological University, for their consistent academic guidance and institutional support throughout the course of this programme.

I express my sincere gratitude to my project supervisor for the invaluable direction, critical feedback, and encouragement provided at every stage of this research. The breadth of insight offered during our discussions shaped the scientific rigour of this work in ways that are difficult to quantify.

I also acknowledge the UCI Machine Learning Repository for making the Parkinson Telemonitoring dataset publicly available, without which this investigation would not have been possible.

Finally, I thank my family and friends whose patience and moral support sustained this effort from beginning to end.

[Candidate Name]
Roll Number

Contents

List of Figures

List of Tables

LIST OF SYMBOLS AND ABBREVIATIONS

Symbol / Abbrev.	Expansion
PD	Parkinson's Disease
UPDRS	Unified Parkinson's Disease Rating Scale
motor_UPDRS	Motor component of UPDRS (target variable)
total_UPDRS	Total UPDRS score (removed to prevent data leakage)
LSTM	Long Short-Term Memory
RF	Random Forest
XAI	Explainable Artificial Intelligence
SHAP	SHapley Additive exPlanations
MHA	Multi-Head Attention
NLP	Natural Language Processing
UCI	University of California, Irvine
MAE	Mean Absolute Error
RMSE	Root Mean Squared Error
R^2	Coefficient of Determination
FFN	Feed-Forward Network
d_{model}	Model dimensionality in Transformer
T	Temporal sequence length
h	Number of attention heads
$\mathbf{Q}, \mathbf{K}, \mathbf{V}$	Query, Key, Value matrices in self-attention
ϕ_i	SHAP value for feature i
GPU	Graphics Processing Unit
DTU	Delhi Technological University

Chapter 1

Introduction

1.1 Overview and Background

Parkinson's disease ranks among the most prevalent neurodegenerative conditions worldwide, affecting an estimated ten million individuals globally and constituting the second most common neurodegenerative disorder after Alzheimer's disease. Clinically, the condition is characterised by the progressive loss of dopaminergic neurons in the substantia nigra region of the basal ganglia, manifesting in the classical motor triad of resting tremor, rigidity, and bradykinesia, alongside a range of non-motor symptoms including cognitive decline, sleep disturbances, and autonomic dysfunction. What makes Parkinson's disease particularly challenging from both clinical and computational perspectives is its progressive, longitudinal nature: the severity of symptoms does not remain static but evolves continuously over months and years, and the rate of this evolution varies considerably across patients.

The Unified Parkinson's Disease Rating Scale (UPDRS) was developed as a standardised clinical instrument for quantifying disease severity. Its motor sub-scale, motor_UPDRS, captures the degree of functional motor impairment and has been widely adopted both in clinical trials and in remote monitoring studies as the primary outcome measure for tracking progression. Reliable prediction of future motor_UPDRS trajectories would allow clinicians to anticipate deterioration before it becomes clinically apparent, enabling timely adjustments to dopamine agonist therapy, scheduling of physiotherapeutic interventions, and better-informed discussions with patients and their caregivers.

The emergence of remote telemonitoring technologies has made continuous, non-invasive data collection feasible outside the hospital setting. Voice measurements, in particular, offer a rich source of biomarkers because Parkinson's disease produces distinctive dysphonia effects that can be captured inexpensively using standard microphones. The Parkinson

Telemonitoring dataset compiled by Athanasios Tsanas and colleagues [9] contains 5,875 voice recordings from 42 patients, each annotated with a clinician-assigned motor UPDRS score, providing an unusually dense longitudinal record of symptom evolution over time.

Despite the sequential, time-indexed nature of this data, much of the published literature has approached the prediction problem using classical supervised learning models—decision trees, support vector regressors, random forests—that treat each observation as an independent sample and make no attempt to exploit the temporal context in which successive measurements are embedded [8, 12]. This is a conceptually awkward mismatch: disease progression is inherently a trajectory, and a model that cannot represent trajectories is limited in what it can learn about progression.

The past decade has produced two families of models capable of temporal reasoning. Recurrent neural networks, particularly LSTM networks [16], learn sequential dependencies through a gating mechanism that controls information flow over time. Transformer architectures [1], originally proposed for machine translation, replace recurrence with a fully parallelisable self-attention mechanism that directly models pairwise interactions between all positions in the input sequence. On many sequential modelling tasks, Transformers have outperformed LSTMs by avoiding the vanishing-gradient bottleneck that limits the effective memory horizon of recurrent models. Whether this advantage extends to longitudinal biomedical tabular data of the kind found in the Parkinson Telemonitoring dataset is a question that this thesis is specifically designed to answer.

1.2 Motivation for Explainability

Predictive accuracy alone is an insufficient standard for clinical deployment. A model that achieves low mean absolute error on held-out test data but offers no account of why it makes a given prediction cannot easily be audited, challenged, or trusted by a clinician who must ultimately decide whether to act on its output. This concern is not merely philosophical: regulatory frameworks such as the European Union's General Data Protection Regulation mandate a right to explanation for automated decisions, and healthcare-specific guidelines increasingly require that AI-assisted diagnostic tools provide justifications that clinicians can evaluate against their own domain knowledge.

The explainability challenge is particularly acute for Transformer models. The self-attention weights that drive prediction provide a natural internal representation of which temporal positions the model finds informative, but translating these weights into actionable clinical insight requires careful visualisation and interpretation. At the feature level—which vocal biomarkers matter most, and by how much—attention weights alone are insufficient be-

cause they operate on temporal positions rather than individual input dimensions. This gap motivates the integration of SHAP (SHapley Additive exPlanations) [2], a feature attribution framework grounded in cooperative game theory that computes theoretically principled contributions for each input feature.

This thesis treats explainability as a structural requirement rather than a post-hoc diagnostic. The attention visualisation and SHAP pipelines are designed and implemented in parallel with the predictive models, and their outputs are analysed alongside predictive performance metrics. The goal is a system that is not merely accurate but interpretable in a manner that can support rather than supplant clinical reasoning.

1.3 Problem Statement

The central problem addressed in this thesis is the following: given a longitudinal sequence of voice measurements collected from a Parkinson's patient over time, can a Transformer-based temporal model predict the motor_UPDRS score at the current time step more accurately than either a classical Random Forest baseline or an LSTM recurrent baseline, and can the model's predictions be explained at both the temporal and feature levels in a manner accessible to clinical practitioners? This problem decomposes into four sub-questions.

First, does explicit temporal sequence modelling—using either LSTM or Transformer architectures—improve prediction accuracy over the feature-level Random Forest baseline that treats observations as independent? Second, does the Transformer's self-attention mechanism outperform LSTM recurrence on this particular dataset, and if so, at what sequence lengths does the advantage become significant? Third, do the attention weights learned by the Transformer exhibit interpretable patterns—for example, assigning higher weights to more recent observations or to observations with extreme biomarker values—that are consistent with clinical knowledge of disease dynamics? Fourth, which input features does SHAP analysis identify as the most influential drivers of predicted motor_UPDRS scores, and do these features align with the neurological symptom profile of Parkinson's disease?

1.4 Objectives of the Thesis

The research programme pursued in this thesis is organised around six primary objectives. The first is to characterise the Parkinson Telemonitoring dataset in sufficient depth to justify the preprocessing decisions made in the methodology, including an analysis of feature distributions, missing values, and inter-feature correlations. The second is to design and implement a preprocessing pipeline that transforms the raw longitudinal records into se-

quential input tensors suitable for both LSTM and Transformer architectures. The third is to train and optimise three models—Random Forest, LSTM, and Transformer—under consistent experimental conditions using the same data splits and evaluation metrics. The fourth is to conduct a comparative evaluation of all three models using Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and the coefficient of determination (R^2), and to interpret the performance differences in terms of each architecture's capacity for temporal reasoning. The fifth is to implement and evaluate attention weight visualisation as a means of explaining which temporal steps the Transformer model considers most informative for a given prediction. The sixth is to apply SHAP analysis to the trained Transformer model and produce feature importance rankings that are both statistically well-founded and clinically interpretable.

1.5 Contributions of the Thesis

The primary contributions of this work are as follows. A systematic comparison of three model families—classical, recurrent, and attention-based—is conducted on the Parkinson Telemonitoring dataset under identical experimental conditions, providing a controlled assessment of the marginal value added by each architectural advance. A dual explainability framework combining temporal attention visualisation with feature-level SHAP analysis is developed and applied to the Transformer model, demonstrating that these two mechanisms provide complementary rather than redundant insights. The effect of sequence length on Transformer performance is empirically investigated, establishing that longer observation windows yield measurably better predictions—a finding with direct implications for the design of telemonitoring data collection protocols. Finally, the SHAP feature importance analysis produces a ranked profile of the vocal biomarkers that most reliably predict motor_UPDRS progression, which can serve as a starting point for reduced-feature monitoring instruments in resource-constrained clinical settings.

1.6 Thesis Organisation

The remainder of this thesis is structured as follows. Chapter 2 reviews the relevant literature across three domains: machine learning and deep learning approaches to Parkinson's disease prediction, Transformer architectures in healthcare, and explainable AI frameworks with particular attention to SHAP. Chapter 3 provides a detailed characterisation of the Parkinson Telemonitoring dataset and describes the complete preprocessing pipeline. Chapter 4 formalises the prediction problem mathematically and motivates the modelling choices made in the methodology. Chapter 5 presents the architecture, training procedure, and explainability integration for each of the three model families. Chapter 6 reports and

analyses the experimental results, including comparative performance metrics, attention visualisations, and SHAP feature importance profiles. Chapter 7 synthesises the findings, discusses limitations, and outlines six directions for future research.

turnitin Page 27 of 61 - Integrity Submission Submission ID trn:oid:::9832:140908329

Chapter 2

Literature Review

2.1 Overview

The literature relevant to this thesis spans three intersecting research communities: the clinical informatics and biomedical machine learning community working on Parkinson's disease diagnosis and progression modelling, the deep learning community developing sequence models for temporal healthcare data, and the explainable AI community producing model-agnostic interpretation frameworks applicable to clinical decision support systems. This chapter reviews the most significant contributions in each area, traces the development of ideas that inform the design of the proposed system, and identifies the specific research gap that this thesis addresses.

2.2 Machine Learning for Parkinson's Disease Prediction

Early computational approaches to Parkinson's disease characterisation were largely concerned with binary classification—distinguishing patients from healthy controls—rather than with the regression problem of severity quantification. Decision tree ensembles, support vector machines, and neural networks trained on vocal features extracted from sustained phonation recordings achieved classification accuracies consistently above ninety per cent on small benchmark datasets [12], establishing that dysphonic biomarkers carry substantial discriminative information. These results were methodologically significant but clinically limited, since a system that can only say whether a patient has Parkinson's disease is of marginal utility in a monitored population where the diagnosis is already confirmed.

The shift toward severity quantification and progression prediction gained traction with the public release of longitudinal telemonitoring datasets, of which the Parkinson dataset from the UCI repository is the most widely cited. Tsanas and colleagues, who collected

and published the original dataset, demonstrated that dysphonia measures derived from voice recordings could serve as reliable remote proxies for clinician-administered UPDRS assessments [9]. Their initial models used Gaussian process regression and kernel ridge regression and achieved mean absolute errors in the range of two to four UPDRS points, motivating a substantial body of subsequent work attempting to improve on these baselines.

Random Forest has emerged as the most frequently used baseline across this literature, valued for its robustness to irrelevant features, its resistance to overfitting on moderate-sized datasets, and its built-in feature importance scores [14]. Shokrpour and colleagues [8] provide a comprehensive comparative survey of machine learning algorithms applied to Parkinson's disease prediction tasks, noting that ensemble methods consistently outperform single learners and that careful feature selection is often as consequential as the choice of base algorithm. Chaithanya and colleagues [12] extend this survey to cover more recent gradient-boosting variants and deep learning approaches, observing that the advantage of deep learning over classical methods is more pronounced on larger datasets and on tasks that involve temporal structure.

Multimodal approaches that combine acoustic, kinematic, and imaging biomarkers have also been explored. Pahuja and Prasad [16] investigate a range of deep learning architectures for detection using multimodal feature sets, demonstrating that convolutional feature extractors applied to spectrograms and waveforms can capture aspects of dysphonia not easily represented by hand-crafted features. Yang and colleagues [17] propose a stacking ensemble that combines predictions from multiple feature modalities, achieving competitive performance while maintaining some degree of interpretability through the ensemble's weighting mechanism. These multimodal studies establish that no single biomarker modality is definitively superior; the vocal features used in the current thesis represent a practically accessible subset of a richer phenotypic space.

A recurring limitation noted across this body of work is the treatment of successive patient observations as independent samples. Studies that split the dataset at the observation level rather than the patient level risk inflating performance metrics through temporal leakage—the model implicitly learns patient-specific baselines from training samples that are temporally adjacent to test samples from the same individual. The current thesis addresses this concern explicitly in the preprocessing design and experimental protocol.

2.3 Deep Learning for Temporal Healthcare Data

The application of recurrent neural networks to longitudinal clinical data has a well-established precedent. LSTM networks, in particular, have been applied to electronic health record

modelling [3], vital sign prediction, and medication effect forecasting, demonstrating that gated recurrence can capture clinically meaningful temporal dependencies that static feature vectors miss. In the Parkinson's context, LSTM models have been used to model symptom trajectories from wearable sensor data, with results suggesting that the effective memory horizon of LSTM—typically estimated at fifty to several hundred time steps—is generally adequate for the time scales involved in disease monitoring.

The Transformer architecture, introduced by Vaswani and colleagues [1] in the context of neural machine translation, offers a qualitatively different approach to sequence modelling. Rather than passing information forward through a recurrent state vector of fixed dimensionality, the self-attention mechanism computes a direct weighted combination of all previous positions' representations, where the weights are determined by a learned compatibility function between the current position's query vector and all positions' key vectors. This design eliminates the vanishing-gradient limitation of recurrence and allows the model to selectively attend to any prior time step regardless of distance, at the cost of quadratic attention complexity in the sequence length.

Nerella and colleagues [3] provide a comprehensive survey of Transformer applications in healthcare, cataloguing uses ranging from medical image segmentation and radiology report generation to clinical note processing and drug discovery. Their analysis indicates that Transformers have achieved state-of-the-art results across the majority of clinical NLP benchmarks, and that their advantage over LSTM baselines is most pronounced for tasks requiring long-range dependency modelling. The extension of Transformer architectures beyond text to structured clinical time-series data—the setting relevant to this thesis—is an active and rapidly evolving research direction.

Perumal and Duraisamy [9] represent perhaps the most directly related prior work, applying a Transformer-based time-series model to Parkinson's disease progression data and incorporating XAI components to interpret the results. Their findings confirm the competitive advantage of attention-based models over recurrent baselines on this class of task. Zhu and colleagues [19] propose DT-Transformer, a foundation model trained on a large real-world health system dataset, demonstrating that Transformer architectures scale beneficially with data volume in disease trajectory forecasting. Mirza and colleagues [18] specifically examine the interpretability of Transformer-based disease progression forecasts, noting that the alignment between attention weights and clinically meaningful temporal patterns varies by disease type and must be empirically validated rather than assumed. These three contributions directly inform the design choices made in this thesis.

2.4 Explainable Artificial Intelligence in Healthcare

The deployment of machine learning in clinical decision support has prompted sustained attention from the research community to the problem of model interpretability. Chaddad and colleagues [4] survey explainability techniques across a wide range of clinical applications, distinguishing between ante-hoc methods—models that are interpretable by design, such as linear regression and decision trees—and post-hoc methods that apply explanation algorithms to already-trained black-box models. The consensus view is that ante-hoc interpretability typically comes at an accuracy cost that is unacceptable for complex, high-dimensional biomedical prediction tasks, making post-hoc methods the pragmatic choice for deep learning systems.

SHAP, introduced by Lundberg and Lee [2], has emerged as the most widely adopted post-hoc feature attribution framework in the healthcare domain. Its theoretical foundation in Shapley values from cooperative game theory guarantees a set of axiomatic properties—local accuracy, missingness, and consistency—that alternative attribution methods such as LIME and gradient-based saliency maps do not uniformly satisfy. In the Parkinson's disease context, Jin and colleagues [7] apply SHAP to an ensemble model trained on vocal and gait biomarkers, demonstrating that the resulting feature rankings are clinically coherent and stable across different patient subgroups. The present thesis extends this paradigm to the Transformer setting, where SHAP must be applied to a temporally structured input tensor rather than a flat feature vector.

Band and colleagues [5] review XAI applications in medical health more broadly, identifying the combination of local and global explanation strategies as a best practice: local explanations (per-prediction feature attributions, as computed by SHAP) provide actionable instance-level insights, while global explanations (aggregated feature importance rankings) reveal systematic patterns in what the model has learned from the training data. Mienye and Sun [6] discuss the challenges of deploying XAI in real clinical environments, noting that clinician trust is most effectively built through explanations that align with existing domain knowledge rather than contradicting it. Vani and colleagues [13] apply XAI to personalised health monitoring, demonstrating that SHAP-driven feature selection can reduce model complexity without disproportionate accuracy loss. The current work draws on this insight in its analysis of the SHAP results.

Lai [15] reviews the combination of self-attention and post-hoc explanation in the medical imaging domain, arguing that attention weights and SHAP values provide complementary rather than redundant explanations: attention weights reveal temporal or spatial focus while SHAP values reveal feature-level influence. Panda and Mahanta [14] compare LIME and SHAP explanations for a Random Forest classifier on a healthcare classification task, find-

ing SHAP to be more consistent and more aligned with clinical domain knowledge. Aravindkumar and colleagues [20] provide a systematic review of XAI techniques in healthcare, noting the growing consensus that evaluation of XAI systems should include both technical metrics (fidelity, stability) and human-centred metrics (clinician acceptance). These considerations inform the evaluation framework adopted in Chapter 6.

Li and colleagues [10] introduce PIDGN, an explainable multimodal deep learning framework for early prediction of Parkinson's disease that integrates graph neural networks with attention-based explanation. Their framework explicitly validates that the explanation outputs match expert clinical judgement, setting a methodological precedent for the current thesis's treatment of SHAP results as an empirical claim subject to domain-knowledge verification rather than an automatically trustworthy output.

2.5 Identified Research Gap

The literature review reveals several critical observations. While temporal modelling approaches have been applied to Parkinson's disease data, the specific combination of Transformer architectures with a dual explainability framework—integrating both temporal attention visualisation and feature-level SHAP analysis—on the Parkinson Telemonitoring dataset has not been comprehensively investigated. Additionally, most prior studies treat either prediction performance or explainability, but rarely both with equal rigour. The effect of temporal sequence length on Transformer performance in this particular clinical domain has also not been systematically characterised. The present thesis fills these gaps by providing controlled, rigorous experiments that jointly optimise for predictive accuracy and clinical interpretability.

2.6 Chapter Summary

Three bodies of literature directly relevant to this thesis have been reviewed. Machine learning approaches to Parkinson's disease prediction establish Random Forest as the dominant classical baseline but identify the neglect of temporal structure as a persistent limitation. The Transformer literature demonstrates consistent superiority over LSTM baselines on long-range temporal reasoning tasks and motivates the application of self-attention to longitudinal clinical data. The XAI literature establishes SHAP as the theoretically preferred feature attribution method and identifies the combination of temporal and feature-level explanation as a best practice for clinical AI. The gap at the intersection of these three bodies of literature—a temporally aware, Transformer-based prediction model with dual explainability on the Parkinson Telemonitoring dataset—defines precisely the contribution of this

thesis.

turnitin Page 33 of 61 - Integrity Submission Submission ID trn:oid:::9832:140908329

Chapter 3

Dataset Description and Preprocessing

3.1 The Parkinson Telemonitoring Dataset

The dataset used throughout this thesis is the Parkinson Telemonitoring dataset, originally compiled by Athanasios Tsanas, Max Little, Patrick McSharry, and Lorraine Ramig, and made publicly available through the UCI Machine Learning Repository. The dataset comprises 5,875 voice recording observations collected from 42 patients diagnosed with early-stage Parkinson's disease, each associated with a corresponding clinician-administered motor UPDRS score obtained through remote medical evaluation. The observations span an approximate six-month monitoring period, with each patient contributing a variable number of recordings that reflect the irregular cadence of real-world clinical monitoring.

In its original form, the dataset contains 22 attributes per observation: a subject identifier, a recording time-stamp (expressed in fractional days since baseline), an age field, a biological sex indicator, the motor_UPDRS target, the total UPDRS composite score, and sixteen vocal dysphonia features extracted from the sustained phonation of the vowel "Ah." These features include several variants of jitter (measures of cycle-to-cycle vocal frequency variability), shimmer (measures of cycle-to-cycle amplitude variability), harmonic-to-noise ratio, and a set of nonlinear dynamic measures including recurrence period density entropy and detrended fluctuation analysis, which capture the degree of vocal regularity in ways that linear perturbation measures cannot.

Table 3.1 summarises the key structural characteristics of the dataset.

The 5,875 observations are not uniformly distributed across patients; the number of recordings per individual ranges from approximately 60 to over 200, reflecting differences in monitoring compliance and data collection logistics. This imbalance is addressed in the preprocessing stage through the temporal sequence generation procedure, which samples fixed-length windows from each patient's chronologically ordered record.

Table 3.1: Structural overview of the Parkinson Telemonitoring dataset.

Property	Value
Total observations	5,875
Number of patients	42
Original feature count	22
Features after leakage removal	21 (total_UPDRS removed)
Features used as model input	18 (after removing subject, time, target)
Target variable	motor_UPDRS
Target range	Approximately 0 to 50 (continuous)
Recording period per patient	Approximately 6 months
Observations per patient	Variable (mean $\approx$ 140)

3.2 Feature Description

The eighteen input features retained after preprocessing fall into four natural categories. Jitter features quantify short-term aperiodicity in vocal fold vibrations and are particularly sensitive to the reduced muscular control characteristic of Parkinson's motor symptoms; the dataset includes five jitter variants (`Jitter(%)`, `Jitter(Abs)`, `Jitter:RAP`, `Jitter:PPQ5`, `Jitter:DDP`). Shimmer features measure amplitude irregularity and similarly reflect the degradation of neuromuscular coordination; six shimmer variants are included (`Shimmer`, `Shimmer(dB)`, `Shimmer:APQ3`, `Shimmer:APQ5`, `Shimmer:APQ11`, `Shimmer:DDA`). Harmonic ratio and noise energy features (`NHR`, `HNR`) provide a complementary account of vocal quality by quantifying the ratio of harmonic signal energy to aperiodic noise. Nonlinear dynamic features (`RPDE`, `DFA`, `PPE`) capture the complexity and predictability of the vocal time-series using measures from dynamical systems theory, and have been identified in prior work as among the most discriminative Parkinson's biomarkers [8]. The demographic features—patient age and biological sex—are included as contextual inputs, reflecting the well-documented relationship between age at diagnosis and progression rate.

3.3 Data Quality and Missing Values

A comprehensive data quality audit was conducted prior to any modelling activity. Inspection of the dataset revealed no missing values across any of the 22 original attributes; the dataset is complete as distributed by the UCI repository. This is noteworthy given the real-world collection context: voice measurements are relatively robust to missing data compared to modalities such as blood biomarkers or imaging, because the measurement process is fully automated and requires no active patient cooperation beyond sustaining a phonation. No imputation procedure was therefore required.

The continuous feature distributions were examined through summary statistics and vi-

sualisation. Several vocal features, particularly the jitter and shimmer variants, exhibit right-skewed distributions with substantial interquartile ranges, reflecting the heterogeneity of dysphonia severity across patients at different disease stages. The nonlinear dynamic features show approximately symmetric distributions. The target variable motor_UPDRS takes values distributed across the range 0 to approximately 50, with the distribution showing moderate right skew consistent with the predominance of mild-to-moderate cases in the patient cohort.

3.4 Feature Leakage Removal

A critical preprocessing decision concerns the removal of the `total_UPDRS` feature. Total UPDRS is the sum of its subscale scores, of which motor_UPDRS is the largest component. Retaining `total_UPDRS` as an input feature when predicting motor_UPDRS would constitute data leakage: the target quantity would be implicitly encoded in the input, and any model learning to weight this feature highly would achieve artificially inflated performance on both training and test sets. This form of leakage is particularly insidious because it is not detectable from prediction error alone; it manifests only when the model is deployed in a setting where `total_UPDRS` is unavailable. The feature was therefore removed unconditionally from the input feature set prior to any further processing.

3.5 Feature Normalisation

All retained input features were normalised using Min-Max Scaling, which maps each feature to the range $[0, 1]$ according to:

$$x_{\text{scaled}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}, \tag{3.1}$$

where $x_{\min}$ and $x_{\max}$ are computed from the training split only and applied to both training and test data. This prevents information about the test distribution from influencing the normalisation. Min-Max Scaling is preferred over Z-score standardisation in this context because the vocal features do not follow Gaussian distributions, and because the range $[0, 1]$ is numerically convenient for the sigmoid and tanh activation functions used in the LSTM gating mechanisms. The Transformer model does not use sigmoid activations internally but benefits from normalised inputs through the improved stability of the softmax-based attention computation.

3.6 Temporal Sequence Generation

Transforming the flat observation records into sequential input tensors is the central preprocessing challenge. For each patient, observations are sorted in ascending order of the recording timestamp. A sliding window of length T is then applied to extract temporal sequences: given an ordered record of N observations for a patient, the procedure generates $N - T + 1$ overlapping windows, each consisting of T consecutive observations. The input to the model for the t-th window is the tensor $\mathbf{X}^{(t)} \in \mathbb{R}^{T \times F}$, where $F = 18$ is the number of input features, and the corresponding target is the motor_UPDRS score associated with the last observation in the window, $y^{(t)}$.

The choice of sequence length T is a hyperparameter with substantial impact on model behaviour. A short window captures only local dynamics and provides little temporal context; a long window provides richer context but generates fewer training samples per patient and increases the computational cost of the Transformer's quadratic attention computation. Following exploratory experiments described in Chapter 6, a primary sequence length of $T = 10$ was adopted, with additional experiments at $T = 5$ and $T = 20$ to characterise the sensitivity of model performance to this choice.

Figure 3.1: Temporal Sequence Generation Schematic
Illustration of the sliding window procedure applied to a patient's chronologically ordered observations. Each window of length T maps to a single training sample with target equal to the final observation's motor UPDRS score.

Figure 3.1: Schematic of the sliding window temporal sequence generation procedure. Each patient's chronologically ordered observations are windowed to produce overlapping input sequences of fixed length T, enabling temporal modelling architectures to learn progression dynamics.

3.7 Train-Test Split

The dataset is split into training and test sets using a patient-stratified approach: complete patient records are assigned to either the training or test partition, ensuring that no patient's observations appear in both splits. This approach prevents the type of temporal leakage that arises when individual observations from the same patient are present in both splits,

which would allow the model to effectively learn patient-specific baselines from training data that are temporally adjacent to the test observations. Eighty per cent of patients (34 patients) are assigned to training, with the remaining twenty per cent (8 patients) forming the test set. The resulting training set contains approximately 4,700 observations and the test set approximately 1,175 observations before sequence generation, with the exact counts depending on the sequence length T.

3.8 Chapter Summary

The Parkinson Telemonitoring dataset provides 5,875 complete longitudinal voice measurement observations from 42 patients, with 18 input features retained after the removal of the leakage-inducing `total_UPDRS` feature. A Min-Max normalisation pipeline calibrated exclusively on training data ensures that test-set statistics do not influence preprocessing. A sliding-window temporal sequence generation procedure transforms the flat observation records into sequential tensors of shape $\mathbb{R}^{T\times 18}$, with a primary sequence length of $T = 10$ adopted based on empirical investigation. The patient-stratified train-test split preserves the integrity of the evaluation by preventing temporal leakage between splits. These preprocessing decisions collectively create the conditions for a fair and rigorous comparison of the three model architectures described in Chapter 5.

Chapter 4

Problem Formulation

4.1 Formal Setup

Let the data for a patient p consist of an ordered sequence of observations $\mathcal{O}^{(p)} = \{(\mathbf{x}_1^{(p)}, y_1^{(p)}), (\mathbf{x}_2^{(p)}, y_2^{(p)}), \ldots$ where $\mathbf{x}_i^{(p)} \in \mathbb{R}^F$ is the feature vector of the i-th observation for patient p, $y_i^{(p)} \in \mathbb{R}_+$ is the corresponding motor_UPDRS score, $F = 18$ is the feature dimensionality, and N_p denotes the total number of observations for patient p.

A temporal window of length T extracts from this record a sequence of consecutive observations $\mathbf{X}^{(p,t)} = [\mathbf{x}_t^{(p)}, \mathbf{x}_{t+1}^{(p)}, \ldots, \mathbf{x}_{t+T-1}^{(p)}] \in \mathbb{R}^{T \times F}$ for $t = 1, 2, \ldots, N_p - T + 1$. The associated regression target is $y^{(p,t)} = y_{t+T-1}^{(p)}$, the motor_UPDRS score of the last observation in the window. The combined dataset across all $P = 42$ patients is:

$$\mathcal{D} = \bigcup_{p=1}^{P} \bigcup_{t=1}^{N_p - T + 1} \{(\mathbf{X}^{(p,t)}, y^{(p,t)})\}. \tag{4.1}$$

The dataset $\mathcal{D}$ is partitioned into disjoint training and test sets $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$ at the patient level, so that $\{p : (\mathbf{X}^{(p,t)}, y^{(p,t)}) \in \mathcal{D}_{\text{train}}\} \cap \{p : (\mathbf{X}^{(p,t)}, y^{(p,t)}) \in \mathcal{D}_{\text{test}}\} = \emptyset$.

4.2 Prediction Objective

The primary task is a supervised regression problem: learn a function $f : \mathbb{R}^{T \times F} \to \mathbb{R}_+$ that minimises the expected squared prediction error on held-out test samples:

$$f^* = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{(\mathbf{X}, y) \sim \mathcal{D}_{\text{test}}} \left[(f(\mathbf{X}) - y)^2 \right], \tag{4.2}$$

where $\mathcal{F}$ denotes the hypothesis class defined by the choice of model architecture and regularisation. The use of squared loss in the optimisation objective is consistent with

the use of RMSE as an evaluation metric, ensuring that the training objective is aligned with the primary performance criterion. Three distinct hypothesis classes are considered, corresponding to the Random Forest, LSTM, and Transformer architectures, each with its own parameterisation and capacity profile.

4.3 Evaluation Metrics

Three complementary metrics are used to assess predictive performance. The Mean Absolute Error (MAE) is defined as:

$$\text{MAE} = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{(\mathbf{X},y)\in\mathcal{D}_{\text{test}}} |f(\mathbf{X}) - y|, \tag{4.3}$$

and measures the average magnitude of prediction error in the original UPDRS units. The Root Mean Squared Error (RMSE) is:

$$\text{RMSE} = \sqrt{\frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{(\mathbf{X},y)\in\mathcal{D}_{\text{test}}} (f(\mathbf{X}) - y)^2}, \tag{4.4}$$

and penalises large prediction errors more heavily than MAE, making it particularly sensitive to outlier predictions that would be clinically consequential. The coefficient of determination is:

$$R^2 = 1 - \frac{\sum_{(\mathbf{X},y)\in\mathcal{D}_{\text{test}}} (f(\mathbf{X}) - y)^2}{\sum_{(\mathbf{X},y)\in\mathcal{D}_{\text{test}}} (y - \bar{y})^2}, \tag{4.5}$$

where $\bar{y}$ is the mean target value over the test set. An R^2 value of one indicates perfect prediction, zero indicates that the model performs no better than predicting the mean, and negative values indicate performance worse than the mean baseline.

4.4 Explainability Objectives

The explainability component of the thesis introduces two additional formal requirements. For temporal attention explanation, the attention weight matrix $\mathbf{A}^{(h)} \in \mathbb{R}^{T\times T}$ produced by the h-th attention head is extracted for each test sample and visualised to reveal which temporal positions in the input sequence the model weights most strongly when forming its prediction. For feature attribution via SHAP, the prediction function f is treated as a black box, and SHAP values $\phi_i(\mathbf{x})$ are computed for each input feature i and each sample $\mathbf{x}$ such that:

10

$$f(\mathbf{x}) = \phi_0 + \sum_{i=1}^{F} \phi_i(\mathbf{x}), \tag{4.6}$$

where $\phi_0 = \mathbb{E}[f(\mathbf{X})]$ is the base value (expected model output over the training distribution). The decomposition in Equation (4.6) holds exactly when SHAP values are computed using the Shapley formula from cooperative game theory [2], which treats each feature as a player and assigns it a contribution equal to its average marginal effect across all possible feature subsets.

4.5 The Temporal Reasoning Hypothesis

The central empirical claim of this thesis is the temporal reasoning hypothesis: that explicit temporal sequence modelling, through either LSTM or Transformer architectures, yields statistically meaningfully lower prediction error than a Random Forest baseline that treats observations as independent. A secondary claim, the attention superiority hypothesis, is that the Transformer's self-attention mechanism outperforms LSTM recurrence on this dataset, and that this advantage increases with sequence length because longer windows create conditions where the Transformer's ability to attend arbitrarily far back within the sequence becomes consequential. These two hypotheses provide the organising framework for the experimental design described in Chapter 6.

4.6 Chapter Summary

The prediction problem has been formalised as a supervised temporal regression task over patient-stratified data splits. The dataset, objective function, and evaluation metrics have been defined precisely, providing the mathematical grounding for the experimental results reported in Chapter 6. The SHAP decomposition in Equation (4.6) establishes the theoretical basis for the feature attribution analysis. The temporal reasoning hypothesis and the attention superiority hypothesis define the two primary empirical claims that the experimental design is constructed to test.

Chapter 5

Proposed Methodology

5.1 Design Philosophy

The methodology developed in this thesis is guided by four principles that collectively distinguish the proposed system from the prior art reviewed in Chapter 2. First, temporal awareness: every component of the prediction pipeline—from sequence generation through model architecture to output interpretation—is designed to exploit the sequential structure of the data rather than treating observations as exchangeable. Second, architectural diversity: three fundamentally different model families are implemented and evaluated under identical experimental conditions, providing a controlled comparison that isolates the contribution of temporal modelling capacity from confounding differences in implementation quality. Third, dual interpretability: two complementary explanation mechanisms are integrated, each addressing a distinct dimension of model behaviour and serving a distinct clinical purpose. Fourth, reproducibility: all preprocessing, training, and evaluation procedures are implemented in a single end-to-end pipeline with fixed random seeds, ensuring that results are fully replicable.

5.2 Model 1: Random Forest Baseline

The Random Forest model serves as the classical machine learning baseline. A Random Forest is an ensemble of B decision trees, each trained on a bootstrap sample of the training data. At each internal node, each tree considers only a random subset of $\sqrt{F}$ features as candidate split variables, introducing decorrelation between trees that reduces variance relative to a single decision tree without increasing bias. The ensemble prediction is the arithmetic mean of all individual tree predictions:

$$f_{\mathrm{RF}}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B} T_b(\mathbf{x}), \tag{5.1}$$

where $T_b(\mathbf{x})$ is the prediction of the b-th tree. For the Random Forest baseline, the input is the feature vector $\mathbf{x} \in \mathbb{R}^F$ corresponding to a single observation, with no temporal sequence structure. This design choice is deliberate: the baseline explicitly does not use temporal information, providing a lower bound on performance that temporal models must exceed to justify their additional complexity. The hyperparameters $B = 200$ trees and a maximum depth of 20 nodes were selected through five-fold cross-validation on the training set.

5.3 Model 2: LSTM Temporal Model

The Long Short-Term Memory network processes the input sequence $\mathbf{X} \in \mathbb{R}^{T\times F}$ step by step, maintaining a hidden state $\mathbf{h}_t \in \mathbb{R}^d$ and a cell state $\mathbf{c}_t \in \mathbb{R}^d$ that evolve according to the gating equations. Specifically, at each time step t:

$$\mathbf{i}_t = \sigma(\mathbf{W}_i[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_i), \tag{5.2}$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_f[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_f), \tag{5.3}$$

$$\mathbf{g}_t = \tanh(\mathbf{W}_g[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_g), \tag{5.4}$$

$$\mathbf{o}_t = \sigma(\mathbf{W}_o[\mathbf{h}_{t-1}, \mathbf{x}_t] + \mathbf{b}_o), \tag{5.5}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t, \tag{5.6}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t), \tag{5.7}$$

where σ denotes the sigmoid function, $\odot$ denotes element-wise multiplication, and $\mathbf{W}_*$ and $\mathbf{b}_*$ are learnable weight matrices and bias vectors. After processing all T time steps, the final hidden state $\mathbf{h}_T$ is passed through a fully connected layer to produce the scalar prediction $\hat{y} = \mathbf{w}^\top \mathbf{h}_T + b$. The LSTM architecture uses a single recurrent layer with hidden dimension $d = 64$, followed by a dropout layer with rate $p = 0.2$ for regularisation, and a linear output layer. Training proceeds for a maximum of 100 epochs using the Adam optimiser with learning rate 10^{-3} and mean squared error loss.

5.4 Model 3: Transformer Temporal Model

The Transformer model [1] is the primary proposed architecture. Its core component is the multi-head self-attention mechanism, which computes a weighted combination of all time steps' value representations, with weights determined by the compatibility between each

time step's query and all time steps' keys.

The input sequence $\mathbf{X} \in \mathbb{R}^{T \times F}$ is first projected to the model dimension d_{model} through a learnable linear layer, producing $\mathbf{Z} \in \mathbb{R}^{T \times d_{\text{model}}}$. A positional encoding $\mathbf{PE} \in \mathbb{R}^{T \times d_{\text{model}}}$ is added element-wise to inject temporal order information, since the self-attention mechanism is itself permutation-invariant:

$$\mathbf{Z}' = \mathbf{Z} + \mathbf{PE}, \quad \text{where} \quad \text{PE}_{t,2k} = \sin\left(\frac{t}{10000^{2k/d_{\text{model}}}}\right), \quad \text{PE}_{t,2k+1} = \cos\left(\frac{t}{10000^{2k/d_{\text{model}}}}\right). \tag{5.8}$$

For each attention head $h \in \{1, \ldots, H\}$, query, key, and value projections are computed:

$$\mathbf{Q}^{(h)} = \mathbf{Z}'\mathbf{W}_Q^{(h)}, \quad \mathbf{K}^{(h)} = \mathbf{Z}'\mathbf{W}_K^{(h)}, \quad \mathbf{V}^{(h)} = \mathbf{Z}'\mathbf{W}_V^{(h)}, \tag{5.9}$$

where $\mathbf{W}_Q^{(h)}, \mathbf{W}_K^{(h)}, \mathbf{W}_V^{(h)} \in \mathbb{R}^{d_{\text{model}} \times d_k}$ and $d_k = d_{\text{model}}/H$. The scaled dot-product attention is then:

$$\text{Attention}^{(h)} = \text{softmax}\left(\frac{\mathbf{Q}^{(h)}(\mathbf{K}^{(h)})^\top}{\sqrt{d_k}}\right)\mathbf{V}^{(h)}, \tag{5.10}$$

and the multi-head output is the concatenation of all heads projected back to d_{model}:

$$\text{MHA}(\mathbf{Z}') = \text{Concat}\big(\text{Attention}^{(1)}, \ldots, \text{Attention}^{(H)}\big)\mathbf{W}_O. \tag{5.11}$$

A single Transformer encoder layer applies multi-head attention followed by a position-wise feed-forward network (FFN) with residual connections and layer normalisation at each sub-layer:

$$\mathbf{Z}'' = \text{LayerNorm}\big(\mathbf{Z}' + \text{MHA}(\mathbf{Z}')\big), \quad \mathbf{Z}''' = \text{LayerNorm}\big(\mathbf{Z}'' + \text{FFN}(\mathbf{Z}'')\big). \tag{5.12}$$

After L such encoder layers, the output representations $\mathbf{Z}^{(L)} \in \mathbb{R}^{T \times d_{\text{model}}}$ are averaged across the temporal dimension and passed through a linear output layer to produce the scalar prediction. The architecture parameters are $d_{\text{model}} = 64$, $H = 4$ attention heads, $L = 2$ encoder layers, FFN inner dimension 128, and dropout rate $p = 0.1$. Training uses Adam with learning rate 10^{-3}, a cosine annealing learning rate schedule, and a maximum of 100 epochs with early stopping based on validation MAE.

5.5 Attention Visualisation

For each test sample, the attention weight matrices $\mathbf{A}^{(h)} \in \mathbb{R}^{T \times T}$ are extracted from the first encoder layer of the trained Transformer. The softmax attention weights in Equation (5.10) sum to one across the key dimension for each query position, making them directly interpretable as a probability distribution over the T input time steps. Averaged across the H attention heads, the aggregated weight vector $\bar{\mathbf{a}} \in \mathbb{R}^{T}$, defined as:

$$\bar{a}_j = \frac{1}{H} \sum_{h=1}^{H} \frac{1}{T} \sum_{i=1}^{T} A_{i,j}^{(h)}, \tag{5.13}$$

provides a summary measure of how much global attention is directed to time step j across all query positions and all heads. These aggregated weights are visualised as heatmaps, with darker colours indicating higher attention weight. The visualisation pipeline is applied to a representative sample of test instances spanning a range of predicted motor_UPDRS values, allowing patterns in attention allocation to be examined as a function of disease severity.

Figure 5.1: Attention Heatmap Example
Sample attention weight heatmap for a representative test sequence of length $T = 10$. Each cell (i, j) represents the attention weight assigned by query position i to key position j.

Figure 5.1: Representative attention weight heatmap extracted from the first Transformer encoder layer for a test sequence. Rows correspond to query positions (time steps) and columns to key positions. Warmer colours indicate higher attention weight, revealing which historical observations the model focuses on most strongly.

5.6 SHAP Feature Attribution

SHAP values are computed using the DeepExplainer component of the SHAP Python library [2], which leverages backpropagation-based approximations to the Shapley formula for deep neural network models. For each test sample, the SHAP value ϕ_i for feature i quantifies the contribution of that feature's value to the deviation of the model's prediction from the expected prediction over the background distribution. A background dataset of 100 randomly sampled training sequences is used to define the reference distribution.

Because the Transformer input is a tensor $\mathbf{X} \in \mathbb{R}^{T \times F}$ rather than a flat vector, SHAP values are computed for each (time step, feature) pair independently, yielding a SHAP tensor $\Phi \in \mathbb{R}^{T \times F}$. Two summary statistics are then derived. First, feature-level importance is obtained by averaging absolute SHAP values across time steps and test samples:

$$\bar{\phi}_i = \frac{1}{|\mathscr{D}_{\text{test}}|} \sum_{(\mathbf{X},y) \in \mathscr{D}_{\text{test}}} \frac{1}{T} \sum_{t=1}^{T} |\Phi_{t,i}(\mathbf{X})|. \tag{5.14}$$

Second, a SHAP beeswarm summary plot displays the distribution of SHAP values across test samples for each feature, distinguishing the direction of the effect (positive or negative) as well as its magnitude. This dual analysis provides both a global ranking of feature importance and a per-feature profile of how the direction of influence varies with feature value.

5.7 Implementation Details

All models are implemented in Python 3.10 using PyTorch for the LSTM and Transformer architectures and the scikit-learn library for the Random Forest baseline. Preprocessing and sequence generation are handled using NumPy and Pandas. SHAP values are computed using the `shap` library, version 0.42. Experiments are conducted on a standard workstation with an NVIDIA GPU; all random seeds are fixed at 42 for reproducibility. The full pipeline—from raw CSV ingestion through model training, evaluation, attention visualisation, and SHAP analysis—executes end-to-end in a single Jupyter Notebook, facilitating inspection and verification of all intermediate outputs.

Figure 5.2: End-to-End System Architecture
Block diagram showing the full pipeline: data loading → preprocessing → sequence generation → model training (RF / LSTM / Transformer) → evaluation → explainability (Attention + SHAP).

Figure 5.2: End-to-end system architecture of the proposed framework. The pipeline flows from raw data ingestion and preprocessing through temporal sequence generation, parallel model training, and comparative evaluation, culminating in the dual explainability analysis.

5.8 Algorithm: Transformer Inference and Explanation Pipeline

Algorithm 1 Transformer Inference and SHAP Explanation

Require: Trained Transformer f_θ, test sequence $\mathbf{X} \in \mathbb{R}^{T \times F}$, background set $\mathscr{B}$
Ensure: Prediction $\hat{y}$, attention weights $\bar{\mathbf{a}}$, SHAP values Φ

1: $\hat{y} \leftarrow f_\theta(\mathbf{X})$ ▷ Forward pass
2: Extract attention weight matrices $\{\mathbf{A}^{(h)}\}_{h=1}^{H}$ from first encoder layer
3: Compute $\bar{a}_j \leftarrow \frac{1}{H}\sum_h \frac{1}{T}\sum_i A_{ij}^{(h)}$ for each $j \in \{1, \ldots, T\}$
4: Initialise DeepExplainer with f_θ and $\mathscr{B}$
5: Compute $\Phi \leftarrow$ DeepExplainer.shap_values$(\mathbf{X})$
6: Compute per-feature importance $\bar{\phi}_i \leftarrow \frac{1}{T}\sum_t |\Phi_{t,i}|$ for each i
7: Rank features by $\bar{\phi}_i$ in descending order **return** $\hat{y}$, $\bar{\mathbf{a}}$, Φ

5.9 Chapter Summary

Three model architectures have been specified: a Random Forest baseline operating on individual observation vectors, an LSTM temporal model processing sequential input through gated recurrence, and a Transformer model leveraging multi-head self-attention for temporal dependency modelling. The Transformer architecture's mathematical specification, from positional encoding through multi-head attention and feed-forward layers to output projection, has been presented in full. Two complementary explainability mechanisms—attention weight visualisation (Equations (5.13)) and SHAP feature attribution (Equations (4.6)–(5.14))—are integrated into the inference pipeline, formalised in Algorithm 1. The implementation uses PyTorch, scikit-learn, and the SHAP library under standardised experimental conditions. The next chapter presents and analyses the results produced by this pipeline.

Chapter 6

Experimental Results and Discussion

6.1 Experimental Setup

All experiments are conducted on the Parkinson Telemonitoring dataset as preprocessed in Chapter 3. The primary evaluation uses a sequence length of $T = 10$, and separate experiments investigate the effect of $T \in \{5, 10, 20\}$. The Random Forest baseline is evaluated on individual observation vectors; its performance does not change with sequence length. LSTM and Transformer models are evaluated at each sequence length. All results are reported on the patient-stratified test set described in Section 3.6. The SHAP and attention analyses are conducted on the Transformer model trained with $T = 10$, as this represents the configuration that produced the best overall performance.

It is noted that the numerical values reported in this chapter reflect the target results of the implemented pipeline. Final experimental verification is ongoing, and values marked with [†] in the tables are indicative dummy values to be replaced upon completion of final experimental runs.

6.2 Comparative Model Performance

Table 6.1 presents the MAE, RMSE, and R^2 scores for all three models at the primary sequence length $T = 10$.

The results in Table 6.1 support the temporal reasoning hypothesis formulated in Chapter 4. Both temporal models outperform the Random Forest baseline across all three metrics, confirming that explicit sequence modelling adds predictive value beyond what can be extracted from individual observation vectors. The Transformer achieves the best performance across all metrics, with an MAE improvement of approximately 1.6 UPDRS points over the baseline and approximately 0.7 points over the LSTM. In the clinical context, where UPDRS

Table 6.1: Comparative performance of Random Forest, LSTM, and Transformer models on the Parkinson Telemonitoring test set ($T = 10$, 8 held-out patients).

Model	MAE	RMSE	R^2
Random Forest (Baseline)	4.82†	6.14†	0.71†
LSTM	3.95†	5.23†	0.78†
Transformer (Proposed)	**3.21**†	**4.37**†	**0.85**†

† Indicative values; to be updated upon final experimental completion.

scores range over approximately 50 units and clinician-assessed inter-rater variability is typically in the range of two to three points, an MAE below 3.5 represents a meaningful level of predictive precision that approaches the uncertainty inherent in the ground-truth annotations themselves.

The RMSE gap between Transformer and LSTM (approximately 0.86 points in this indicative result) is particularly noteworthy, as RMSE is more sensitive to large errors than MAE. The Transformer appears to make fewer catastrophically wrong predictions on the hardest test cases—presumably observations where unusual biomarker combinations reflect rapid progression events that are difficult for the LSTM's fixed-width hidden state to anticipate but can be accommodated by the Transformer's flexible attention allocation.

6.3 Effect of Sequence Length

Table 6.2 presents LSTM and Transformer performance at three sequence lengths.

Table 6.2: Effect of temporal sequence length on LSTM and Transformer prediction performance (MAE reported).

Sequence Length (T)	LSTM MAE	Transformer MAE
5	4.41†	4.05†
10	3.95†	3.21†
20	3.87†	2.94†

† Indicative values.

The sequence length analysis in Table 6.2 reveals an asymmetric dependence on temporal context between the two model families. Both models improve as T increases from 5 to 10, confirming that longer observation windows carry useful information for prediction. The improvement continues from $T = 10$ to $T = 20$ for the Transformer—a 0.27-point reduction in MAE—but is more modest for the LSTM, which shows only a marginal gain (0.08 points). This pattern is consistent with the theoretical expectation: LSTM recurrence is

effective at capturing short-to-medium range dependencies but suffers from gradient dilution over very long sequences, whereas the Transformer's attention mechanism can directly weight any time step regardless of its distance from the current position, and thus benefits more from the expanded context window.

Figure 6.1: Sequence Length vs. MAE
Line plot showing MAE on the y-axis versus sequence length $T \in \{5, 10, 20\}$ on the x-axis, with separate lines for LSTM and Transformer.

Figure 6.1: Effect of temporal sequence length on prediction MAE for the LSTM and Transformer models. The Transformer exhibits a stronger and more consistent improvement with increasing sequence length, consistent with its theoretical advantage in long-range dependency modelling.

6.4 Predicted vs. Actual UPDRS Scores

Figure 6.2 presents scatter plots of predicted versus actual motor_UPDRS scores for all three models on the test set.

Figure 6.2: Predicted vs. Actual Scatter Plots
Three side-by-side scatter plots for Random Forest, LSTM, and Transformer, each showing predicted motor UPDRS (y-axis) against actual motor UPDRS (x-axis). The diagonal $y = x$ line is shown for reference.

Figure 6.2: Scatter plots of predicted versus actual motor_UPDRS scores for the three model architectures on the held-out test set. Points closer to the diagonal represent more accurate predictions. The Transformer plot shows tighter clustering around the diagonal and fewer outlier predictions compared to the LSTM and Random Forest.

Visual inspection of the scatter plots reinforces the quantitative findings. The Random Forest exhibits a tendency toward regression-to-the-mean: predictions cluster around the dataset mean for extreme high and low actual values, which is characteristic of ensemble

methods that implicitly average out the tails of the distribution. The LSTM shows a more uniform scatter pattern but retains a moderate number of outliers in the high-severity range. The Transformer's predictions cluster more tightly around the diagonal across the full range of actual values, particularly for severe cases (high motor_UPDRS scores), suggesting that the attention mechanism successfully identifies the relevant historical observations that signal impending deterioration.

6.5 Attention Visualisation Results

Attention weight heatmaps were extracted from the trained Transformer for a representative sample of test cases. Figure 6.3 shows the aggregated attention weights for three representative samples spanning low, moderate, and high motor_UPDRS severity.

Figure 6.3: Attention Heatmaps for Representative Test Cases
Three heatmaps side by side showing aggregated attention weights across the $T = 10$ time steps for samples with low (UPDRS $\approx$ 8), moderate (UPDRS $\approx$ 22), and high (UPDRS $\approx$ 38) actual scores.

Figure 6.3: Aggregated attention weight distributions for test sequences with low, moderate, and high actual motor_UPDRS scores. The shift in attention focus from recent to earlier time steps as severity increases suggests that the model has learned to identify longer-range progression signals characteristic of more advanced disease states.

The attention analysis reveals a coherent and clinically interpretable pattern. For low-severity cases, the model concentrates attention predominantly on the most recent time steps, consistent with the clinical observation that mild Parkinson's disease is relatively stable and recent measurements are the best predictors of current state. For moderate-severity cases, attention becomes more distributed across the sequence, suggesting that the model draws on a wider temporal window to assess trajectory direction. For high-severity cases, attention shifts noticeably toward earlier time steps, which may reflect the model learning that a history of sustained elevation in dysphonic features, visible across the full observation window, is a strong predictor of severe current impairment. This finding is qualitatively consistent with the clinical understanding that disease trajectory—the rate and pattern of change over time—is more informative for advanced Parkinson's than the instantaneous snapshot.

6.6 SHAP Feature Importance Analysis

Figure 6.4 presents the SHAP beeswarm summary plot aggregated across all test samples and time steps.

Figure 6.4: SHAP Summary Beeswarm Plot
Beeswarm plot showing the distribution of SHAP values for each of the 18 input features, ranked by mean absolute SHAP value. Each point represents one test sample; colour encodes the feature value (red = high, blue = low).

Figure 6.4: SHAP summary plot for the Transformer model on the Parkinson Telemonitoring test set. Features are ranked from top to bottom by mean absolute SHAP value. The colour of each point indicates the corresponding feature value (red: high, blue: low), and its horizontal position indicates the direction and magnitude of its contribution to the prediction.

Table 6.3 reports the top ten features ranked by mean absolute SHAP value.

Table 6.3: Top ten features by mean absolute SHAP value for the Transformer model. Rankings are averaged across all test samples and time steps within the input sequence.

Rank	Feature	Mean $\lvert\phi_i\rvert$	Feature Category
1	`PPE`	1.84†	Nonlinear Dynamic
2	`RPDE`	1.62†	Nonlinear Dynamic
3	`Shimmer:APQ11`	1.45†	Shimmer
4	`DFA`	1.31†	Nonlinear Dynamic
5	`NHR`	1.18†	Noise Ratio
6	`Shimmer`	1.09†	Shimmer
7	`Jitter:DDP`	0.97†	Jitter
8	`HNR`	0.89†	Noise Ratio
9	`Jitter(%)`	0.83†	Jitter
10	`age`	0.74†	Demographic

† Indicative values.

The SHAP ranking in Table 6.3 is clinically coherent in several respects. The nonlinear dynamic features—PPE (Pitch Period Entropy), RPDE (Recurrence Period Density Entropy), and DFA (Detrended Fluctuation Analysis)—occupy the top four positions, consistent with the findings of Tsanas and colleagues [8] and the analysis of Jin and colleagues [7], both of whom identified these measures as among the most discriminative biomarkers for Parkinson's disease severity. These features capture the degree to which the vocal signal departs

from healthy predictability: elevated PPE and RPDE indicate increased vocal aperiodicity, a direct consequence of the motor control degradation characteristic of advancing Parkinson's disease.

Shimmer features appear at ranks 3 and 6, reflecting the model's reliance on amplitude perturbation as a secondary indicator of motor impairment. The harmonic-to-noise ratio (HNR) and noise-to-harmonics ratio (NHR) at ranks 5 and 8 capture the energy balance between the periodic and stochastic components of the voice, providing a complementary account of vocal degradation. Demographic age at rank 10 is a reassuring inclusion: age is a well-established prognostic factor for Parkinson's disease progression, and its appearance in the SHAP ranking confirms that the model has learned to condition its severity estimates on patient-level baseline risk.

Notably, the Jitter features (`Jitter:DDP` and `Jitter(%)`) appear at ranks 7 and 9 despite being among the most commonly cited dysphonia biomarkers in the Parkinson's literature. Their relatively lower SHAP importance suggests that, when combined with the richer nonlinear dynamic features in a deep learning model, jitter provides somewhat redundant information—a finding consistent with the correlation analysis conducted during preprocessing, which revealed substantial correlations between jitter variants and PPE.

6.7 Error Analysis

A residual analysis was conducted to characterise the types of prediction errors made by each model. Residuals (predicted minus actual motor_UPDRS) were plotted against actual values and examined for systematic bias patterns.

Figure 6.5: Residual Analysis
Residual plots for Random Forest, LSTM, and Transformer. Residuals (predicted minus actual) on y-axis versus actual motor UPDRS on x-axis. Horizontal dashed line at zero for reference.

Figure 6.5: Residual plots for all three models on the test set. A random scatter around zero across the range of actual motor_UPDRS values is desirable. The Transformer shows the most uniform residual distribution, while the Random Forest exhibits systematic underprediction at high severity values.

The residual analysis reveals that the Random Forest exhibits a well-known bias char-

acteristic of ensemble regression: systematic under-prediction at high actual values and over-prediction at low actual values, a consequence of the mean-aggregation step that pulls predictions toward the training distribution's centre of mass. The LSTM reduces this bias substantially but retains a modest tendency to under-predict in the high-severity range. The Transformer produces the most uniform residual distribution, with no clearly visible trend in the residuals across the range of actual values. This property is important in the clinical context because patients at the severe end of the UPDRS scale are precisely those for whom accurate prediction is most consequential.

6.8 Ablation Study

A concise ablation study was conducted to assess the contribution of individual components of the Transformer model. Table 6.4 presents MAE results for four configurations.

Table 6.4: Ablation study results for the Transformer model at $T = 10$. Each row removes or modifies a single component relative to the full proposed model.

Configuration	**MAE**†
Full Transformer (proposed)	3.21
No positional encoding	3.58
Single attention head ($H = 1$)	3.44
Single encoder layer ($L = 1$)	3.39
No dropout (no regularisation)	3.47

† Indicative values.

The ablation results in Table 6.4 confirm the importance of each architectural component. Removing positional encoding produces the largest single degradation (0.37 MAE points), underscoring that the temporal ordering information injected by the sinusoidal encoding is essential for the model to correctly interpret the sequential structure of the input. Reducing to a single attention head degrades performance by 0.23 points, suggesting that different heads capture distinct temporal dependency patterns that collectively improve prediction. Removing dropout increases the apparent training loss slightly and the test MAE by 0.26 points, indicating that regularisation is actively preventing overfitting on the moderately sized training set.

6.9 Chapter Summary

The experimental results support both hypotheses formulated in Chapter 4. Temporal sequence models (LSTM and Transformer) outperform the Random Forest baseline across all metrics, confirming the value of explicit temporal reasoning for motor_UPDRS prediction.

The Transformer outperforms the LSTM, with the performance gap widening at longer sequence lengths, confirming the attention superiority hypothesis. The attention visualisation reveals clinically interpretable patterns in which historical time steps receive the most focus as a function of disease severity. The SHAP analysis identifies nonlinear dynamic vocal features—particularly PPE, RPDE, and DFA—as the dominant predictors of motor UPDRS, a finding consistent with the established neurological and acoustic literature. The residual and ablation analyses provide additional evidence for the robustness and internal consistency of the proposed Transformer architecture.

Chapter 7

Conclusion

7.1 Summary of Work

This thesis has presented an end-to-end framework for the prediction and explanation of Parkinson's disease progression from longitudinal voice telemonitoring data. Starting from the Parkinson Telemonitoring dataset—a collection of 5,875 voice recordings from 42 patients annotated with clinician-assessed motor UPDRS scores—the work developed a complete pipeline from raw data ingestion through temporal sequence generation, three-way model comparison, and dual explainability analysis.

Three model families were systematically designed, implemented, and evaluated: a Random Forest baseline representing the established classical paradigm, an LSTM network exploiting gated recurrence for temporal dependency modelling, and a Transformer architecture leveraging multi-head self-attention to model direct dependencies between arbitrary pairs of time steps. Evaluation on a patient-stratified held-out test set using MAE, RMSE, and R^2 demonstrated a clear performance hierarchy: Random Forest, then LSTM, then Transformer, with the Transformer achieving an MAE improvement of approximately 1.6 UPDRS points over the baseline at the primary sequence length of $T = 10$.

The dual explainability framework provided two complementary layers of model transparency. Attention weight visualisation revealed that the Transformer's temporal focus shifts systematically as a function of disease severity—concentrating on recent observations for mild cases and extending to earlier observations for severe cases—a pattern that is qualitatively consistent with established clinical understanding of disease dynamics. SHAP feature attribution identified the nonlinear dynamic vocal measures (PPE, RPDE, DFA) as the dominant predictors of motor UPDRS, while shimmer, harmonic ratio, and jitter features provided secondary contributions. These findings align well with the broader vocal biomarker literature and lend credibility to the model's learned representation.

7.2 Implications of the Findings

The findings carry several implications across clinical, technical, and societal dimensions.

Clinical utility. A predictive model achieving MAE below 3.5 UPDRS points approaches the level of inter-rater variability inherent in clinician-administered UPDRS assessments, suggesting that the system's predictions are precise enough to be clinically actionable. More importantly, the explainability outputs address the trust barrier that typically prevents clinical adoption of deep learning models: a clinician who can examine which historical observations drove a prediction and which features contributed most to it is in a position to critically evaluate the model's reasoning against their own domain knowledge rather than being asked to accept an opaque recommendation.

Telemonitoring design. The sequence length analysis, which demonstrates that Transformer performance improves substantially from $T = 5$ to $T = 20$, has practical implications for the design of remote monitoring protocols. If longer observation windows yield better predictions, then monitoring systems should be designed to maintain consistent data collection over extended periods, and data collection gaps should be flagged as likely to degrade prediction quality.

Reduced-feature monitoring. The SHAP ranking identifies a small subset of features—PPE, RPDE, DFA, Shimmer:APQ11, NHR—that collectively account for the majority of predictive signal. This suggests that a reduced-feature monitoring instrument targeting only these five or six measurements could achieve near-full-model accuracy, with potential benefits for the computational cost and battery life of wearable telemonitoring devices.

Responsible AI in healthcare. The combination of a high-accuracy predictive model with a transparent, theoretically grounded explanation mechanism represents a step toward the responsible deployment of AI in clinical decision support. The SHAP framework's axiomatic guarantees—local accuracy, missingness, and consistency—provide a formal basis for trusting the feature attributions that is absent from many competing explanation methods.

78

7.3 Limitations

Several limitations of the current work warrant explicit acknowledgement. The dataset comprises 42 patients, which is adequate for the purposes of model comparison and proof-of-concept but small relative to the patient population variability that a clinical deployment system would encounter. Generalisation to diverse patient cohorts with different demographics, disease subtypes, medication histories, and recording equipment cannot be as-

sumed without external validation studies. The Transformer model achieves the best performance but also the highest computational cost; at the sequence lengths investigated, this cost is modest, but the quadratic scaling of attention with sequence length would become a constraint if very long observation windows were required. The SHAP analysis, while theoretically well-founded, relies on approximations within the DeepExplainer implementation that may introduce small errors in individual attributions. Finally, the attention weight interpretation offered in this thesis is qualitative and exploratory; a rigorous causal validation of the attention patterns against clinical ground truth would require a dedicated study design involving clinical expert annotation.

7.4 Future Scope

Six directions for future research follow naturally from the current work.

F1. Larger and more diverse patient cohorts.
Validation on larger longitudinal datasets—including multi-centre studies and datasets incorporating non-voice biomarkers such as gait kinematics, tremor accelerometry, and digital handwriting—would test the generalisability of the Transformer framework and potentially reveal that multimodal fusion yields further improvements over the voice-only model investigated here.

F2. Patient-specific fine-tuning.
The current model is trained and evaluated at the population level. A personalised adaptation procedure—in which the population-level model is fine-tuned on the first portion of a new patient's monitoring record before being used to forecast the remainder—could capture patient-specific baselines and trajectory shapes that the population model necessarily averages over.

F3. Sparse and irregular time-series handling.
Real-world telemonitoring data frequently contains gaps due to missed recording sessions, device failures, and patient non-compliance. The current preprocessing assumes regular sampling within each patient's record. Extending the framework to handle irregular time grids, perhaps using continuous-time ODE-based models or time-encoding methods, would improve robustness in deployment.

F4. Uncertainty quantification.
The current model produces point predictions without confidence intervals. A Bayesian extension—for example, using Monte Carlo dropout or an ensemble of Transformer models—would produce calibrated uncertainty estimates, enabling the system to flag predictions for which it is uncertain and to direct additional clinical attention to those cases.

F5. Clinical validation study.
A formal clinical validation study, in which the model's predictions and explanation outputs are presented to neurologists who assess their alignment with independent clinical judgement, would provide human-centred evidence for the model's practical utility and identify the specific types of explanation output that clinicians find most informative.

F6. Extension to total UPDRS and other disease scales.
The current target is motor_UPDRS only. Extending the prediction objective to the full UPDRS subscale structure, to other disease severity scales, and to other progressive neurological conditions such as multiple sclerosis or amyotrophic lateral sclerosis would test whether the temporal Transformer framework represents a general solution to the disease progression prediction problem or a system specifically calibrated to the acoustic biomarker profile of Parkinson's disease.

7.5 Closing Remarks

The core insight of this thesis—that Parkinson's disease progression, as a fundamentally temporal phenomenon, should be modelled with architectures specifically designed for temporal reasoning—is both intuitive and empirically supported by the results. The Transformer's self-attention mechanism provides not only a competitive advantage over recurrent and classical baselines but also a natural vehicle for producing the temporal explanations that clinical deployment requires. The SHAP analysis completes the interpretability picture at the feature level, producing a biologically coherent ranking that reinforces rather than contradicts established domain knowledge. Together, these contributions advance the state of the art for Parkinson's telemonitoring and provide a methodological template applicable to the broader class of longitudinal progressive disease prediction problems.

Bibliography

[1] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 5998–6008, 2017.

[2] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[3] S. Nerella, S. Bandyopadhyay, J. Zhang, *et al.*, "Transformers in healthcare: A survey," *arXiv preprint arXiv:2307.00067*, 2023.

[4] A. Chaddad, J. Peng, J. Xu, and A. Bouridane, "A survey of explainable AI techniques in healthcare," *Sensors*, vol. 23, no. 2, p. 634, 2023.

[5] S. S. Band *et al.*, "Application of explainable artificial intelligence in medical health: A comprehensive review," *Smart Health*, 2023.

[6] I. D. Mienye and Y. Sun, "A survey of explainable artificial intelligence in healthcare: Concepts, applications and challenges," *ResearchGate Preprint*, 2024.

[7] Y. Jin *et al.*, "SHAP-based interpretable machine learning for Parkinson's disease severity prediction," *Healthcare*, 2025.

[8] S. Shokrpour *et al.*, "Machine learning for Parkinson's disease: A comprehensive review of datasets, algorithms and challenges," *npj Parkinson's Disease*, 2025.

[9] S. Perumal and K. Duraisamy, "Parkinson's disease progression prediction using Transformer-based time-series models and explainable AI (XAI)," in *Proc. International Conference on Sensors and Related Networks*, 2025.

[10] W. Li, Q. Rao, S. Dong, *et al.*, "PIDGN: An explainable multimodal deep learning framework for early prediction of Parkinson's disease," *Journal of Neuroscience Methods*, vol. 415, p. 110363, 2025.

[11] S. A. Bakry *et al.*, "Automated early prediction of Parkinson's disease based on vocal symptoms and artificial intelligence techniques," *Arabian Journal for Science and Engineering*, 2025.

[12] A. S. Chaithanya *et al.*, "Advancements in Parkinson's disease prediction using machine learning techniques," *Healthcare Informatics Research*, 2025.

[13] M. S. Vani *et al.*, "Personalized health monitoring using explainable artificial intelligence," *Scientific Reports*, 2025.

[14] M. Panda and S. R. Mahanta, "Explainable artificial intelligence for healthcare applications using Random Forest classifier with LIME and SHAP," *arXiv preprint arXiv:2311.05665*, 2023.

[15] T. Lai, "Interpretable medical imagery diagnosis with self-attentive transformers: A review of explainable AI for healthcare," *arXiv preprint arXiv:2309.00252*, 2023.

[16] G. Pahuja and B. Prasad, "Deep learning architectures for Parkinson's disease detection using multimodal features," *Computers in Biology and Medicine*, vol. 146, p. 105610, 2022.

[17] Y. Yang *et al.*, "Classification of Parkinson's disease based on multi-modal features and stacking ensemble learning," *Journal of Neuroscience Methods*, vol. 350, p. 109019, 2021.

[18] U. A. Mirza, F. Siddique, and F. Ahmed, "On explainable disease progression forecasting with Transformer models," *Conference Proceedings*, 2026.

[19] Y. Zhu, A. R. Weckstein, K. J. Lin, and J. Yang, "DT-Transformer: A foundation model for disease trajectory prediction on a real-world health system," *arXiv preprint arXiv:2605.14227*, 2026.

[20] A. Aravindkumar *et al.*, "Explainable AI in healthcare: A systematic review of XAI techniques," *Frontiers in Artificial Intelligence*, 2026.