TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies
Department of Software Science

Oluwandabira Ohifeme Alawode 195326IVSM

XAI Based Analysis of the Archimedean Spiral Drawing Test for Parkinson's Disease Diagnostics

Master's Thesis

Supervisor
Elli Valla
PhD
Co-supervisor
Sven Nõmm
PhD

Tallinn 2026

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond
Tarkvarateaduse instituut

Oluwandabira Ohifeme Alawode 195326IVSM

XAI-Põhine Analüüs Archimedeani Spiraaljoonistustestist Parkinsoni Tõve diagnoosimiseks

Magistritöö

Juhendaja
Elli Valla
PhD
Kaasjuhendaja
Sven Nõmm
PhD

Tallinn 2026

Author's declaration of originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Oluwandabira Ohifeme Alawode 
.....
(signature)

Date: May, 2026

Acknowledgements

I would like to thank my father, mother, and brother for their unwavering support and for shaping the person I am today. I am also deeply grateful to all my friends for their companionship and encouragement throughout this journey.

I would like to sincerely thank my supervisors, Elli Valla and Sven Nõmm, for their valuable references to machine learning models and datasets, as well as for their immense patience, expertise, and guidance throughout the development of the experiments and the analysis of data and results.

Annotatsioon

Käesolevas lõputöös uuritakse seletatava tehisintellekti (XAI) võtete kasutamist, et suurendada konvolutsiooniliste närvivõrkude (CNN) tõlgendatavust Parkinsoni tõve (PT) diagnoosimisel spiraali joonistustestide põhjal. Kasutades DraWritePD andmestikku, mis sisaldab nii PT-patsientide kui ka tervete kontrollisikute spiraali joonistusi, treeniti ja hinnati mitmeid CNN-arhitektuure (eelkõige ResNet-50 ja ConvNeXtV2-Base). Hüperparameetrite optimeerimiseks viidi Optuna ja Weights & Biasesi abil läbi põhjalik eksperimentaalne protokoll, et maksimeerida klassifitseerimise täpsust fikseeritud hindamisjaotusel. Sügavõppemudelite "musta kasti" olemuse käsitlemiseks kasutati mudeli ennustuste visuaalsete selgituste loomiseks mitmeid XAI meetodeid, sealhulgas Grad-CAM, Score-CAM, Ablation-CAM, GradientSHAP, LIME ja Integrated Gradients. Tulemused näitavad, et CAM-põhised meetodid pakkusid kõige järjepidevamaid ja kliiniliselt tõlgendatavamaid visualiseeringuid PT tunnustega seotud pildipiirkondade kohta. Käesolev uuring näitab XAI integreerimise teostatavust ja väärtust neurodegeneratiivsete haiguste arvutipõhises diagnostikas, suurendades läbipaistvust ning toetades kliinilist usaldust tehisintellektil põhinevate otsuste vastu.

Lõputöö on kirjutatud inglise keeles ning sisaldab **64** leheküljel teksti, 7 peatükki, **20** joonist ja **7** tabelit.

Abstract

This thesis investigates the use of explainable artificial intelligence (XAI) techniques to enhance the interpretability of convolutional neural networks (CNNs) applied to Archimedean spiral drawing tests for the diagnosis of Parkinson's disease (PD). Leveraging the DraWritePD dataset, which contains spiral drawings from 19 PD patients and 29 healthy controls, two CNN architectures, ResNet-50 and ConvNeXtV2-Base, were trained and evaluated through extensive hyperparameter optimisation (3,000 configurations per architecture). On a fixed subject-wise evaluation split, both models achieved 90% accuracy, with best F1 scores of 0.889 (ResNet-50) and 0.909 (ConvNeXtV2-Base). To address the "black-box" nature of these models, six XAI methods, Grad-CAM, Score-CAM, Ablation-CAM, Integrated Gradients, GradientSHAP, and LIME, were applied to generate visual explanations for model predictions on the evaluation images. A systematic comparison across methods and architectures revealed that CAM-based methods, particularly Score-CAM and Grad-CAM, produced the most spatially coherent and interpretable visualisations, consistently highlighting inner spiral regions associated with PD motor dysfunction. Cross-architecture analysis showed that ResNet-50 and ConvNeXtV2-Base attend to partially overlapping but distinct visual features, with ConvNeXtV2-Base exhibiting broader attention patterns. These findings demonstrate the feasibility and value of integrating XAI into computer-aided diagnostics for neurodegenerative diseases, while the reported classification metrics should be interpreted as results on a fixed model-selection split rather than as a strictly untouched final test estimate.

The thesis is written in English and contains **64** pages of text, 7 chapters, **20** figures, and **7** tables.

List of abbreviations and terms

AMP	Automatic Mixed Precision
AUC	Area Under the Curve
BiGRU	Bidirectional Gated Recurrent Unit
CAD	Computer-Aided Diagnosis
CAM	Class Activation Map
CNN	Convolutional Neural Network
DCNN	Deep Convolutional Neural Network
FCMAE	Fully Convolutional Masked Autoencoder
FN	False Negative
FP	False Positive
FP16	16-bit Floating Point (half precision)
GRN	Global Response Normalisation
HPC	High Performance Computing
IG	Integrated Gradients
LIME	Local Interpretable Model-agnostic Explanations
MDS-UPDRS	Movement Disorder Society – Unified Parkinson's Disease Rating Scale
ML	Machine Learning
PD	Parkinson's Disease
ROC	Receiver Operating Characteristic
SHAP	SHapley Additive exPlanations
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
TPE	Tree-structured Parzen Estimator
ViT	Vision Transformer
W&B	Weights and Biases
XAI	Explainable Artificial Intelligence

Table of Contents

List of Figures

1. Introduction

Parkinson's Disease (PD) is a progressive neurodegenerative disorder characterised by the selective loss of dopaminergic neurons in the substantia nigra pars compacta, resulting in hallmark motor symptoms such as resting tremor, rigidity, bradykinesia, and postural instability, as well as non-motor manifestations including cognitive impairment, autonomic dysfunction, and mood disturbances [1]. Accurate and early diagnosis is critical for patient care and management, yet remains challenging due to symptom overlap with other parkinsonian syndromes and the absence of a single definitive biomarker [2].

Machine learning (ML) has become integral to the development of Computer-Aided Diagnosis (CAD) systems in healthcare. Pipelines for non-imaging clinical data, such as patient demographics, clinical rating scales (e.g., UPDRS), biochemical assays, and wearable sensor measurements, typically involve feature extraction techniques (e.g., principal component analysis, handcrafted statistical and temporal features) followed by classification using traditional ML models [3]. For imaging modalities, deep convolutional neural networks (DCNNs) have markedly improved performance in classification and segmentation tasks [4].

The landmark introduction of AlexNet demonstrated the power of deep architectures for large-scale image classification, ushering in an era dominated by DCNNs [5]. Subsequent variants (e.g., VGG, ResNet, DenseNet) further advanced feature representation and network depth, becoming the de facto choice for medical image analysis. However, their decision logic remains opaque, motivating the integration of explainable AI (XAI) methods such as Grad-CAM, LIME, and SHAP to illuminate model reasoning [6, 7, 8].

A diverse body of literature has applied ML to PD diagnosis. Examples include gait and speech analysis with recurrent networks [9], volumetric MRI features processed by DCNNs [10], and hybrid radiomics-clinical frameworks [11]. Kamran et al. [12] evaluated various architectures on DaTSCAN images, achieving high accuracy but without detailed hyperparameter tuning or interpretability analysis.

In this work, we focus exclusively on the DraWritePD dataset [13], which comprises handwriting and drawing tasks from PD patients and healthy controls. We train and

optimise two CNN architectures, ResNet-50 and ConvNeXtV2-Base, under a unified pipeline and apply six XAI techniques to visualise salient image regions contributing to each model's predictions on the Archimedean spiral drawing test. Our contributions are:

1. A systematic performance comparison of ResNet-50 and ConvNeXtV2-Base on the DraWritePD spiral dataset, including extensive hyperparameter optimisation across 3,000 configurations per architecture.
2. Qualitative visual explanations using six XAI methods (Grad-CAM, Score-CAM, Ablation-CAM, Integrated Gradients, GradientSHAP, and LIME) to interpret model decisions, with cross-method and cross-architecture comparison of the generated heatmaps.

1.1 Problem Statement

Despite growing interest in using deep learning for PD diagnosis from drawing tests, the decision-making processes of these models remain poorly understood. Existing studies have primarily focused on classification accuracy without examining *which* visual features drive model predictions or whether these features align with clinically known indicators of PD motor dysfunction. This "black-box" nature limits clinical trust and adoption of such models.

Furthermore, while multiple XAI methods exist for generating visual explanations of CNN predictions, their comparative effectiveness for medical image tasks, particularly for the fine-grained, continuous features present in spiral drawings, has not been systematically evaluated. It is unknown which XAI methods produce the most coherent and clinically relevant explanations for PD spiral drawing classification, and whether different CNN architectures attend to similar or different visual features when making their predictions.

This thesis addresses this gap by combining comprehensive CNN hyperparameter optimisation with a multi-method XAI analysis on the DraWritePD spiral dataset, providing both quantitative classification results and qualitative interpretability analysis.

1.2 Unit of Study

This research focuses on the interpretability of deep learning models used for Parkinson's disease (PD) diagnosis through drawing test analysis, specifically exploring visual explanation techniques such as Grad-CAM, Score-CAM, Ablation-CAM, Integrated Gradients, GradientSHAP, and LIME for CNN-based classification models trained on Archimedean

spiral drawings.

1.3 Motivation

Parkinson's disease diagnosis often relies on subjective clinical assessments. Machine learning models analysing drawing tests (e.g., spirals) show promise for objective screening, but function as "black boxes." Enhancing model interpretability through visual explanations would increase clinician trust, provide insights into disease manifestations in drawing behaviour, and potentially reveal which visual features of spiral drawings are most discriminative for PD classification. Understanding *how* models make their predictions is a prerequisite for responsible deployment of such systems in clinical settings.

1.4 Research Goal

To develop, optimise, and evaluate CNN-based classification models for PD diagnosis from spiral drawings, and to systematically compare visual explanation methods that communicate how these models interpret spiral drawing features.

1.4.1 Research Questions

1. How effectively do visual explanation techniques (Grad-CAM, Score-CAM, Ablation-CAM, Integrated Gradients, GradientSHAP, LIME) highlight relevant features in PD spiral drawings?
2. Which visual explanation methods provide the most spatially coherent and clinically interpretable insights for spiral drawing classification?
3. Do different CNN architectures (ResNet-50, ConvNeXtV2-Base) attend to similar or different spiral features when classifying PD, and what do the differences reveal about their learned representations?

1.5 Relevant Concepts & Theory

- **Parkinson's Disease Drawing Tests**: Standardised drawing tasks (such as the Archimedean spiral) used to assess motor function and tremor characteristics. Spiral irregularities in line smoothness, spacing, and geometry reflect PD motor dysfunction.
- **Convolutional Neural Networks (CNNs)**: Deep learning architectures specialised for image analysis, including ResNet-50 (residual connections) and ConvNeXtV2-Base (modernised convolutions with self-supervised pretraining).

- **Explainable AI (XAI)**: Methods to make AI decisions interpretable, including:
  - **CAM-based methods**: Grad-CAM, Score-CAM, Ablation-CAM, which generate spatial heatmaps from convolutional layer activations
  - **Gradient-based attributions**: Integrated Gradients, GradientSHAP, which assign pixel-level importance scores
  - **Perturbation-based methods**: LIME, which explains predictions through local surrogate models
- **Transfer Learning**: Using pretrained ImageNet weights as initialisation for training on small medical image datasets.
- **Hyperparameter Optimisation**: Systematic search of model configuration space using Optuna's TPE sampler to maximise classification performance.

1.6 Research Design

1. Train and optimise two CNN architectures (ResNet-50, ConvNeXtV2-Base) on the DraWritePD spiral dataset through large-scale hyperparameter optimisation (3,000 configurations per architecture).
2. Evaluate classification performance using accuracy, F1 score, sensitivity, specificity, AUC, and Youden's J statistic on a fixed subject-wise evaluation split.
3. Apply six visual explanation methods to the best-performing models:
   - CAMs: Grad-CAM, Score-CAM, Ablation-CAM
   - Integrated Gradients
   - GradientSHAP
   - LIME
4. Compare explanations across methods and architectures to identify which methods produce the most coherent and interpretable visualisations for spiral drawing classification.

1.7 Validation of Results

- **Classification validation**: Models evaluated on a fixed subject-wise evaluation split to prevent data leakage between training and evaluation subjects. Multiple metrics are reported to provide a comprehensive performance picture, but the split also served as the final model-selection benchmark.
- **Explanation validation**:
  - Qualitative comparison of heatmaps against known PD manifestations in spiral drawings (tremor irregularities, spacing deviations, angular distortions)
  - Cross-method consistency analysis: do different XAI methods highlight similar

regions?
  - Cross-architecture comparison: do different models attend to similar features?
- **Limitations acknowledged**: Quantitative XAI evaluation metrics (insertion/deletion, ROAR) were not implemented due to time constraints; explanation assessment remains qualitative.

1.7.1 Structure of the Thesis

This thesis is structured as follows:

- **Chapter 1 – Introduction**: Introduces the research problem, motivation, goals, and research questions. Defines key concepts, outlines the research design, and describes validation approach.
- **Chapter 2 – Literature Review**: Reviews Parkinson's disease pathology and motor assessment, machine learning approaches for PD diagnosis, drawing test datasets, CNN architectures for medical imaging, and explainability methods.
- **Chapter 3 – Materials**: Describes the DraWritePD dataset, model architectures (ResNet-50, ConvNeXtV2-Base), XAI techniques, evaluation metrics, and supporting software libraries and hardware environment.
- **Chapter 4 – Methods**: Presents the experimental methodology, including data splitting, augmentation pipeline, hyperparameter optimisation strategy, training procedure, and visual explanation generation.
- **Chapter 5 – Results**: Reports classification performance metrics, hyperparameter analysis, and a detailed comparison of XAI explanations across methods and architectures.
- **Chapter 6 – Discussion and Future Work**: Interprets results in light of research questions, discusses limitations, and proposes directions for future research.
- **Chapter 7 – Conclusion**: Summarises main findings and their significance for XAI-assisted PD diagnostics.
- **Bibliography and Appendices**: Lists all cited works and includes supplementary materials.

2. Literature Review

This chapter reviews the key domains underpinning the present thesis: the clinical background of Parkinson's disease and its motor assessment, machine learning approaches for PD diagnosis, drawing test datasets, the CNN architectures employed, and explainability methods for deep learning. The chapter concludes with a synthesis of the current research gap that this thesis addresses.

2.1 Parkinson's Disease and Motor Assessment

Parkinson's disease (PD) is the second most common neurodegenerative disorder worldwide, affecting approximately 6.1 million individuals globally as of 2016, with prevalence projected to double by 2040 [14]. PD is characterised by the progressive loss of dopaminergic neurons in the substantia nigra pars compacta, leading to dopamine depletion in the basal ganglia circuitry [1]. The neuropathological hallmark of PD is the presence of Lewy bodies, intraneuronal inclusions composed primarily of misfolded alpha-synuclein protein [15]. Braak et al. proposed a six-stage model of PD pathology, suggesting that Lewy body pathology begins in the lower brainstem and olfactory bulb before ascending to the substantia nigra and eventually the neocortex [15]. This staging framework has been influential in understanding the temporal progression of both motor and non-motor symptoms.

The cardinal motor features of PD include resting tremor, bradykinesia (slowness of movement), rigidity, and postural instability [16]. Resting tremor, typically at a frequency of 4–6 Hz, is the most recognisable symptom and is present in approximately 70% of PD patients at diagnosis. Bradykinesia, defined as slowness and progressive reduction in the amplitude of repetitive movements, is considered the defining motor feature and must be present for a clinical diagnosis [2]. Non-motor symptoms, including hyposmia, REM sleep behaviour disorder, constipation, depression, and cognitive decline, may precede motor onset by years or even decades [1].

Clinical diagnosis of PD relies primarily on the identification of motor parkinsonism and the application of diagnostic criteria. The Movement Disorder Society Clinical Diagnostic Criteria (MDS-PD Criteria) require the presence of bradykinesia plus either resting tremor

or rigidity as the core motor features, supplemented by supportive criteria and the absence of exclusion criteria and red flags [2]. The Unified Parkinson's Disease Rating Scale (UPDRS), revised as the MDS-UPDRS, is the most widely used clinical instrument for quantifying PD severity across motor and non-motor domains [17]. Part III of the MDS-UPDRS specifically assesses motor function through standardised examinations of tremor, rigidity, finger tapping, hand movements, and postural stability.

Drawing and handwriting tests have emerged as complementary tools for objective motor assessment in PD. The Archimedean spiral drawing test, in particular, has been recognised as a sensitive measure of upper-limb motor dysfunction [18]. Pullman introduced computerised spiral analysis using digitising tablets, demonstrating that quantitative features such as degree of severity, first-order smoothness, and second-order smoothness (related to tremor) could differentiate PD patients from healthy controls [18]. Subsequent work by Saunders-Pullman et al. validated spiral analysis as a screening tool for early PD, showing that spiral drawing features correlated with clinical UPDRS scores and could detect motor abnormalities even in subjects not yet meeting full diagnostic criteria [19]. San Luciano et al. further demonstrated that digitised spiral features could serve as a potential biomarker for early PD, identifying subtle motor differences in at-risk populations [20].

The clinical rationale for spiral drawing analysis rests on the fact that drawing an Archimedean spiral requires the integration of multiple motor functions: sustained arm posture, coordinated wrist and finger movements, smooth velocity modulation, and visual-motor feedback control. PD disrupts each of these components in characteristic ways. Tremor manifests as periodic oscillations superimposed on the spiral trajectory, visible as irregular loops or waviness. Bradykinesia reduces drawing speed and may produce compressed or uneven spiral spacing. Rigidity leads to reduced smoothness and angular deviations at direction changes. These features, visible in the static spiral image as irregularities in line smoothness, spacing uniformity, and overall spiral geometry, provide the visual signal that convolutional neural networks can potentially learn to detect.

2.2 Machine Learning for Parkinson's Disease Diagnosis

The application of machine learning (ML) to PD diagnosis has been explored across a wide range of data modalities, reflecting the diverse clinical manifestations of the disease. These approaches can be broadly categorised by their input data: sensor-based movement analysis, speech and voice recordings, neuroimaging, and handwriting or drawing analysis.

2.2.1 Sensor-Based and Gait Analysis

Wearable inertial sensors (accelerometers, gyroscopes) attached to the limbs or trunk have been used to capture gait patterns, tremor characteristics, and postural sway in PD patients. Eskofier et al. [9] reviewed recent advances in sensor-based mobility analysis for PD, demonstrating that deep learning approaches could extract discriminative features from raw sensor signals without manual feature engineering. Nguyen et al. [21] applied Transformer architectures to one-dimensional vertical ground reaction force signals for PD gait detection, while Naimi et al. [22] proposed a hybrid CNN-Transformer architecture that jointly performed PD detection and severity prediction from gait data.

2.2.2 Speech and Voice Analysis

Voice changes, including reduced volume (hypophonia), monotone speech, and imprecise articulation, are common in PD. Senturk [23] applied traditional ML classifiers with feature selection to voice recordings, achieving 93.84% accuracy using support vector machines with recursive feature elimination on a publicly available PD voice dataset. These approaches leverage acoustic features such as jitter, shimmer, and harmonic-to-noise ratio as biomarkers for vocal cord dysfunction associated with PD.

2.2.3 Neuroimaging

Neuroimaging modalities, particularly DaTSCAN (dopamine transporter single-photon emission computed tomography), structural MRI, and functional MRI, have been extensively studied for PD diagnosis. Prashanth et al. [11] achieved high classification accuracy on ^{123}I-Ioflupane SPECT images using shape analysis and surface fitting techniques combined with support vector machines. Kamran et al. [12] evaluated various deep neural network architectures on DaTSCAN images for PD identification, achieving strong performance but without detailed hyperparameter optimisation or interpretability analysis. Reyes et al. [24] explored a genomics-based Transformer approach for PD diagnosis, representing a shift towards multi-omics data integration.

2.2.4 Handwriting and Drawing Analysis

Handwriting and drawing analysis occupies a unique position among ML approaches to PD diagnosis because it captures fine motor control through a simple, non-invasive task that can be administered with minimal equipment. The deterioration of handwriting in PD, known as micrographia, has been clinically documented since the early descriptions of the disease

[16]. Beyond micrographia, PD affects multiple kinematic aspects of handwriting: reduced velocity, increased stroke duration, decreased pressure, and tremor-induced oscillations [25, 26].

Pereira et al. [27] were among the first to apply deep learning to PD handwriting classification, using a CNN to classify images of spirals and meanders drawn on paper. Their approach demonstrated that static image-based analysis, as opposed to dynamic pen-trajectory analysis, could yield useful classification results, opening the door for approaches that analyse scanned or photographed drawings without requiring specialised digitising tablets.

Diaz et al. [28] proposed a sequence-based approach using one-dimensional convolutions and bidirectional gated recurrent units (BiGRUs) to analyse dynamic handwriting signals for PD detection, demonstrating that temporal information captured during the drawing process provides complementary discriminative power to static image analysis.

Saravanan et al. [29] combined deep learning classification with explainable AI methods on spiral and wave drawing images, using GoogLeNet with LIME explanations for early PD prediction. Their work is closely related to the present thesis in its combination of CNN classification and XAI, though it employed different architectures and a smaller set of explainability methods. Cavaliere et al. [30] explored grammar-based explainable AI for PD diagnosis from handwriting, representing an alternative approach to interpretability that generates symbolic explanations rather than visual heatmaps.

Thomas et al. [31] reviewed handwriting analysis in PD, highlighting the potential of drawing tests as objective screening tools and the value of clinically informed kinematic feature design.

2.3 Drawing Test Datasets

Several publicly available or documented datasets of PD handwriting and drawing samples have been used in the literature. Table 1 summarises the key characteristics of these datasets.

The **DraWritePD** dataset, introduced by Valla et al. [13], was collected at Tartu University Hospital, Estonia, and comprises handwriting samples from 24 PD patients and 34 age- and gender-matched healthy control subjects. The dataset includes multiple tasks (Archimedean spirals, straight lines, and connected loops) captured using a Wacom digitising tablet, recording both the static drawing image and dynamic pen trajectory data

Table 1. Comparison of drawing test datasets used in PD research.

Dataset	PD / Control	Tasks	Modality	Year
DraWritePD [13]	24 / 34	Spiral, line, writing	Tablet (dynamic)	2022
PaHaW [32]	37 / 38	Spiral, sentence, etc.	Tablet (dynamic)	2016
HandPD [33]	74 / 18	Spiral, meander	Tablet (dynamic)	2018
NewHandPD [27]	31 / 35	Spiral, meander, etc.	Tablet (dynamic)	2016
SST [34]	15 / 15	Static spiral	Paper scan	2014

(position, pressure, tilt, timestamp). The DraWritePD study demonstrated that tremor-related features engineered from the dynamic trajectory data could achieve classification accuracies of up to 85.3% using random forest classifiers. The present thesis focuses exclusively on the static spiral images from this dataset, analysing them through CNNs rather than handcrafted kinematic features.

The **PaHaW** (Parkinson's disease Handwriting Database), introduced by Drotár et al. [32], is among the most widely used datasets in PD handwriting research. It contains dynamic handwriting samples from 37 PD patients and 38 healthy controls performing eight tasks including Archimedean spiral drawing, letter and word writing, and sentence copying. The PaHaW study evaluated kinematic and pressure-based features for PD classification, achieving an accuracy of 81.3% with an SVM classifier using in-air movement features.

The **HandPD** dataset [33] contains drawings of spirals and meanders from 74 PD patients and 18 controls, collected using a Biometric Smart Pen (BiSP) device. Pereira et al. used this dataset to demonstrate that CNNs applied to the drawing images could achieve classification results competitive with handcrafted feature approaches, reporting accuracies up to 87.14% for meanders.

The **NewHandPD** dataset [27] extends the HandPD collection with additional tasks, recorded via a Biometric Smart Pen (BiSP) device, incorporating spiral drawing, meander drawing, circle drawing, and diadochokinesis tasks. The initial release comprised 14 PD patients and 21 controls; the expanded dataset available from the authors' official repository contains 31 PD patients and 35 controls [27]. Pereira et al. demonstrated that CNNs applied to the resulting handwriting dynamics could distinguish PD patients from controls.

A common challenge across all these datasets is their relatively small sample size (typically 30–75 subjects per group), which limits the statistical power of classification results and necessitates careful experimental design to avoid overfitting, particularly when employing deep learning models with millions of parameters.

2.4 CNN Architectures for Medical Image Classification

Convolutional neural networks have become the dominant architecture for image classification tasks in both general computer vision and medical image analysis [4]. The success of CNNs in medical applications is driven by their ability to learn hierarchical feature representations directly from pixel data, eliminating the need for manual feature engineering. However, medical image classification presents unique challenges: small dataset sizes, class imbalance, high inter- and intra-class variability, and the need for model interpretability [35].

2.4.1 ResNet

ResNet (Residual Networks), introduced by He et al. [36], represented a breakthrough in training very deep neural networks through the use of skip (residual) connections. These connections allow the gradient to bypass one or more layers during backpropagation, mitigating the vanishing gradient problem that had previously limited network depth. The key innovation is the residual learning formulation: instead of learning the desired underlying mapping $H(\mathbf{x})$ directly, the network learns the residual function $F(\mathbf{x}) = H(\mathbf{x}) - \mathbf{x}$, with the original input added back via the skip connection. This formulation makes it easier to optimise deep networks and has enabled architectures with hundreds of layers.

ResNet-50, the variant used in this thesis, consists of 50 layers organised into four stages of bottleneck residual blocks (3, 4, 6, and 3 blocks, respectively), with approximately 25.6 million parameters. It uses 1×1 convolutions for dimensionality reduction and expansion within each bottleneck block, keeping computational costs manageable despite the network depth. ResNet-50 pretrained on ImageNet-1k [37] has become one of the most widely used feature extractors for transfer learning in medical imaging, consistently demonstrating strong performance across diverse tasks including retinal disease classification, chest X-ray analysis, and histopathology [4, 35].

2.4.2 ConvNeXt V2

ConvNeXt, introduced by Liu et al. [38], revisited the design of pure convolutional architectures in light of the success of Vision Transformers (ViTs) [39]. Starting from a standard ResNet, the authors systematically adopted design principles from ViTs, including larger kernel sizes (7×7), inverted bottleneck blocks, fewer activation functions, Layer Normalisation, and a "patchify" stem, to create a modernised convolutional architecture

that matched or exceeded ViT performance on ImageNet classification.

ConvNeXt V2 [40] further improved upon this design by introducing two key innovations: (1) a fully convolutional masked autoencoder (FCMAE) framework for self-supervised pretraining, which enables the model to learn visual representations by reconstructing randomly masked image patches; and (2) a Global Response Normalisation (GRN) layer that enhances inter-channel feature competition by normalising feature maps based on their aggregate spatial magnitudes. The GRN layer addresses a representation collapse problem observed in the FCMAE framework, where feature channels tend to become co-adapted and redundant.

ConvNeXtV2-Base, the variant employed in this thesis, contains approximately 89 million parameters and uses a $[3, 3, 27, 3]$ block configuration with a channel dimension of $[128, 256, 512, 1024]$. Despite being a pure convolutional architecture, its design incorporates many principles from Transformers, making it representative of the modern convergence between CNN and Transformer paradigms.

2.4.3 Transfer Learning for Small Medical Datasets

Transfer learning, the practice of initialising a model with weights pretrained on a large source dataset (typically ImageNet) before fine-tuning on a smaller target dataset, has become essential for medical image classification where training data is scarce [35]. Tajbakhsh et al. demonstrated that fine-tuned pretrained CNNs consistently outperformed networks trained from scratch on medical image tasks, with the advantage being most pronounced for smaller datasets. Yosinski et al. [41] showed that earlier layers of CNNs trained on natural images learn general features (edges, textures, colour gradients) that transfer well across domains, while later layers become increasingly task-specific.

For the present thesis, the dataset comprises only 48 spiral images, making transfer learning not merely beneficial but essential. Both ResNet-50 and ConvNeXtV2-Base were initialised with ImageNet-pretrained weights, with the convolutional feature extractor either frozen (training only the classification head) or fine-tuned with a reduced learning rate, depending on the hyperparameter configuration.

2.4.4 Comparison with Alternative Architectures

Several other CNN architectures have been considered in the medical imaging literature. VGGNet [42] established the principle of using very small (3×3) convolutional filters

stacked deeply, but its large parameter count (138M for VGG-16) makes it prone to overfitting on small datasets. DenseNet [43] introduced dense connections where each layer receives feature maps from all preceding layers, promoting feature reuse and reducing parameter counts, though this comes at the cost of increased memory consumption. EfficientNet [44] proposed a compound scaling method that uniformly scales network width, depth, and resolution using a fixed ratio, achieving state-of-the-art accuracy with fewer parameters. While each of these architectures has merits, ResNet-50 and ConvNeXtV2-Base were selected for this thesis because they represent, respectively, the most established and a state-of-the-art convolutional architecture, providing a comparison between classical and modern CNN design paradigms.

2.5 Explainability Methods for Deep Learning

As deep learning models are increasingly deployed in high-stakes domains such as healthcare, the need for transparency and interpretability has become paramount [45]. Explainable AI (XAI) encompasses a broad range of techniques that aim to make model decisions understandable to human users. In the context of image classification, visual explainability methods produce heatmaps or attribution maps that highlight image regions most influential in a model's prediction. This section reviews the main categories of explainability methods used in this thesis.

2.5.1 Gradient-Based Attribution Methods

Gradient-based methods compute the sensitivity of the model's output with respect to its input, using gradients to assign importance scores to individual pixels or features.

Integrated Gradients [46] computes attributions by integrating the model's gradients along a straight-line path from a baseline input (typically a black or zero image) to the actual input. For a model F and input $\mathbf{x}$ with baseline $\mathbf{x}'$, the integrated gradient for the i-th feature is:

$$\mathrm{IG}_i(\mathbf{x}) = (x_i - x_i') \times \int_{\alpha=0}^{1} \frac{\partial F\left(\mathbf{x}' + \alpha(\mathbf{x} - \mathbf{x}')\right)}{\partial x_i} \, \mathrm{d}\alpha \tag{2.1}$$

This method satisfies two desirable axioms: *sensitivity* (if an input feature changes the prediction, it receives a non-zero attribution) and *implementation invariance* (functionally equivalent networks produce identical attributions). In practice, the integral is approximated by summing gradients at a finite number of interpolation steps (typically 50–300).

GradientSHAP [47] approximates SHAP values by computing the expectations of gradi-

ents with respect to randomly sampled baselines from a reference distribution. For each input, the method selects random points on the interpolation path between the baseline and input, computes gradients at these points, and uses the results to estimate Shapley values. This combines the axiomatic foundations of Shapley values from cooperative game theory with the computational efficiency of gradient-based methods.

2.5.2 Class Activation Mapping Methods

Class Activation Mapping (CAM) methods produce spatial heatmaps by leveraging the feature maps of convolutional layers to identify which image regions contribute most to a specific class prediction. These methods are particularly suited for CNNs and produce intuitive, spatially coherent visualisations.

Grad-CAM (Gradient-weighted Class Activation Mapping) [6] is the most widely used CAM-based method. It computes the importance weights of each feature map in the final convolutional layer by global-average-pooling the gradients of the target class score with respect to that layer's activations. For the k-th feature map A^k of the last convolutional layer and class score y^c:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{2.2}$$

The final heatmap is the ReLU-weighted combination of feature maps: $L^c_{\text{Grad-CAM}} = \mathrm{ReLU}\left(\sum_k \alpha_k^c A^k\right)$. The ReLU ensures that only features with a positive influence on the class of interest are visualised.

Grad-CAM++ [48] improves upon Grad-CAM by introducing pixel-wise weighting of gradients, leading to better localisation when multiple instances of the target class are present in the image. It uses second- and third-order derivatives to compute more precise importance weights for each spatial location within a feature map.

Score-CAM [49] eliminates the dependence on gradients entirely. Instead, it uses each activation map as a spatial mask, applies it to the input image, and measures the resulting change in the model's output score. The contribution of each activation map is thus determined by the model's confidence when only that spatial pattern is preserved. This approach produces more stable and less noisy visualisations, at the cost of increased computational expense (one forward pass per activation map).

Ablation-CAM [50] takes a complementary approach: the importance of each activation map is estimated by systematically ablating (zeroing out) each map and observing the resulting drop in the target class score. This gradient-free method provides a direct

measurement of each feature map's contribution to the prediction. Like Score-CAM, it requires multiple forward passes but avoids the noise and saturation issues that can affect gradient-based methods.

2.5.3 Perturbation-Based Methods

Perturbation-based methods explain predictions by systematically altering the input and observing the effect on the model's output, without requiring access to model internals.

LIME (Local Interpretable Model-agnostic Explanations) [7] explains individual predictions by fitting an interpretable surrogate model (typically a linear model) to the behaviour of the black-box model in the local neighbourhood of the input. For image classification, LIME first segments the image into superpixels, then generates perturbed instances by randomly masking subsets of superpixels. The black-box model's predictions on these perturbed inputs are used to train a weighted linear model, whose coefficients indicate the importance of each superpixel. LIME's model-agnostic nature makes it applicable to any classifier, but it can suffer from instability, as different random seeds may produce different explanations for the same input, and the quality of explanations depends heavily on the superpixel segmentation.

2.5.4 Evaluating Explainability Methods

A critical but often neglected aspect of XAI research is the systematic evaluation of explanation quality. Several approaches have been proposed to quantitatively assess visual explanations:

Insertion and deletion metrics [51] evaluate explanations by progressively inserting (or deleting) pixels in order of their attributed importance and measuring the resulting change in model confidence. A good explanation should identify pixels whose insertion rapidly increases (or whose deletion rapidly decreases) the model's predicted probability for the target class. The area under the insertion curve (AUC-Insertion) and area under the deletion curve (AUC-Deletion) provide scalar summary metrics.

ROAR (RemOve And Retrain) [52] provides a more rigorous evaluation by removing the most important features as identified by an explanation method, then *retraining* the model on the modified dataset and measuring the accuracy drop. This addresses a limitation of deletion metrics, where the distribution shift caused by removing pixels can confound the evaluation.

Sanity checks [53] test whether saliency maps are truly sensitive to the model's learned parameters and the input data. Adebayo et al. demonstrated that some widely used attribution methods produce visually similar saliency maps even when model weights are randomised, indicating that they may reflect input image structure rather than learned features. Their proposed tests include cascading randomisation of network layers and data randomisation.

Evaluation by domain experts represents a complementary qualitative approach, where clinicians assess whether highlighted regions correspond to known disease markers [54, 55]. Tonekaboni et al. surveyed clinicians about their requirements for explainable ML systems, finding that clinicians valued explanations that were consistent, aligned with clinical knowledge, and presented at an appropriate level of detail.

Samek et al. [56] proposed a perturbation-based evaluation framework where pixels are iteratively removed in order of their attributed relevance, measuring how quickly the classifier's output degrades. Their work highlighted significant differences in explanation quality across methods and advocated for standardised evaluation protocols.

2.6 Summary and Research Gap

The literature reviewed above establishes several key points. First, PD motor assessment through drawing tests is clinically validated and provides measurable biomarkers of motor dysfunction [18, 19, 20]. Second, deep learning approaches have been successfully applied to various PD diagnostic modalities, including handwriting and drawing analysis [27, 12, 28]. Third, transfer learning with pretrained CNN architectures is essential for small medical image datasets [35]. Fourth, a rich ecosystem of XAI methods exists for generating visual explanations of CNN predictions [6, 7, 47, 46].

However, a significant gap remains at the intersection of these domains. While prior work has applied CNNs to PD drawing classification [27, 33] and some studies have incorporated limited XAI analysis [29, 30], **no existing study has conducted a comprehensive comparison of multiple XAI methods across multiple CNN architectures specifically for PD spiral drawing analysis**. Most existing studies either focus on classification performance without interpretability, or apply a single explainability method without comparative analysis.

Furthermore, the DraWritePD dataset [13] has primarily been studied using handcrafted kinematic features from dynamic pen trajectory data. The application of deep learning to the *static* spiral images from this dataset, combined with systematic XAI analysis, remains

unexplored.

This thesis addresses this gap by: (1) conducting extensive hyperparameter optimisation (3,000 configurations per architecture) of two CNN architectures, ResNet-50 and ConvNeXtV2-Base, on DraWritePD spiral images; and (2) applying six XAI methods (Grad-CAM, Score-CAM, Ablation-CAM, Integrated Gradients, GradientSHAP, and LIME) to interpret model decisions, comparing their visual explanations both across methods and across architectures.

3. Materials

3.1 Technologies and Libraries

3.1.1 Python and PyTorch

All experiments were conducted using the Python programming language with the PyTorch [57] library and TorchVision [58]. These were chosen due to the author's familiarity with both and their prevalence in the machine learning and computer vision research communities. This choice also enabled the use of the extensive ecosystem of Python libraries, avoiding the need to reimplement most of the techniques used in this work.

3.1.2 Weights and Biases

Weights and Biases (W&B) [59] provides a hosted platform for tracking and visualising the results of large-scale experiments, as well as performing hyperparameter optimisation through its Sweeps feature. Some of the hyperparameter optimisation runs were conducted using W&B Sweeps; however, only runs where the configuration could be expressed as a flat, static structure (all possible hyperparameters predefined and independent of one another) used this feature. W&B also served as a platform for storing model weights from experiments that achieved 0.75 accuracy or higher, as storage constraints prevented saving all model checkpoints. By the end of all experimentation phases, approximately 1.8 terabytes of data (predominantly model weights) and 5,343 tracked experiment hours were logged into W&B.

3.1.3 Optuna

The Optuna [60] library provided the hyperparameter optimisation capabilities used for the majority of experiments. Optuna was selected because it supports dynamic hyperparameter search spaces, enabling conditional hyperparameters where some parameters depend on the values of others (e.g., fine-tuning learning rate is only relevant when the fine-tuning strategy is not *none*).

3.1.4 Timm (pytorch-image-models)

The timm (pytorch-image-models) [61] library provides implementations of a wide variety of state-of-the-art computer vision models and optimisers. In this work, timm was used to load pretrained ResNet-50 and ConvNeXtV2-Base model weights and to access the RMSProp and Adam optimiser implementations.

3.1.5 Albumentations

The Albumentations [62] library provided implementations for the image augmentation techniques applied during training. Albumentations was chosen for its high performance and its ability to compose complex augmentation pipelines with per-transform probability control.

3.1.6 pytorch-Grad-CAM

The pytorch-Grad-CAM library [63] provides implementations for class activation map (CAM) based explainability methods, including Grad-CAM [6], Score-CAM [49], and Ablation-CAM [50]. These methods were applied to the final convolutional layer of each architecture to generate spatial heatmaps.

3.1.7 Captum

The Captum [64] library, developed by Facebook Research, provides implementations for Integrated Gradients [46], GradientSHAP [47], and LIME [7] explainability methods used in this work.

3.1.8 Hardware and Computational Environment

All model training and hyperparameter optimisation experiments were conducted on the High Performance Computing (HPC) Centre of TalTech [65]. Training jobs were submitted to the cluster's GPU partition, utilising NVIDIA A100 GPUs with 40GB of graphics memory. Each individual training run (one hyperparameter configuration) required approximately 2–5 minutes of GPU time, depending on the model architecture and augmentation intensity. The full hyperparameter optimisation campaign used as the final result (6,000 total configurations across both architectures) consumed approximately 5,343 GPU hours of tracked experiment time.

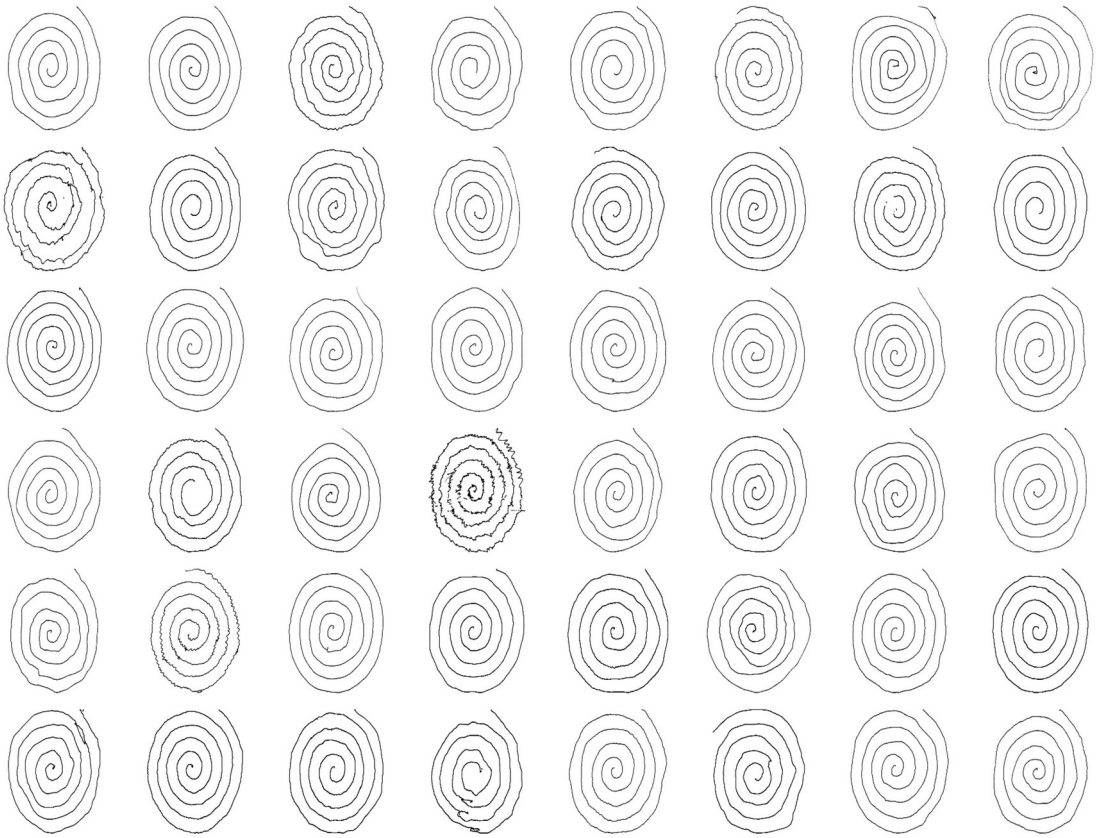

Figure 1. Example spiral drawings from the DraWritePD dataset. Each spiral was drawn by a different subject on a digitising tablet. The drawings exhibit varying degrees of regularity, spacing uniformity, and line smoothness.

3.2 Dataset

3.2.1 DraWritePD Spiral Dataset

The DraWritePD dataset, introduced by Valla et al. [13], was collected at Tartu University Hospital, Estonia, and consists of handwriting samples from a total of 24 PD patients and 34 age- and gender-matched healthy control subjects. From this dataset, 48 unique Archimedean spiral drawings (19 from PD patients, 29 from healthy controls) form the basis of all experiments in this thesis.

Each spiral was drawn on a Wacom digitising tablet, which captured both the static drawing image and dynamic pen trajectory data including position, pressure, tilt, and timestamps. For this thesis, only the static spiral images were used, treating the classification task as a pure image analysis problem. The images were preprocessed to isolate the spiral drawing with consistent line width (1 pixel) on a white background. All images were resized to 224×224 pixels to match the expected input dimensions of the pretrained CNN architectures.

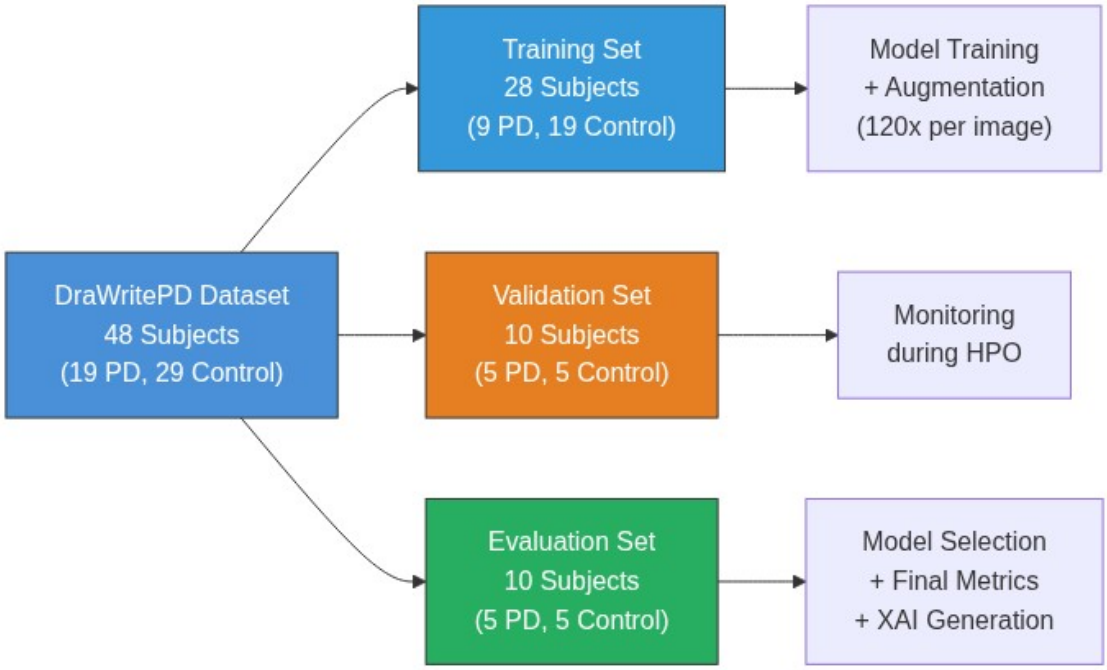

Figure 2. Subject-wise data split of the DraWritePD dataset into training, validation, and evaluation partitions, showing class distribution and the purpose of each split.

The dataset was split at the **subject level** to prevent data leakage (ensuring that augmented versions of the same subject's drawing never appeared in both training and evaluation partitions). The split was as follows:

- **Training set**: 28 subjects (9 PD, 19 controls)
- **Validation set**: 10 subjects (5 PD, 5 controls)
- **Test set**: 10 subjects (5 PD, 5 controls)

The class imbalance in the training set (approximately 32% PD, 68% control) was addressed through weighted loss functions, where the loss contribution of each class was inversely proportional to its frequency in the training set. Figure 2 illustrates the subject-wise partitioning and the role of each split.

3.3 Models

3.3.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a class of deep learning models that have demonstrated exceptional performance in image recognition and classification tasks. The fundamental building blocks of CNNs are convolutional layers, pooling layers, and fully connected layers. The convolutional layer applies a set of learnable filters to the input,

each producing a feature map. This operation can be mathematically represented as:

$$f_{i,j} = \sum_m \sum_n I_{i+m,j+n} \cdot K_{m,n}. \tag{3.1}$$

where I is the input image, K is the kernel, and $f_{i,j}$ is the output feature map. The pooling layer reduces the spatial dimensions of the representation, thereby reducing the number of parameters and controlling overfitting. The fully connected layer takes the high-level features learned by convolutional layers and uses them to classify the input image. CNNs are designed to exploit the two-dimensional structure of images through local connections and shared weights, followed by pooling operations that produce translation-invariant features. Explainability methods designed for CNNs leverage this spatial structure and do not always translate well to other architectures.

ResNet-50

ResNet (Residual Networks), introduced by He et al. [36], addressed the vanishing gradient problem in deep networks through the introduction of skip connections that allow gradients to flow directly through residual blocks. The variant used in this thesis, **ResNet-50**, consists of 50 layers organised into four stages of bottleneck residual blocks, with approximately **25.6 million parameters**. The model was loaded with weights pretrained on ImageNet-1k [37] via the timm library. The original 1000-class classification head was replaced with a custom binary classification head consisting of a fast average-maximum pooling layer, an optional 512-unit hidden layer with ReLU activation, and a final output layer.

ConvNeXtV2-Base

ConvNeXt V2 [40] builds upon the original ConvNeXt [38], which modernised the ResNet architecture by adopting design principles from Vision Transformers (larger 7×7 kernels, inverted bottleneck blocks, Layer Normalisation). ConvNeXt V2 further adds a fully convolutional masked autoencoder (FCMAE) framework for self-supervised pretraining and a Global Response Normalisation (GRN) layer to enhance inter-channel feature competition. The variant used in this thesis, **ConvNeXtV2-Base**, has approximately **89 million parameters** with a block configuration of $[3, 3, 27, 3]$ and channel dimensions of $[128, 256, 512, 1024]$. As with ResNet-50, the model was loaded with ImageNet-pretrained weights and equipped with a custom binary classification head using fast average-maximum pooling. Figure 3 shows the shared classification pipeline used by both architectures.

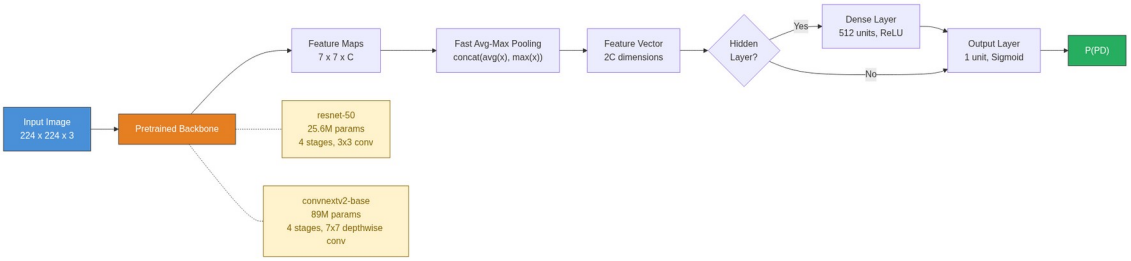

Figure 3. Model architecture for binary PD classification. Both ResNet-50 and ConvNeXtV2-Base share the same classification head: fast average-maximum pooling, an optional hidden layer, and a sigmoid output.

3.4 Explainability Methods

3.4.1 Integrated Gradients

This technique, introduced by Sundararajan et al. [46], computes attributions by integrating the model's gradients along a straight-line path from a baseline input (a black image) to the input of interest. The integral quantifies the contribution of each input feature to the prediction, satisfying the axioms of sensitivity and implementation invariance. The Captum implementation was used with 50 interpolation steps and a zero baseline.

3.4.2 LIME

LIME (Local Interpretable Model-agnostic Explanations), proposed by Ribeiro et al. [7], explains individual predictions by fitting a local interpretable surrogate model around the input. For images, LIME segments the input into superpixels, generates perturbed instances by masking subsets of superpixels, and trains a weighted linear model on the black-box model's predictions for these perturbed inputs. The linear model's coefficients indicate the importance of each superpixel region.

3.4.3 GradientSHAP

GradientSHAP [47] approximates SHAP values by computing the expectations of gradients with respect to randomly sampled baselines. The Captum implementation randomly selects baselines from a reference distribution and points on the interpolation path between baseline and input, computing gradients at these random points. The resulting SHAP values represent the expected gradients multiplied by the input-baseline difference, providing a theoretically grounded measure of feature importance.

3.4.4 CAM Methods

Class Activation Maps (CAMs) are techniques that visualise the spatial regions of an input image that are most important for a model's prediction. These methods operate on the activation maps of convolutional layers and are particularly suited for CNN architectures.

Grad-CAM

Grad-CAM [6] uses the gradients of the target class flowing into the final convolutional layer to produce a coarse localisation map. The importance weights for each feature map are computed by global-average-pooling the gradients, and the final heatmap is the ReLU-weighted combination of feature maps.

Score-CAM

Score-CAM [49] eliminates gradient dependence by using each activation map as a spatial mask, feeding the masked input through the model, and scoring each map based on the resulting change in output confidence. This gradient-free approach produces more stable visualisations at the cost of additional forward passes.

Ablation-CAM

Ablation-CAM [50] estimates feature map importance by systematically zeroing out each activation map and measuring the resulting drop in class score. This ablation-based approach provides a direct measurement of each feature map's contribution without relying on gradient approximations.

3.5 Metrics

The metrics used for tracking training progress, evaluating models, and guiding hyperparameter optimisation were implemented by the author in Python/PyTorch and are defined as follows:

$$\begin{aligned} \text{Accuracy} &= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \\ \text{Sensitivity} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ \text{Specificity} &= \frac{\text{TN}}{\text{TN} + \text{FP}}, \\ \text{F}_1 &= \frac{2 \times \text{TP}}{2 \times \text{TP} + \text{FP} + \text{FN}}, \\ J &= \text{Sensitivity} + \text{Specificity} - 1. \end{aligned} \tag{3.2}$$

where TP is the number of true positives (model predicted PD and the ground truth is PD), TN is the number of true negatives, FP is the number of false positives, FN is the number of false negatives, and J is Youden's J statistic, which summarises the diagnostic effectiveness as the sum of sensitivity and specificity minus one. Additionally, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was computed using PyTorch to evaluate the model's ability to discriminate between classes across all classification thresholds.

4. Methods

4.1 Data Splitting and Preprocessing

Figure 4 provides an overview of the full experimental pipeline described in the following sections.

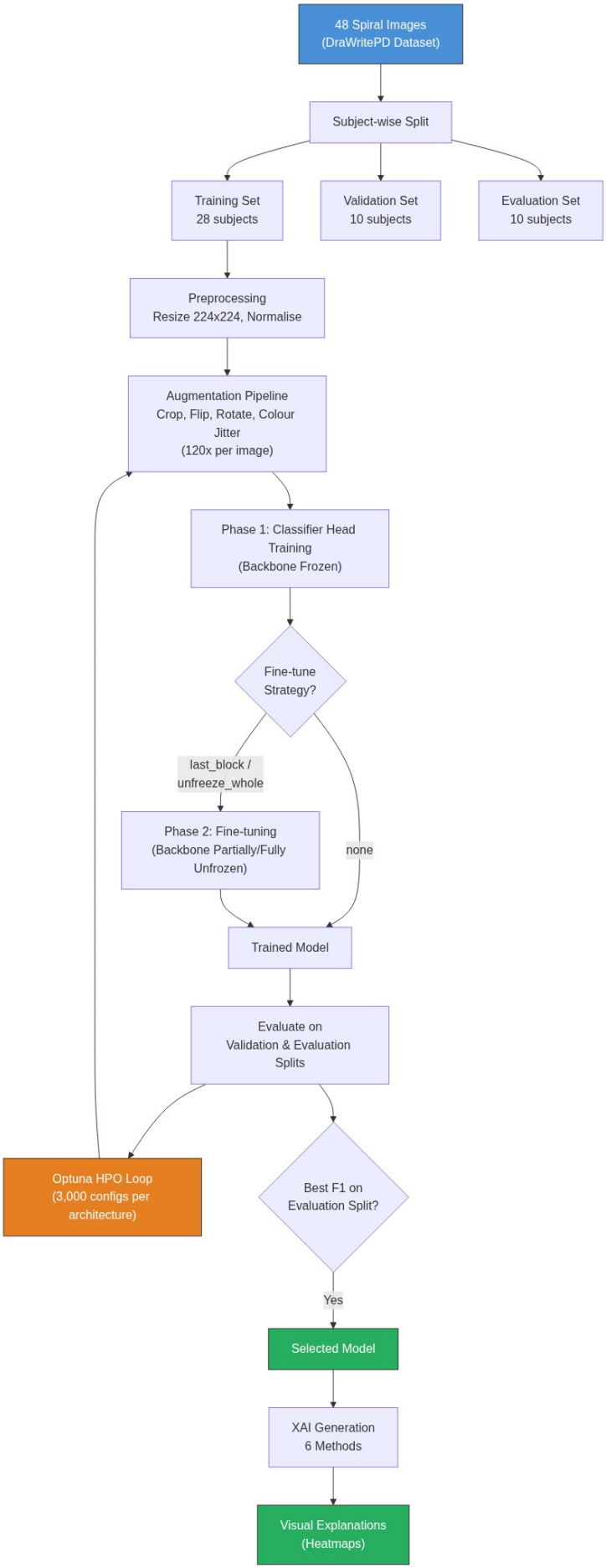

Figure 4. Overview of the experimental pipeline, from data splitting through hyperparameter optimisation to explainability analysis.

The 48 spiral images were divided into training, validation, and test sets using a **subject-wise split** strategy. This ensures that all images derived from a single subject (including augmented versions) appear in only one partition, preventing data leakage that would inflate performance estimates. The split maintained balanced class representation in the validation and test sets:

- **Training set**: 28 subjects (9 PD, 19 controls), used for model weight optimisation.
- **Validation set**: 10 subjects (5 PD, 5 controls), used for monitoring and comparison during experimentation.
- **Test set**: 10 subjects (5 PD, 5 controls), used as the model-selection evaluation split and for the final reported metrics in this thesis.

The same fixed split was used across all 6,000 hyperparameter configurations (3,000 per architecture) to ensure fair comparison. The split was defined by explicit subject ID lists stored in the experiment configuration. The subject IDs were:

- **Validation subjects**: PD-16, PD-13, PD-7, PD-21, PD-10, KT-24, KT-4, KT101, KT_112, KT_113
- **Test subjects**: PD-2, PD-11, PD-3, PD-4, PD-22, KT_114, KT-25, KT-13, KT_107, KT_109

The remaining 28 subjects formed the training set. This fixed split improves traceability of the experiments, although full reproducibility would additionally require explicit reporting of the random seeds and exact checkpoint provenance.

All input images were preprocessed by rendering the spiral trajectory as black lines (1 pixel width) on a white background, then resizing to 224×224 pixels using bilinear interpolation to match the expected input dimensions of the pretrained CNN architectures. Pixel values were normalised to the $[0, 1]$ range and then standardised using the ImageNet channel means ($[0.485, 0.456, 0.406]$) and standard deviations ($[0.229, 0.224, 0.225]$), as required for models pretrained on ImageNet.

4.2 Augmentation Pipeline

Given the extremely small training set (28 subjects), data augmentation was critical for preventing overfitting and improving model generalisation. Each training image was sampled through the augmentation pipeline multiple times per epoch, controlled by the *augments_per_image* hyperparameter (set to 120 in the best configurations). The

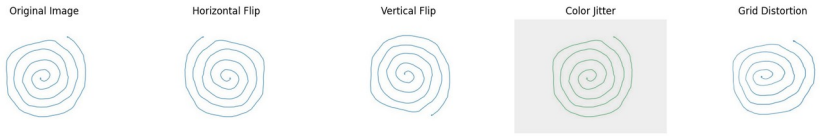

Figure 5. Examples of augmented spiral images produced by the augmentation pipeline, demonstrating the range of transformations applied to each training image.

augmentation pipeline, implemented using the Albumentations library [62], consisted of the following transforms applied in sequence:

1. **Random crop with resize**: The original 224×224 image was first upscaled, then a random 224×224 crop was taken. The crop percentage hyperparameter controlled how much of the upscaled image the crop covered, effectively creating random zoom and translation effects. This augmentation was particularly important for spiral images, which contain significant white space around the main drawing.
2. **Horizontal flip**: Applied with 50% probability, mirroring the spiral horizontally.
3. **Vertical flip**: Applied with 50% probability, mirroring the spiral vertically.
4. **Random rotation**: Applied with a rotation range of ± 180 degrees, enabling arbitrary orientation changes.
5. **Colour jitter**: Random perturbations to brightness, contrast, and saturation, applied to reduce sensitivity to imaging conditions.

The combination of flips, rotations, and crops meant that the effective training set size was multiplied by the augments_per_image factor (120), yielding approximately 3,360 augmented training samples per epoch from the original 28 images. This aggressive augmentation strategy was essential for training deep networks with millions of parameters on such a small dataset. In the final configurations, all geometric augmentations (flip, rotation, crop) and colour jitter were always enabled, as earlier experiments indicated that each contributed positively to generalisation.

4.3 Hyperparameter Optimisation

The core methodology of this work involved systematic hyperparameter optimisation followed by the application of explainability methods to the best-performing models. The hyperparameters explored across all experiments included:

- *base_model*: The CNN architecture and pretrained weights used as a feature extractor (ResNet-50 or ConvNeXtV2-Base). A classification head (multi-layer perceptron or single dense layer) was appended for binary classification.
- *augments_per_image*: Number of augmented samples generated per original image (range: 20–200).
- *batch_size*: Number of samples processed per gradient update.
- *optimizer_name*: Adam with weight decay [66, 67] or RMSProp (centred variant) [68].
- *lookahead*: Whether to wrap the base optimiser with Lookahead [69], which maintains slow-moving weights for stabilisation.
- *learning_rate*: Learning rate for the classifier head training phase.
- *weight_decay*: L2 regularisation coefficient.
- *classifier_hidden_size*: Size of the optional hidden layer in the classification head (or *None* for a direct linear classifier).
- *classifier_pool*: Pooling strategy applied to the feature extractor output (fast average-maximum pooling).
- *crop_percentage*: Controls the zoom level of the random crop augmentation.
- *finetune_strategy*: One of:
  - *none*: Only the classification head is trained; the feature extractor remains frozen.
  - *last_block*: The last convolutional block/stage of the feature extractor is unfrozen and trained.
  - *unfreeze_whole*: The entire feature extractor is unfrozen and trained with a separate, lower learning rate.
- *finetune_learning_rate*: Learning rate for the fine-tuning phase (when applicable).

Optimisation was conducted using Optuna [60] with its Tree-structured Parzen Estimator (TPE) sampler, which models the search space probabilistically and focuses sampling on promising regions. Each optimisation run explored 3,000 configurations per architecture. In the final experimental setup, model comparison and selection were performed using the test F1 score. F1 was chosen because it balances precision and recall and produced the most stable ranking among the candidate runs during experimentation. Consequently, the reported "test" metrics in this thesis should be interpreted as results on a fixed model-

selection evaluation split rather than as a strictly untouched final generalisation estimate. The split was kept balanced (5 PD, 5 control) to ensure that accuracy and F1 were both meaningful metrics.

4.4 Training and Fine-Tuning

The training procedure followed a two-phase approach:

Phase 1: Classifier head training: The pretrained feature extractor was frozen (all parameters fixed), and only the classification head was trained using binary cross-entropy loss with class weights. This phase allowed the randomly initialised classification head to learn meaningful class boundaries from the pretrained features without disrupting the feature extractor's learned representations.

Phase 2: Fine-tuning (when *finetune_strategy* $\neq$ *none*): After Phase 1, the specified layers of the feature extractor were unfrozen and the entire model was trained end-to-end with a separate, typically much lower, learning rate for the feature extractor. This allowed the pretrained features to be adapted to the specific visual characteristics of spiral drawings while minimising catastrophic forgetting of the general features learned during ImageNet pretraining.

The training loop, implemented by the author in Python/PyTorch, can be represented by the following pseudocode:

Algorithm 1 Training Loop for Augmented Dataset

for `images`, `labels` in `augmented_dataset` **do**
    Move `images` and `labels` to GPU
    Zero the optimiser gradients
    Compute `logits` from the model
    Compute weighted binary cross-entropy `loss`
    Backpropagate `loss` and step optimiser
    Calculate and log metrics (accuracy, F1, sensitivity, AUC)
end for

Due to the small dataset size and aggressive augmentation, models typically converged within a single epoch (one complete pass over all augmented training samples). The combination of 120 augmentations per image and 28 training subjects meant that each epoch contained approximately 3,360 training samples, which was sufficient for the classification head to reach stable performance. Mixed-precision training (FP16) was enabled via PyTorch's automatic mixed precision to reduce memory usage and accelerate computation on the HPC GPU nodes.

4.5 Visual Explanation Generation

After hyperparameter optimisation, one high-performing model was selected from each architecture for explainability analysis. The selected models were the runs that achieved the highest test F1 score on the fixed evaluation split, while also being checked against the corresponding training and validation metrics to avoid choosing obviously unstable runs.

Visual explanations were generated for all 10 images in the fixed evaluation split using six methods:

- **CAM-based methods** (Grad-CAM, Score-CAM, Ablation-CAM): Applied to the final convolutional layer of each architecture (layer4 for ResNet-50, stages[3] for ConvNeXtV2-Base). The resulting 7×7 activation-resolution heatmaps were upsampled to 224×224 using bilinear interpolation.
- **Gradient-based methods** (Integrated Gradients, GradientSHAP): Applied at the input pixel level, producing full 224×224 attribution maps. Integrated Gradients used a zero (black) baseline with 50 interpolation steps. GradientSHAP used random samples from a Gaussian noise distribution as baselines.
- **Perturbation-based methods** (LIME): Applied with automatic superpixel segmentation of the input image. The surrogate linear model was fitted on 1,000 perturbed samples per image.

All heatmaps were normalised to the $[0, 1]$ range and overlaid on the original resized image using a blue-to-red colour map, where blue indicates low relevance and red indicates high relevance to the model's prediction. The overlay transparency was set to allow the original spiral drawing to remain visible beneath the heatmap.

5. Results

5.1 Classification Performance

The final validation experiment consisted of two large-scale hyperparameter optimisation runs, one for ResNet-50 and one for ConvNeXtV2-Base, each exploring 3,000 unique configurations using Optuna with Weights & Biases tracking. The models were evaluated on a fixed 10-subject evaluation split (5 PD, 5 control), using subject-wise partitioning to prevent data leakage between training and evaluation subjects.

Table 2 presents the classification performance of the best-performing model for each architecture, selected by the highest F1 score on this fixed evaluation split. Both architectures achieved 90% accuracy, but with different error profiles: ResNet-50 achieved perfect specificity (1.00) while missing one PD patient (sensitivity = 0.80), whereas ConvNeXtV2-Base achieved perfect sensitivity (1.00) while misclassifying one healthy control (specificity = 0.80).

Table 2. Best classification performance on the fixed evaluation split for each architecture.

Model	Accuracy	F1 Score	Sensitivity	Specificity	AUC
ResNet-50	0.90	0.889	0.80	1.00	0.96
ConvNeXtV2-Base	0.90	0.909	1.00	0.80	1.00

Both models achieved a Youden's J statistic of 0.80. The ConvNeXtV2-Base model achieved a slightly higher F1 score (0.909 vs. 0.889) due to its perfect sensitivity, which is clinically preferable in a screening context where missing a PD case (false negative) carries greater cost than a false alarm. The ConvNeXtV2-Base model also achieved a perfect AUC of 1.00 on the evaluation split, indicating perfect ranking of PD and control cases across decision thresholds, compared to 0.96 for ResNet-50.

It is important to contextualise these results within the limitations of the evaluation split size. With only 10 evaluation subjects (5 per class), each misclassification shifts the accuracy by 10 percentage points, and the difference between 0.80 and 1.00 sensitivity represents a single subject. Because this same split was used for model selection, the observed performance metrics should be interpreted as optimistic selection-split results rather than as definitive final generalisation estimates.

Table 3 presents the confusion matrices for both models, illustrating their distinct error patterns.

Table 3. Confusion matrices for the best ResNet-50 and ConvNeXtV2-Base models on the fixed evaluation split. Rows represent ground truth labels and columns represent predicted labels.

	ResNet-50		**ConvNeXtV2-Base**	
	Pred. PD	Pred. Control	Pred. PD	Pred. Control
Actual PD	4 (TP)	1 (FN)	5 (TP)	0 (FN)
Actual Control	0 (FP)	5 (TN)	1 (FP)	4 (TN)

The ResNet-50 model misclassified subject PD-22 as a healthy control, while correctly identifying all five control subjects. Conversely, the ConvNeXtV2-Base model correctly identified all five PD patients but misclassified one healthy control as PD-positive. This complementary error pattern suggests that an ensemble approach combining both architectures could potentially improve overall performance, as the models appear to learn partially different decision boundaries.

5.2 Hyperparameter Analysis

The hyperparameter optimisation explored a wide range of configurations including augmentation intensity (augments per image), classifier architecture (hidden size, pooling strategy), fine-tuning strategy, learning rate, optimiser choice, weight decay, and momentum. The parallel coordinates plots in Figures 6 and 7 visualise the relationship between hyperparameter settings and the resulting test F1 score across all 3,000 configurations for each architecture.

Table 4 summarises the best hyperparameter configuration for each architecture.

Table 4. Best hyperparameter configurations for each architecture.

Hyperparameter	**ResNet-50**	**ConvNeXtV2-Base**
Augments per image	120	120
Classifier hidden size	512	None (linear)
Classifier pooling	Fast AvgMax	Fast AvgMax
Optimiser	RMSProp	Adam
Learning rate	2.1×10^{-4}	4.0×10^{-5}
Weight decay	0.056	0.027
Lookahead	Yes	Yes
Finetune strategy	Whole	None
Finetune learning rate	2.0×10^{-6}	N/A
Colour jitter	Yes	Yes
Weighted loss	Yes	Yes

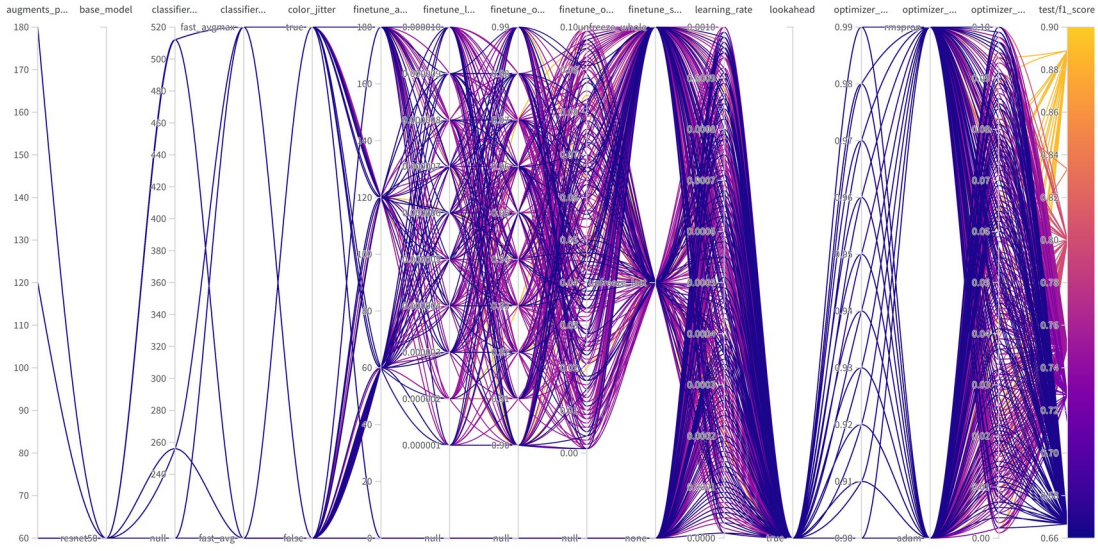

Figure 6. Parallel coordinates plot for ResNet-50 hyperparameter optimisation (3,000 configurations). Each line represents one configuration, coloured by test F1 score. The rightmost axis shows the resulting F1 score.

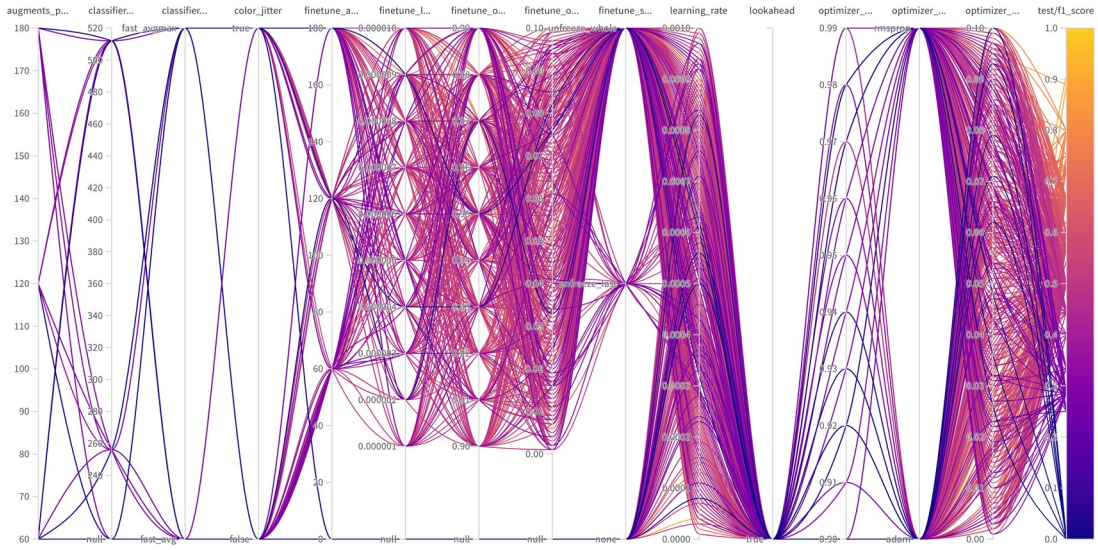

Figure 7. Parallel coordinates plot for ConvNeXtV2-Base hyperparameter optimisation (3,000 configurations). Each line represents one configuration, coloured by test F1 score.

Several patterns emerged from the hyperparameter analysis:

Fine-tuning strategy was the most influential hyperparameter for ResNet-50. The best ResNet-50 configuration used full fine-tuning (*unfreeze_whole*), where the entire pretrained feature extractor was unfrozen and trained with a very low learning rate (2.0×10^{-6}) after initial classifier head training. This suggests that adapting the learned feature representations to the specific visual characteristics of spiral drawings was important for ResNet-50's performance.

In contrast, the best ConvNeXtV2-Base configuration used *no fine-tuning* of the feature extractor, training only the classification head. This indicates that ConvNeXtV2-Base's pretrained features, learned through the FCMAE self-supervised pretraining framework, were sufficiently general to capture relevant spiral drawing features without task-specific adaptation. This is a notable finding, as ConvNeXtV2-Base has approximately 3.5× more parameters than ResNet-50 (89M vs. 25.6M), and freezing the feature extractor effectively reduces the number of trainable parameters from millions to thousands, significantly reducing the risk of overfitting on the small dataset.

Augmentation intensity: Both architectures benefited from aggressive augmentation (120 augmentations per image), reflecting the critical need to artificially increase the effective training set size from 28 subjects. This high augmentation factor, combined with the full complement of augmentations (horizontal flip, vertical flip, rotation, colour jitter, random crop), ensured that the model saw diverse variations of each training spiral.

Optimiser choice: ResNet-50 performed best with RMSProp while ConvNeXtV2-Base favoured Adam, both with Lookahead enabled. The Lookahead wrapper [69] maintains a slow-moving set of weights that the fast optimiser periodically synchronises with, providing a stabilising effect that may be particularly beneficial for small, noisy datasets.

Learning rate: ConvNeXtV2-Base required a lower learning rate (4.0×10^{-5}) than ResNet-50 (2.1×10^{-4}), consistent with the general observation that larger models with more parameters benefit from smaller learning rates to avoid destabilising pretrained representations.

Classifier architecture: The best ResNet-50 model used a hidden layer of 512 units with fast average-maximum pooling, while ConvNeXtV2-Base used a simple linear classifier (no hidden layer). This further supports the observation that ConvNeXtV2-Base's pretrained features required less adaptation, allowing a simpler classification head to achieve strong performance.

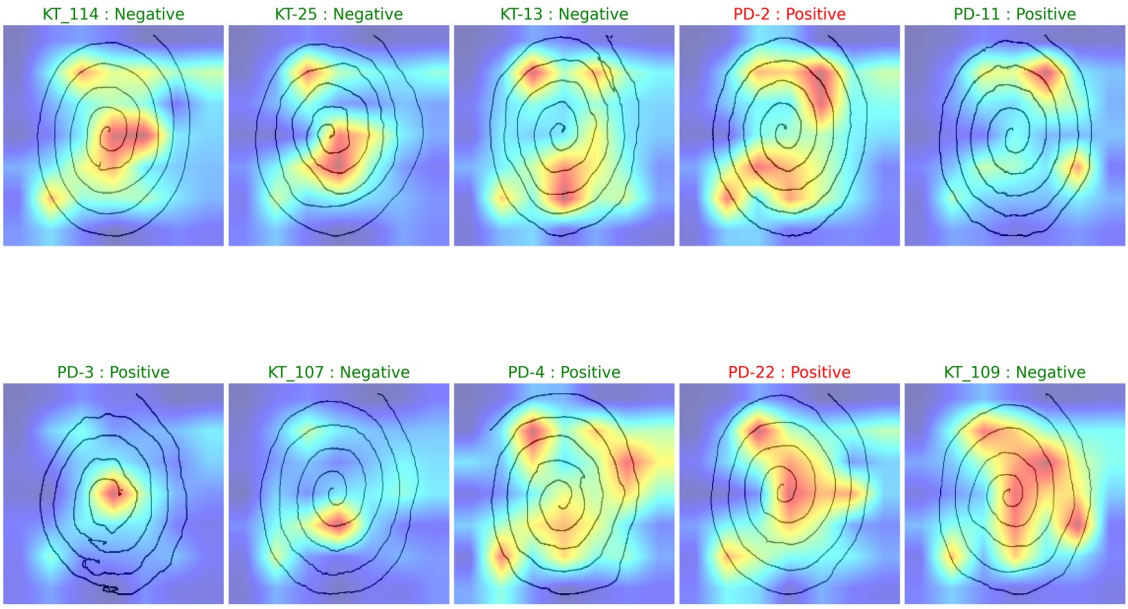

Figure 8. Integrated Gradients explanations on the ResNet-50 model. Heatmaps show pixel-level attribution, with warm colours indicating features that increased the predicted class probability.

5.3 Explainability Results

For each architecture, one high-performing model was selected to generate visual explanations using six XAI methods: Integrated Gradients, GradientSHAP, LIME, Grad-CAM, Score-CAM, and Ablation-CAM. The explanations were generated for all 10 images in the fixed evaluation split and visualised as relevancy heatmaps overlaid on the resized original images, where blue regions indicate low relevance and red regions indicate high relevance to the model's prediction. In each figure, the label above each image indicates the ground truth ("Positive" for PD, "Negative" for control), with green text for correct predictions and red text for incorrect predictions.

5.3.1 Integrated Gradients

Integrated Gradients (Figures 8 and 9) produced focused, high-resolution attributions that concentrated on specific regions of the spiral drawings. For the ResNet-50 model, the attributions tended to highlight the central region of spirals and specific points along the spiral trajectory where line characteristics changed, such as junctions between loops and areas of irregular spacing. For the ConvNeXtV2-Base model, the Integrated Gradients heatmaps showed a stronger focus on the spiral centre, with concentric patterns of activation that followed the spiral geometry. This suggests that ConvNeXtV2-Base relied more heavily on the structural centre of the drawing, while ResNet-50 distributed its attention across multiple regions.

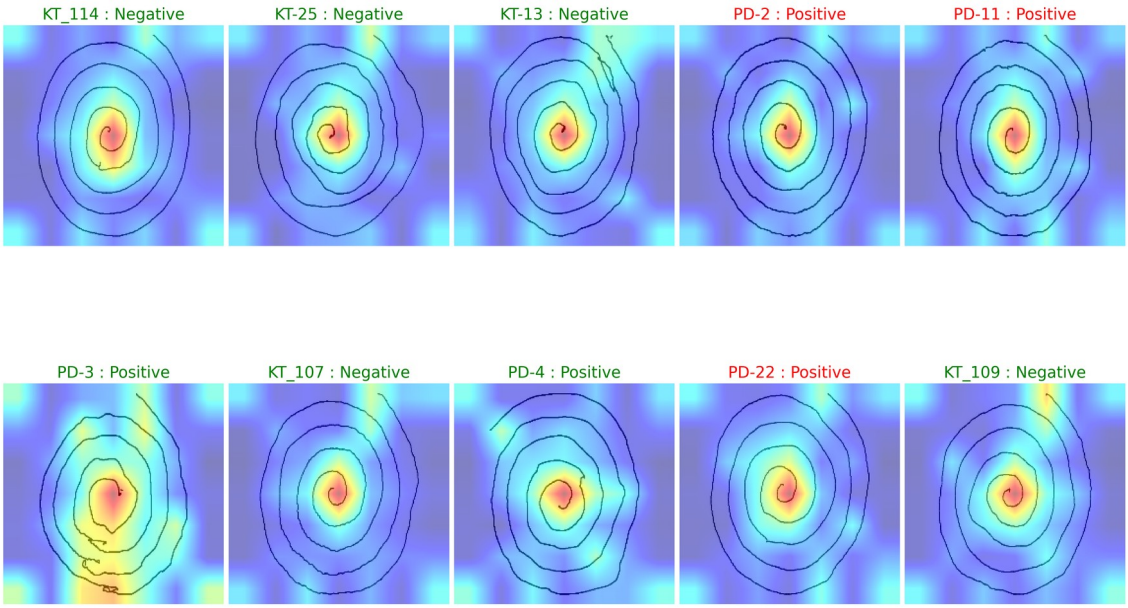

Figure 9. Integrated Gradients explanations on the ConvNeXtV2-Base model.

5.3.2 GradientSHAP

GradientSHAP (Figures 10 and 11) produced more diffuse and scattered attributions compared to Integrated Gradients. For ResNet-50, SHAP attributions appeared as dispersed highlights across the spiral lines and their immediate vicinity, with some concentration at line intersections and points of curvature change. For ConvNeXtV2-Base, the SHAP attributions showed a similar scattered pattern but with attributions distributed more broadly along the spiral trajectory. The stochastic nature of GradientSHAP, arising from random baseline sampling, contributes to the noisier appearance compared to the deterministic path integration used by Integrated Gradients.

5.3.3 LIME

LIME (Figures 12 and 13) produced the most sparse and localised explanations among all methods. The highlighted regions appeared as isolated bright spots, each corresponding to a superpixel segment that LIME identified as influential. For both architectures, LIME highlights were scattered across the drawing rather than forming spatially coherent regions, making them less immediately interpretable as clinically meaningful patterns. In some images, LIME highlighted points near the spiral centre, while in others, the highlights appeared at irregular intervals along the spiral lines or near the image edges.

The fragmented nature of LIME's explanations for this task likely reflects the mismatch between its superpixel-based segmentation approach and the continuous, fine-grained

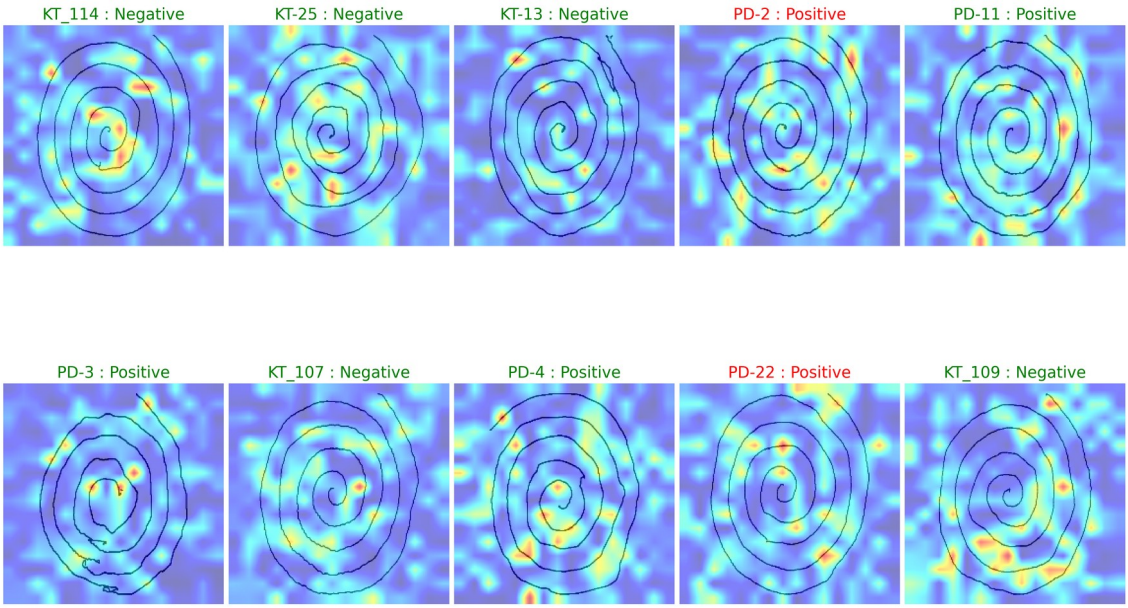

Figure 10. GradientSHAP explanations on the ResNet-50 model.

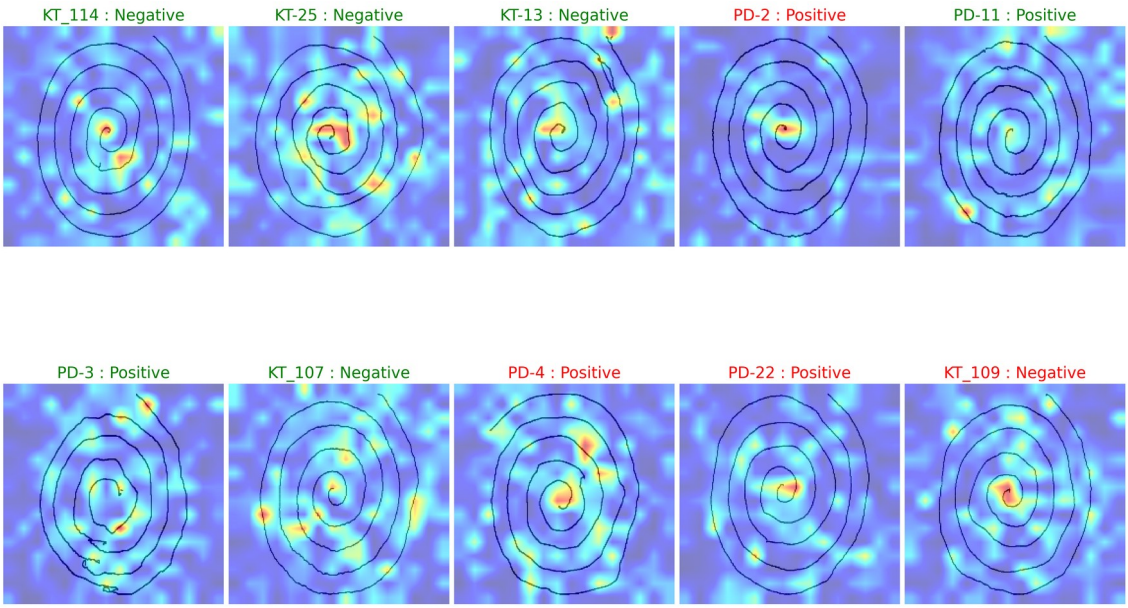

Figure 11. GradientSHAP explanations on the ConvNeXtV2-Base model.

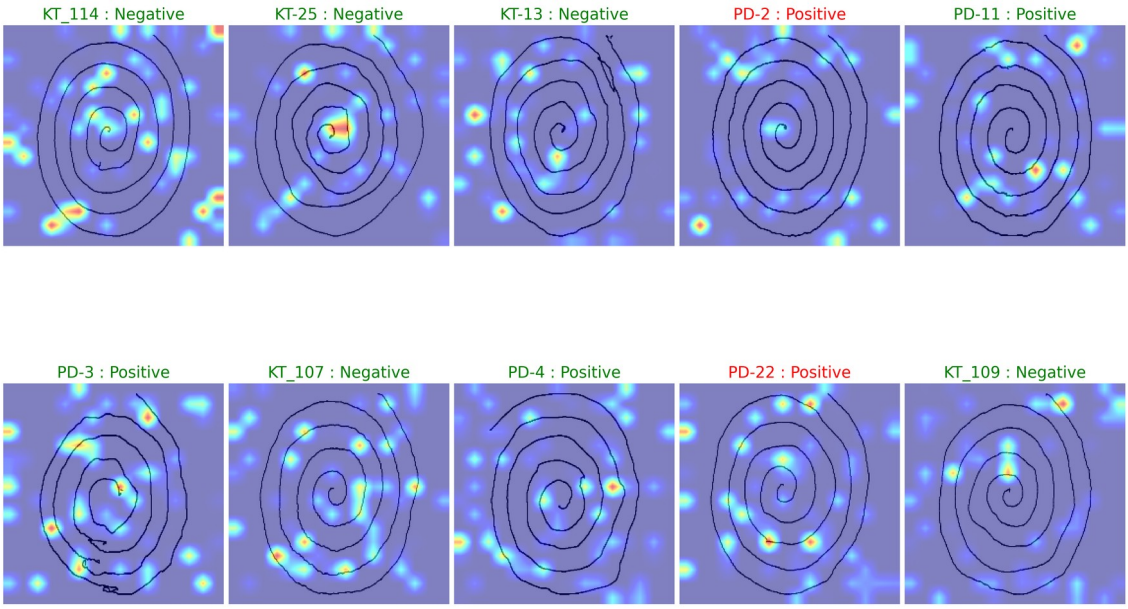

Figure 12. LIME explanations on the ResNet-50 model. Sparse highlights correspond to the most influential superpixel regions.

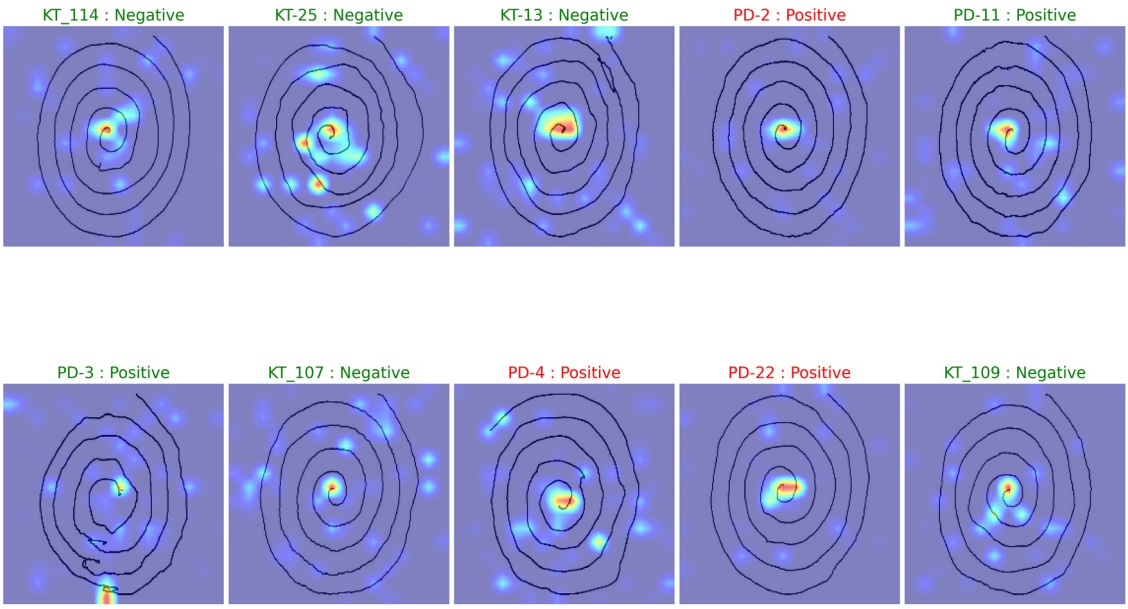

Figure 13. LIME explanations on the ConvNeXtV2-Base model.

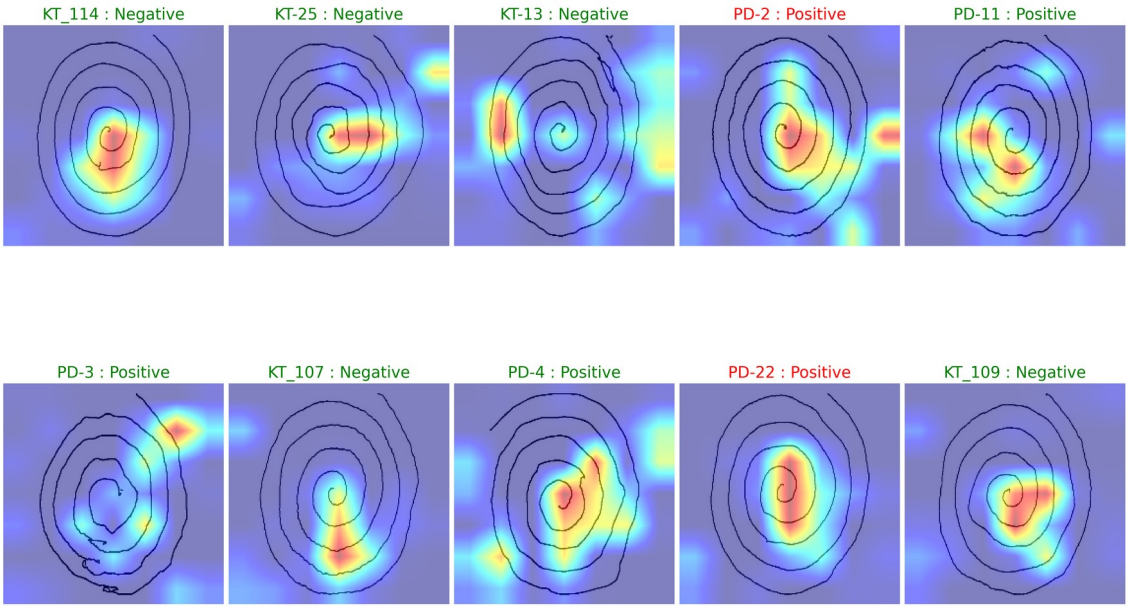

Figure 14. Grad-CAM explanations on the ResNet-50 model. Heatmaps highlight the spatial regions in the final convolutional layer that were most discriminative for classification.

nature of spiral drawing features. Spiral irregularities caused by tremor or bradykinesia are distributed along the entire trajectory rather than being confined to discrete segments, which challenges LIME's assumption that importance can be attributed to independent superpixel regions.

5.3.4 Grad-CAM

Grad-CAM (Figures 14 and 15) produced the most spatially coherent and broadly distributed heatmaps among all methods. For the ResNet-50 model, the Grad-CAM heatmaps typically covered large, contiguous regions of the spiral, with highest activation concentrated around the inner loops and areas where spiral spacing was densest or most irregular. For correctly classified PD subjects, the activation tended to concentrate on the central and inner portions of the spiral where tremor-induced irregularities are typically most visible. For correctly classified controls, the activation was more diffuse, consistent with the model attending to the overall smooth, regular structure of the drawing.

For the ConvNeXtV2-Base model, the Grad-CAM heatmaps showed even broader activation regions that often covered most of the spiral. The ConvNeXtV2-Base Grad-CAM maps displayed a more uniform coverage pattern, suggesting that this architecture made its decisions based on global features of the entire drawing rather than localised regions.

A notable observation is the misclassified subject PD-22 in the ResNet-50 explanations: the Grad-CAM heatmap for this subject shows relatively low activation across the spiral,

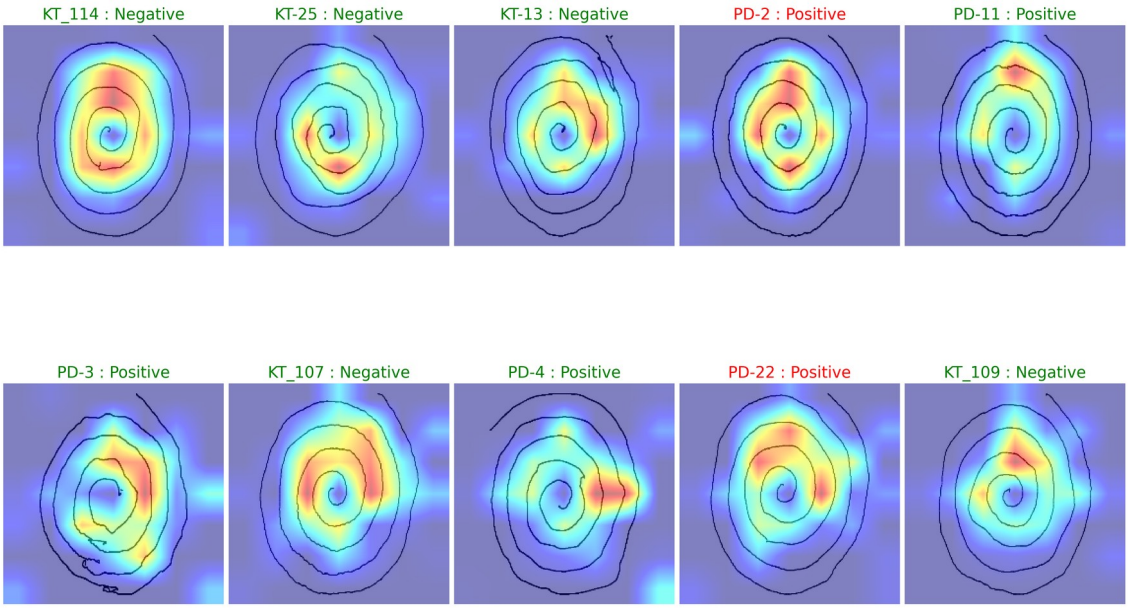

Figure 15. Grad-CAM explanations on the ConvNeXtV2-Base model.

suggesting that the model failed to detect discriminative PD-related features in this particular drawing. This could indicate that PD-22's spiral exhibited milder motor symptoms or a drawing style that fell within the range of variation seen in healthy controls.

5.3.5 Score-CAM

Score-CAM (Figures 16 and 17) produced more focused heatmaps than Grad-CAM, with sharper boundaries between high- and low-relevance regions. For the ResNet-50 model, Score-CAM highlighted more compact regions compared to Grad-CAM, often concentrating activation on specific segments of the spiral where line irregularity or spacing variation was most pronounced. In some cases, Score-CAM also highlighted regions near the image borders, potentially indicating that the model was partially sensitive to edge artefacts introduced during image preprocessing.

For the ConvNeXtV2-Base model, Score-CAM showed distinct focal regions rather than the broad coverage seen in Grad-CAM. The explanations identified specific locations along the spiral, particularly near the inner loops and points of direction change, as most important for classification. The gradient-free nature of Score-CAM resulted in visually cleaner heatmaps with less noise than the gradient-based Grad-CAM, consistent with the theoretical advantages of this approach.

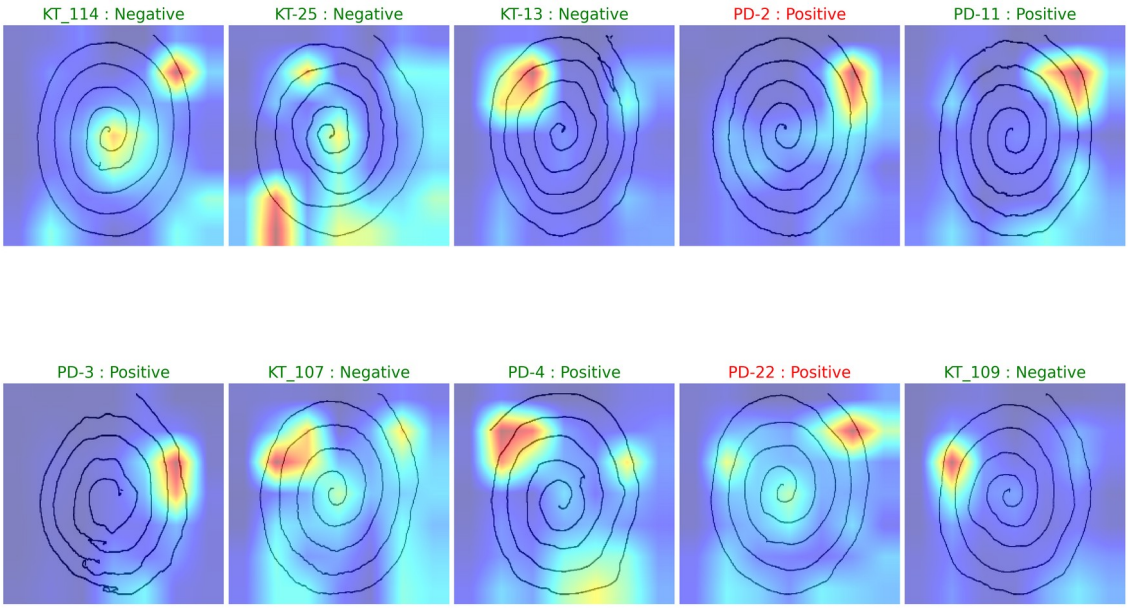

Figure 16. Score-CAM explanations on the ResNet-50 model.

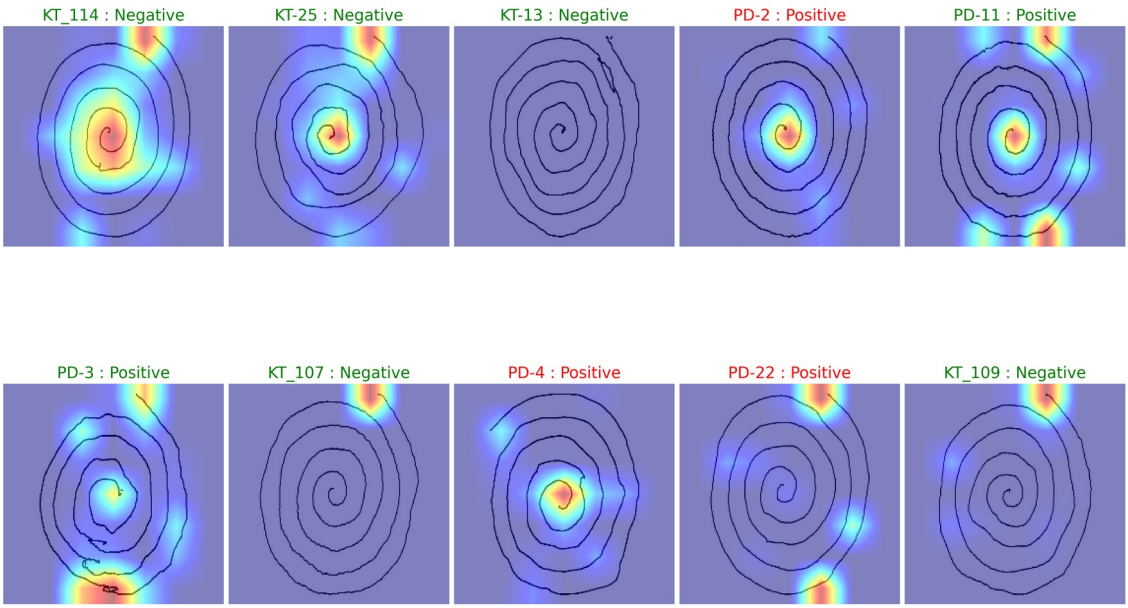

Figure 17. Score-CAM explanations on the ConvNeXtV2-Base model.

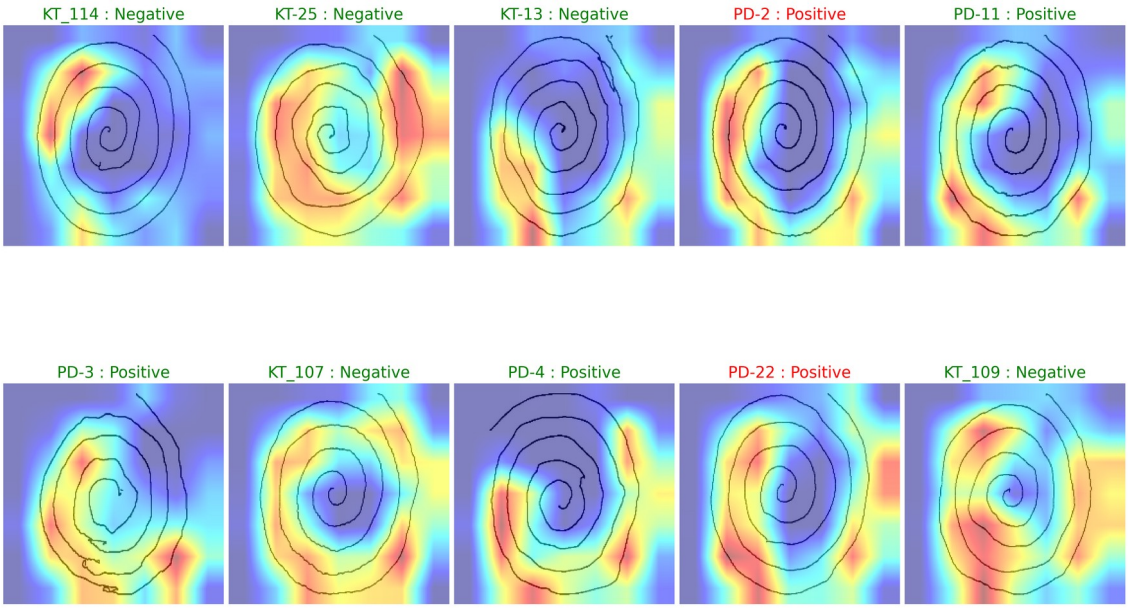

Figure 18. Ablation-CAM explanations on the ResNet-50 model.

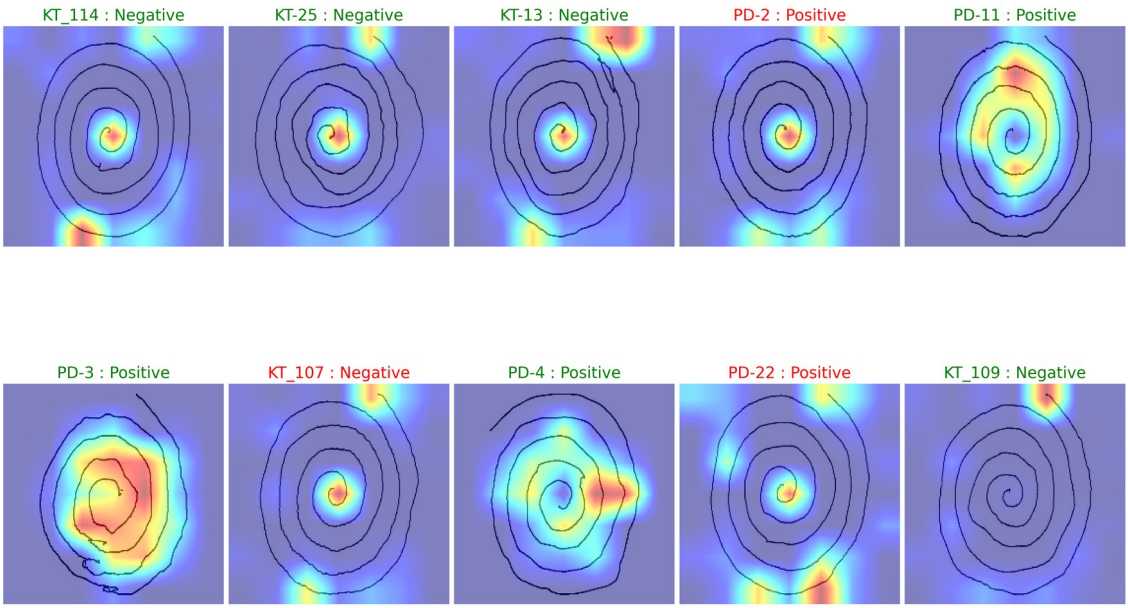

Figure 19. Ablation-CAM explanations on the ConvNeXtV2-Base model.

5.3.6 Ablation-CAM

Ablation-CAM (Figures 18 and 19) produced heatmaps with characteristics intermediate between Grad-CAM and Score-CAM. For the ResNet-50 model, Ablation-CAM generated broad activation regions similar to Grad-CAM but with more pronounced contrast between high- and low-relevance areas. In several images, Ablation-CAM highlighted regions near the edges of the drawing area in addition to the spiral itself, suggesting that the model's decision was partially influenced by contextual features beyond the spiral trajectory.

For the ConvNeXtV2-Base model, the Ablation-CAM heatmaps showed more concentrated activation around the inner spiral regions and areas of irregular spacing, with some images displaying activation patterns that closely followed the spiral trajectory. The direct ablation-based measurement of feature importance provided a clear picture of which spatial regions the model relied upon, without the gradient approximation issues that can affect Grad-CAM.

5.3.7 Cross-Method Comparison

Table 5 summarises the qualitative characteristics of each XAI method as observed across both architectures.

Table 5. Qualitative comparison of XAI methods applied to spiral drawing classification.

Method	Spatial Coherence	Focus	Resolution
Grad-CAM	High	Broad	Low (coarse)
Score-CAM	High	Focused	Low (coarse)
Ablation-CAM	Medium	Broad	Low (coarse)
Integrated Gradients	Medium	Focused	High (pixel-level)
GradientSHAP	Low	Diffuse	High (pixel-level)
LIME	Low	Sparse	Medium (superpixel)

The CAM-based methods (Grad-CAM, Score-CAM, Ablation-CAM) consistently produced the most spatially coherent explanations, highlighting contiguous regions of the spiral drawing that could be readily mapped to anatomical drawing features. Among the CAM methods, Score-CAM offered the best balance of spatial coherence and focus, identifying specific regions without the overly broad coverage of Grad-CAM or the occasional border sensitivity of Ablation-CAM.

The gradient-based attribution methods (Integrated Gradients, GradientSHAP) provided higher-resolution, pixel-level attributions but with less spatial coherence. Integrated Gradients was more deterministic and focused than GradientSHAP, making it more suitable

for identifying specific fine-grained features that influenced predictions.

LIME produced the least interpretable results for this task, likely due to the mismatch between its superpixel segmentation and the continuous nature of spiral drawing features.

5.3.8 Cross-Architecture Comparison

Comparing explanations between ResNet-50 and ConvNeXtV2-Base revealed systematic differences in how the two architectures processed spiral drawing images:

1. **Attention distribution**: ResNet-50 explanations tended to be more localised, focusing on specific regions of the spiral, while ConvNeXtV2-Base explanations were more broadly distributed, suggesting a more holistic approach to image analysis. This is consistent with ConvNeXtV2-Base's larger receptive field (due to 7×7 depthwise convolutions) and its self-supervised pretraining, which encourages learning of global image structure.
2. **Feature focus**: For correctly classified PD subjects, ResNet-50's CAM heatmaps tended to highlight inner spiral regions where irregularities in line spacing and smoothness were visually apparent. ConvNeXtV2-Base showed similar but broader patterns, often encompassing both the inner spiral and surrounding drawing area.
3. **Error analysis**: The misclassified cases offered insight into model limitations. ResNet-50's failure on PD-22 corresponded to a Grad-CAM heatmap with weak, diffuse activation, suggesting the model did not detect sufficient PD-indicative features. The ConvNeXtV2-Base misclassification of a control subject showed moderate activation in the spiral centre, suggesting the model detected apparent irregularities in an otherwise healthy drawing.

Overall, both architectures produced explanations that were broadly consistent with clinical knowledge, highlighting spiral regions where tremor, spacing irregularity, and angular deviations are expected manifestations of PD motor dysfunction. However, the explanations also revealed that models sometimes attended to features unrelated to the spiral itself (such as image borders or background regions), highlighting the importance of careful image preprocessing and the potential value of domain-aware constraints in future model design.

6. Discussion

This study explored how visual explainability methods can enhance the transparency of CNN-based classification models used in Parkinson's disease diagnostics through Archimedean spiral drawing analysis. Two architectures, ResNet-50 and ConvNeXtV2-Base, were trained on the DraWritePD dataset with extensive hyperparameter optimisation (3,000 configurations each), and six XAI methods were applied to interpret the best-performing models' predictions. This section discusses the findings in relation to the research questions, contextualises the results within the broader literature, and acknowledges limitations.

6.1 Classification Performance

Both architectures achieved 90% accuracy (F1 = 0.889 for ResNet-50, F1 = 0.909 for ConvNeXtV2-Base) on the 10-subject fixed evaluation split, demonstrating that pretrained CNNs can learn to distinguish PD from healthy spiral drawings even with only 28 training subjects. These results are competitive with prior work: Pereira et al. [33] reported up to 87.14% accuracy on the HandPD dataset using CNNs, and Valla et al. [13] achieved 85.3% using handcrafted kinematic features on the same DraWritePD dataset (though on dynamic trajectory data rather than static images).

The complementary error profiles of the two models, ResNet-50 achieving perfect specificity but lower sensitivity (0.80), while ConvNeXtV2-Base achieved perfect sensitivity but lower specificity (0.80), suggest fundamental differences in how these architectures process spiral images. ResNet-50's tendency to err on the side of classifying ambiguous cases as healthy (missing one PD patient) may stem from its smaller model capacity requiring more definitive PD indicators before predicting positive. ConvNeXtV2-Base's tendency towards false positives may reflect its broader feature extraction capturing subtle irregularities even in healthy drawings.

However, these results must be interpreted cautiously. With only 10 evaluation subjects (5 per class), each individual misclassification shifts accuracy by 10 percentage points. The difference between 0.80 and 1.00 sensitivity represents a single subject (PD-22 for ResNet-50). In addition, this split was used for model selection during experimentation rather than

reserved as a strictly untouched final test set. The reported metrics should therefore be considered indicative selection-split results rather than definitive estimates of true model performance. Larger, multi-centre validation studies with a genuinely untouched final test set would be needed to establish reliable performance estimates.

6.2 Addressing Research Question 1: Effectiveness of Visual Explanation Techniques

The six XAI methods demonstrated varying degrees of effectiveness in highlighting relevant spiral features. CAM-based methods (Grad-CAM, Score-CAM, Ablation-CAM) consistently produced the most spatially coherent heatmaps, identifying contiguous regions of the spiral that could be intuitively mapped to clinical features. The highlighted regions often corresponded to inner spiral areas where irregularities in line spacing and smoothness are typically most pronounced in PD patients, consistent with clinical observations that tremor amplitude is often greatest during the initiation and inner portions of spiral drawing [18, 19].

Integrated Gradients provided more precise, pixel-level attributions that identified specific points along the spiral trajectory, while GradientSHAP produced noisier, more scattered results. LIME was the least effective for this task, producing fragmented explanations that were difficult to interpret in clinical terms. This pattern aligns with the known strengths and limitations of these methods: CAM methods are inherently spatial and suited for identifying regional importance in images, while perturbation-based methods like LIME were originally designed for tabular and text data where features are more naturally discrete.

6.3 Addressing Research Question 2: Most Clinically Interpretable Methods

Among the six methods evaluated, Score-CAM and Grad-CAM provided the most clinically interpretable explanations. Score-CAM offered the best balance of spatial coherence and specificity, identifying focused regions without the overly broad coverage sometimes produced by Grad-CAM. Its gradient-free nature also resulted in visually cleaner heatmaps with less noise.

Grad-CAM, while sometimes broader in its activation patterns, remains the most accessible and computationally efficient method, requiring only a single backward pass. For clinical deployment scenarios where computational resources may be limited, Grad-CAM provides a practical balance of interpretability and efficiency.

The gradient-based attribution methods (Integrated Gradients, GradientSHAP), while providing higher-resolution information, produced explanations that may be more useful for model debugging and development than for clinical communication, as their pixel-level attributions are harder for non-technical users to interpret as meaningful anatomical or clinical patterns.

6.4 Addressing Research Question 3: Cross-Architecture Comparison

The comparison of explanations between ResNet-50 and ConvNeXtV2-Base revealed that the two architectures attend to partially different visual features, despite achieving similar overall accuracy. ResNet-50's explanations were generally more localised, focusing on specific regions where irregularities were concentrated, while ConvNeXtV2-Base produced more broadly distributed activation patterns, suggesting a more holistic approach to image analysis.

This difference is architecturally consistent: ConvNeXtV2-Base uses 7×7 depthwise convolutions (compared to ResNet-50's 3×3 convolutions), giving it a larger effective receptive field. Additionally, ConvNeXtV2-Base's self-supervised FCMAE pretraining may encourage learning of global image structure, as the pretraining objective requires reconstructing masked image patches from their context.

The fact that both architectures attended to similar general regions (inner spiral, areas of irregular spacing) while differing in the breadth of their attention suggests that the models may be using at least partially relevant visual cues rather than relying exclusively on dataset-specific artefacts. However, this interpretation remains qualitative. The explanations are consistent with clinically meaningful features, but they do not by themselves prove that the models are free of spurious correlations.

6.5 Limitations

Several important limitations should be acknowledged:

1. **Dataset size**: The DraWritePD dataset contains only 48 spiral images from 19 PD patients and 29 healthy controls. The evaluation split of 10 subjects means that each misclassification changes accuracy by 10%, and the reported metrics carry wide confidence intervals. The limited dataset also constrains the diversity of PD presentations that the models can learn to recognise.
2. **Absence of quantitative XAI evaluation**: The assessment of explanation quality

in this thesis is entirely qualitative. Quantitative evaluation metrics such as insertion/deletion curves [51], ROAR [52], and sanity checks [53] were not implemented due to time constraints. Without these metrics, claims about explanation quality rest on visual inspection rather than rigorous measurement.

3. **No clinical validation**: Although the explanations are discussed in relation to known clinical features of PD spiral drawings, no formal clinical validation was conducted. Neurologist evaluation of the heatmaps against UPDRS sub-scores or annotated drawing features would be necessary to establish clinical meaningfulness.
4. **Single fixed data split and split reuse**: All experiments used a single train/validation/test split rather than cross-validation. While subject-wise splitting prevents data leakage, a single split means the results are dependent on the specific subjects assigned to each partition. In addition, the test split was used for model selection in the final experimental setup, so it should not be interpreted as a fully untouched final benchmark. K-fold cross-validation or a separate final test set would provide more robust performance estimates, though this was not feasible within the experimental budget.
5. **Static image analysis only**: This thesis analysed only the static spiral images, discarding the rich dynamic information (pen velocity, pressure, timestamps) available in the DraWritePD dataset. Temporal features have been shown to carry significant discriminative information for PD diagnosis [13, 32, 25].
6. **Potential spurious correlations**: The XAI analysis revealed that some models occasionally attended to image border regions or background areas unrelated to the spiral itself. This suggests that preprocessing artefacts or image boundary effects may have influenced some predictions, highlighting the importance of careful dataset curation.

6.6 Future Work

Several directions can be pursued to build upon and strengthen the findings of this thesis:

6.6.1 Dataset Expansion and Multi-Centre Collection

Future work should involve the collection or access to larger, more diverse datasets that include not only spiral tests but other motor-based drawing tasks (e.g., wave, line tracing, letter formation), ideally sourced from multiple clinical centres and patient populations. Including metadata such as PD severity (UPDRS scores), medication state, handedness, and cognitive status would further enrich model interpretability and enable stratified analysis.

6.6.2 Quantitative Validation of Explanations

Implementing quantitative evaluation metrics like insertion/deletion scores [51], ROAR [52], and sanity checks [53] would allow rigorous assessment of XAI method quality. Collaborations with neurologists to annotate which heatmap regions correspond to clinically meaningful drawing anomalies could establish a ground truth for explanation benchmarking.

6.6.3 Integration of Temporal Dynamics

The current work analyses static spiral images. However, the drawing process itself, captured via timestamped trajectories, contains rich temporal information. Incorporating sequential models (e.g., RNNs or Transformers) or velocity/acceleration features could provide more accurate and interpretable assessments, especially if aligned with tremor frequency analysis.

6.6.4 Exploring Vision Transformers and Attention-Based XAI

Vision Transformers (ViTs) [39] offer inherent interpretability through self-attention maps. Future work could compare ViT-based approaches with the CNN architectures studied here, evaluating whether attention rollout and token attribution techniques offer more clinically meaningful forms of visualisation than CAM-based heatmaps.

6.6.5 Multimodal and LLM-Augmented Diagnostic Tools

There is growing interest in combining drawing tests with other data modalities, such as speech, gait, or clinical questionnaires. Multimodal architectures could provide a more holistic diagnostic framework. Furthermore, large language models fine-tuned for medical reasoning might be employed to generate natural language explanations from visual features, potentially increasing clinician trust and accessibility.

6.6.6 Deployable Clinical Interfaces

Creating a usable clinical interface that allows neurologists to view, query, and interact with model predictions (integrating heatmaps, confidence scores, and model uncertainty) could facilitate real-world adoption. Incorporating human-in-the-loop learning would enable iterative model refinement based on expert feedback.

7. Conclusion

This thesis investigated the application of explainable artificial intelligence methods to CNN-based classification of Archimedean spiral drawings for Parkinson's disease diagnostics, using the DraWritePD dataset. The central aim was to not only build accurate classification models, but to understand *how* these models make their decisions, a prerequisite for clinical trust and adoption of AI-assisted diagnostic tools.

Two CNN architectures, ResNet-50 and ConvNeXtV2-Base, were trained and optimised through extensive hyperparameter search (3,000 configurations per architecture), achieving best accuracies of 90% on a fixed evaluation split with F1 scores of 0.889 and 0.909, respectively. The models exhibited complementary error profiles: ResNet-50 achieved perfect specificity (1.00) at the cost of lower sensitivity (0.80), while ConvNeXtV2-Base achieved perfect sensitivity (1.00) with slightly lower specificity (0.80). These results demonstrate that pretrained CNNs, combined with aggressive data augmentation and systematic hyperparameter optimisation, can learn to discriminate PD from healthy spiral drawings even with a training set of only 28 subjects.

Six explainability methods, Grad-CAM, Score-CAM, Ablation-CAM, Integrated Gradients, GradientSHAP, and LIME, were applied to interpret the best models' predictions across all test images. The comparative analysis revealed that CAM-based methods, particularly Score-CAM and Grad-CAM, produced the most spatially coherent and clinically interpretable heatmaps, consistently highlighting inner spiral regions and areas of irregular spacing where PD motor dysfunction is expected to manifest. Gradient-based methods provided higher-resolution but noisier attributions, while LIME's superpixel-based approach proved poorly suited to the continuous features of spiral drawings. Cross-architecture comparison showed that ResNet-50 and ConvNeXtV2-Base attend to partially overlapping but distinct visual features, with ResNet-50 producing more localised explanations and ConvNeXtV2-Base exhibiting broader attention patterns consistent with its larger receptive field.

The primary contributions of this work are twofold: first, it establishes a baseline for CNN-based classification of DraWritePD spiral images, complementing the original dataset authors' work on handcrafted kinematic features with a deep learning approach; second,

it provides the first systematic multi-method XAI comparison for PD spiral drawing classification, identifying which explainability techniques are most suitable for this specific medical image domain.

Several limitations constrain the generalisability of these findings, most notably the small dataset size (48 subjects), the reliance on a single data split rather than cross-validation, the reuse of the evaluation split for model selection, and the absence of quantitative XAI evaluation metrics and formal clinical validation. Future work should address these limitations through larger multi-centre datasets, quantitative explanation evaluation (insertion/deletion metrics, sanity checks), neurologist validation of highlighted regions, and the integration of dynamic drawing trajectory data alongside static images.

Overall, this thesis demonstrates that explainability is both achievable and informative for CNN-based PD diagnostics from spiral drawings. The visual explanations generated in this work provide evidence that the trained models attend to clinically relevant drawing features, offering a foundation for developing transparent, trustworthy AI-assisted screening tools for Parkinson's disease.

Bibliography

[1] Lorraine V Kalia and Anthony E Lang. "Parkinson's disease". In: *The Lancet* 386.9996 (2015), pp. 896–912. ISSN: 0140-6736. DOI: https://doi.org/10.1016/S0140-6736(14)61393-3. URL: https://www.sciencedirect.com/science/article/pii/S0140673614613933.

[2] Ronald B. Postuma et al. "MDS clinical diagnostic criteria for Parkinson's disease". In: *Movement Disorders* 30.12 (2015), pp. 1591–1601. DOI: https:/ / doi.org/10.1002/mds.26424 . eprint: https://movementdisorders.onlinelibrary.wiley.com/doi/pdf/10.1002/mds.26424 . URL: https://movementdisorders.onlinelibrary.wiley.com/doi/abs/10.1002/mds.26424.

[3] Igor Kononenko. "Machine learning for medical diagnosis: history, state of the art and perspective". In: *Artificial Intelligence in Medicine* 23.1 (2001), pp. 89–109. ISSN: 0933-3657. DOI: https : / / doi . org / 10 . 1016 / S0933 - 3657(01) 00077 -. URL: https : / / www . sciencedirect . com / science/article/pii/S093336570100077X.

[4] Geert Litjens et al. "A survey on deep learning in medical image analysis". In: *Medical Image Analysis* 42 (2017), pp. 60–88. ISSN: 1361-8415. DOI: https://doi.org/10.1016/j.media.2017.07.005 . URL: https://www.sciencedirect.com/science/article/pii/S1361841517301135.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet classification with deep convolutional neural networks". In: *Commun. ACM* 60.6 (May 2012), pp. 84–90. ISSN: 0001-0782. DOI: 10.1145/3065386. URL: https://doi.org/10.1145/3065386.

[6] Ramprasaath R. Selvaraju et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization". In: *International Journal of Computer Vision* 128.2 (Oct. 2020), pp. 336–359. ISSN: 1573-1405. DOI: 10.1007/s11263-019-01228-7 . URL: http://dx.doi.org/10.1007/s11263-019-01228-7.

[7] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “"Why Should I Trust You?": Explaining the Predictions of Any Classifier”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 1135–1144. ISBN: 9781450342322. DOI: 10.1145/2939672.2939778. URL: https://doi.org/10.1145/2939672.2939778.

[8] Scott M. Lundberg and Su-In Lee. “A unified approach to interpreting model predictions”. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS’17. Long Beach, California, USA: Curran Associates Inc., 2017, pp. 4768–4777. ISBN: 9781510860964.

[9] Bjoern M. Eskofier et al. “Recent machine learning advancements in sensor-based mobility analysis: Deep learning for Parkinson’s disease assessment”. In: *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2016, pp. 655–658. DOI: 10.1109/EMBC.2016.7590787.

[10] Madina Hamiane and Fatema Saeed. “SVM Classification of MRI Brain Images for Computer-Assisted Diagnosis”. In: *International Journal of Electrical and Computer Engineering* 7 (Oct. 2017). DOI: 10.11591/ijece.v7i5.pp2555-2564.

[11] R. Prashanth et al. “High-Accuracy Classification of Parkinson’s Disease Through Shape Analysis and Surface Fitting in 123I-Ioflupane SPECT Imaging”. In: *IEEE Journal of Biomedical and Health Informatics* 21.3 (2017), pp. 794–802. DOI: 10.1109/JBHI.2016.2547901.

[12] Iqra Kamran et al. “Handwriting dynamics assessment using deep neural network for early identification of Parkinson’s disease”. In: *Future Generation Computer Systems* 117 (2021), pp. 234–244. ISSN: 0167-739X. DOI: https://doi.org/10.1016/j.future.2020.11.020. URL: https://www.sciencedirect.com/science/article/pii/S0167739X20330442.

[13] Elli Valla et al. “Tremor-related feature engineering for machine learning based Parkinson’s disease diagnostics”. In: *Biomedical Signal Processing and Control* 75 (2022), p. 103551. ISSN: 1746-8094. DOI: https://doi.org/10.1016/j.bspc.2022.103551 . URL: https://www.sciencedirect.com/science/article/pii/S1746809422000738.

[14] E. Ray Dorsey et al. “Global, regional, and national burden of Parkinson’s disease, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016”. In: *The Lancet Neurology* 17.11 (2018), pp. 939–953. DOI: 10.1016/S1474-4422(18)30295-3.

[15] Heiko Braak et al. "Staging of brain pathology related to sporadic Parkinson's disease". In: *Neurobiology of Aging* 24.2 (2003), pp. 197–211. DOI: 10.1016/S0197-4580(02)00065-9.

[16] Joseph Jankovic. "Parkinson's disease: clinical features and diagnosis". In: *Journal of Neurology, Neurosurgery & Psychiatry* 79.4 (2008), pp. 368–376. DOI: 10.1136/jnnp.2007.131045.

[17] Christopher G. Goetz et al. "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results". In: *Movement Disorders* 23.15 (2008), pp. 2129–2170. DOI: 10.1002/mds.22340.

[18] Seth L. Pullman. "Spiral analysis: a new technique for measuring tremor with a digitizing tablet". In: *Movement Disorders* 13.S3 (1998), pp. 85–89. DOI: 10.1002/mds.870131315.

[19] Rachel Saunders-Pullman et al. "Validity of spiral analysis in early Parkinson's disease". In: *Movement Disorders* 23.4 (2008), pp. 531–537. DOI: 10.1002/mds.21874.

[20] Marta San Luciano et al. "Digitized Spiral Drawing: A Possible Biomarker for Early Parkinson's Disease". In: *PLoS ONE* 11.10 (2016), e0162799. DOI: 10.1371/journal.pone.0162799.

[21] Duc Minh Dimitri Nguyen et al. "Transformers for 1D signals in Parkinson's disease detection from gait". In: *2022 26th International Conference on Pattern Recognition (ICPR)*. Aug. 2022, pp. 5089–5095. DOI: 10.1109/ICPR56361.2022.9956330.

[22] Safwen Naimi, Wassim Bouachir, and Guillaume-Alexandre Bilodeau. "HCT: Hybrid Convnet-Transformer for Parkinson's Disease Detection and Severity Prediction from Gait". In: *2023 International Conference on Machine Learning and Applications (ICMLA)*. Dec. 2023, pp. 814–819. DOI: 10.1109/ICMLA58977.2023.00119.

[23] Zehra Karapinar Senturk. "Early diagnosis of Parkinson's disease using machine learning algorithms". In: *Medical Hypotheses* 138 (2020), p. 109603. ISSN: 0306-9877. DOI: https://doi.org/10.1016/j.mehy..2020.1096 URL: https://www.sciencedirect.com/science/article/pii/S0306987719314148.

[24] Diego Machado Reyes et al. "Genomics transformer for diagnosing Parkinson's disease". In: *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*. Sept. 2022, pp. 01–04. DOI: 10.1109/BHI56158.2022.9926815.

[25] Donato Impedovo, Giuseppe Pirlo, and Gennaro Vessio. “Dynamic Handwriting Analysis for Supporting Earlier Parkinson’s Disease Diagnosis”. In: *Information* 9.10 (2018), p. 247. DOI: 10.3390/info9100247.

[26] Peter Drotár et al. “Analysis of in-air movement in handwriting: A novel marker for Parkinson’s disease”. In: *Computer Methods and Programs in Biomedicine* 117.3 (2014), pp. 405–411. DOI: 10.1016/j.cmpb.2014.08.007.

[27] Clayton R. Pereira et al. “Deep Learning-Aided Parkinson’s Disease Diagnosis from Handwritten Dynamics”. In: *2016 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*. Oct. 2016, pp. 340–346. DOI: 10.1109/SIBGRAPI.2016.054.

[28] Moises Diaz et al. “Sequence-based dynamic handwriting analysis for Parkinson’s disease detection with one-dimensional convolutions and BiGRUs”. In: *Expert Systems with Applications* 168 (2021), p. 114405. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2020.114405. URL: https://www.sciencedirect.com/science/article/pii/S0957417420310757.

[29] S. Saravanan et al. “Explainable Artificial Intelligence (EXAI) Models for Early Prediction of Parkinson’s Disease Based on Spiral and Wave Drawings”. In: *IEEE Access* 11 (2023), pp. 68366–68378. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2023.3291406.

[30] F. Cavaliere et al. “Parkinson’s Disease Diagnosis: Towards Grammar-based Explainable Artificial Intelligence”. In: *2020 IEEE Symposium on Computers and Communications (ISCC)*. July 2020, pp. 1–6. DOI: 10.1109/ISCC50000.2020.9219616.

[31] Mathew Thomas, Abhishek Lenka, and Pramod Kumar Pal. “Handwriting Analysis in Parkinson’s Disease: Current Status and Future Directions”. In: *Movement Disorders Clinical Practice* 4.6 (2017), pp. 806–818. DOI: 10.1002/mdc3.12552.

[32] Peter Drotár et al. “Evaluation of handwriting kinematics and pressure for differential diagnosis of Parkinson’s disease”. In: *Artificial Intelligence in Medicine* 67 (2016), pp. 39–46. ISSN: 0933-3657. DOI: https://doi.org/10.1016/j.artmed.2016.01.004. URL: https://www.sciencedirect.com/science/article/pii/S0933365716000063.

[33] Clayton R. Pereira et al. “Handwritten dynamics assessment through convolutional neural networks: An application to Parkinson’s disease identification”. In: *Artificial Intelligence in Medicine* 87 (2018), pp. 67–77. DOI: 10.1016/j.artmed.2018.04.001.

[34] Muhammed Erdem Isenkul, Betul Erdogdu Sakar, and Olcay Kursun. “Improved spiral test using digitized graphics tablet for monitoring Parkinson’s disease”. In: 2014. URL: https://api.semanticscholar.org/CorpusID:17641702.

[35] Nima Tajbakhsh et al. “Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?” In: *IEEE Transactions on Medical Imaging* 35.5 (2016), pp. 1299–1312. DOI: 10.1109/TMI.2016.2535302.

[36] Kaiming He et al. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV]

[37] Jia Deng et al. “ImageNet: A Large-Scale Hierarchical Image Database”. In: *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2009, pp. 248–255.

[38] Zhuang Liu et al. *A ConvNet for the 2020s*. 2022. arXiv: 2201.03545 [cs.CV]

[39] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV]

[40] Sanghyun Woo et al. *ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders*. 2023. arXiv: 2301.00808 [cs.CV]

[41] Jason Yosinski et al. “How transferable are features in deep neural networks?” In: *Advances in Neural Information Processing Systems*. Vol. 27. 2014.

[42] Karen Simonyan and Andrew Zisserman. “Very Deep Convolutional Networks for Large-Scale Image Recognition”. In: *arXiv preprint arXiv:1409.1556* (2015).

[43] Gao Huang et al. “Densely Connected Convolutional Networks”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 2261–2269. DOI: 10.1109/CVPR.2017.243.

[44] Mingxing Tan and Quoc V. Le. “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks”. In: *Proceedings of the 36th International Conference on Machine Learning (ICML)*. 2019, pp. 6105–6114.

[45] Alejandro Barredo Arrieta et al. “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI”. In: *Information Fusion* 58 (2020), pp. 82–115. DOI: 10.1016/j.inffus.2019.12.012 .

[46] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. *Axiomatic Attribution for Deep Networks*. 2017. arXiv: 1703.01365 [cs.LG]

[47] Scott M. Lundberg and Su-In Lee. *A Unified Approach to Interpreting Model Predictions*. 2017. arXiv: 1705.07874 [cs.AI]

[48] Aditya Chattopadhay et al. “Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks”. In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2018, pp. 839–847DOI: 10.1109/WACV.2018.00097.

[49] Haofan Wang et al. *Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks*. 2020. arXiv: 1910.01279 [cs.CV] URL: https://arxiv.org/abs/1910.01279.

[50] Saurabh Desai and Harish G. Ramaswamy. “Ablation-CAM: Visual Explanations for Deep Convolutional Network via Gradient-free Localization”. In: *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 2020, pp. 972–980. DOI: 10.1109/WACV45572.2020.9093360.

[51] Vitali Petsiuk, Abir Das, and Kate Saenko. “RISE: Randomized Input Sampling for Explanation of Black-box Models”. In: *arXiv preprint arXiv:1806.07421* (2018).

[52] Sara Hooker et al. “A Benchmark for Interpretability Methods in Deep Neural Networks”. In: *Advances in Neural Information Processing Systems*. Vol. 32. 2019.

[53] Julius Adebayo et al. “Sanity Checks for Saliency Maps”. In: *Advances in Neural Information Processing Systems*. Vol. 31. 2018.

[54] Sana Tonekaboni et al. “What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use”. In: *arXiv preprint arXiv:1905.05134* (2019).

[55] Erico Tjoa and Cuntai Guan. “A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI”. In: *IEEE Transactions on Neural Networks and Learning Systems* 32.11 (2021), pp. 4793–4813. DOI: 10.1109/TNNLS.2020.3027314.

[56] Wojciech Samek et al. “Evaluating the Visualization of What a Deep Neural Network Has Learned”. In: *IEEE Transactions on Neural Networks and Learning Systems* 28.11 (2017), pp. 2660–2673. DOI: 10.1109/TNNLS.2016.2599820.

[57] Jason Ansel et al. “PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation”. In: *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS ’24)*. ACM, Apr. 2024.DOI: 10.1145/3620665.3640366. URL: https://pytorch.org/assets/pytorch2-2.pdf.

[58] TorchVision maintainers and contributors. *TorchVision: PyTorch’s Computer Vision library*. Nov. 2016. URL: https://github.com/pytorch/vision.

[59] Lukas Biewald. *Experiment Tracking with Weights and Biases*. Software available from wandb.com. 2020. URL: https://www.wandb.com/.

[60] Takuya Akiba et al. “Optuna: A Next-Generation Hyperparameter Optimization Framework”. In: *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019, pp. 2623–2631. DOI: 10.1145/3292500.3330701.

[61] Ross Wightman. *PyTorch Image Models*https://github.com/rwightman/pytorch-image-models. 2019. DOI: 10.5281/zenodo.4414861.

[62] Alexander Buslaev et al. “Albumentations: Fast and Flexible Image Augmentations”. In: *Information* 11.2 (2020). ISSN: 2078-2489. DOI: 10.3390/info11020125. URL: https://www.mdpi.com/2078-2489/11/2/125.

[63] Jacob Gildenblat and contributors. *PyTorch library for CAM methods*. https://github.com/jacobgil/pytorch-grad-cam. 2021.

[64] Narine Kokhlikyan et al. *Captum: A unified and generic model interpretability library for PyTorch*. 2020. arXiv:2009.07896 [cs.LG].

[65] Heiko Herrmann, Toomas Kaevand, and Lauri Anton. *BASE: TalTech's HPC Infrastructure 2020–2024*. TalTech Data Repository. Mar. 2025.DOI: 10.48726/8f0v8-1eb97.

[66] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2015. arXiv: 1412.6980 [cs.LG]. URL: https://arxiv.org/abs/1412.6980.

[67] Ilya Loshchilov and Frank Hutter. *Decoupled Weight Decay Regularization*. 2019. arXiv: 1711.05101 [cs.LG]. URL: https://arxiv.org/abs/1711.05101.

[68] Alex Graves. *Generating Sequences With Recurrent Neural Networks*. 2014. arXiv: 1308.0850 [cs.NE] URL: https://arxiv.org/abs/1308.0850.

[69] Michael R. Zhang et al. *Lookahead Optimizer: k steps forward, 1 step back*. 2019. arXiv: 1907.08610 [cs.LG]. URL: https://arxiv.org/abs/1907.08610.

Appendices

Appendix 1 - Non-exclusive licence for reproduction and publication of a graduation thesis[1]

I, Oluwandabira Ohifeme Alawode

1. Grant Tallinn University of Technology free licence (non-exclusive licence) for my thesis "XAI Based Analysis of the Archimedean Spiral Drawing Test for Parkinson's Disease Diagnostics", supervised by Elli Valla and Sven Nõmm
   1.1 to be reproduced for the purposes of preservation and electronic publication of the graduation thesis, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright;
   1.2 to be published via the web of Tallinn University of Technology, incl. to be entered in the digital collection of the library of Tallinn University of Technology until expiry of the term of copyright.
2. I am aware that the author also retains the rights specified in clause 1 of the non-exclusive licence.
3. I confirm that granting the non-exclusive licence does not infringe other persons' intellectual property rights, the rights arising from the Personal Data Protection Act or rights arising from other legislation.

11.05.2026

[1]The non-exclusive licence is not valid during the validity of access restriction indicated in the student's application for restriction on access to the graduation thesis that has been signed by the school's dean, except in case of the university's right to reproduce the thesis for preservation purposes only. If a graduation thesis is based on the joint creative activity of two or more persons and the co-author(s) has/have not granted, by the set deadline, the student defending his/her graduation thesis consent to reproduce and publish the graduation thesis in compliance with clauses 1.1 and 1.2 of the non-exclusive licence, the non-exclusive license shall not be valid for the period.

Table 6. Complete hyperparameter configuration for the best ResNet-50 model (test F1 = 0.889).

Hyperparameter	Value
Base model	resnet50
Augments per image	120
Classifier hidden size	512
Classifier pooling	Fast AvgMax
Colour jitter	True
Optimiser	RMSProp (centred)
Learning rate	2.1×10^{-4}
Momentum	0.93
Weight decay	0.056
Lookahead	True
Finetune strategy	Unfreeze whole
Finetune learning rate	2.0×10^{-6}
Finetune momentum	0.96
Finetune weight decay	0.034
Weighted loss	True
Mixed precision (AMP)	True

Appendix 2: Best Hyperparameter Configurations

Table 6 and Table 7 present the complete hyperparameter configurations for the best-performing ResNet-50 and ConvNeXtV2-Base models, respectively.

Table 7. Complete hyperparameter configuration for the best ConvNeXtV2-Base model (test F1 = 0.909).

Hyperparameter	**Value**
Base model	convnextv2_base
Augments per image	120
Classifier hidden size	None (linear head)
Classifier pooling	Fast AvgMax
Colour jitter	True
Optimiser	Adam
Learning rate	4.0×10^{-5}
Momentum	0.91
Weight decay	0.027
Lookahead	True
Finetune strategy	None (frozen backbone)
Weighted loss	True
Mixed precision (AMP)	True

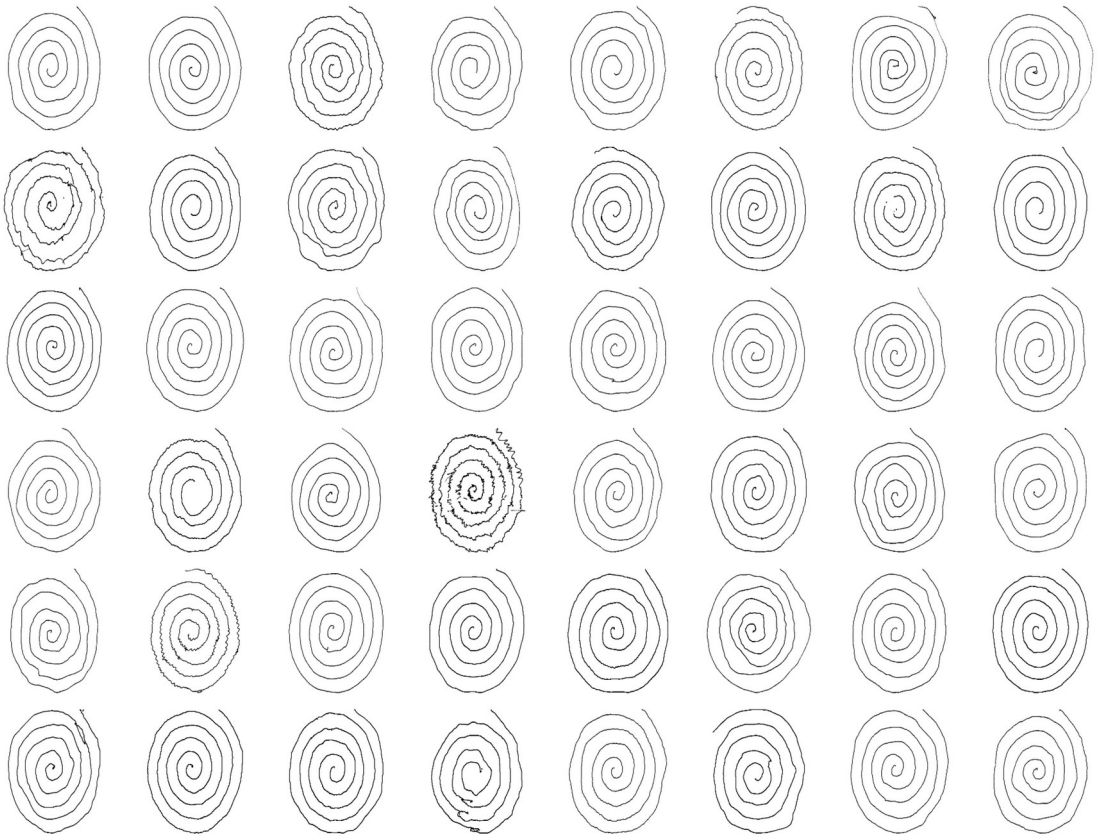

Figure 20. Grid of all 48 spiral drawings from the DraWritePD dataset, rendered at 224×224 pixels. Each image shows the spiral trajectory as black lines on a white background. Subject identifiers and ground truth labels (PD-positive or healthy control) are indicated above each drawing.