Journal of Computer Science and Frontier Technologies
ISSN Print: 3104-4204
ISSN Online: 3104-4212
CD ISSN: 3104-4220
https://doi.org/10.63313/JCSFT.9052

HierFuse: Hierarchical Gated Fusion Network for Parkinson's Disease Screening Based on Hand Tremor Characteristics

LiHao Zhao
College of Computer Science and Technology, Qingdao University, Qingdao 266071, China
Email: zlh_qud@163.com

How to cite this paper: Zhao, L. H. (2026). HierFuse: Hierarchical Gated Fusion Network for Parkinson's Disease Screening Based on Hand Tremor Characteristics. *Journal of Computer Science and Frontier Technologies*, 2(3), 114-124. ISSN Print: 3104-4204; ISSN Online: 3104-4212.
https://doi.org/10.63313/JCSFT.9052
Published: 2026-03-18

Copyright © 2026 by author(s) and Erytis Publishing Limited.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/

Abstract

Parkinson's disease (PD) impairs motor control, making handwriting analysis a valuable non-invasive biomarker. This paper presents HierFuse, a hierarchical gated multimodal fusion network for PD screening on the NewHandPD dataset, which provides paired images and sensor signals from drawing tasks. HierFuse integrates temporal 1D and spatial 2D convolutional branches through cross-scale alignment, shared attentive compression, and progressive gated fusion with hidden propagation. Five-fold cross-validation shows state-of-the-art performance on Spiral (acc 0.912, F1 0.908) and Meander tasks (acc 0.929, F1 0.927), outperforming baselines. Ablations validate component contributions, with hierarchical fusion and hidden propagation most critical, while gating shows task-specific effects. Results demonstrate that hierarchical multimodal fusion effectively leverages handwriting dynamics for accurate PD detection.

Keywords

Multimodal; Convolutional Neural Network; Parkinson's Disease

1. Introduction

Parkinson's disease (PD) is a progressive neurodegenerative disorder primarily affecting the motor system, with hallmark symptoms including resting tremor, bradykinesia, rigidity, and postural instability[1]. Clinically, PD diagnosis relies predominantly on medical history and neurological examination based on established criteria such as the United Kingdom Parkinson's Disease Society Brain Bank criteria[2] or the Movement Disorder Society (MDS) clinical diagnostic criteria[3]. These criteria emphasize the presence of bradykinesia plus at least one of rigidity or resting tremor, along with supportive features and the absence of red flags. However, early-stage diagnosis remains challenging due to subtle and overlapping symptoms, resulting in misdiagnosis rates as high as 20--30% even

among specialists[4]. Objective, non-invasive biomarkers are therefore urgently needed to support early and accurate detection.

Handwriting and drawing tasks, particularly the drawing of Archimedes spirals and meander patterns, have emerged as promising non-invasive biomarkers for capturing PD-related motor impairments such as hand tremor, micrographia, and kinematic abnormalities[5]. The NewHandPD dataset provides a comprehensive resource for this purpose, consisting of static images of hand-drawn spirals and meanders together with synchronously recorded dynamic signals (e.g., pen pressure, tilt, azimuthal angle, and positional coordinates over time) from both healthy controls and PD patients[6,7].

While unimodal approaches leveraging either static images[8] or dynamic signals[9] alone have demonstrated encouraging performance through convolutional neural networks, recurrent models, or transfer learning, they often fail to fully exploit the complementary information between spatial-visual patterns in drawings and temporal-kinematic characteristics in handwriting dynamics. Multimodal fusion has shown superior potential in integrating these heterogeneous sources for improved diagnostic accuracy[10,11]. Nevertheless, most existing methods perform fusion at a single scale or late stage, limiting their ability to capture hierarchical and multi-resolution interactions between modalities.

In this paper, we propose HierFuse, a novel hierarchical gated multimodal fusion network tailored for PD detection on the NewHandPD dataset. By progressively integrating features from image and signal branches using gated multimodal units across multiple scales, HierFuse effectively combines static visual cues with dynamic tremor signals, yielding more robust and discriminative representations. Extensive experiments demonstrate that our approach achieves state-of-the-art performance on the NewHandPD dataset while offering enhanced interpretability through modality-specific gating.

2. Related Work

Parkinson's disease (PD) manifests motor symptoms including resting and action tremors, bradykinesia, and rigidity, which prominently affect fine motor control during handwriting and drawing tasks such as Archimedean spirals or waves. These tasks elicit visible distortions and kinematic irregularities, serving as non-invasive biomarkers for PD detection across datasets like HandPD[12], NewHandPD[6], PaHaW[5], and others.

2.1. Single-Modal Image-Based Approaches

Initial efforts employed hand-engineered features from static images of spirals and meanders, quantifying deviations from templates or tremor-induced irregularities via metrics like histograms of oriented gradients or radius variations[7,13]. Deep learning has since prevailed, with transfer-learned CNNs such as EfficientNetV2, VGG19, and ResNet achieving accuracies over 95% on spiral and wave images by

automatically capturing morphological distortions[14,15,16]. These image-based methods offer accessibility through simple scanning or photography, excelling in cases with pronounced visual impairments while struggling with subtle early-stage manifestations where drawings remain visually near-normal, often leading to reduced sensitivity, overfitting on limited datasets, and limited interpretability.

2.2. Single-Modal Signal-Based Approaches

Datasets providing on-line recordings, such as NewHandPD and PaHaW, enable analysis of temporal kinematics including pressure, velocity, and acceleration via sequence models like 1D-CNNs fused with BiGRUs[9]. Hybrid architectures incorporating recurrent units have demonstrated high sensitivity to tremor frequency and velocity fluctuations. Signal-based techniques capture dynamic irregularities imperceptible in static images, facilitating potential early detection and real-time monitoring, though they necessitate specialized digitizing hardware, exhibit vulnerability to recording noise, and face challenges in normalizing variable drawing speeds.

2.3. Multimodal Approaches

Integrating complementary static images and dynamic signals has yielded enhanced robustness, with frameworks employing dual-head networks, attention-based fusion, or Siamese architectures processing modalities in parallel[11,17,18]. Spatio-temporal fusions and spectrogram representations of signals combined with images have improved generalization across disease stages. While multimodal strategies leverage both morphological and kinematic cues for superior diagnostic reliability, prior implementations frequently rely on simplistic concatenation or late fusion, limiting adaptive integration of cross-modal interactions and increasing computational demands, particularly with scarce aligned datasets.

In contrast, the proposed hierarchical gated multimodal fusion network introduces cross-scale gated units to progressively and adaptively integrate features from image and signal branches, selectively amplifying complementary information across multiple abstraction levels while mitigating noise and redundancy; this design overcomes the constraints of existing fusions, promoting heightened sensitivity for early-stage detection, improved generalization, and greater interpretability relative to single-modal or conventional multimodal baselines.

3. Method

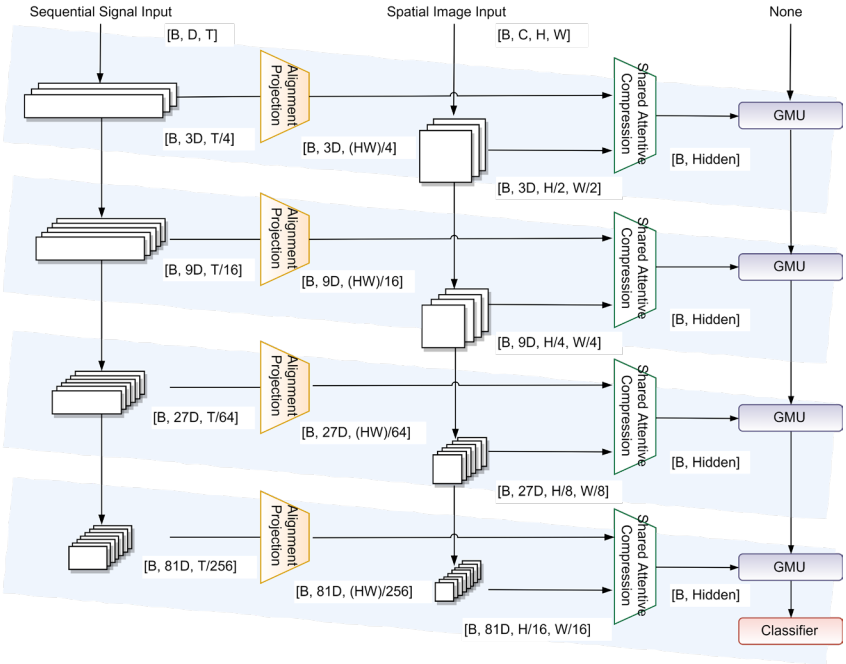

Figure 1. The architecture of the proposed HierFuse model. Dual branches employ 1D and 2D convolutions to extract hierarchical features from sensor signals and handwriting images, respectively. Corresponding features are aligned via projection and progressively fused using Gated Multimodal Units (GMUs) with layered attentive compression and prior hidden states. The final fused representation is fed to a classifier.

The overall architecture of proposed **HierFuse** model (as illustrated in Fig.1) consists of dual branches: a temporal branch employing 1D convolutions to extract hierarchical features from long sensor signal sequences, and a spatial branch using 2D residual convolutions to process handwriting images. Features from corresponding hierarchical levels are first aligned through linear projection to match spatial dimensions, enabling effective cross-modal compression. These aligned features are then compressed into a shared latent space using a shared attentive compressor, followed by progressive fusion via **Gated Multimodal Units (GMUs)** [19] that incorporate hidden states from previous levels. This hierarchical fusion mechanism allows the model to gradually integrate complementary information across scales while preserving modality-specific details. Finally, for prediction, the fully fused representation is passed to a classifier head consisting of linear projections, layer normalization, GELU activations, and dropout.

3.1. Temporal Branch

The temporal branch processes long sequential sensor signals using a stack of **TemporalCNN** modules to capture multi-scale temporal patterns.

Each TemporalCNN module employs a multi-branch 1D convolutional structure

with different kernel sizes (e.g., 3, 5, 7) to extract diverse temporal receptive fields:

$$\mathbf{f}_k = \mathrm{Conv1d}_k(\mathbf{x};\theta_k), \quad k \in \{3,5,7\} \tag{1}$$

where $\mathbf{x} \in \mathbb{R}^{B\times C_{\text{in}}\times T}$ is the input signal. The branch outputs are concatenated and combined with a residual connection from a max-pooled and 1×1 -convolved pathway:

$$\mathbf{o} = \mathrm{Concat}(\mathbf{f}_3,\mathbf{f}_5,\mathbf{f}_7) + \mathrm{Conv1d}_{1\times1}(\mathrm{MaxPool}(\mathbf{x})) \tag{2}$$

followed by layer normalization and GELU activation. Downsampling is achieved implicitly through strided convolutions (stride=4), progressively reducing the length of temporal feature while expanding the channel depth across hierarchical levels.

3.2. Spatial Branch

The spatial branch extracts hierarchical features from handwriting images using a series of SpatioCNN modules, each built upon a residual block[20].

A SpatioCNN module applies two 3×3 convolutions with batch normalization and GELU activation, optionally including a 1×1 shortcut convolution for channel and spatial dimension adjustment (stride=2 for downsampling):

$$\mathbf{y} = \mathrm{GELU}\Big(\mathbf{x} + \mathrm{Conv}_{3\times3}(\mathrm{GELU}(\mathrm{BatchNorm}(\mathrm{Conv}_{3\times3}(\mathbf{x}))))\Big) \tag{3}$$

This design, inspired by ResNet bottlenecks, efficiently captures spatial patterns in images while progressively downsampling from the input resolution (e.g., 224×224) to coarser feature maps.

3.3. Alignment Linear Projection

To enable cross-modal interaction at each hierarchical level, temporal features are reshaped and aligned with the corresponding spatial feature map dimensions via linear projection. For a temporal feature map $\mathbf{f}_t \in \mathbb{R}^{B\times C\times L}$, a linear layer projects the temporal dimension L to match the flattened spatial resolution $H\times W$ of the corresponding image feature map:

$$\mathbf{f}_{t'} = \mathrm{Linear}_{L\to HW}(\mathbf{f}_t^{\top})^{\top} \in \mathbb{R}^{B\times(HW)} \tag{4}$$

This alignment ensures dimensional compatibility prior to attentive compression, facilitating meaningful modality interaction.

3.4. Shared Attentive Compression

To facilitate effective cross-modal fusion, features from corresponding hierarchical levels are compressed into a unified latent space using the **Shared Attentive Compressor**.

For temporal features $\mathbf{f}_t \in \mathbb{R}^{B\times C\times L}$ and spatial features $\mathbf{f}_s \in \mathbb{R}^{B\times C\times H\times W}$, channel-wise attention is applied separately:

$$\mathbf{p}_t = \sum_{l=1}^{L} \text{softmax}(\mathbf{w}_t^\top \mathbf{f}_t[:,:,l]) \cdot \mathbf{f}_t[:,:,l] \tag{5}$$

$$\mathbf{p}_s = \sum_{n=1}^{HW} \text{softmax}(\mathbf{w}_s^\top \mathbf{f}_s[:,:,n]) \cdot \mathbf{f}_s[:,:,n] \tag{6}$$

where $\mathbf{w}_t, \mathbf{w}_s \in \square^{C}$ are learnable attention vectors. The pooled vectors $\mathbf{p}_t, \mathbf{p}_s \in \square^{B\times C}$ are then projected through a **shared** linear layer:

$$\mathbf{v}_t = \text{Proj}(\mathbf{p}_t), \quad \mathbf{v}_s = \text{Proj}(\mathbf{p}_s) \tag{7}$$

followed by layer normalization, dropout, and GELU activation. This parameter-sharing strategy ensures consistent representation spaces across modalities, promoting aligned and comparable compressed features for subsequent fusion.

4. Experiment

4.1. Dataset

This study utilizes the NewHandPD dataset[6], a publicly available handwritten dynamics dataset designed for Parkinson's disease (PD) detection. The data was collected at the Botucatu Medical School of São Paulo State University, Brazil, comprising 66 participants (35 healthy controls and 31 PD patients) [21].

Each participant completed 12 handwriting tasks using a BiSP smart pen, comprising four spirals, four meanders, two circular movements (on paper and in air), and two diadochokinesia tasks. This yielded nine static images per subject (264 total), alongside dynamic signals from the pen's four sensors capturing six features: microphone, fingrip pressure, axial pressure, and tri-axial tilt and acceleration. NewHandPD offers a more balanced class distribution, supporting multimodal-based analyses for PD detection.

4.2. Implementation Details

We employ 5-fold cross-validation for model evaluation. Input images are resized to 224×224 pixels and normalized using ImageNet statistics (mean values of 0.485, 0.456, 0.406 and standard deviations of 0.229, 0.224, 0.225). During training, we apply color jitter (brightness, contrast, saturation, and hue perturbations) as data augmentation, while validation only involves resizing and normalization. For the signal branch, raw sensor data are first normalized per channel using z-score, then linearly interpolated to a fixed length of 10,000 samples, and finally scaled to [0, 1] via min-max normalization. The model is designed with a hidden dimension of 1024. It is trained for 30 epochs with a batch size of 8, using the Adam optimizer (learning rate of 0.001, weight decay of 0.01) and a step learning rate scheduler that decays the learning rate by a factor of 0.8 every 5 epochs. All experiments are conducted

with a fixed random seed of 800 to ensure reproducibility.

4.3. Evaluation Metrics

We evaluate model performance using standard metrics for binary classification (Parkinson's disease vs. healthy controls): accuracy, precision, recall, F1-score, and Cohen's kappa coefficient. All metrics are computed on the test set of each fold during five-fold cross-validation, and the final results are reported as the average across the five folds. Accuracy measures the overall proportion of correctly classified subjects. Precision quantifies the accuracy of positive predictions (identifying Parkinson's disease), while recall indicates the sensitivity in detecting actual positive cases (patients with Parkinson's). The F1-score provides a harmonic mean of precision and recall, offering a balanced assessment. Cohen's kappa evaluates the agreement between predictions and ground truth, accounting for the possibility of chance agreement.

4.3. Results on NewHandPD Dataset

We evaluate the proposed HierFuse model and baseline methods on the NewHandPD dataset for Parkinson's disease classification. For the baseline methods (SVM, DT, ResNet, and ConvNeXt), the input is constructed by preprocessing the image and signal data and concatenating them into a unified tensor (or feature vector) as a simple multimodal representation. All results are reported as the average of five-fold cross-validation on the test fold of each split. Table 1 summarizes the performance of different methods across three handwriting tasks: Spiral, Meander, and Circle.

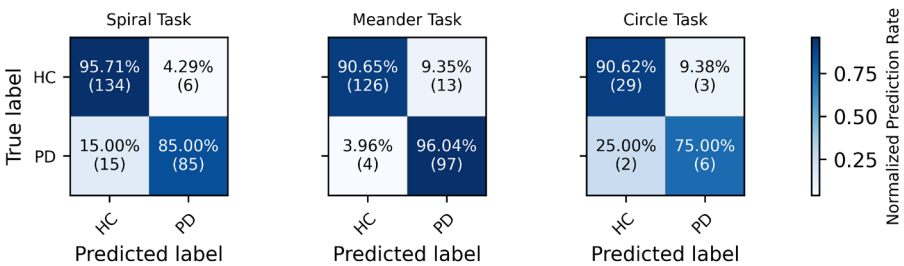

Figure 2. Confusion matrices on the NewHandPD dataset.

Table 1. Performance comparison of different methods on the NewHandPD dataset.

Handwriting Task	Method	Accuracy	Precision	Recall	F1	Kappa
Spiral Task	SVM	0.883	0.890	0.884	0.881	0.765
	DT	0.785	0.780	0.806	0.796	0.477
	ResNet [20]	0.717	0.730	0.718	0.712	0.434
	ConvNeXt [22]	0.904	0.903	**0.915**	0.904	0.767
	HierFuse (ours)	**0.912**	**0.917**	0.904	**0.908**	**0.817**
Meander Task	SVM	0.866	0.871	0.866	0.866	0.828
	DT	0.881	0.892	0.902	0.897	0.496
	ResNet [20]	0.637	0.649	0.663	0.629	0.279
	ConvNeXt [22]	0.901	0.903	0.914	0.901	0.777
	HierFuse (ours)	**0.929**	**0.929**	**0.931**	**0.927**	**0.855**
Circle Task	SVM	0.590	0.617	0.622	0.583	0.207
	DT	0.740	0.743	0.738	0.737	0.477
	ResNet [20]	0.761	0.806	0.766	0.754	0.524
	ConvNeXt [22]	0.825	0.829	0.823	0.822	**0.800**
	HierFuse (ours)	**0.875**	**0.833**	**0.882**	**0.843**	0.614

For the Spiral task, our HierFuse model achieves the highest scores on most metrics, with an accuracy of 0.912, precision of 0.917, F1-score of 0.908, and Cohen's kappa of 0.817, outperforming the best baseline ConvNeXt (accuracy 0.904, kappa 0.767). The recall of our method (0.904) is slightly lower than that of ConvNeXt (0.915), indicating a minor trade-off between precision and recall.

In the Meander task, HierFuse consistently surpasses all compared methods across every metric. It attains an accuracy of 0.929, precision of 0.929, recall of 0.931, F1-score of 0.927, and kappa of 0.855, demonstrating robust classification capability.

For the Circle task, HierFuse again yields the best results in accuracy (0.875), precision (0.833), recall (0.882), and F1-score (0.843). However, its kappa (0.614) is lower than that of ConvNeXt (0.800), suggesting that the agreement beyond chance is more favorable for ConvNeXt in this particular task.

Fig.2 presents the confusion matrices of the proposed HierFuse model on the three handwriting tasks. For the Spiral task, the model correctly classifies 95.71% of healthy controls (HC) and 85.00% of Parkinson's disease (PD) patients, with misclassifications mainly occurring as HC being mistaken for PD (15.00%) and PD being mistaken for HC (4.29%). In the Meander task, HC are identified with 90.65% accuracy, while PD classification shows a lower rate, with notable confusion between the two classes. For the Circle task, HC achieve 90.62% correct classification, and PD reach 75.00%, indicating a moderate level of misclassification where PD are more frequently misidentified as HC (29.00%) compared to the reverse. The model performs best on the Spiral and Meander tasks, with the Circle task exhibiting higher confusion between the two groups.

Overall, the proposed HierFuse model achieves superior or competitive performance on the NewHandPD dataset, especially in the Spiral and Meander tasks,

validating its effectiveness for Parkinson's disease classification based on handwriting dynamics.

4.3. Ablation Experiment

Table 2. The accuracy results of ablation experiments.

	Spiral	*Meander*	*Circle*	Mixed
w/o_image_branch	0.854	0.838	0.825	0.836
w/o_signal_branch	0.842	0.871	0.825	0.864
w/o_hidden_propagation	0.892	0.912	0.900	0.900
w/o_hierarchical_fusion	0.871	0.846	0.850	0.848
w/o_gating	0.896	0.908	**0.975**	0.898
HierFuse (full model)	**0.912**	**0.929**	0.875	**0.911**

We conduct ablation experiments to investigate the contribution of each key component in the proposed HierFuse model. The variants include removing the image branch (w/o_image_branch), removing the signal branch (w/o_signal_branch), removing hidden state propagation (w/o_hidden_propagation), removing hierarchical fusion (w/o_hierarchical_fusion), and replacing the gating mechanism with element-wise addition (w/o_gating). Performance is evaluated on each individual handwriting task (Spiral, Meander, Circle) as well as a mixed task where samples from all three tasks are combined for classification.

As shown in Table 2, the full model achieves the highest accuracy on Spiral (0.912), Meander (0.929), and the mixed task (0.911). On the Circle task, however, the full model (0.875) is outperformed by the w/o_gating variant (0.975), suggesting that simple addition may be more effective than the gated fusion for this specific task, possibly due to different data characteristics. Among the components, removing hidden propagation and hierarchical fusion consistently degrades performance across most settings, highlighting their importance for multimodal integration. The removal of either modality branch also leads to notable drops, confirming the complementary nature of image and signal data. Overall, the ablation results validate the effectiveness of the proposed architectural choices, while also revealing task-specific variations in component contributions.

5. Conclusion

We proposed HierFuse, a hierarchical gated multimodal network that progressively integrates image and signal features for PD detection on NewHandPD. The model achieves superior performance on Spiral and Meander tasks, with ablation studies confirming the importance of hierarchical fusion and hidden propagation. The gating mechanism shows task-specific effects, with simple addition proving more effective for the Circle task. Overall, HierFuse effectively leverages multimodal handwriting data for PD screening, and future work may explore adaptive fusion strategies tailored to task characteristics.

Acknowledgements

This work was not supported by any external grants. The author thanks the maintainers of the publicly available datasets used in this study.

References

[1] Jankovic J. Parkinson's disease: clinical features and diagnosis[J]. Journal of neurology, neurosurgery & psychiatry, 2008, 79(4): 368-376.

[2] Hughes A J, Daniel S E, Kilford L, et al. Accuracy of clinical diagnosis of idiopathic Parkinson's disease: a clinico-pathological study of 100 cases[J]. Journal of neurology, neurosurgery & psychiatry, 1992, 55(3): 181-184.

[3] Li J, Jin M, Wang L, et al. MDS clinical diagnostic criteria for Parkinson's disease in China[J]. Journal of neurology, 2017, 264(3): 476-481.

[4] Schrag A, Ben-Shlomo Y B, Quinn N. How valid is the clinical diagnosis of Parkinson's disease in the community?[J]. Journal of Neurology, Neurosurgery & Psychiatry, 2002, 73(5): 529-534.

[5] Drotár P, Mekyska J, Rektorová I, et al. Evaluation of handwriting kinematics and pressure for differential diagnosis of Parkinson's disease[J]. Artificial intelligence in Medicine, 2016, 67: 39-46.

[6] Pereira C R, Weber S A T, Hook C, et al. Deep learning-aided Parkinson's disease diagnosis from handwritten dynamics[C]//2016 29th SIBGRAPI conference on graphics, patterns and images (SIBGRAPI). Ieee, 2016: 340-346.

[7] Pereira C R, Pereira D R, Rosa G H, et al. Handwritten dynamics assessment through convolutional neural networks: An application to Parkinson's disease identification[J]. Artificial intelligence in medicine, 2018, 87: 67-77.

[8] San Luciano M, Wang C, Ortega R A, et al. Digitized spiral drawing: A possible biomarker for early Parkinson's disease[J]. PloS one, 2016, 11(10): e0162799.

[9] Diaz M, Moetesum M, Siddiqi I, et al. Sequence-based dynamic handwriting analysis for Parkinson's disease detection with one-dimensional convolutions and BiGRUs[J]. Expert Systems with Applications, 2021, 168: 114405.

[10] Skaramagkas V, Pentari A, Kefalopoulou Z, et al. Multi-modal deep learning diagnosis of Parkinson's disease—a systematic review[J]. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2023, 31: 2399-2423.

[11] Zhao A, Wu H, Chen M, et al. A Spatio-Temporal Siamese Neural Network for Multimodal Handwriting Abnormality Screening of Parkinson's Disease[J]. International Journal of Intelligent Systems, 2023, 2023(1): 9921809.

[12] Pereira C R, Pereira D R, Silva F A, et al. A new computer vision-based approach to aid the diagnosis of Parkinson's disease[J]. Computer methods and programs in biomedicine, 2016, 136: 79-88.

[13] Folador J P, Santos M C S, Luiz L M D, et al. On the use of histograms of oriented gradients for tremor detection from sinusoidal and spiral handwritten drawings of people with Parkinson's disease[J]. Medical & biological engineering & computing, 2021, 59(1): 195-214.

[14] Huang Y, Chaturvedi K, Nayan A A, et al. Early Parkinson's disease diagnosis through hand-drawn spiral and wave analysis using deep learning techniques[J]. Information, 2024, 15(4): 220.

[15] Aldhyani T H H, Al-Nefaie A H, Koundal D. Modeling and diagnosis Parkinson disease by using hand drawing: deep learning model[J]. AIMS Mathematics, 2024, 9(3): 6850-6877.

[16] Farhah N. Utilizing deep learning models in an intelligent spiral drawing classification system for Parkinson's disease classification[J]. Frontiers in Medicine, 2024, 11: 1453743.

[17] Bi X, Liu J, Meng F, et al. Diagnosis of Parkinson's disease by Archimedes spiral dynamic

handwriting analysis based on Dual-head Multimodal Networks[J]. Biomedical Signal Processing and Control, 2026, 112: 108448.

[18] Sahin E, Akman Aydin E. A Multimodal Convolutional Neural Network Model for Parkinson's Disease Diagnosis Based on Fused Handwriting Dynamics Signals[J]. Journal of Medical Systems, 2025, 49(1): 132.

[19] Arevalo J, Solorio T, Montes-y-Gómez M, et al. Gated multimodal units for information fusion[J]. arXiv preprint arXiv:1702.01992, 2017.

[20] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[21] Zhao A, Liu Y, Yu X, et al. Artificial intelligence-enabled detection and assessment of Parkinson's disease using multimodal data: A survey[J]. Information Fusion, 2025, 121: 103175.

[22] Liu Z, Mao H, Wu C Y, et al. A convnet for the 2020s[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 11976-11986.