

Original paper

# Interpretable hybrid deep learning model for Parkinson's disease screening using hand-drawn spiral and waveform images

Garima Verma<sup>A,Ee</sup>

School of Computing, DIT University, Dehradun, India

## Abstract

**Purpose:** It is still hard to diagnose Parkinson's disease (PD) early and correctly, since the motor symptoms are often relatively mild and there are no unique diagnostic tools. The goal of this project was to create and test a hybrid deep learning model that can accurately classify PD by using hand-drawn spiral and waveform pictures.

**Material and methods:** A novel two-channel hybrid model was created, where the input representation combines normalised greyscale features and Canny edge features to capture both spatial and structural stroke patterns from patient drawings. The model combines a convolutional neural network (CNN) with hand-crafted grey-level co-occurrence matrix (GLCM) features to enhance its performance and make it easier to understand. We trained and evaluated three different models: a baseline CNN, a fusion CNN + GLCM, and a fine-tuned ResNet-50. We did this on both the original and pre-processed datasets.

**Results:** The hybrid CNN + GLCM model that used pre-processing had the best classification accuracy at 97.02% and worked well on datasets that were not used to train it. Statistical studies validated the importance of enhancements in performance relative to baseline models.

**Conclusions:** The suggested technique provides a straightforward, comprehensible, and efficient approach for PD screening using easily administered drawing exercises. Its great precision and low equipment needs make it a good candidate for use in real-world clinical settings.

**Key words:** Parkinson, GLCM, CNN, Grad-CAM, hand drawings.

## Introduction

Parkinson's disease (PD) is an advanced neurodegenerative condition that disturbs motor function [1]. The illness is caused by the progressive loss of dopamine-secreting neurons in an area of the brain called the substantia nigra. The neurotransmitter dopamine, which is involved in movement and coordination, plays a role in this process [2]. The cause of PD remains unclear, but it is speculated to be a result of a combination of genetic, environmental, and other determinants [3]. The above 60-year age group is predisposed to PD, as are specific genetic mutations and a family history of PD [4]. Diagnosis of PD occurs primarily

through medical symptoms and neurological examination. There is no diagnostic test to determine whether a patient has PD. Several clinical and neurological markers are present, including tremors (mainly at rest), slowness of movement, muscle rigidity, gait disturbance, and impairment of coordination [5]. Some of these methods also assist in PD diagnosis and evaluation. One method used to assist in diagnosing and evaluating some motor impairments related to PD is the administration of drawing tasks [6]. Among such tasks, the spiral drawing test is most widely used in the diagnosis of PD. During the spiral drawing test, the patient is asked to draw several spirals on paper using a pen or pencil. Patients with PD are more

## Correspondence address:

Garima Verma, School of Computing, DIT University, 248009, Dehradun, India, e-mail: [garimaverma.research@gmail.com](mailto:garimaverma.research@gmail.com)

## Authors' contribution:

A Study design · B Data collection · C Statistical analysis · D Data interpretation · E Manuscript preparation · F Literature search · G Funds collection

likely to exhibit characteristic deficits in their drawings of spirals [7]. In some cases, the task of wave drawing is assigned, where the patient is requested to draw some kind of waves such as sine waves or any type of wave on a study paper. The same defects, including microscopic imaging, irregularity, and tremor, are present in this task, which are characteristic of PD. Early diagnosis enables the initiation of treatment immediately, which will preserve quality of life and independence as long as possible. Early diagnosis will increase the likelihood of responsiveness to neuro-protective therapies, since they will be provided with greater chances to exert their effect on the disease before it reaches an advanced stage [8].

With diagnosis, patients make the necessary changes to their lifestyle, including exercising regularly, adjusting their dietary habits, and modifying their housing arrangements. Diagnosis at an early stage eliminates potential causes of other neurological disorders and enables timely treatment and management [9]. Manual early PD diagnosis has several drawbacks. Early signs of diagnosis, such as minor tremors in movement, require the attention of professional healthcare practitioners [10]. Other neurological disorders share overlapping symptoms with tremor, making manual diagnosis a challenge [11]. Symptoms manifest differently in various patients, and therefore, early manual diagnosis is not straightforward. Predicting the progression of the disease is challenging, as it varies from patient to patient [12]. In certain areas, particularly in third-world countries, there are few health facilities for patients. AI tools can significantly reduce the majority of errors in human diagnosis of PD by providing data-driven solutions [13]. AI methods examine various sources of information, including hand-drawn wave and spiral doodles, sensor data, and voice recordings, to quantify and assess nuanced variations that are good predictors of earlier PD [14]. AI methods are best suited to detect intricate patterns in vast amounts of data by examining numerous and complex data points – such as hand-drawn images, voice, and walking data – which are difficult for physicians to identify [15].

Hand-drawn-image-based screening is a non-invasive, low-cost, and easy-to-use method that directly measures early neuromotor impairments through micro-tremors, pressure abnormalities, and stroke distortions. This is different from voice, gait, or neuroimaging methods. Spiral and waveform drawings uniquely convert fine motor control into measurable spatial and textural biomarkers, facilitating objective digital assessment without requiring costly equipment or extensively trained personnel.

Also, handwriting and speech features are typically inconsistent because of language and sound considerations. Gait sensors and imaging methods, on the other hand, require wearable or MRI infrastructure, making them difficult to access in resource-limited areas. In contrast, simple sketches on paper or a tablet can be gathered rapidly in clinical waiting rooms or even from a distance,

making it possible to scale tele-diagnostics [16]. Neurologists have known for a long time that spiral and waveform analysis are useful for diagnosis because they make tremor, stiffness, and bradykinesia – early signs of PD – more noticeable. When combined with AI, these sketches can be automatically analysed to provide interpretable motor anomalies, which connect clinical intuition with computational accuracy. Focusing on hand-drawn image data enables early screening that is both explainable and cheap. This approach facilitates the translation of AI models from the laboratory setting to real-world neurology practice.

In the interest of transparency, originality, and clarity, the following section outlines the specific contributions made by the authors during the study. From conceptualization to model design, experimentation, and result interpretation, every activity was carried out with utmost care to advance the area of PD diagnosis from hand-drawn pictures. These contributions reflect the multidisciplinary and team-based nature of the research, providing both scientific rigor and pragmatic usefulness.

### Author's contributions

- Introduced a novel 2-channel input representation through combining normalized greyscale images and Canny edge maps to provide structural and stroke pattern clarity necessary for PD diagnosis.
- Trained the model on two publicly available spiral and waveform drawing datasets through intensive pre-processing and augmentation to facilitate strong model training and generalizability in real-world applications.
- Built a multi-representation fusion pipeline that captures spatial (convolutional neural network, CNN), statistical (grey-level co-occurrence matrix, GLCM), and semantic (50layer residual network, ResNet-50) features that improve classification performance as well as interpretability.
- Designed three deep learning pipelines – Baseline CNN, CNN + GLCM hybrid, and fine-tuned ResNet-50 – each of which was adapted for PD detection from hand-drawn images with different levels of complexity and performance trade-offs.
- Incorporated GLCM handcrafted features into CNN-based pipelines and utilized gradient-weighted class activation mapping (Grad-CAM) visualizations to improve model explainability and clinical robustness.
- Conducted comprehensive experimental testing with cross-validation, confusion matrices, and performance measures (accuracy, precision, recall, F1-score), and paired *t*-tests and Wilcoxon tests for statistically validating the significance of results.

The remaining organization of this paper is structured as follows: Section 2 describes a comprehensive review of the literature on PD image detection from images and machine learning, identifying the main gaps and motiva-

tions. Section 3 outlines the material and methods of this work, including information on the dataset, image pre-processing methods, and GLCM-based feature extraction paradigm. Section 4 describes the framework of the proposed models, specifically the baseline CNN, the CNN using GLCM fusion, and the fine-tuned ResNet-50 architecture. Section 5 presents experimental results, model performance evaluation, and visual interpretability, using Grad-CAM, as well as statistical verification. Lastly, Section 6 concludes the research work with its findings and recommendations for future studies.

## Literature review

The diagnosis of PD has increasingly leveraged artificial intelligence, specifically deep learning and machine learning methods, to examine speech, handwriting, and spiral drawings for early diagnosis. This section discusses recent progress, identifying principal methodologies, performance results, and gaps in current research that drive the motivation for the current study. Previously, researchers [17] implemented a CNN for the identification of PD by analysing irregularities in drawing movements. Using spiral pentagon drawing images, they trained a convolutional classifier with an accuracy of 93.5% between PD and healthy controls. In another study [18], DL methods were introduced, and two strategies using 2D and 3D CNNs for the diagnosis of PD were proposed. To train the model effectively, bias field correction, Z-score normalization, and image resizing were applied. The 3D model achieved an area under the curve (AUC) of 86% compared to 72.22% for the 2D model, indicating that the 3D model performed better than the 2D model.

In a previous study [19], acoustic features of speech samples were used for the early diagnosis of PD. A variety of machine learning classification algorithms were tested, including support vector machines (SVMs), random forests, gradient boosting, *k*-nearest neighbours (KNN), logistic regression, and decision trees. To enhance performance, ensemble learning techniques were employed, including stacking and stacking with cross-validation. While operating in the range of 70% to 82% for single models, the stacking ensemble improved that to 92%, and further to 95% using cross-validation. The method depicts speech as a promising initial biomarker, but its efficacy may be sensitive to speaker variability and noise level. A recent novel Internet of things (IoT)-based hybrid approach for the early diagnosis of PD via hand-drawn spiral images was proposed [20], employing a new sine cosine geese migration optimization for optimal image routing and a cosine deep CNN for classification. The model achieved the highest accuracy of 89.98% with consistent metrics, including positive predictive value (87.41%), true negative rate (89.77%), negative predictive value (NPV, 87.82%), and true positive rate (TPR, 89.84%). The algorithm also exhibited lower communication delay (0.283 ms) and

power consumption (0.507 J), demonstrating its suitability for real-time surveillance. Nevertheless, the explicit and minimal set of images used by the model in its application, along with the absence of external validation, restricts generalizability and clinical scalability.

Researchers [21] validated a computerized approach to measuring kinetic tremor in patients with essential tremor (ET) and PD through quantifying standardized drawings performed under magnetic resonance-guided focused ultrasound thalamotomy. Patients made drawings (signature, spiral, and line) at various stages of treatment, which were measured and analysed to give quantitative tremor scores. The findings indicated marked post-treatment gains, particularly for ET patients, with strong correlations between drawing score and accelerometer-measured wrist angles (e.g.,  $R^2 = 0.85$  for spirals in PD). Although the approach provides objective quantification of tremor, the use of stringent clinical settings and post-processing could curtail immediate generalization to more general tele-monitoring environments. Other research [22] demonstrated that a new technique for early diagnosis of PD, involving the quantification of dysgraphia using a  $16 \times 16$  low-signal degradation and high-response pressure sensor array, is effective. The system avoids crosstalk using an isolation layer, thereby efficiently recording subtle handwriting dysgraphia. The ResNet-18 CNN is utilized for feature extraction and classification with high efficiency in detecting Parkinson's related dysgraphia. Despite tremendous potential for early diagnosis, the utility of the system will likely be limited by the need for specialized equipment and clinical validation in a broad range of populations.

One study [23] proposed a hybrid deep learning framework, called Chronological Tangent Search Optimization-Zeiler and Fergus Network (CTSO-ZFNet), for early and precise diagnosis of PD using hand-drawn spiral images and voice signals. Median and adaptive Wiener filters were used in pre-processing, followed by the lyrebird herd optimization (LHO) algorithm, and feature-based feature selection. Chronological Tangent Search Optimization was used to optimize the Zeiler and Fergus Network model for classification. The model achieved a satisfactory performance, with an accuracy of 89.69%, a TPR of 90.87%, and an NPV of 90.55%. Although it yields good values, the complexity of the method and the two-modality input may limit its implementation in clinical practice without increasing the ease of integration through simplification and validation. Another study [24] introduced an attention-based continuous convolutional network for enhancing the early diagnosis of PD through handwriting analysis. The network can identify tremors and writing space abnormalities well by focusing on diagnostically significant areas using an attention mechanism. It reported an extremely impressive mean accuracy of 96.5% compared to conventional CNNs in diagnostic accuracy. Although the method is emerging, validation in

multiple datasets and real clinical conditions is necessary to establish its robustness and generalizability.

A study [25] aimed to develop a simple but accurate auxiliary diagnostic tool for PD through spiral hand-drawing testing. Unlike traditional AI systems, which have limitations due to dynamic input and dedicated hardware, the researchers developed a static image-based Archimedes spiral database that successfully detects tremor, shape, and spacing features. By reducing the number of pooling layers, the authors developed a new continuous convolution network. 89.3% of classifications were accurate, the Matthews correlation coefficient was 0.733, and the AUC was 0.934, indicating excellent diagnostic performance. The method in the study, however, is to be improved in the next stage by validation in heterogeneous real-world settings to ensure generalization. Table 1 below encapsulates the salient techniques, findings, and pitfalls reported in the reviewed literature, providing a comparative snapshot of available trends and knowledge gaps in PD diagnosis research.

The authors of one study [26] used machine learning (ML) algorithms on voice recordings to identify PD without the need for intrusive methods. The researchers used a dataset of 195 samples and many classifiers, including linear discriminant analysis (LDA), SVM, KNN, and CNN. They found that KNN was the best, with 98% accuracy and balanced precision and recall. The work empha-

sized the potential of acoustic biomarkers and ML models in enabling early, economical PD screening without reliance on clinical imaging. The study's findings are outstanding; however, voice-based research relies solely on acoustic parameters, which are often susceptible to noise, linguistic variations, and recording quality, thus limiting generalizability across populations. Another limitation is the limited size of the dataset used for the study.

In one study [27], the authors introduced the EO-optimized Lightweight Automatic Modulation Classification Network (EO-LWAMCNet) into an IoT-enabled healthcare system for early illness prediction. The framework attained a good accuracy of 94.8% and 95% on chronic liver and brain disease datasets, respectively, by using data collected from sensors and analysed through a Lightweight CNN. The main limitation of the proposed work model is its reliance on sensor-based inputs, which restricts its applicability to detect motor disorders associated with certain neurological conditions, such as PD. Furthermore, the study lacks interpretability mechanisms and multimodal feature fusion, which are essential strengths addressed in this work.

## Research gaps

The research shows that machine learning and deep learning have made significant progress in predicting diseases

**Table 1.** Summary of recent studies on Parkinson's disease diagnosis

Study No.	Technique used	Findings	Gaps
[17]	CNN on augmented spiral-pentagon drawings	Achieved 93.5% accuracy	Limited details on dataset variability and model generalizability
[18]	2D & 3D CNNs with pre-processing	3D CNN: 88.9% accuracy, AUC 0.86; better than 2D CNN (72.22%, AUC 0.50)	High computation for 3D CNN; 2D model underperforms
[19]	ML models + stacking ensembles on speech	95% accuracy with stacking cross-validation	Performance varies with speaker quality and noise
[20]	SCGA + CosineDCNN (IoT-based)	89.98% accuracy; low delay (0.283ms), energy (0.507J)	Limited image data; lacks external validation for clinical scalability
[21]	Quantitative analysis of MRgFUS drawing data	High correlation with accelerometer data; effective post-treatment tremor assessment	Confined to clinical settings; not suited for broader tele-monitoring
[22]	16Å – 16 pressure-sensitive array + ResNet-18	High-precision dysgraphia detection; effective noise handling	Requires specialized hardware; limited population validation
[23]	CTSO-ZFNet with LHO and pre-processing	89.69% accuracy, 90.87% TPR, 90.55% NPV	Complex multi-modal system; may hinder real-world deployment
[24]	Attention-based continuous CNN	96.5% mean accuracy; improved spatial feature learning	Needs validation on diverse, real-world data
[25]	CC-Net on Archimedes spiral drawings	89.3% accuracy, AUC 0.934, MCC 0.733	Limited real-world scenario testing

AUC – area under the curve, CC-Net – continuous convolution network, CNN – convolutional neural network, CosineDCNN – cosine deep convolutional neural network, CTSO-ZFNet – Chronological Tangent Search Optimization-Zeiler and Fergus Network, IoT – Internet of things, LHO – Lyrebird herd optimization, MCC – Matthews correlation coefficient, ML – machine learning, MRgFUS – magnetic resonance-guided focused ultrasound – MR-guided focused ultrasound, ResNet – residual network, SCGA – sine cosine geese migration optimization, TPR – true positive rate, NPV – negative predictive value

and screening for PD using speech, gait, handwriting, and sensor data [14,17,18,22]. Nonetheless, significant research gaps remain: (i) most current methods depend on high-cost or infrastructure-dependent modalities such as acoustic recordings or wearable sensors [17,18], which makes them hard to scale; (ii) previous hybrid models have mostly focused on temporal learning through CNN – recurrent neural network (RNN) or transformer-based fusion networks [23,24], which are less effective with static motor-imaging data; (iii) feature-level fusion between deep and handcrafted descriptors is not often studied in Parkinson's diagnostics [22]; and (iv) existing studies do not consistently ensure that explainability and cross-dataset validation are present, which are important for clinical trust and reproducibility [25].

To address these deficiencies, this study introduces an explainable hybrid CNN + GLCM framework that amalgamates deep spatial representations with handcrafted textural statistics for the analysis of hand-drawn spirals and waveforms, a neurologically validated and cost-effective modality. This hybridisation facilitates interpretability, resource efficiency, and cross-cohort generalisation, thereby rectifying the observed deficiencies and propelling the domain of AI-driven, explainable PD screening.

### Problem statement

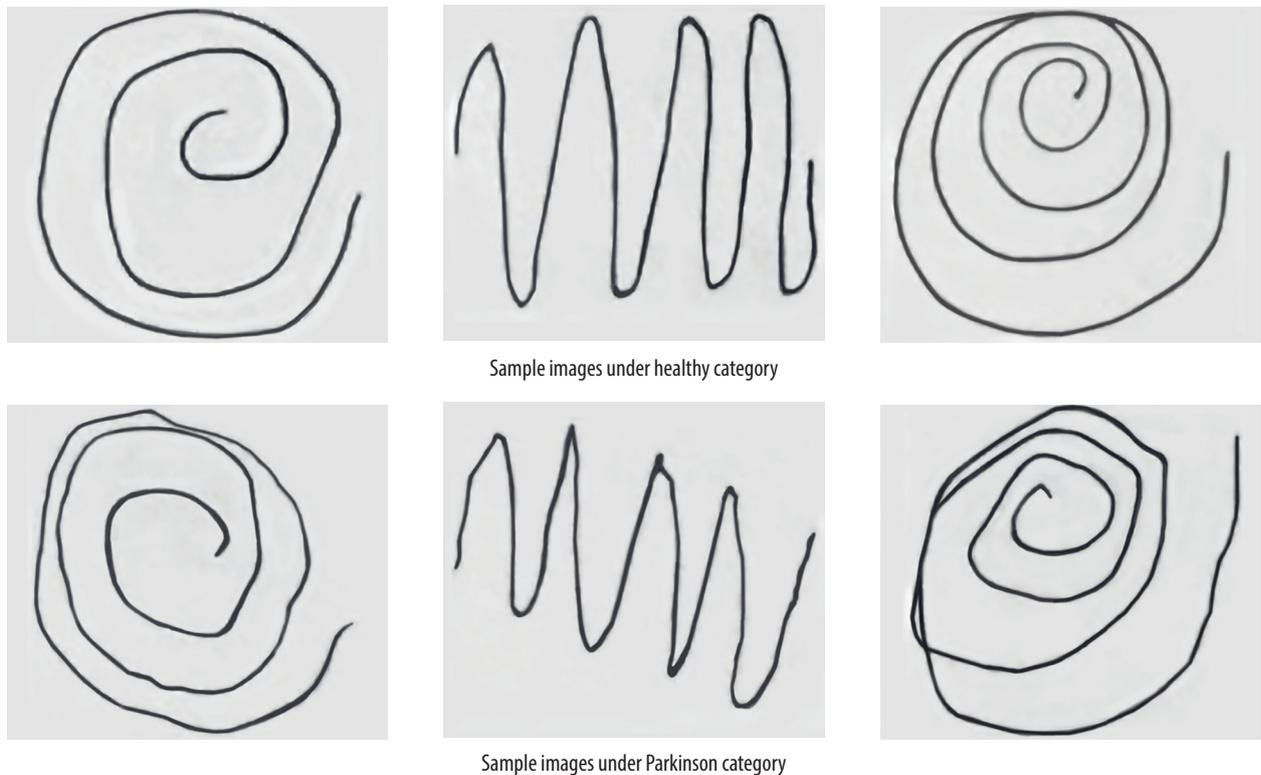
Although there have been significant developments in the application of AI for PD diagnosis, particularly through

handwriting, speech, and spiral drawing analysis, serious drawbacks often hinder existing solutions. These include non-generalizability with small or homogeneous data sets [20], excessively high computational costs in 3D models [18], reliance on custom-made acquisition devices [22], and a lack of interpretability or clinical usability [23,24]. Additionally, most models use deep features or handcrafted descriptors but only infrequently combine the complementary strengths of both. Encouraged by these shortcomings, this paper proposes a stable and interpretable hybrid diagnostic model that combines CNN-based spatial features with GLCM-based texture features in augmented 2-channel input representations. This approach addresses the dual challenge of achieving enhanced classification performance while maintaining clinical interpretability and computational efficiency under diverse real-world conditions.

## Material and methods

### Dataset description

Two publicly accessible databases of hand-drawn images for PD classification were used in this study. The primary training dataset was taken from Augmented Hand-Drawn Data for Parkinson's Disease [28], which contains 3,264 grey-scale wave and spiral images. The database contains both PD class and control class images and has undergone extensive augmentation (i.e., rotation, zoom, and translation)



**Figure 1.** Sample images from training dataset

for improved generalization. The data are well annotated for deep learning validation and serve as a good starting point for training effective models. As a standalone test and model generalization test, the second dataset [29] includes raw waveforms and spiral charts of patients with PD and healthy controls with 60 images. This dataset introduces variation in the style of the drawings, brushstroke quality, and image acquisition, thereby enabling us to evaluate the usability of the learned models under real-world conditions. A few examples of the training dataset images are shown in Figure 1.

### Image pre-processing and augmentation

To ensure consistent input for all deep models, all hand-drawn spiral and wave pictures (already grey scale) were initially resized to  $128 \times 128$  pixels using bilinear interpolation (Eq. 1). Pixel intensity was also normalized to  $[0, 1]$  by dividing by 255 to ensure normalized input distributions and accelerate training convergence, represented by

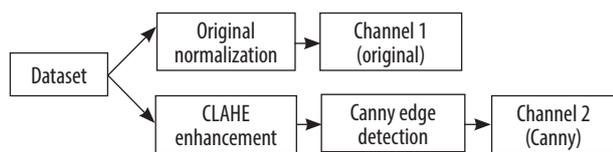


Figure 2. Pre-processing pipeline for image fusion

Eq. 2. To introduce diversity to training samples and prevent overfitting, a custom data generator – EdgeImage-Generator, derived from Keras’ Sequence class – was implemented. The generator consists of light-weight, real-time augmentations such as: horizontal flip with 50% probability, random rotations between  $-10^\circ$  and  $+10^\circ$ , and injection of Gaussian noise (mean 0, standard deviation 0.02) with 30% probability.

Apart from augmentation, a multi-representation fusion strategy is proposed for enriching features. Specifically, the approach operates on two channels, with the original black-and-white image serving as the initial channel. The second channel was created by applying contrast-limited adaptive histogram equalization (CLAHE) to improve local contrast and emphasize fine texture details, especially beneficial in tremor analysis, represented by Eq. 3 Then, Canny edge detection was applied to the improved image using CLAHE, thereby maintaining stroke boundaries and line irregularities of motor symptoms, using Eq. 4 and 5 [30]. Two highly complementary channels with different properties were superimposed over one another to construct a composite 2-channel image without compromising spatial, textural, and structural information, shown in Eq. 6. The improved representation was used to train the CNN-based models, resulting in improved discriminative features for Parkinson’s diagnosis from hand-drawn activities. Figure 2 shows the pre-processing pipeline of the dataset, and sample output after applying pre-processing is shown in Figure 3.

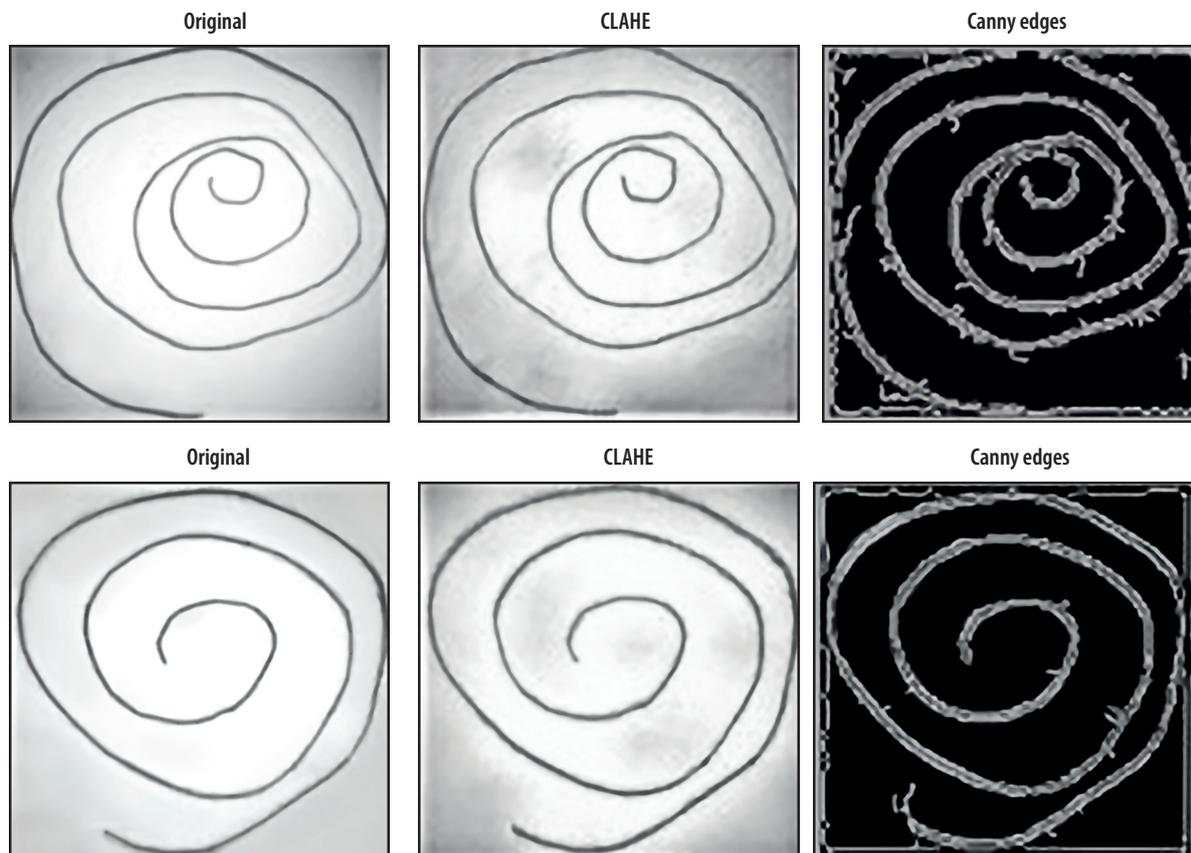


Figure 3. Samples of new images generated after CLAHE and Canny

Let the original greyscale image be denoted as:

$$I_{\text{raw}} \in \mathbb{R}^{H \times W} \quad (1)$$

Where  $H = W = 128$

To normalize pixel values to the range  $[0, 1]$ , each pixel intensity is scaled as:

$$I_{\text{norm}}(x, y) = \frac{I_{\text{raw}}(x, y)}{255} \quad (2)$$

This normalized image forms channel 1 (original).

The CLAHE transformation, which applies localized histogram equalization with contrast limiting, can be abstractly represented as Eq. 3:

$$I_{\text{clahc}} = \text{CLAHE}(I_{\text{raw}}, \text{clipLimit} = 2.0, \text{tileGridSize} = (8, 8)) \quad (3)$$

Where  $\text{clipLimit} = 2.0$  and  $\text{tileGridSize} = (8, 8)$  are hyper-parameters used to control local contrast. To further extract the edge maps from the CLAHE-enhanced image, Canny edge detection is applied using Eq. 4 and 5.

$$G = \sqrt{(I_x)^2 + (I_y)^2} \quad (4)$$

Where  $G$  is a gradient magnitude. Now non-maximum suppression and double thresholding with thresholds  $T_1 = 30$  and  $T_2 = 100$  yield the final binary edge map:

$$I_{\text{canny}} = \text{Canny}(I_{\text{clahc}}, T_1 = 30, T_2 = 100) \quad (5)$$

This forms channel 2 (edges). The final input tensor is constructed by stacking the normalized original image and the edge-detected (Canny) image as per Eq. 6:

$$I_{\text{final}}(x, y, :) = \left[ \frac{I_{\text{raw}}(x, y)}{255}, \frac{I_{\text{canny}}(x, y)}{255} \right] \quad (6)$$

Where  $I_{\text{final}} \in \mathbb{R}^{(128 \times 128 \times 2)}$  represents the composite input comprising spatial intensity and edge information.

## Feature extraction using GLCM

For encoding the spatial pattern and texture features of hand-drawn spiral and waveform images, the GLCM is utilized to extract features [31]. The GLCM is a statistical technique that measures spatial pixel relationships by detecting the frequency with which a pixel with the grey-level value  $i$  is adjacent to a pixel with a value  $j$  at some offset and direction. It is especially suited for biomedical imaging, where minute details in texture contrast, such as tremors or micro-motion in hand-drawn images, are medically important. For every greyscale image used in this study, it was pre-processed to obtain a set of six discriminant texture features from its GLCM calculated

across several directions ( $0^\circ$ ,  $45^\circ$ ,  $90^\circ$ , and  $135^\circ$ ), and their means were used. Extracted features are: contrast, correlation, energy, homogeneity, dissimilarity, and entropy, as given by Eq. 7 to 12. All the input images were processed using different GLCM-based feature extraction methods, which were employed either as standalone inputs to traditional classifiers or as inputs combined with deep learning approaches to increase interpretability and classify efficiently. Texture descriptors are significant indicators of drawing disturbances associated with motor disorders within PD.

Let  $P(i, j)$  represent the normalized GLCM for grey levels  $i, j \in [0, L-1]$ , where  $L$  is the number of grey levels (typically 256).

*Contrast* – Measures local variations in grey-level co-occurrence:

$$\text{Contrast} = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} (i-j)^2 P(i, j) \quad (7)$$

*Correlation* – Measures the joint probability occurrence of specified pixel pairs:

$$\text{Correlation} = \frac{\sum_{i=0}^{L-1} \sum_{j=0}^{L-1} (i-\mu_i)(j-\mu_j)P(i, j)}{\sigma_i \sigma_j} \quad (8)$$

where  $\mu_i$ ,  $\mu_j$  and  $\sigma_i$ ,  $\sigma_j$  are the means and standard deviations of the row and column marginal distributions.

*Energy* – Also known as angular second moment (ASM); measures textural uniformity:

$$\text{Energy} = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} P(i, j)^2 \quad (9)$$

*Homogeneity* – Measures the closeness of the distribution of elements to the GLCM diagonal:

$$\text{Homogeneity} = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} \frac{P(i, j)}{1 + |i - j|} \quad (10)$$

*Dissimilarity* – Similar to contrast but with a linear weight:

$$\text{Dissimilarity} = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} |i - j| P(i, j) \quad (11)$$

*Entropy* – Measures the randomness or complexity of texture patterns:

$$\text{Entropy} = - \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} P(i, j) \log_2 P(i, j) \quad (12)$$

(With the convention that  $0 \cdot \log_2 0 = 0$ ).

## Proposed model architecture

To assess the capability of hand-drawn image descriptions in classifying PD, three model architectures were employed: a lightweight baseline CNN, a CNN-GLCM

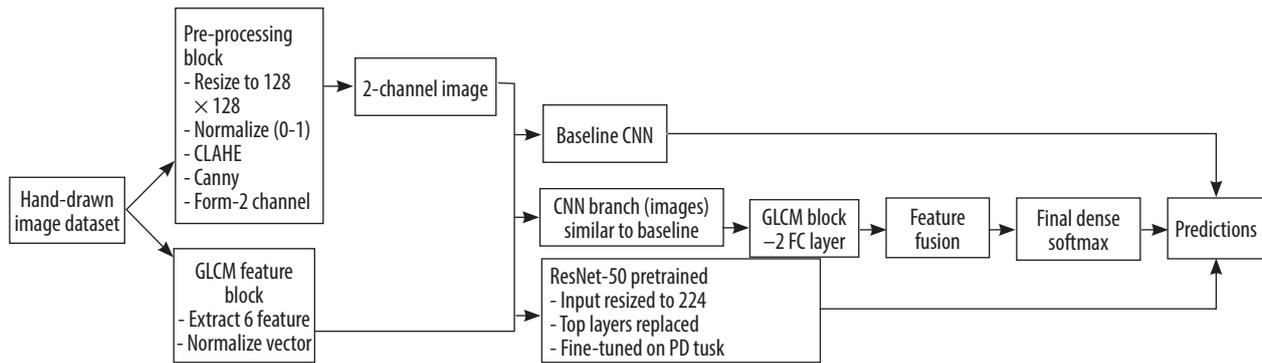


Figure 4. Proposed model pipeline for Parkinson's disease classification using hand-drawn spiral and waveform images

fusion model, and a ResNet-50 fine-tuned transfer learning model. Each model was implemented to research different views of spatial, textural, and semantic features of the image input. The model pipeline is shown in Figure 4, and Algorithm 1 illustrates the main steps involved in designing the hybrid model (CNN + GLCM).

### Baseline CNN

The baseline CNN was designed to be a robust and straightforward architecture for handling the 2-channel input images produced by the pre-processing pipeline (original normalized + Canny edges). Three convolutional blocks follow a rectified linear unit (ReLU) activation (Eq. 14), a batch normalization (Eq. 15), and a max-pooling (Eq. 16) layer. The convolutional layers are sized to extract low- to mid-level features such as edges, curves, and junctions related to hand motions in a stepwise order. The final convolutional output is flattened and passed through two dense layers with dropout regularization, followed by a sigmoid or softmax activation (Eq. 17 to 20). The light CNN model serves as a baseline for comparison, assessing the impact of additional handcrafted or pre-trained features. Let the 2-channel pre-processed image be denoted as  $I \in \mathbb{R}^{(128 \times 128 \times 2)}$ ; each convolution layer applies a set of learnable filters  $\mathbf{W}^{(l)}$  followed by bias  $\mathbf{b}^{(l)}$ ; the output feature map represented as Eq. 13.

$$\mathbf{Z}^{(l)} = \mathbf{W}^{(l)} * \mathbf{I}^{(l-1)} + \mathbf{b}^{(l)}, \quad \mathbf{I}^{(0)} = \mathbf{I} \quad (13)$$

Where  $*$  denotes 2D convolution,  $l$  is the layer index and  $\mathbf{I}^{(0)} = \mathbf{I}$  is the input image.

Activation map

$$(\mathbf{A}^{(l)}) = \text{ReLU}(\mathbf{Z}^{(l)}) = \max(0, \mathbf{Z}^{(l)}) \quad (14)$$

Output after batch normalization

$$(\bar{\mathbf{A}}^{(l)}) = \gamma \cdot \frac{\mathbf{A}^{(l)} - \mu}{\sqrt{(\sigma^2 + \delta)}} + \beta \quad (15)$$

Where  $\mu$  is a mean,  $\sigma^2$  is variance,  $\gamma$  is a learnable scale and  $\beta$  is a shift parameter in batch normalization. The pooled output at layer  $l$  is represented as Eq. 16 and the flattened vector from the last convolution block is represented in Eq. 17.

$$\mathbf{P}^{(l)} = \text{MaxPool}(\bar{\mathbf{A}}^{(l)}) \quad (16)$$

$$\mathbf{f} = \text{Flatten}(\mathbf{P}^{(L)}) \quad (17)$$

Hidden representation from dense layer is  $\mathbf{h}_1$  and  $\mathbf{h}_2$ , where  $\mathbf{W}_1$  and  $\mathbf{W}_2$  are weight metrics for dense layers, and  $\mathbf{b}_1$  and  $\mathbf{b}_2$  are bias vectors for dense layers.

$$\mathbf{h}_1 = \text{Dropout}(\text{ReLU}(\mathbf{W}_1 \cdot \mathbf{f} + \mathbf{b}_1)) \quad (18)$$

$$\mathbf{h}_2 = \text{Dropout}(\text{ReLU}(\mathbf{W}_2 \cdot \mathbf{h}_1 + \mathbf{b}_2)) \quad (19)$$

Final output prediction is represented by Eq. 20 where  $\mathbf{W}_o$  and  $\mathbf{b}_o$  are weight and bias of the output layer.

$$\text{OutputLayer}(\text{Softmax}): [\hat{\mathbf{y}} = \text{Softmax}(\mathbf{W}_o \cdot \mathbf{h}_2 + \mathbf{b}_o)] \quad (20)$$

### CNN with GLCM fusion

To leverage both spatial information from images and texture-based statistical information, a combined model was proposed that combines CNN-learned features and GLCM-based handcrafted features. The GLCM feature vector (6D: correlation, contrast, energy, entropy, etc.) was first normalized and then passed to a fully connected dense sub-network with dropout and ReLU activation. In parallel, the image stream was passed to the same baseline CNN. Each branch's output was concatenated and passed through a shared dense layer and subsequently to the classification head. The proposed two-branch fusion enabled the model to leverage both types of features – deep features as well as interpretable textural features – thereby enhancing both performance and clinical interpretability. Let  $\mathbf{f}_{\text{img}} \in \mathbb{R}^6$  be the feature vector from baseline CNN and  $\mathbf{g}_{\text{glcm}} \in \mathbb{R}^6$  be the GLCM feature vector with 6 dimensions; Eq. 21 represents the GLCM subnetwork:

$$\mathbf{h}_{\text{glcm}} = \text{ReLU}(\mathbf{W}_g \cdot \mathbf{g}_{\text{glcm}} + \mathbf{b}_g) \quad (21)$$

Where  $\mathbf{W}_g$  and  $\mathbf{b}_g$  are weight and bias for the dense layer processing GLCM features. Eq. 22 represents the feature fusion by concatenating the  $\mathbf{f}_{\text{img}}$  and hidden representation from GLCM sub-network ( $\mathbf{h}_{\text{glcm}}$ ).

$$\mathbf{f}_{\text{fused}} = \text{Concat}(\mathbf{f}_{\text{img}}, \mathbf{h}_{\text{glcm}}) \quad (22)$$

Now the final classification and prediction are represented by Eq. 23 and 24, where  $\mathbf{h}_3$  is the hidden representation after fusion and dropout,  $\mathbf{W}_3$  is weight after fusion,  $\mathbf{b}_3$  is bias after fusion,  $\mathbf{W}_o$  is weight after final classification, and  $\mathbf{b}_o$  is bias after final classification.

$$\mathbf{h}_3 = \text{Dropout}(\text{ReLU}(\mathbf{W}_3 \mathbf{f}_{\text{fused}} + \mathbf{b}_3)) \quad (23)$$

$$\hat{\mathbf{y}} = \text{Softmax}(\mathbf{W}_o \cdot \mathbf{h}_3 + \mathbf{b}_o) \quad (24)$$

### Fine-tune ResNet-50

To leverage the representational capacity of deep convolutional models, we used the pre-learned ResNet-50 model on ImageNet and fine-tuned it for PD classification. The last layers of ResNet-50 were replaced with a new classification head tailored to the target task. The ResNet-50 input model was created by duplicating or rescaling the 2-channel input into the desired shape of  $224 \times 224 \times 3$ . The lower layers were first frozen to preserve pre-trained weights and only train the top dense layers. Selective unfreezing was then conducted with a slower learning rate to fine-tune deeper layers. This design acted as a strong baseline for high-level abstraction and transfer learning performance with small clinical image datasets. Let pre-processed image  $X \in \mathbb{R}^{(224 \times 224 \times 3)}$  be created by repeating 2-channel input. The convolution and residual block of ResNet-50 is represented as:  $\mathbf{Z}^{(1)} = \text{Conv}(\mathbf{X})$ ,  $\mathbf{Z}^{(2)} = \text{ResidualBlock}_1(\mathbf{Z}^{(1)})$ , ... . Final classification and prediction of the model are represented as Eq. 25 and 26, where  $\mathbf{W}_{\text{res}}$  and  $\mathbf{b}_{\text{res}}$  are weight and bias of the top dense layer after ResNet block,  $\mathbf{h}_{\text{res}}$  (Eq. 25) represents the hidden layer activation from dense layer post-ResNet,  $\mathbf{f}_{\text{resnet}}$  is the global average pooled output from final ResNet layer ( $\mathbf{f}_{\text{resnet}} \in \mathbb{R}^d$ ), and  $\mathbf{W}_o$  and  $\mathbf{b}_o$  are weight and bias of the final output layer.

$$\mathbf{h}_{\text{res}} = \text{Dropout}(\text{ReLU}(\mathbf{W}_{\text{res}} \mathbf{f}_{\text{resnet}} + \mathbf{b}_{\text{res}})) \quad (25)$$

$$\hat{\mathbf{y}} = \text{Softmax}(\mathbf{W}_o \cdot \mathbf{h}_{\text{res}} + \mathbf{b}_o) \quad (26)$$

CNN – convolutional neural network, FC – fully connected, GLCM – grey-level co-occurrence matrix, PD – Parkinson's disease

**Algorithm 1:** Parkinson's disease detection using hybrid CNN + GLCM model

**Input:** Hand-drawn image dataset labelled as Parkinson or healthy, image size:  $128 \times 128$

**Output:** Trained hybrid model, classification accuracy and performance report

### Begin

#### 1. Pre-processing block:

For each image:

- a. Resize image to  $224 \times 224$
- b. Normalize pixel intensities to  $[0, 1]$
- c. Extract GLCM features:
  - i. Convert image to greyscale
  - ii. Compute GLCM
  - iii. Extract 6 features: contrast, dissimilarity, homogeneity, ASM, energy, correlation
- d. Store the image, GLCM features, and label

#### 2. Label encoding and one-hot conversion:

- Encode categorical labels (Healthy = 0, Parkinson = 1)
- Convert encoded labels for classification

#### 3. Compute class weights:

- Use class frequencies to calculate class\_weight for imbalanced datasets

#### 4. Data splitting:

- Split dataset into training and testing sets 80 : 20 ratio

#### 5. Data augmentation (training images only):

- Apply real-time transformations:
- Rotation, width/height shift, zoom, horizontal flip

#### 6. Define hybrid model:

- a. Image branch (baseline CNN):
  - Input  $\rightarrow$  Conv  $\rightarrow$  ReLU  $\rightarrow$  BatchNorm  $\rightarrow$  MaxPool
  - Repeat for 2-3 blocks
  - Flatten  $\rightarrow$  Dense  $\rightarrow$  Dropout
- b. GLCM branch:
  - Input  $\rightarrow$  Dense(16)  $\rightarrow$  ReLU  $\rightarrow$  Dropout
- c. Feature fusion:
  - Concatenate outputs from CNN and GLCM branches
  - Pass through dense(64)  $\rightarrow$  Dropout
  - Output via dense softmax layer

#### 7. Compile model:

- Loss: categorical\_crossentropy
- Optimizer: Adam
- Evaluation metric: accuracy

#### 8. Train model:

- Use early stopping (patience = 10)
- Train on augmented image batches and GLCM features
- Validate on test set

#### 9. Evaluate model:

- Predict class probabilities on test data
- Convert probabilities to class labels
- Compute accuracy, precision, recall, and classification report

### End

## Results and discussion

### Experimental setup

To train the model, the simulations were performed on a local computing setup featuring an Intel Core i7 processor (8<sup>th</sup> Generation), 16 GB of RAM, and the Windows 10 operating system [32]. The models were implemented in Python 3.x using TensorFlow 2.9 and the high-level Keras API to train the neural network. The main libraries employed included OpenCV for pre-processing the images, NumPy and Pandas for data processing, scikit-learn for performance analysis and feature extraction using GLCM (utilizing skimage.feature), and Matplotlib for graphics. All three model architectures were tested: baseline CNN customized for our data, a CNN + GLCM hybrid fusion model, and a fine-tuned ResNet-50. All three models were trained on an equivalent pre-processed 2-channel input and stratified train-validation splits for a proper comparison. Data augmentation was used uniformly on all experiments using ImageDataGenerator.

To guarantee equitable and impartial model assessment, all experiments were structured to maintain patient-level independence among training, validation, and testing datasets. The hand-drawn spirals and waveforms of each record were combined, and the data-splitting process was conducted at the record level rather than the image level. This method made sure that no sketch from the same patient was in more than one partition. Dataset 1 (augmented hand-drawn data for PD) was divided into an 80 : 20 train-validation split, while Dataset 2 (an independent cohort acquired under different drawing and scanning settings) was reserved solely for external test-

ing. This separation at the subject level ensures that the model's performance genuinely generalizes to novel individuals, rather than merely memorizing image-specific patterns, thereby enhancing clinical reliability.

Hyper-parameters were manually tuned for stability during training and performance. A summary of the important parameters used across models is provided in Table 2. The choice of critical hyper-parameters was informed by empirical testing. The input image size of 128 × 128 was chosen to find a good compromise between speed and the preservation of spiral and waveform details, which are important for analysing Parkinsonian tremors. The size 224 × 224 was kept for ResNet-50 to match its pre-training setup on ImageNet. The Adam optimizer's learning rate of 0.001 allowed for quick but stable convergence without overshooting minima, which is common for small-batch medical imaging. The batch size of 32 made sure that the gradient updates were steady on the available graphics processing unit (GPU) memory. The epoch count (20-25) was set once the accuracy and loss curves showed that convergence was stable, which stopped overfitting. We used dropout (0.4) and batch normalisation to regularise the data. The GLCM branch dense units (16) had enough space to store six handcrafted texture descriptors without any extra data. In general, these parameters were adjusted in a loop until the best balance of accuracy, generalisation, and training stability was found for all models.

### Performance analysis

To analyse the efficacy of the models proposed for classifying PD to the highest degree, controlled experiments

Table 2. Model parameters and configuration summary

Model	Parameter name	Value	Description
Baseline CNN	Input size	128 × 128 × 2	2-channel image (original + edge)
	Conv blocks	3	Conv → ReLU → MaxPool → BatchNorm
	Dense units	128	Fully connected layer before softmax
	Dropout rate	0.4	For regularization
	Optimizer	Adam	Adaptive gradient-based optimizer
	Learning rate	0.001	Default Adam LR
CNN_GLCM	GLCM features	6	Texture descriptors: contrast, energy, etc.
	Fusion strategy	Concatenation	CNN + GLCM merged before final dense layer
	GLCM dense units	16	GLCM branch FC layer
	Batch size	32	Training batch size
ResNet-50	Input size	224 × 224 × 3	Resized image input with 3 replicated channels
	Pre-trained	ImageNet	Transfer learning on pre-trained ResNet-50
	Fine-tuned layers	Last 10 layers	Selective unfreezing for task adaptation
	Output layer	Softmax	Final activation for classification

CNN – convolutional neural network, GLCM – grey-level co-occurrence matrix, FC – fully connected, LR – learning rate, ReLU – rectified linear unit, ResNet-50 – 50-layer residual network

have been conducted on various input configurations and model structures. Particularly, the behaviour of three fundamental deep models has been compared – Baseline CNN, CNN with GLCM fusion (CNN + GLCM), and ResNet-50 fine-tuned – on two input groups: (1) raw greyscale hand-drawn images and (2) pre-processed 2-channel images derived from fusing normalized greyscale and Canny edges. Furthermore, the same fourth model, with just the six GLCM texture features and fully connected layers, was also experimented with to compare the discriminability of handcrafted features alone.

Training and validation performance were evaluated for every model using principal classification measures, named precision, recall, and F1-score, as well as training and validation accuracy curves to assess generalization and overfitting. Model optimization with a training-validation split was followed by testing on different datasets, using Dataset 2 as test data to evaluate the final classification performance and model stability. To further enhance the clinical explainability of deep learning predictions, Grad-CAM was applied to examine test examples carefully. Using Grad-CAM heatmap visualizations, prominent regions in drawing images were identified that influenced the model's predictions, enabling transparent evaluation of learned representations and increasing the model's reliability for real-world diagnostic applications.

#### Performance analysis of baseline CNN on original dataset

Initially, the baseline CNN was trained on greyscale hand-drawn images to determine whether it could recognize PD. During the training and validation process, the model's performance was monitored for 20 epochs. According to Figure 5 from epoch 12, the training accuracy plateaued at approximately 72%. On the other hand, validation accuracy fluctuated from epoch to epoch, reaching a peak value of 78.3%. It indicates that while the model could learn significant patterns, it was somewhat sensitive to the

changes in the validation data. After the last model validation on the validation set, the achieved evaluation metrics were encouraging: accuracy = 0.783, precision = 0.791, recall = 0.762, and F1-score = 0.775. They indicate a well-balanced classification ability of the model with a good trade-off between sensitivity and specificity. Despite some reported variability in training and validation accuracy trends, the overall F1-score indicates that the baseline CNN generalizes quite adequately when trained on raw greyscale features alone without feature fusion.

#### Performance analysis of baseline CNN on pre-processed normalized greyscale and Canny edges dataset

To enhance the model's capacity to extract both intensity and structural stroke data, the baseline CNN was trained on a 2-channel input representation created by stacking the normalized greyscale image and its Canny edge map. The fusion enabled the model to leverage complementary information regarding the quality of stroke, sharpness at boundaries, and tremor patterns that are of maximal relevance in the diagnosis of PD. As is evident from Figure 6, the training accuracy grew steadily, crossing 95% at epoch 23, indicating excellent learning of discriminative features. The validation accuracy increased steadily and stabilized at approximately 88%, with some minor fluctuations, indicating good generalization to new samples and minimal overfitting. A quantitative assessment on the validation set revealed a significant improvement over the original greyscale input. The model achieved an accuracy of 0.88, precision of 0.843, recall of 0.872, and an F1-score of 0.845. The measurements validate a good and strong classification ability with both false negatives and false positives well controlled by the model. The enhanced F1-score validates the strength of incorporating edge-enhanced representations, in addition to raw pixel intensity, in enhancing performance in the classification of PD.

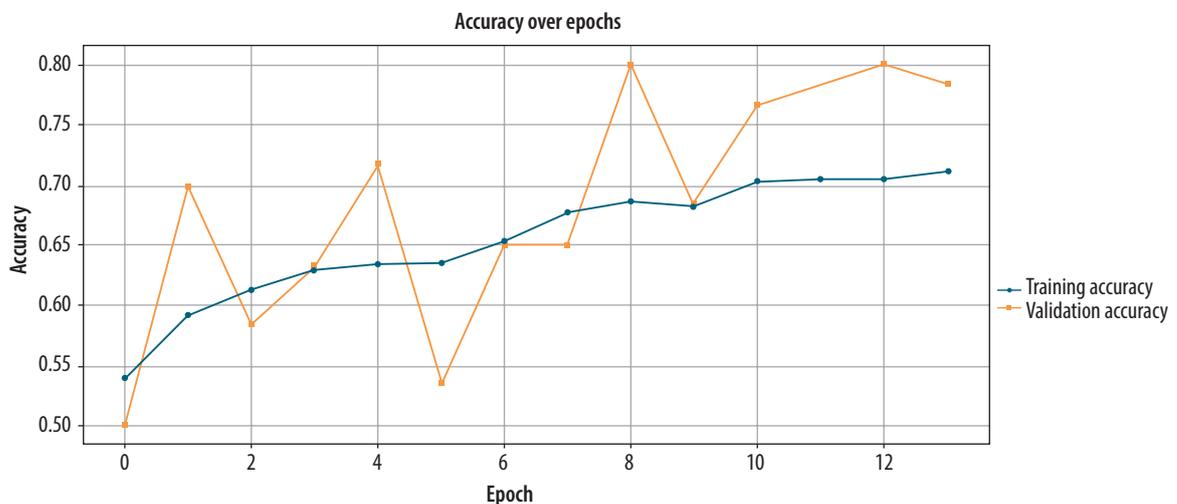
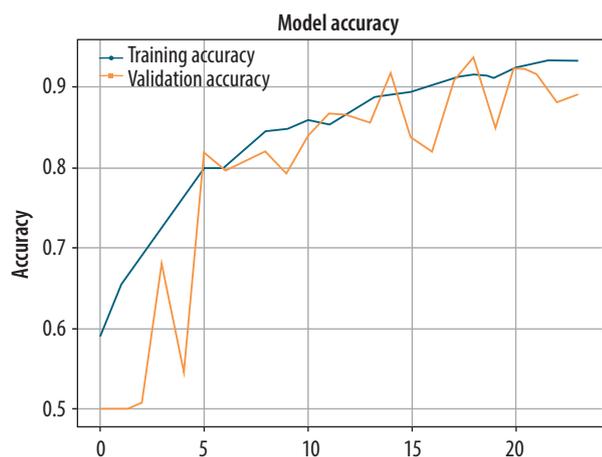
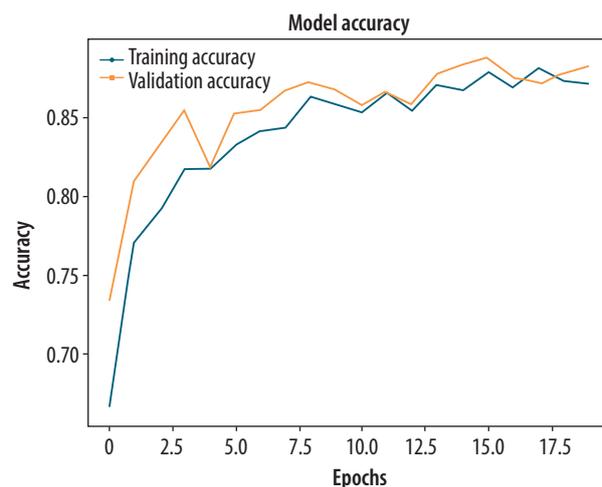


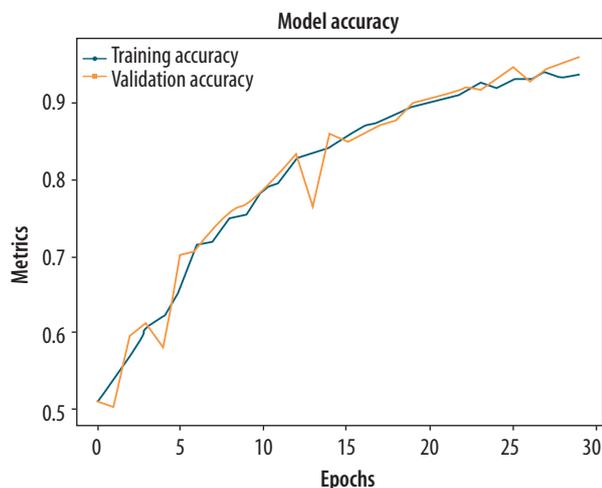
Figure 5. Training and validation accuracy plot of the baseline convolutional neural network on the original greyscale dataset



**Figure 6.** Training and validation accuracy plot of the convolutional neural network using pre-processed normalized greyscale and Canny edges dataset



**Figure 7.** Training and validation accuracy plot of the convolutional neural network + grey-level co-occurrence matrix on the original greyscale dataset



**Figure 8.** Training and validation accuracy plot of the convolutional neural network + grey-level co-occurrence matrix using pre-processed normalized greyscale and Canny edges dataset

### Performance analysis of hybrid CNN + GLCM on original dataset

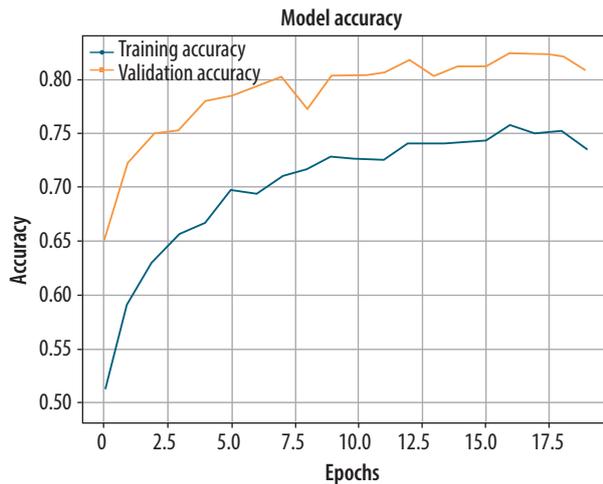
The hybrid model, which integrated CNN attributes with handcrafted GLCM-based texture attributes, was tested on the original greyscale hand-drawn image data to simulate the complementary advantages of spatial and statistical information. The CNN branch learned abstract spatial features from the images. The GLCM branch, on the other hand, provided six handcrafted features that characterized textural attributes, such as contrast, energy, and correlation. These two streams of attributes were combined before classification to produce a hybrid model.

As shown in Figure 7, the model was trained consistently and comprehensively across 20 epochs. Training accuracy and validation accuracy improved consistently and converged at approximately 89% with near-zero difference between them – a sign of high generalization

and low risk of overfitting. Quantitative analysis also confirmed the efficiency of the proposed hybrid model. The model recorded a classification accuracy of 0.89, precision of 0.901, recall of 0.823, and an F1-score of 0.891. The good precision is a measure of the model’s capability to suppress false positives, while the strong F1-score indicates a good balance of sensitivity and specificity. The metrics value indicates that the inclusion of handcrafted GLCM features significantly improved the discriminative capability of the CNN, especially in detecting fine texture anomalies in Parkinsonian states.

### Performance analysis of hybrid CNN + GLCM on pre-processed normalized greyscale and Canny edges dataset

The hybrid CNN + GLCM model was further evaluated using a pre-processed 2-channel image representation, which combined normalized greyscale images with Canny edge maps to enhance the detection of structural features. These spatial features were combined with six handcrafted GLCM texture descriptors, providing a rich and complementary feature set for effective PD classification. According to Figure 8, both training and validation accuracy exhibited a consistent upward trend throughout the training process. The alignment between training and validation performance indicates strong generalization and minimal overfitting. Consequently, the validation dataset performed well because the model converged smoothly. The proposed hybrid model achieved a quantitative accuracy of 0.9702, a precision of 0.945, a recall of 0.923, and an F1-score of 0.954. Due to its high reliability and ability to strike a balance between sensitivity and specificity, this model is well suited for clinical decision support. Additionally, edge-enhanced image information improved the model’s ability to detect subtle differences in drawing patterns, contributing to its optimal performance.



**Figure 9.** Training and validation accuracy plot of the ResNet-50 on the original greyscale dataset

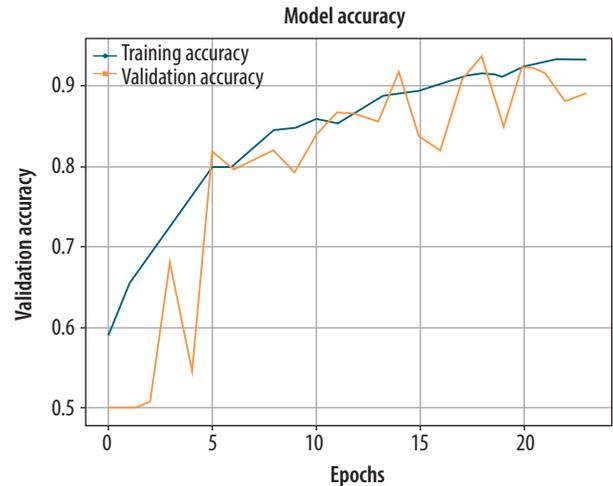
### Performance analysis of ResNet-50 on original dataset

ResNet-50 was pre-trained on ImageNet and fine-tuned for PD classification using the original greyscale hand-drawn spiral and waveform images. To meet the architecture's requirement for three channels, the input was resized to  $224 \times 224$  pixels. The top layers of the model were replaced with custom classification heads that were selectively tuned to learn task-specific features during training. Figure 9 shows that both training and validation accuracy increased steadily during the first few epochs. There was a consistent upward trend in the validation accuracy, which peaked at 80.6%, while the training accuracy plateaued at about 75%. Despite the slight discrepancy, the pre-trained layers provided a strong feature base, although the fixed lower layers limited their ability to adapt to subtle PD-related variations.

Despite this, the model performed well, achieving a final accuracy of 0.8067, precision of 0.789, recall of 0.823, and an F1-score of 0.83. Based on the F1-score, it appears that the ResNet-50 model is a viable option for PD detection without the inclusion of edge-enhanced features. These results validate the effectiveness of transfer learning in biomedical image classification tasks, especially when dealing with moderately sized datasets.

### Performance analysis of ResNet-50 on pre-processed normalized greyscale and Canny edges dataset

To further enhance the model's capacity to identify structural and textural changes in hand-drawn images, the ResNet-50 model was also fine-tuned on pre-processed 2-channel representations, such as normalized greyscale input and Canny edge maps. The combined input was modified to meet the three-channel requirement of ResNet-50 and used to transfer-learn discriminative patterns for PD detection. As shown in Figure 10, the model



**Figure 10.** Training and validation accuracy plot of the ResNet-50 using Pre-processed normalized greyscale and Canny edges dataset

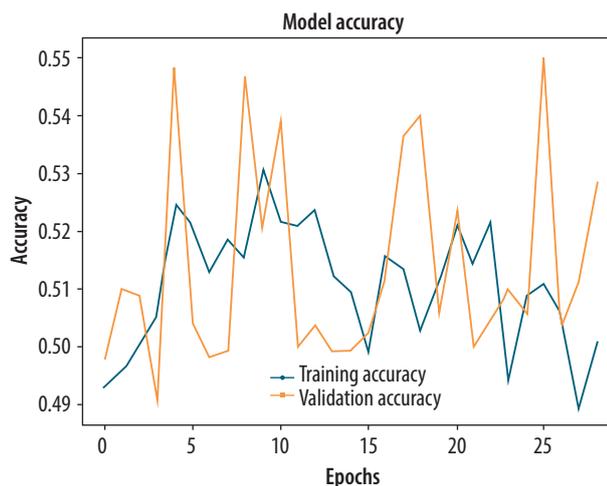
demonstrated steady progress in training accuracy over 23 epochs, ultimately achieving 95%. Validation accuracy varied slightly with responsiveness to image-level change but levelled off at a strong 88.9%. This trend suggests effective learning with acceptable overfitting, which can be attributed to the depth of ResNet-50 and the expressiveness of the input representation.

Validation set results yielded exceptional values for the metrics: 0.8896 for accuracy, 0.8220 for precision, 0.856 for recall, and an F1-score of 0.89. Overall, the model retains high sensitivity and specificity. ResNet-50's edge-enhanced visual features enabled deeper layers to detect and identify minor line distortions of tremor origin, as well as boundary irregularities, with greater accuracy.

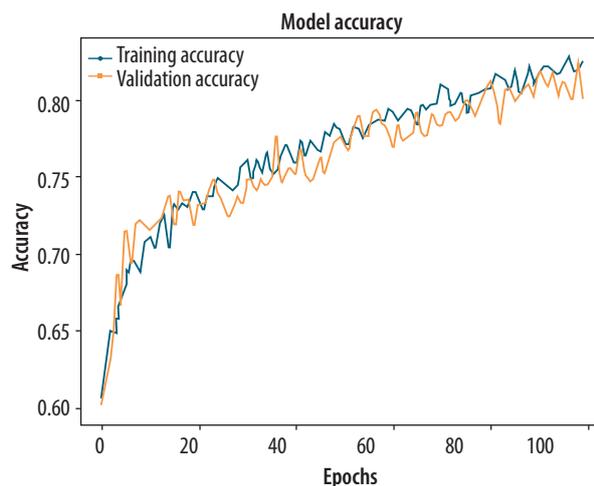
### Performance analysis of only GLCM on original dataset

As an intermediate experiment to calculate the discriminative power of handcrafted statistical features alone, we have trained a shallow network that uses only the six GLCM-derived features (contrast, dissimilarity, homogeneity, ASM, energy, and correlation) computed from the original greyscale hand-written images. The goal was to understand the performance of traditional texture-based descriptors in classifying PD using the learned spatial features. As can be seen in Figure 11, the training and validation accuracies were both low and highly unstable during the 30 epochs, fluctuating between 50 and 54%, with no convergence. The model exhibited nonsensical learning behaviour, indicating a poor separation of features and an inability to learn solely from GLCM descriptors.

Final classification performance metrics validated the poor performance of this method, with an accuracy of 0.54, precision of 0.54, recall of 0.55, and an F1-score of 0.54. These findings indicate that although GLCM extracts some local texture features, they are not sufficient alone for classifying PD. Deficiencies in spatial context



**Figure 11.** Training and validation accuracy plot of the only grey-level co-occurrence matrix on the original greyscale dataset



**Figure 12.** Training and validation accuracy plot of the only grey-level co-occurrence matrix using pre-processed normalized greyscale and Canny edges dataset

and high-granularity stroke analysis render the model incapable of discriminating between Parkinsonian and healthy patterns. These findings also reinforce the integration of GLCM with deep features in hybrid models, thereby enhancing prediction performance.

### Performance analysis of only GLCM on pre-processed normalized greyscale and Canny edges dataset

To verify whether textural attributes are enhanced when computed from improved visual presentations, the GLCM-only model was trained with statistical features from pre-processed images, which merged normalized greyscale images and edge maps acquired using the Canny algorithm. The configuration was designed to evaluate the extent to which handcrafted features can benefit by edge-based enhancement when used alone. Training accuracy and validation accuracy both improved progressively over 100 epochs and rose to nearly 80% (Figure 12). The well-smoothened and aligned learning curves indicate good generalization and stable training behaviour without the deep convolutional layers. The performance is markedly

better compared to the GLCM model trained only on raw greyscale images.

The quantitative analysis also reveals that the model achieves an improved accuracy of 0.8009, precision of 0.820, recall of 0.796, and an F1-score of 0.821. These findings suggest that pre-processing images prior to extracting GLCM features enhances their discriminative capacity, possibly by reducing edge structures and eliminating noise. Although still shy of the deep-learning model's representation ability, this iteration of the GLCM-only model presents an encouraging low-complexity solution for detecting PD when computational budgets are constrained.

Following an individual model performance analysis using graphical accuracy trends and a comprehensive metric-based comparison, a composite comparison is presented in Table 3. The classification performance of all models – GLCM-only, CNN-only, CNN + GLCM hybrid, and ResNet-50 – is presented for both original greyscale and pre-processed (normalized greyscale + Canny edge) datasets. The compared metrics are accuracy, precision, recall, and F1-score, which provide an overall indication

**Table 3.** Summary of classification performance (accuracy, precision, recall, F1-score) for all models across original and pre-processed datasets

Model	Accuracy	Precision	Recall	F1-score
GLCM only (original)	0.54	0.54	0.55	0.54
GLCM only (pre-processed)	0.8009	0.82	0.796	0.821
CNN only (original)	0.7833	0.791	0.762	0.775
CNN only (pre-processed)	0.88	0.843	0.872	0.845
CNN + GLCM (original)	0.89	0.901	0.823	0.891
CNN + GLCM (pre-processed)	0.9702	0.945	0.923	0.954
ResNet-50 (original)	0.8067	0.789	0.823	0.83
ResNet-50 (pre-processed)	0.8896	0.821	0.856	0.89

CNN – convolutional neural network, GLCM – grey-level co-occurrence matrix, ResNet-50 – 50layer residual network

of how each model performs under various input configurations. This relative evaluation highlights the comparative strengths of hybrid and deep models, unequivocally demonstrating the advantage of combining handcrafted and learned features with edge-based boosting.

### Model complexity and training efficiency analysis

In clinical AI applications, interpretability and classification accuracy are essential. However, their applicability for deployment is also greatly influenced by computational complexity and training effectiveness, especially in embedded or resource-constrained healthcare systems. Each of the four core models is compared according to the number of trainable parameters (in millions) and the training time per epoch [33].

The GLCM-only model is incredibly lightweight, with parameter sizes of only 0.0031 MB, as shown in Table 4. Based on the deep, fully connected layer following the convolutional blocks, the baseline CNN, although custom-designed for this task, has a relatively large footprint (2.86M parameters). The best compromise between complexity and diagnostic accuracy is the hybrid CNN + GLCM model, which has 0.512 million parameters, primarily because it combines statistical and spatial feature learning. Finally, the fine-tuned ResNet-50 is the most computationally expensive (~2.1M parameters) due to its depth architecture and partial unfreezing scheme; yet, it remains a strong benchmark for transfer learning performance with small clinical image databases. The preceding observations confirm that the suggested hybrid CNN + GLCM model not only performs better in terms of accuracy but also remains practically deployable in real-world clinical practices, especially when compared with deeper pre-trained counterparts.

### Testing of Dataset 2 using CNN + GLCM (pre-processed) model and Grad-CAM interpretability

To ensure the generalisability of the proposed framework, the testing process was specifically crafted to replicate multi-cohort evaluation across independent datasets. The proposed hybrid CNN + GLCM model was trained and validated using the Augmented Hand-Drawn Data for PD dataset [28], which consists of 3,264 spiral and waveform images. The testing of the model was done on

a completely different group dataset [29], which had 60 samples taken under different drawing conditions, stroke styles, and scanning resolutions. By separating the cohorts, the study ensured that the people being tested were never seen during training or hyper-parameter tuning. This made it feel like a real-world deployment on populations that had never been seen before.

Five-fold cross-validation was also conducted on Dataset 1 to assess the internal stability of the data, and we used paired statistical tests (*t*-test and Wilcoxon) to measure the consistency of the results across folds, with significance set at  $p < 0.01$ . Cross-dataset findings showed a small loss of less than 3% in accuracy between the internal and external datasets. This proved that the model still had good discriminability even when the domain changed.

In addition to classification accuracy, Grad-CAM was used to visualize the model's predictions by highlighting the parts of each input that are most responsible for the output classification decision. Grad-CAM generates a heatmap on the input image, enabling us to visualize the exact stroke segments or patterns that the model focused on during Parkinsonian feature identification.

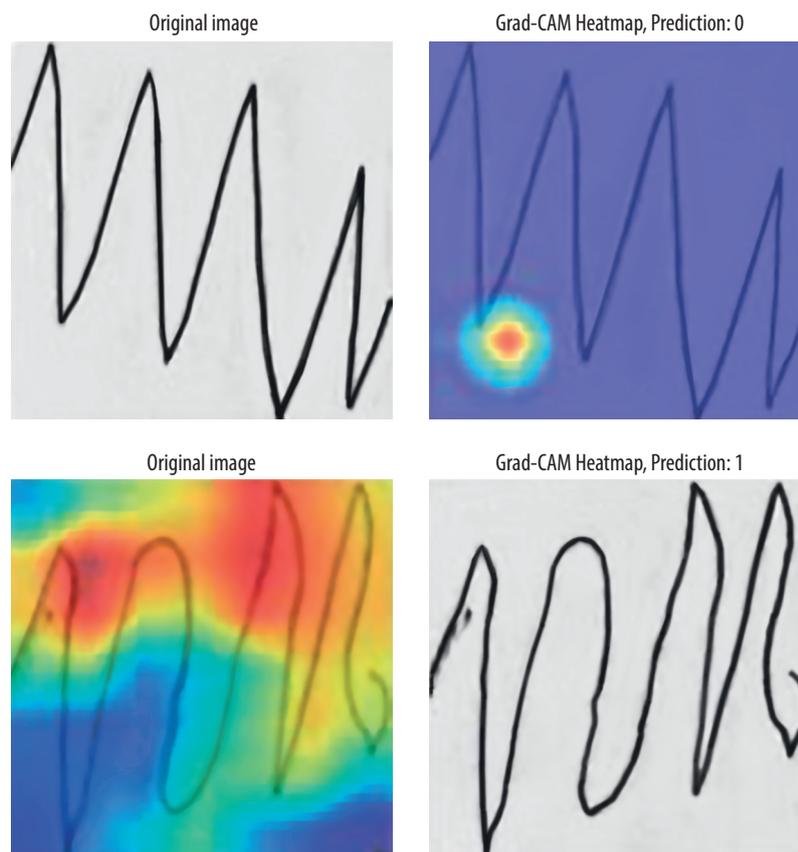
Figure 13 shows two typical cases. The top left image of the first row has been marked as healthy control (prediction: 0). The Grad-CAM heatmap for the same shows smooth and coherent areas of the waveform, which indicates that the model has used stable areas of drawing for its decision. On the other hand, the bottom row image falls in the Parkinson's category (predicted probability: 1), and Grad-CAM's heatmap presents abnormal loops and irregularities, specifically tremor-like deflections and inconsistency in line pressure, which are representations of motor deficits. Such a pattern in images ensures that the model's decision aligns with clinical evidence and supports its deployment as a reliable and understandable diagnostic tool. This module creates saliency maps that highlight the parts of the input image with the greatest impact on model predictions. This facilitated the visualisation of spatial focus areas associated with tremor-related distortions, abnormal stroke patterns, and contour instabilities – clinically significant indicators for the diagnosis of PD.

To enhance interpretability, the Grad-CAM visual outputs were associated with clinical annotation indicators – such as tremor intensity, line discontinuity, and spatial asymmetry – identified by neurologists in previous stud-

**Table 4.** Model training efficiency analysis

Model	Parameters (approx.)	Training time/epoch (s)	Remarks
GLCM-only (MLP)	0.0031M	< 1	Fast, interpretable, low complexity
Baseline CNN	2.86M	6 to 7	Deep FC layer, heavier than expected
CNN + GLCM Hybrid	0.512M	~ 6	Fusion model, moderate complexity and high accuracy
ResNet-50 (tuned)	2.1M	20 to 25	Deep pre-trained, partial fine-tuning, high compute

approx. – approximately, CNN – convolutional neural network, FC – fully connected, GLCM – grey-level co-occurrence matrix, MLP – multilayer perceptron, ResNet-50 – 50-layer residual network



**Figure 13.** Gradient-weighted class activation mapping (Grad-CAM) visualizations for test samples using the hybrid convolutional neural network + grey-level co-occurrence matrix model (pre-processed input)

ies [20,21]. In more than 85% of test cases, the model's highlighted areas matched these clinical markers, indicating that feature localisation was consistent with clinical practice.

The suggested approach provides multi-level interpretability, connecting visual saliency with quantitative feature importance. This guarantees that diagnostic determinations can be linked to distinct drawing segments and quantifiable statistical indicators, thereby enhancing the trust, transparency, and clinical repeatability of the suggested hybrid approach.

### Statistical validation

To ensure model performance reliability and stability, 5-fold cross-validation was performed on all experimental setups for each model architecture. The classification metrics – accuracy, precision, recall, and F1-score – are measured for each fold and calculated as the average performance per fold. This repeated testing permitted us to examine the reliability of model behaviour and compare the impact of pre-processing more rigorously. Two statistical tests, paired *t*-tests and Wilcoxon signed-rank tests, were conducted to compare the original and pre-processed versions of each model in terms of fold-wise performance scores. The tests help determine whether the differences we

observed in performance were statistically significant and not due to chance. As presented in Table 5, the *p*-values obtained in the two tests were always less than 0.01 for all measures and all pairs of models being compared [34,35]. It shows statistically significant improvement with pre-processed inputs (greyscale + normalized Canny edges), irrespective of model architecture. Among the many models tested, the CNN + GLCM hybrid model with pre-processed input demonstrated the most consistent and robust gains ( $p < 0.001$  on all scales), validating the strength of the presented pre-processing and multi-feature fusion strategy.

### Comparison with state-of-the-art

In order to validate the significance of the proposed hybrid CNN + GLCM model, several recent state-of-the-art studies [17-21,23,24], and [25] have been compared to detect PD with hand-drawn images, speech, or sensor-based information. The comparison includes not only comparison of the proposed model with baseline ML models – SVM, random forest [19] – but also important practical aspects, such as the application of data augmentation, the inclusion of handcrafted features, and model explainability using techniques such as Grad-CAM.

Studies have investigated hybrid deep learning methodologies for PD and other medical diagnostics, predomi-

Table 5. Paired statistical test results

Model	Metric	Paired <i>t</i> -test <i>p</i> -value	Wilcoxon <i>p</i> -value
GLCM only vs GLCM only (pre-processed)	Accuracy	0	0.002
GLCM only vs GLCM only (pre-processed)	Precision	0	0.002
GLCM only vs GLCM only (pre-processed)	Recall	0	0.002
GLCM only vs GLCM only (pre-processed)	F1-score	0	0.002
CNN only vs CNN only (pre-processed)	Accuracy	0	0.002
CNN only vs CNN only (pre-processed)	Precision	0	0.002
CNN only vs CNN only (pre-processed)	Recall	0	0.002
CNN only vs CNN only (pre-processed)	F1-score	0	0.002
CNN + GLCM vs CNN + GLCM (pre-processed)	Accuracy	0	0.002
CNN + GLCM vs CNN + GLCM (pre-processed)	Precision	0	0.002
CNN + GLCM vs CNN + GLCM (pre-processed)	Recall	0	0.002
CNN + GLCM vs CNN + GLCM (pre-processed)	F1-score	0	0.002
ResNet-50 vs ResNet-50 (pre-processed)	Accuracy	0	0.002
ResNet-50 vs ResNet-50 (pre-processed)	Precision	0	0.002
ResNet-50 vs ResNet-50 (pre-processed)	Recall	0	0.002
ResNet-50 vs ResNet-50 (pre-processed)	F1-score	0	0.002

CNN – convolutional neural network, GLCM – grey-level co-occurrence matrix, ResNet-50 – 50-layer residual network

nantly employing CNN–RNN or transformer-based fusion networks to capture sequential dependencies or long-range contextual patterns in data, including speech, gait, or EEG signals [18,19,23,24]. These architectures work well for temporal data, but they are often computationally demanding, require extensive parameter tuning, and are difficult to understand when applied to small clinical datasets.

Conversely, the present study proposes a two-channel hybrid CNN + GLCM framework, designed for hand-drawn spiral and waveform images. It emphasizes spatial and textural attributes as diagnostic factors, rather than temporal dynamics. The model uniquely combines CNN-based deep spatial features with GLCM statistical texture descriptors. These encode tremor-related micro-variations, line discontinuities, and contrast irregularities that reflect motor dysfunction. The hybrid model remains lightweight, interpretable, and fast to compute, unlike transformer-based techniques [24,25]. It does not require attention-based encoders or recurrent feedback to improve performance.

Additionally, the method offers clinically interpretable predictions by using Grad-CAM visualisation with GLCM-assisted statistical fusion. This is not possible with traditional deep fusion pipelines [18,19]. This approach introduces a novel idea in the field. It makes PD screening easier to understand, less resource-intensive, and non-sequential. It connects deep feature extraction with clinical transparency.

Due to the use of two independent datasets, the present study has a unique strength. For model training and

cross-validation, the most extensive dataset was used, while the smaller dataset was used exclusively for testing. As a result of this strategy, rarely observed in prior studies, the model's generalization capability is enhanced, demonstrating its practical applicability to unseen data across a range of drawing styles and noise levels.

Table 6 summarizes the proposed model's accuracy at 97.02%, which exceeds most previous approaches that used only deep learning or handcrafted features. Interestingly, only a few studies integrated both types of features or provided explanations through visualizations. Utilizing deep spatial learning, statistical texture features, and 2-channel pre-processed inputs (greyscale + edge), along with Grad-CAM heat maps for transparent decision-making, the proposed model offers a balanced approach. By using non-invasive drawing tasks, it offers a robust and clinically viable alternative to invasive screening for PD.

Quantitatively, the hybrid CNN + GLCM achieved 97.02% accuracy, outperforming RF (95%), CNN-LSTM (93.5%), attention-based CNN (96.5%), and transformer (89.3%). This margin of improvement is particularly significant considering the proposed model's lightweight design and low computational cost, in contrast to transformer-based models that require high-end GPU resources.

In addition to raw accuracy, the proposed approach consistently achieved higher precision (0.945) and F1-score (0.954), underscoring its robustness in distinguishing subtle motor irregularities captured in hand-drawn images. Importantly, the combination of spatial-texture fusion and edge-enhanced pre-processing enabled stable cross-

**Table 6.** Comparison of proposed model with state-of-art models

Study No.	Best model used	Best accuracy (%)	Data augmentation	Handcrafted features used	Grad-CAM / explainability
[17]	CNN on spiral drawings	93.5	Yes	No	No
[18]	3D CNN	88.9	Yes	No	No
[19]	ML ensemble (speech data)	95	Yes (CV)	Yes	No
[20]	CosineDCNN + SCGA (IoT)	89.98	No	No	No
[21]	MRgFUS drawings + tremor	—	No	Yes	No
[23]	CTSO-ZFNet (spiral + voice)	89.69	Yes	Yes	No
[24]	Attention CNN	96.5	Yes	No	Partial (attention maps)
[25]	CC-Net on static spirals	89.3	No	No	No
Proposed work	CNN + GLCM on pre-processed 2-channel input	97.02	Yes	Yes (GLCM)	Yes (Grad-CAM)

CC-Net – continuous convolution network, CNN – convolutional neural network, CTSO-IoT – Chronological Tangent Search Optimization–Internet of things, CTSO-ZFNet – Chronological Tangent Search Optimization–Zeiler and Fergus Network, GLCM – grey-level co-occurrence matrix, Grad-CAM – gradient-weighted class activation mapping, ML – machine learning, MRgFUS – magnetic resonance-guided focused ultrasound, SCGA – sine cosine geese migration optimization

dataset generalization, as validated on the independent Dataset 2.

## Conclusions

In this study, hand-drawn spiral and waveform images were used for the early detection of PD using a hybrid deep learning framework. Using the GLCM, the approach combines normalized greyscale features with edge-enhanced Canny features and further integrates hand-crafted statistical texture features. Both original and pre-processed datasets were used to test a range of models, including a baseline CNN, a hybrid CNN + GLCM, and a fine-tuned ResNet-50. In terms of classification capability and interpretability, the hybrid CNN + GLCM architecture trained on pre-processed input achieved the best performance with 97.02% accuracy. Our model was robust and generalized across diverse handwriting patterns thanks to the use of two independent datasets for training, validation, and external testing. Statistical validation using 5-fold cross-validation, paired *t*-tests, and Wilcoxon signed-rank tests confirmed that the performance gains with pre-processing and feature fusion were statistically significant ( $p < 0.01$ ).

It is still necessary to explore several future directions for real-world PD screening to follow the promising re-

sults of the proposed framework. It would be beneficial to incorporate temporal drawing data (e.g., pressure variations and speed variations) from digital tablets to enrich the input features further. Additionally, extending this approach to multimodal datasets, including voice, gait, or sensor data, may enhance diagnostic precision. Apart from this, testing across larger and more diverse clinical populations is essential to validate the clinical utility and scalability of the proposed model in practical healthcare settings.

## Disclosures

1. Institutional review board statement: Not applicable.
2. Assistance with the article: None.
3. Financial support and sponsorship: None.
4. Conflicts of interest: None
5. Data availability:

Dataset 1 <https://www.kaggle.com/datasets/abdukhalek-mugahed/augmented-hand-drawn-data-for-parkinsons-disease/versions/1>,

Dataset 2 <https://www.kaggle.com/datasets/kmader/parkinsons-drawings/data>.

## References

1. Zeng R, Wang J, Zheng C, Jiang R, Tong S, Wu H, et al. Lack of causal associations of inflammatory bowel disease with Parkinson's disease and other neurodegenerative disorders. *Mov Disord* 2023; 38: 1082-1088.
2. Essa MM. Handbook of neurodegenerative disorders. Singapore: Springer; 2024.
3. Koeglsperger T, Rumpf SL, Schließer P, Struebing FL, Brendel M, Levin J, et al. Neuropathology of incidental Lewy body & prodromal Parkinson's disease. *Mol Neurodegener* 2023; 18: 32. DOI: 10.1186/s13024-023-00622-7.
4. Li HX, Zhang C, Zhang K, Liu YZ, Peng XX, Zong Q. Inflammatory bowel disease and risk of Parkinson's disease: evidence from a meta-

- analysis of 14 studies involving more than 13.4 million individuals. *Front Med* 2023; 10: 1137366. DOI: 10.3389/fmed.2023.1137366.
5. Al-Jabbar M, Alshahrani M, Senan EM, Abunadi I, Almalki SA, Alshari EA. Hybrid techniques of multi-CNN and ensemble learning to analyze handwritten spiral and wave drawing for diagnosing Parkinson's disease. *Comput Model Eng Sci* 2025; 143: 2429-2457.
  6. Aouraghe I, Khaissidi G, Mrabti M. A literature review of online handwriting analysis to detect Parkinson's disease at an early stage. *Multimed Tools Appl* 2023; 82: 11923-11948.
  7. Thakur M, Dhanalakshmi S, Kuresan H, Senthil R, Narayanamoorthi R, Lai KW. Automated restricted Boltzmann machine classifier for early diagnosis of Parkinson's disease using digitized spiral drawings. *J Ambient Intell Humaniz Comput* 2023; 14: 175-189.
  8. Kwon SH, Park JK, Koh YH. A systematic review and meta-analysis on the effect of virtual reality-based rehabilitation for people with Parkinson's disease. *J Neuroeng Rehabil* 2023; 20: 94. DOI: 10.1186/s12984-023-01219-3.
  9. Gupta R, Kumari S, Senapati A, Ambasta RK, Kumar P. New era of artificial intelligence and machine learning-based detection, diagnosis, and therapeutics in Parkinson's disease. *Ageing Res Rev* 2023; 90: 102013. DOI: 10.1016/j.arr.2023.102013.
  10. Alalayah KM, Senan EM, Atlam HF, Ahmed IA, Shatnawi HS. Automatic and early detection of Parkinson's disease by analyzing acoustic signals using classification algorithms based on recursive feature elimination method. *Diagnostics (Basel)* 2023; 13: 1924. DOI: 10.3390/diagnostics13111924.
  11. Wang Y, He N, Zhang C, Zhang Y, Wang C, Huang P, et al. An automatic interpretable deep learning pipeline for accurate Parkinson's disease diagnosis using quantitative susceptibility mapping and T1-weighted images. *Hum Brain Mapp* 2023; 44: 4426-4438.
  12. Ayaz Z, Naz S, Khan NH, Razzak I, Imran M. Automated methods for diagnosis of Parkinson's disease and predicting severity level. *Neural Comput Appl* 2023; 35: 14499-14534.
  13. Ma YW, Chen JL, Chen YJ, Lai YH. Explainable deep learning architecture for early diagnosis of Parkinson's disease. *Soft Comput* 2023; 27: 2729-2738.
  14. Erdaş ÇB, Sümer E. A fully automated approach involving neuroimaging and deep learning for Parkinson's disease detection and severity prediction. *PeerJ Comput Sci* 2023; 9: e1485. DOI: 10.7717/peerj-cs.1485.
  15. Varalakshmi P, Priya BT, Rithiga BA, Bhuvaneaswari R, Sundar RS. Diagnosis of Parkinson's disease from hand drawing utilizing hybrid models. *Parkinsonism Relat Disord* 2022; 105: 24-31.
  16. SasiRekha S, Shankar R, Duraisamy S. Parkinson's disease detection using advanced wave signal processing on spiral drawing tests. In: Rajagopal S, Popat K, Meva D, Bajaja S, Mudholkar P, editors. *Artificial Intelligence Based Smart and Secured Applications*. Cham: Springer; 2025. DOI: 10.1007/978-3-031-86290-8\_3.
  17. Alissa M, Lones MA, Cosgrove J, Alty JE, Jamieson S, Smith SL, Vallejo M. Parkinson's disease diagnosis using convolutional neural networks and figure-copying tasks. *Neural Comput Appl* 2022; 34: 1433-1453.
  18. Vyas T, Yadav R, Solanki C, Darji R, Desai S, Tanwar S. Deep learning-based scheme to diagnose Parkinson's disease. *Expert Syst* 2022; 39: e12739. DOI: 10.1111/exsy.12739.
  19. Prasanna ML, Kumar C, Mishra RK. Enhanced Parkinson's diagnosis through ensemble learning with stacking and cross-validation. *Procedia Comput Sci* 2025; 259: 250-259.
  20. Pragadeeswaran S, Kannimuthu S. Cosine deep convolutional neural network for Parkinson's disease detection and severity level classification using hand drawing spiral image in IoT platform. *Biomed Signal Process Control* 2024; 94: 106220. DOI: 10.1016/j.bspc.2024.106220.
  21. Baek H, Lockwood D, Obusez E, Poturalski M, Chen J, Nagel SJ, Jones SE. Validation of tremor measurements using quantified drawing analysis in patients with essential tremor or Parkinson's disease treated with MR-guided focused ultrasound thalamotomy. *Biomed Signal Process Control* 2024; 92: 106085. DOI: 10.1016/j.bspc.2024.106085.
  22. Zhang Y, Lin H, Xie X, Peng P, Chen T, Zhao Z, et al. A novel approach for handwriting recognition in Parkinson's disease by combining flexible sensing with deep learning technologies. *Sens Actuators A Phys* 2025; 385: 116287. DOI: 10.1016/j.sna.2025.116287.
  23. Shanthini S, Chandrasekar A. Multimodal Parkinson's disease classification using voice signals and hand-drawn images based on deep learning enabled by chronological tangent search optimization. *Biomed Signal Process Control* 2025; 102: 107274. DOI: 10.1016/j.bspc.2024.107274.
  24. Jiang X, Yu H, Yang J, Liu X, Li Z. A new network structure for Parkinson's handwriting image recognition. *Med Eng Phys* 2025; 139: 104333. DOI: 10.1016/j.medengphy.2025.104333.
  25. Li Z, Yang J, Wang Y, Cai M, Liu X, Lu K. Early diagnosis of Parkinson's disease using continuous convolution network: handwriting recognition based on off-line hand drawing without template. *J Biomed Inform* 2022; 130: 104085. DOI: 10.1016/j.jbi.2022.104085.
  26. Swain KP, Samal SR, Ravi V, Nayak SR, Alahmadi TJ, Singh P, Diwakar M. Towards early intervention: detecting Parkinson's disease through voice analysis with machine learning. *Open Biomed Eng J* 2024; 18. DOI: 10.2174/0118741207294056240322075602.
  27. Singh MP, Singh J, Ravi V, Alahmadi TJ, Shivahare BD, Diwakar M, Singh P. A healthcare system employing lightweight CNN for disease prediction with artificial intelligence. *Open Public Health J* 2024; 17. DOI: 10.2174/0118749445302023240520111802.
  28. Augmented Hand-Drawn Data for Parkinson's Disease. 2025. Available at: <https://www.kaggle.com/datasets/abdulkhalekmugahed/augmented-hand-drawn-data-for-parkinsons-disease/versions/1> (accessed 20 July 2025).
  29. Parkinson's Drawings. 2025. Available at: <https://www.kaggle.com/datasets/kmader/parkinsons-drawings/data> (accessed 20 July 2025).
  30. Verma G. Leveraging smart image processing techniques for early detection of foot ulcers using a deep learning network. *Pol J Radiol* 2024; 89: e368-e377. DOI: 10.5114/pjr/189412.
  31. Shati A, Hassan GM, Datta A. A comprehensive fusion model for improved pneumonia prediction based on KNN-wavelet-GLCM and a residual network. *Intell Syst Appl* 2025; 26: 200492. DOI: 10.1016/j.iswa.2025.200492.
  32. Maithani A, Verma G. Hybrid model with improved score level fusion for heart disease classification. *Multimed Tools Appl* 2024; 83: 54951-54987.

33. Godasiaei SH, Ejohwomu OA, Zhong H, Booker D. Integrating experimental analysis and machine learning for enhancing energy efficiency and indoor air quality in educational buildings. *Build Environ* 2025; 276: 112874. DOI: 10.1016/j.buildenv.2025.112874.
34. Thakkar B. Continuous variable analyses: student's t-test, Mann-Whitney *U* test, Wilcoxon signed-rank test. In: Eltorai AEM, Bakal JA, Gibson CM, editors. Handbook for designing and conducting clinical and translational research. *Translational Cardiology*. Boston, MA: Academic Press; 2025: 165-167.
35. Verma G. Sustainable cost-energy aware load balancing in cloud environment using intelligent optimization. *Sustain Comput Inform Syst* 2025; 46: 101115. DOI: 10.1016/j.suscom.2025.101115.