Received: 4 January 2023 | Revised: 10 April 2023 | Accepted: 19 April 2023

DOI: 10.1049/ipr2.12820

IET Image Processing

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH

A coordinate attention enhanced swin transformer for handwriting recognition of Parkinson's disease

Nana Wang[1] | Xuesen Niu[1] | Yiyang Yuan[2] | Yunze Sun[2] | Ran Li[2] | Guoliang You[2] | Aite Zhao[1]

[1]College of Computer Science and Technology, Qingdao University, Qingdao, Shandong, China

[2]Turing Innovation Team, Qingdao University, Qingdao, Shandong, China

Correspondence
Aite Zhao, College of Computer Science and Technology, Qingdao University, Qingdao, Shandong, China.
Email: zhaoaite@qdu.edu.cn

Funding information
Natural Science Foundation of Shandong Province, Grant/Award Number: ZR2021QF084; China Postdoctoral Science Foundation, Grant/Award Number: 2022M711741; National Natural Science Foundation of China, Grant/Award Number: 62106117

Abstract

Diagnosing Parkinson's disease (PD) in its early stages is a significant challenge in medicine. Hand tremors and dysgraphia, which are typical early motor symptoms of PD, can manifest for decades before a formal diagnosis is made. Therefore, handwriting analysis has become an important tool for detecting PD. While many machine learning algorithms have been applied in this area, they struggle to capture the subtle changes in handwriting and must describe features from various perspectives. To address these issues, this paper proposes a Coordinate Attention Enhanced Swin Transformer (CAS Transformer) model for PD handwriting recognition. It establishes the long-term dependence of features on the joint coordinate attention application, which enables the model to more accurately localize the important features of handwriting data and also extract the fuzzy edge features of handwriting images.These characteristics of the CAS Transformer enable it to outperform current advanced deep learning methods in classification, with an accuracy of 92.68% in experiments conducted on two handwritten datasets.

1 | INTRODUCTION

Parkinson's Disease (PD) [1] is the second most prevalent chronic neurodegenerative disease in the world, and it mainly affects the motor nervous system, especially in countries with aging populations such as the European Union and the United States [2]. According to previous studies, the number of people with PD has been continuously increasing worldwide over the past decades. This rapid growth trend is expected to persist over the next few years [3], and PD will become the fastest-growing major neurological disorder. However, there is currently no cure for Parkinson's disease in the world, so early diagnosis of PD is critical to the prospects for drug therapy and to assessing the effectiveness of new drug treatments in the early stages.

To facilitate early diagnosis of Parkinson's disease, the scientific community has proposed four categories of biomarkers: imaging, clinical, genetic, and biochemical [4, 5]. While each biomarker has a different predictive value, accurately assessing their predictive and practical utility requires large-scale clinical studies. The most obvious symptoms of PD in its early stages are slow movement and tremors, limb stiffness, reduced motor function, and abnormal gait [6, 7]. While neuropathological studies have validated that several non-motor symptoms (such as mood, olfactory, sleep, psychiatric behavior, and gastrointestinal) may occur before the onset of motor symptoms and persist throughout the disease, these non-motor symptoms often lack specificity, thus making early identification of Parkinson's disease a challenging task [8]. One of the most obvious and common onset symptoms of PD is tremor at rest, and handedness is the most obvious area of tremor in patients, and assessment of handedness through handwriting is also the easiest and most feasible way for early symptom detection.

Handwriting has been shown in numerous studies to be an effective tool for early PD diagnosis [9–11]. Handwriting is a motor activity that is significantly impacted by Parkinson's disease, and its deterioration may be the first observable sign of the disease. Handwriting is also a complex and highly skilled coordinated activity that requires dynamic interactions between the muscles of the lower arm, wrist, and fingers. In healthy subjects, handwriting movements are automatically coherent and require no other attentional resources, while in patients with Parkinson's disease, handwriting typically exhibits reduced letter size, distorted writing, and jittering of handwriting lines. Therefore, handwriting changes can be a valid biometric feature for assessing the early diagnosis of Parkinson's disease.

Machine learning provides various methods for image classification, which have also been applied to the field of PD diagnosis and have provided great help [12–14]. Mainstream classification methods include support vector machines (SVM) [15], random forests [16], and decision trees [17]. However, none of these methods consider the temporal and spatial variability of handwriting. Long short-term memory (LSTM) and gated recurrent units (GRU) can successfully capture these small variations in handwriting and can better learn the data in handwriting.

In summary, this paper makes the following key contributions:

- For data augmentation, we use CycleGAN [18] to expand the dataset to generate new handwriting images from the handwriting images of healthy subjects and PD patients, further solving the problem of small sample size in the PD handwriting dataset.
- For effective information enhancement, we design a coordinate attention enhanced swin transformer (CAS Transformer) to embed location information into channel attention by coordinate attention [19] to establish long-range dependencies, which in turn can accurately detect the features of blurred edges of handwriting.
- In the feature fusion module, we perform multi-feature fusion on the extracted containing edge position information and other handwriting features, and utilize an ensemble learning algorithm at the decision module to evaluate the fused handwriting features. Experimental results show that our proposed model results in improved overall classification performance and outperforms other models in the literature.

The overall structure of the paper is as follows. Section 2 provides an overview of the related research conducted in recent years on handwriting analysis for diagnosing Parkinson's disease. Section 3 describes the proposed feature fusion model. Section 4 provides the final evaluation results of the model. In Sections 5 and 6, we discuss the final results and future work, respectively.

2 | RELATED WORK

The clinical diagnosis of Parkinson's disease poses a challenge as a result of the absence of dependable biomarkers. Therefore, researchers have been committed to using handwriting as an accurate biomarker for PD detection. In recent years, machine learning and deep neural networks have become increasingly popular in various clinical applications [20], and the development of automated systems for PD detection based on speech, gait, and handwriting data has attracted extensive research interest [21].

Currently, the most widely used method for assessing Parkinson's disease is the use of the Unified Parkinson's Disease Rating Scale (UPDRS) [22]. Using this scale, attending physicians can diagnose Parkinson's disease without the need for specialized instruments. However, the diagnosis based on UPDRS has some limitations. That is the evaluation results of the scale mainly depend on the patient's description, the doctor's visual observation and clinical experience, which may be affected by the doctor's subjective consciousness. Numerous scholars have conducted extensive research on computer-aided technology to facilitate a more convenient and objective diagnosis of Parkinson's disease. In this paper, we will discuss three aspects, machine learning, deep learning, and research based on attention mechanisms.

2.1 | Traditional machine learning methods

Traditional machine learning methods have been evaluated to be effective for handwriting data from paper documents, photographs, and other devices [23]. Traditional machine learning methods can extract unique information from recorded signals, usually using special pens or image digitizing tablets.

Drotar et al. [24, 25] have done several works in PD detection. Their studies on handwriting analysis were all performed on the same dataset, PaHaW. The authors classified PD by selecting features using the Mann-Whitney U-test filter and the Relief algorithm, considering not only movement on the ground but also movement in the air, but both movements seem to carry non-redundant information. In a separate study, the authors fed the extracted kinematic features into an SVM classifier to distinguish between healthy and PD patients, obtaining an accuracy of 79.4% and 84%. The Relief algorithm [26] was used to extract handwritten features, and then the SVM model was adopted to identify PD patients, a 10-fold cross-validation method was finally verified for classification, with an average accuracy of 88.13%. In addition to calculating traditional kinematic and spatio-temporal handwriting parameters such as velocity, acceleration, jitter, stroke height, stroke width, and movement time, the same group used related quantifiers based on entropy, signal energy, and empirical modal decomposition. Using this approach, a prediction accuracy of 98% was achieved for the PaHaW dataset.

In the meantime, Pereira et al. developed the offline HandPD dataset during their research [27], and the authors extracted spatial, tremor, and average relative measurement-related features from the HandPD dataset and input the features into a plain Bayesian, SVM, and optimal path forest (OPF) classifier for handwriting data recognition, and the plain Bayesian classifier achieved an average accuracy of 78.9%.

However, these studies have only analyzed the differences in handwriting globally, ignoring some subtle variations in the handwriting data, such as the size of the handwriting and small variations in the lines. In this paper, we not only consider handwriting globally but also further consider the edge position information of handwriting. In addition, multi-scale feature fusion enables the capturing of subtle variations in handwriting, which can be utilized to differentiate between PD patients and healthy subjects based on handwriting differences.

2.2 | Deep learning methods

On the other hand, with the rapid development of deep learning algorithms in the field of image classification, the use of deep neural networks [28] to analyze handwriting data is gradually developing into a mainstream trend. The advent of deep neural networks has eliminated the need for feature extraction from original data. Instead, we can directly process time series data using neural networks or transform it into images.

Pereira et al. suggested using CNN to solve the problem of PD detection by recording finger grip, axial pressure, tilt, and acceleration captured by a special pen as time series and then converting them to images, achieving 80.19% of the results. Diaz et al. [29] and other researchers have proposed a one-dimensional convolutional and BiGRU classification model to assess the viability of utilizing handwriting information (e.g. velocity, acceleration, and pressure) to identify symptoms of Parkinson's disease. Handwriting features such as dynamic handwriting spectrum and spectra for automatic machine learning (AutoML) were further analyzed by Norazco-Flores et al. [30] Afonso et al. [31] applied recursive graphs to learn the signal data mapped to CNN, and they used the HandPD dataset for their experiments, which yielded an average accuracy of 87%. Although these dynamic features are effective, these methods still require the use of special devices (such as custom electronic pens and digital tablets), and special pens or tablets capable of acquiring several different handwriting styles must be used to obtain accurate results.

Due to the lack of sufficient handwriting sample data, neither deep learning methods nor traditional machine learning can yield accurate experimental results, and transfer learning can be fine-tuned for different source tasks to improve the performance of PD diagnosis.

Naseer [32] used fine-tuning and transfer learning to classify the dataset using the pre-training results of AlexNet and CNN with results as high as 98%. Moetesum et al. [33] improved the classification results by using fusion techniques, using the pre-trained AlexNet as a feature extractor for extracting handwriting, and in addition to acquiring features from the original image, they also used the median of the image features were extracted from the filtered residuals and fully exploited the edge information of the image for classification, achieving 83% accuracy on support vector machine (SVM) based. Kamran [21] proposed an end-to-end deep learning migration method to transfer pre-learned features to handwriting data of PD patients for early recognition of PD using different architectures of CNN. Gazda [34] trained weights and diagnoses offline handwriting Parkinson's disease by multiple fine-tuned convolutional neural networks.

2.3 | Attention mechanism-based methods

Attention mechanisms have been successfully shown to be useful for various computer vision tasks, such as image classification and image segmentation. Many researchers are now combining attention with various network architectures. Without changing the existing network structure, attention mechanisms can capture more useful features that we need for different classification tasks and needs, and assign higher feature weights to important information to improve the final classification performance. Yeqi Fei et al. [35] proposed the Opti-SA model to classify flowers by invoking the attention mechanism module ECAnet and incorporating it into the shufflenet network, which improves the number of parameters, accuracy, and classification speed compared with the network without attention. Park et al. [36] proposed a simple and effective bottleneck attention module (BAM) that can be used to infer attention maps for spatial and channel paths. Zhang et al. [37] designed a stochastic unit to combine the two types of attention mechanisms, but inevitably increased the computational overhead of the algorithmic process. Meanwhile, some researchers [38] combined self-supervised learning, 3DBAM, and 3D stochastic attention mechanisms to enable the model to mitigate the impact of classification on irrelevant features. Qibin Hou [19] proposed a new mobile network attention mechanism, namely coordinate attention, which addresses the location information ignored by other attention and successfully embeds location information in channel attention.

3 | THE PROPOSED METHOD

In order to successfully identify PD patients and healthy subjects, we introduce a feature fusion network for data enhancement, feature extraction, and image classification. We first introduce CycleGAN for data enhancement. Secondly, we describe CAS Transformer and Vision Transformer (ViT) models that can extract features. Finally, we demonstrate the designed ensemble learning model to classify the fused handwriting features.

3.1 | The overall structure

Our proposed handwriting recognition algorithm for PD is shown in Figure 1. In this framework, handwriting data is first input to CycleGAN for preprocessing operations. Next, the generated images are fed into the feature extractor, and the coordinate attention enhanced swin transformer (CAS Transformer) is developed to extract the fuzzy edge information and the location coordinate information of the handwriting data, furthermore simultaneously extracting the location information corresponding to the segmented image. The extracted features are then multi-scale fused and the classification results are characterized by multiple classifiers in the decision layer.

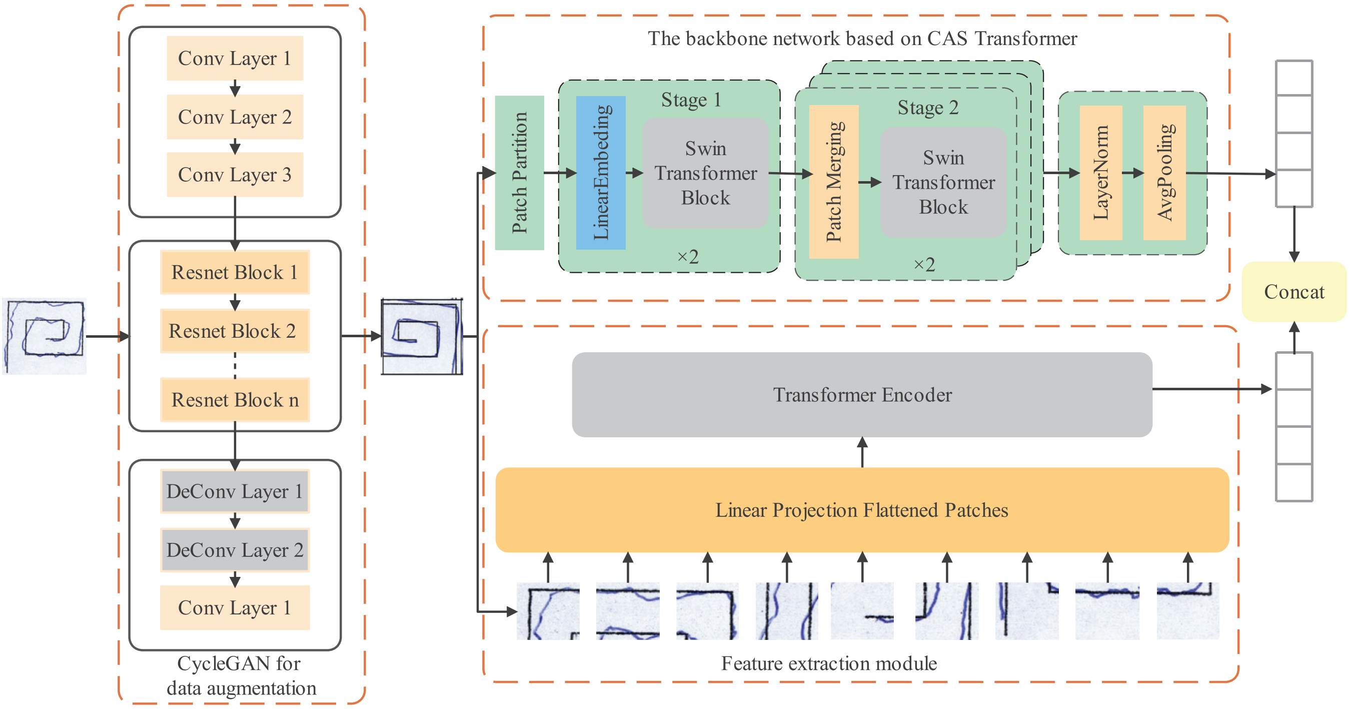


FIGURE 1 The overall architecture of PD detection. This model contains three modules: CycleGAN for data augmentation, a feature extraction module for handwriting feature extraction, and a classification decision module.

First, we formulate the handwriting recognition problem with the sequence of images defined as $X = \{x_i \in R^{W*H}, i = 1, 2, 3, \dots, N\}$ corresponding to the sequence of binary labels, where N denotes the number of inputs of handwriting samples, H represents the height of the input image and W represents the width, while $W * H$ is the dimensionality of the training samples.

3.2 | Data preprocess

The absence of appropriate data presents a significant challenge in PD diagnosis. The main reason is that the images of healthy subjects are easier to collect, while handwriting data of PD patients are not only difficult to obtain, but also small and private. In order to overcome the problem of a small number of samples in the dataset, the most effective method is to expand the dataset. A significant amount of samples can lead to better training of the model and improve its generalization ability.

We first normalize the handwriting data image to remove the noise and enhance the contour emphasis to enhance the features of the handwriting data. Furthermore, to increase the diversity of the Parkinson's disease handwriting dataset, it was augmented. For building large datasets and augmenting the number of training samples, numerous versions of the GAN have been introduced. GAN can generate images similar to training sets, but it can only generate images randomly, and cannot control the appearance of output images, nor can it specify specific images, let alone perform image conversion and other operations. The pix2pix model can convert images, but the pix2pix model requires that data must be in pairs, and in reality, it is difficult to find images in pairs in two areas of life.

The emergence of CycleGAN solves these problems. The application of consistency in CycleGAN reduces the number of random mappings, thus ensuring that the images obtained are no longer random, and realizing the transformation of images put from one domain to another. At the same time, CycleGAN does not make any requirement on whether the data are paired or not, it can use unpaired data for training.

CycleGAN is a neural network that learns the data between two domains (F(x), G(x)) for the transformation function. To learn both F and G functions, CycleGAN uses two traditional GANs, each with a generator network and a discriminator within it. We first pass the handwriting image into the decoder and then symmetrize it along the edges, both above and below, to increase the image resolution. Then use the Residual block module to recover the data to enhance it, use deconvolution, IN regularization and ReLU activation function to recover the image size, and finally enhance the resolution of the image by ReflectionPad2d and recover it to the original size of the image by convolution to effectively resolve the object edge information. Figure 2 shows the image of the dataset after data preprocessing. The CycleGAN loss function is defined as the sum of the two GAN losses and the Cycle consistency loss, and the formula is as follows:

$$Loss = Loss_{GAN}(F, D_y) + Loss_{GAN}(G, D_x) + \lambda Loss_{Cycle}, \quad (1)$$

$$Loss_{Cycle} = E|F(G(y)) - y| + E|G(F(x)) - x|, \quad (2)$$

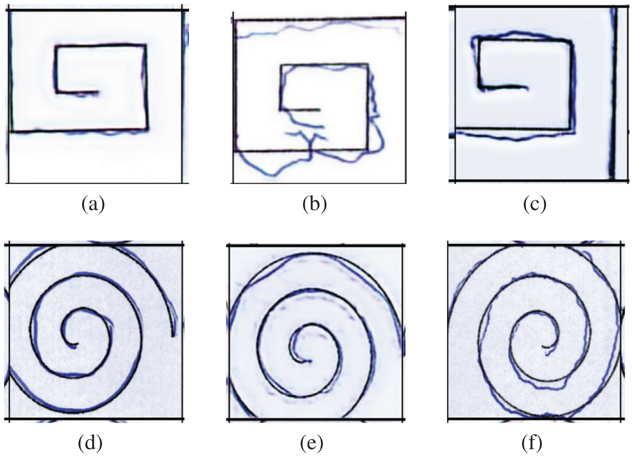


FIGURE 2 Image in the dataset after data augmentation.

$$Loss_{GAN} = E[log(1 - D_x(G(x)))] + E[log(D_x(x))], \quad (3)$$

$$Loss_{GAN} = E[log(1 - D_y(F(y)))] + E[log(D_y(y))], \quad (4)$$

where G and F are generators, D_x and D_y are discriminators, and the λ is weighting factor.

In the current work, we have considerably enhanced the diversity of PD handwriting samples by utilizing the CycleGAN method. Through our experimental results, it can be shown that the generalization ability of the model is enhanced and the classification results are significantly improved.

3.3 | CAS transformer

To obtain a richer feature map of handwriting images, as well as to accurately locate the edge information of handwriting data and establish the dependencies between remote information between features, this paper proposes the coordinate attention enhanced swin transformer (CAS Transformer) as the main backbone network, as depicted in Figure 3. Our network is based on the use of a transformer idea, hierarchical architecture, and moving window attention mechanism to establish the interconnections between features.

The CAS Transformer backbone network consists of four stacked stages and a patch partition module. Based on the layered architecture design, the backbone network introduces coordinate attention (CA) [19] in the patch partition module in order to extract the location information that is ignored in all other algorithms with almost no increase in parameters, which embeds the location information in the handwriting data into the channel attention to enhance the network's ability to extract features. And this method is employed to address the issue of information loss that may occur between patches. In this way, long-distance relationships in one direction can be captured while preserving spatial information in the other direction. The addition of the coordinate attention block increases the sensitivity of the model to information channels, and the global average pooling used in coordinate attention also helps CAS Transformer to capture the global information missing from the convolution.

CA, as the core module of patch partition, can capture relationships at longer distances by following horizontal or vertical directions, while also maintaining accurate and complementary location information, enhancing the extracted valid features, and establishing remote information relationships between individual features. The specific details of the patch partition are shown in Figure 4. First, we split the input H×W×3 handwriting image into non-overlapping equal-sized patches by a 4×4 convolutional layer, at which time the feature dimension is $H/4 \times W/4 \times 48$. The patch-splitting module is then extended to a patch-by-patch embedding by inserting a coordinate attention (CA) block. In the CA block, we do the following for the input H and W: (1) Averaging pooling operations on the input features $X_{in} \in R^{H/4 \times W/4 \times 48}$ in the width and height directions, respectively. Thus, the output of the c-th channel at the height h and the width w can be formulated, respectively, as

$$X_{in}^{h}(h) = \frac{1}{W} \sum_{0 \leq i < W} x_c(h, i), \quad (5)$$

$$X_{in}^{w}(w) = \frac{1}{H} \sum_{0 \leq j < H} x_c(j, w). \quad (6)$$

In this way, we can aggregate the input features X_{in} from both height and width directions and get the feature maps in both directions, $X_{in}^{w} \in R^{48 \times H/4 \times 1}$ and $X_{in}^{h} \in R^{48 \times W/4 \times 1}$. (2) To do the concat of spatial dimension for two multi-channel vectors, compress the number of channels by 1×1 convolution, and obtain the feature map $X_{mid} \in R^{48/r \times 1 \times (W/4 + H/4)}$, where r is the reduction ratio of the control channel. (3) Subsequently, spatial information in vertical and horizontal directions is encoded by BN and Non-linear. (4) Reshape the number of channels of the two directional eigenvectors by 1×1 convolution to transform x_h and x_w into tensor with the same number of channels as the input X_{in}, where x_h and x_w are two tensors in the spatial dimension, respectively. It is then output by the sigmoid function and finally weighted in both directions with the original input information.

After patch partition, the feature map shape undergoes a transformation from [H, W, 3] to [H/4, W/4, 48], and the feature map shape becomes [H/4, W/4, C] by spreading the feature map into a series of d-dimensional patch tokens from 48 to C through the linear embedding layer in stage 1. The first module of patch merging and feature transformation is represented as stage 2, and the resolution is kept as $H/4 \times W/4$. To generate a layered representation, the patch merging layer reduces the number of tokens as the network's depth increases. The output's dimensionality is set to 2C, and the process is repeated twice. Stage 3 and stage 4, respectively, and the output resolution is $H/16 \times W/16$ and $H/32 \times W/32$, respectively. These four stages together produce a layered representation, and the output has the same resolution as a typical convolutional network.

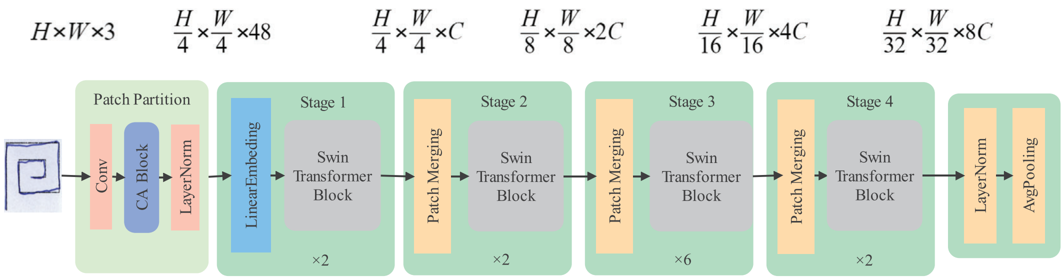


FIGURE 3 The network structure diagram of CAS Transformer.

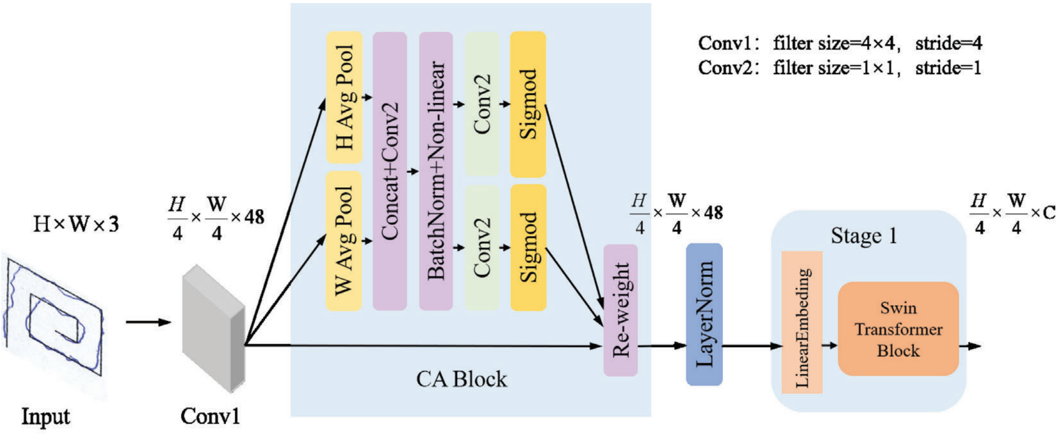

Conv1: filter size=4×4, stride=4
Conv2: filter size=1×1, stride=1
$H \times W \times 3$
$\frac{H}{4} \times \frac{W}{4} \times 48$
Input
Conv1
H Avg Pool
W Avg Pool
Concat+Conv2
BatchNorm+Non-linear
Conv2
Conv2
Sigmod
Sigmod
CA Block
Re-weight
$\frac{H}{4} \times \frac{W}{4} \times 48$
LayerNorm
Stage 1
LinearEmbeding
Swin Transformer Block
$\frac{H}{4} \times \frac{W}{4} \times C$

FIGURE 4 The structure of patch partition module.

3.4 | Vision transformer

Vision Transformer (ViT) replaces traditional convolutional computation with linear mapping in deep neural networks, allowing it to show remarkable performance in numerous image classification tasks, achieving state-of-the-art results. ViT divides an image into a series of non-overlapping patches and then learns the representation between the patches using the multi-headed self-attentiveness in the transformer. Although ViT requires a substantial amount of data and is computationally intensive, ViT is able to obtain global features at a shallow level due to the use of the full transformer structure, thus making the features obtained at the shallow and deep levels more similar to each other. In the shallow layer, it is similar to CNN in some head-attention parts with local windows, and in the deep layer the global windows are used in the head-attention parts, and with a large amount of data, it can learn high-quality intermediate features. In addition, the effect of jump connections in ViT is more significant than in CNNs (ResNet) and has a significant impact on the performance and similarity of features. Based on these points, this paper chooses ViT to extract handwriting data features for Parkinson's disease.

We first cut the image for detection into several pieces to increase the perceptual field of the model, and then convert each piece into a patch by a linear layer, and each patch is followed by a class token, and then embed location information for each token, and add the location information to the two-dimensional data of token by direct summation. The location information is a parameter with the same dimension as the token, and the embedded patches and class tokens are sent to the Transformer Encoder layer for encoding.

3.5 | Feature fusion

When it comes to image classification, it is common practice to fuse features of varying scales to enhance the model's accuracy. For multi-scale feature fusion, the operation of adding or concating is usually performed in the experiment. And different methods have different capabilities for feature extraction. In contrast to the convolution operation in CNN, which has a restricted perceptual field, the transformer's self-attention mechanism can perceive distant information and adaptively modify the perceptual field based on the image content. Therefore transformer is considered more flexible and powerful than CNN and has made more progress in the field of vision [39]. So we adopt the above two approaches containing a transformer module for feature extraction of handwriting.

In this paper, a series of features extracted about handwriting such as edge, position, contour, and handwriting are synthesized by multi-scale fusion to achieve the complementary effect of the advantages between different features and synthesize a feature that is more judgmental than the original input features, and the final experimental results also show that the accuracy is

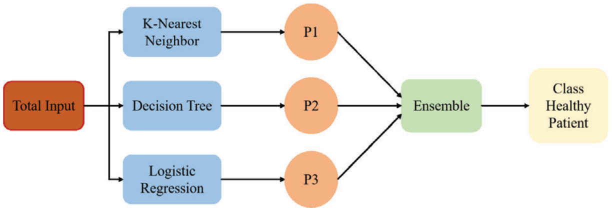


FIGURE 5 The training process of the detection module.

significantly improved when we perform multi-feature fusion and then input.

3.6 | Handwriting recognition

Ensemble learning obtains the final classification results by fusing multiple individual learners based on the predictions of each classifier on the data. The training process of ensemble learning is shown in Figure 5. The selection of individual classifiers is usually required to be weak and simple enough because the models learned by strong learners tend to be similar, and the process of training multiple weak classifiers is to increase the information differences between individuals while maintaining the relative accuracy of each classifier.

Decision trees and neural networks are usually chosen to be the main ones. The ensemble learning in this paper chooses to use LogisticRegression, Decision Tree, and K-Nearest Neighbor as classifiers, and after obtaining the predicted category probabilities of each classifier for the data, the obtained individual category probabilities are weighted and summed to obtain the final classification results. The model is described as follows:

$$H(x) = \frac{1}{T}\sum_{i=1}^{T} h_i(x), \tag{7}$$

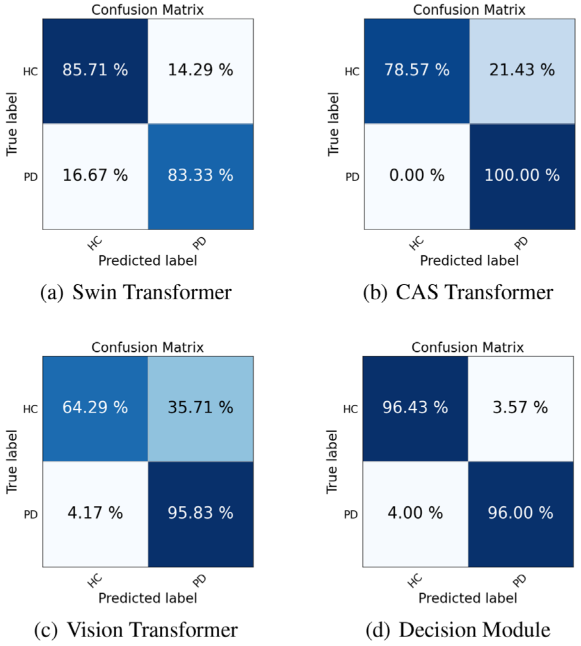
where $h_i(x)$ refers to the h_i classifier prediction sample x, and T refers to the number of classifiers in ensemble learning.

4 | EXPERIMENT

In this section, we present the experimental setup and two datasets, HandPD [27] and NewHandPD [33], and repeatedly train and test all sequences on these datasets to calculate their classification accuracy relative to the real labels to assess the overall performance of our model.

4.1 | Dataset introduction

During our experimental analysis, we performed several experiments to ascertain that our model was efficient for detecting PD. This study was conducted on two public datasets, HandPD and NewHandPD, the largest publicly available handwriting datasets and mainly used to detect Parkinson's disease for automatic system performance.



(a) Swin Transformer (b) CAS Transformer (c) Vision Transformer (d) Decision Module

FIGURE 6 Confusion matrix for the identification of patients with PD.

HandPD [40] is a considerably large dataset that comprises samples from 92 subjects (18/74 healthy subjects/PD) performing both spiral and meander tasks. It contains a total of 736 images, with 368 each for spiral and meander (72/296 healthy subjects/PD).

The NewHandPD dataset comprises data from 66 individuals, including 35 healthy and 31 non-healthy subjects individuals. Among the PD patients, 10 were female and 21 were male. Among the healthy subjects, 17 were female and 18 were male. The handwriting data were collected by the State University of São Paulo, Brazil, which asked the subjects to perform four tasks that are not very important for PD patients, namely drawing "spiral", "meander" and "circle" in a form with a special pen. Each subject was asked to perform the task 12 times, including 4 times for the spiral and meander task and 2 times for the drawing circle task.

4.2 | Experimental settings

In this work, we used a device configuration with a GeForce RTX 2080 graphics card, an Intel Xeon(R) W-2133 CPU, and 31.1 GiB RAM. For the two experimental handwriting datasets, we disrupted and divided the datasets into train and test sets in the ratio of 7:3. A randomly selected 70% of the data is used for training on the experimental network, and the remaining 30% is used for comparison and testing, for a total experimental data volume of 1464 to 6640. In the data preprocessing module, CycleGAN allowed us to normalize the input images and expand the data accordingly. The handwriting image data is input by the feature extraction module at 128*128*3, and the model is trained in CAS Transformer using convolutional kernels of 4*4 and 1*1 size, respectively, choosing a batch

TABLE 1 Performance comparison with state-of-the-art approaches on datasets.

Method	DataSet	Precision (%)	Recall (%)	F1 score (%)	Accuracy (%)
ShuffleNet	HandPD	82.19	83.51	81.74	80.82
	NewHandPD	78.02	80.67	79.63	78.85
EfficientNet	HandPD	88.88	87.47	83.10	87.02
	NewHandPD	86.93	87.65	84.02	86.53
MobileNet	HandPD	88.17	88.67	86.24	85.67
	NewHandPD	89.19	90.20	89.69	88.06
MobileNetV2	HandPD	86.19	89.31	89.55	89.06
	NewHandPD	90.26	91.48	90.14	90.89
MobileViT	HandPD	88.88	87.47	83.10	87.02
	NewHandPD	87.82	89.21	87.49	89.62
ConvNeXt	HandPD	86.79	70.58	74.97	86.30
	NewHandPD	90.30	91.47	90.35	90.38
Vision Transformer	HandPD	82.68	86.24	85.52	82.99
	NewHandPD	82.18	84.99	81.63	82.97
CAS Transformer	HandPD	**87.74**	**88.90**	**97.53**	**88.92**
	NewHandPD	**92.59**	**92.77**	**91.38**	**92.68**
Decision Module	HandPD	**95.30**	**96.28**	**95.42**	**95.45**
	NewHandPD	**97.54**	**98.37**	**97.74**	**97.02**

size of 128 and an initial learning rate of 0.0001. In the decision module, the maximum number of iterations of the weak learning classifier is 100. We evaluated the model using three metrics: accuracy(AC), specificity(SP), and sensitivity(SE). Their definitions are as follows:

$$AC = \frac{TN + TP}{FN + FP + TP + TN}, \tag{8}$$

$$SE = \frac{TP}{FN + TP}, \tag{9}$$

$$SP = \frac{TN}{TN + FP}. \tag{10}$$

In this context, TP represents the number of cases where PD patients were correctly identified, FP represents the number of cases where healthy subjects were incorrectly identified as PD patients, TN represents the number of cases where healthy subjects were correctly identified, and FN represents the number of cases where PD patients were incorrectly identified as healthy subjects.

4.3 | Comparison experiment

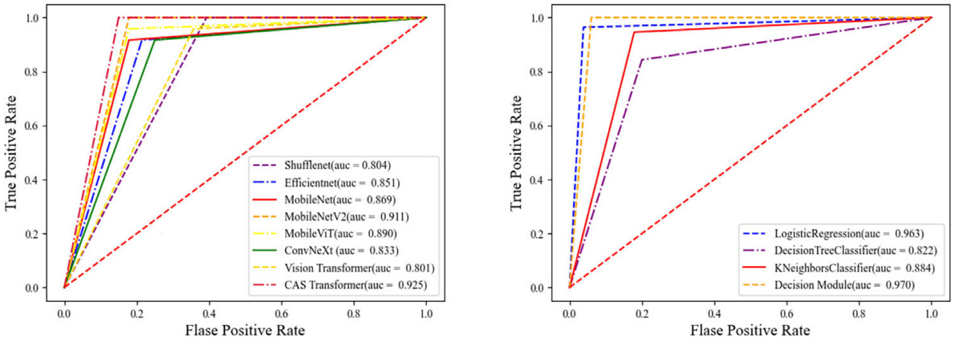
To evaluate the positive effect of our method on Parkinson's disease detection, we trained the Parkinson's disease detection model using the two enhanced datasets as input data for the model. After initializing with pre-trained weights, we evaluated the results of our experiments using several popular methods. The comparison methods that emerged in the experiments are as follows:

Several of the state-of-the-art methods compared in this paper are as follows:

- ShuffleNet [41] is a more efficient convolutional neural network for mobile.
- EfficientNet [42] is a fast and high-precision model that uses depth, width, and input image resolution co-adjustment techniques. It is based on architectural algorithms and includes SE attention.
- MobileNet [43] is a lightweight deep neural network built using depthwise separable convolution, which is an efficient model proposed for mobile and embedded devices.
- MobileNetV2 [44] is a network architecture with an inverse residual structure added to MobileNet's deeply separable convolution. In this paper, MobileNetV2 [19] with the addition of coordinate attention is used.
- MobileViT [45] combines the architecture of CNN and transformer, continuing the lightweight and efficient network and transformer's self-attention mechanism and global vision.
- ConvNeXt [46] is a pure convolutional neural network without transformer.
- ECA-Net [47] is a channel attention model for local cross-channel interaction without dimensionality reduction.
- CrossViT [48] is a new two-branch visual transformer based on cross-attention design.
- CoAtNets [49] naturally unify depthwise convolution and self-attention by simple relative attention.
- Twins-SVT [50] proposes spatially separable self-attention, where attention is first computed separately for local spatial dimensions, and then fusion of grouped attention is performed globally.

TABLE 2 Recognition accuracy of handwriting datasets with attention-mechanism-based methods.

DataSet	ResNet+ECA-Net	CrossViT	CoAtNets	Twins-SVT	CAS Transformer
HandPD	79.93	86.35	87.31	85.40	88.92
NewHandPD	84.14	90.88	92.71	84.7	92.68



(a) ROC curve and AUC for each class

(b) Fusion feature ROC curve and AUC in a multi-classifier

FIGURE 7 ROC curve of CAS Transformer and decision module on the dataset. The AUC value is utilized to assess both the stability and classification effectiveness of each class.

We report the classification results of currently popular deep learning algorithms on HandPD and NewHandPD in Table 1. The CAS Transformer model improves on the four evaluation metrics of F1 score, precision, accuracy and recall, and the other deep learning models on both datasets remain stable experimental results. Meanwhile, the ensemble learning also achieves optimal results on both handwriting datasets with an accuracy of over 95%, which indicates that the handwriting data features fused by the model can better reflect the differences between the handwriting of PD patients and that of healthy individuals. A comparison of the recognition accuracy of the attention mechanism-based methods is shown in Table 2. CAS Transformer achieves a large performance improvement compared to CrossViT, CoAtNets, and Twins-SVT and obtains similar results to the ResNet network with the addition of the ECA-Net module.

The confusion matrix of the classification model is shown in Figure 6. The confusion matrix is utilized to provide an overview of the PD recognition model's performance. The diagonal blue line of the matrix denotes accurate recognition, while the other entries signify erroneous recognition. It is clear from the confusion matrix that CAS Transformer and ensemble learning produced very few erroneous results for PD recognition, indicating that the model's detection of Parkinson's disease was significantly improved when the positional features of the writing data were added to the model and when the multiscale fusion of all features was performed.

We also used both ROC curves and AUC values to compare the performance of the currently more popular image classification techniques, and through Figure 7 we can see that CAS Transformer ranks high, and due to the Transformer, MobileViT, and CAS Transformer achieve similar AUC values.

TABLE 3 Comparison of recognition performance of models before and after dataset augmentation.

	Swin transformer			CAS transformer		
DataSet	AC(%)	SE(%)	SP(%)	AC(%)	SE(%)	SP(%)
HandPD	87.01	86.97	87.21	87.01	86.15	87.83
HandPD+X	**87.57**	**86.24**	**88.29**	**88.92**	**89.63**	**87.90**
NewHandPD	90.25	89.94	89.53	90.25	89.94	90.34
NewHandPD+X	**87.62**	**86.36**	**91.32**	**92.68**	**91.32**	**91.15**

4.4 | Ablation experiment

In this section, in order to demonstrate the effect of the CycleGAN-enhanced dataset and the coordinate attention in the CAS Transformer on the classification performance of the model, we conducted corresponding ablation experiments.

Table 3 summarizes the experimental results of different recognition models before and after augmentation with handwriting datasets, and X in the table represents the dataset after using CycleGAN. From Table 3, we can find that the pre-processed dataset leads to an increase in all metrics of the model, with CAS Transformer achieving 92.68% accuracy, 91.32% sensitivity, and 91.15% specificity on the tested augmented dataset.

To verify the effectiveness of CA in the CAS Transformer, we conducted several experiments on the handwritten dataset. From the experimental results in Table 4, we can see that the CAS Transformer with embedded CA block achieved 92.68% accuracy, a 5.06% improvement over the original

TABLE 4 Ablation experiments on embedded CA using CAS transformer architecture.

Method	DataSet	Precision (%)	Recall (%)	F1 score (%)	Accuracy (%)
Swin transformer	HandPD	88.17	88.64	86.23	87.57
	NewHandPD	97.54	98.37	97.74	87.62
CAS transformer	HandPD	87.74	88.90	97.53	88.92
	NewHandPD	**92.59**	**92.77**	**91.38**	**92.68**

TABLE 5 Computation cost and accuracy of different Transformer variants.

Method	Parameters (M)	GFLOPs (G)	Accuracy (%)
ViT	326.73	17.60	82.97
Swin transformer	186.13	8.77	87.62
CAS transformer	**186.17**	**8.78**	**92.68**

swin transformer, and maintains convergence stability during the experiments.

CAS Transformer was developed by incorporating the CA block into SwinT architecture. As shown in Table 5, after embedding CA, the parameter count escalated from 186.13 MB to 186.17 MB, which is only an increase of 0.04 MB in the number of parameters and a slight increase in GFLOPs. So it is shown that the experimental results that the combination of CA and SwinT improves the performance of the transformer. Finally, we can see that CAS Transformer is superior to other deep learning models, and also shows that the addition of coordinate attention captures cross-channel coordinate information so that the model can more accurately locate and identify the target.

5 | CONCLUSION

In this paper, a novel model is proposed for PD detection. To increase the amount of data to meet the conditions of the deep learning model, we denoise the handwriting image and generate a similar handwriting image through CycleGAN before the feature extraction process. We also embed the coordinate attention (CA) structure into the CAS Transformer to generate handwriting features containing location information. In order to overcome the limitation of single feature analysis, we use deep learning to fuse handwriting data features at multiple scales to generate the most discriminating high-quality feature vector. At the same moment, the traditional classifier is used for PD detection.

Our experimental results show that our proposed model and the approach of multi-scale fusion of features can more easily distinguish PD patients from healthy individuals. With the addition of CA extraction to location-aware features and multi-feature fusion in the framework, the performance of PD detection was significantly improved with an accuracy of 92.68%, which is more accurate and effective than the current mainstream algorithms for assessing Parkinson's disease.

In the future, we hope to combine image-based handwriting and video-based hand movements to create a system that can automatically rate movement disorders. Therefore, we need to record more experimental data from patients as well as healthy controls, and we will adopt these motor data of PD to propose an integrated system of supplementary diagnosis and rehabilitation.

AUTHOR CONTRIBUTIONS

Nana Wang: Writing - original draft, writing - review and editing. Xuesen Niu: Methodology, software, writing - original draft. Yiyang Yuan: Formal analysis, methodology, software, writing - original draft. Yunze Sun: Methodology, software, writing - original draft. Ran Li: Methodology, software, visualization. Guoliang You: Methodology, software, visualization. Aite Zhao: Writing - review and editing.

ACKNOWLEDGEMENTS

This research was supported in part by National Natural Science Foundation of China under Grant No. 62106117, China Postdoctoral Science Foundation under Grant No. 2022M711741, and Natural Science Foundation of Shandong Province under Grant No. ZR2021QF084.

CONFLICT OF INTEREST STATEMENT

The authors declare that there are no conflicts of interest regarding the publication of this paper.

DATA AVAILABILITY STATEMENT

Data available on request from the authors.

ORCID

Nana Wang https://orcid.org/0000-0002-5299-0611

REFERENCES

1. De Stefano, C., Fontanella, F., Impedovo, D., Pirlo, G., di Freca, A.S.: Handwriting analysis to support neurodegenerative diseases diagnosis: A review. Pattern Recognit. Lett. 121, 37–45 (2019)
2. Wingate, J., Kollia, I., Bidaut, L., Kollias, S.: Unified deep learning approach for prediction of parkinson's disease. IET Image Process. 14(10), 1980–1989 (2020)
3. Feigin, V.L., Nichols, E., Alam, T., Bannick, M.S., Beghi, E., Blake, N., et al.: Global, regional, and national burden of neurological disorders, 1990–2016: a systematic analysis for the global burden of disease study 2016. Lancet Neurol. 18(5), 459–480 (2019)
4. Le, W., Dong, J., Li, S., Korczyn, A.D.: Can biomarkers help the early diagnosis of parkinson's disease? Neurosci. Bull. 33(5), 535–542 (2017)
5. Li, T., Le, W.: Biomarkers for parkinson's disease: how good are they? Neurosci. Bull. 36(2), 183–194 (2020)

6. Zesiewicz, T.A., Bezchlibnyk, Y., Dohse, N., Ghanekar, S.D.: Management of early parkinson disease. Clin. Geriatr. Med. 36(1), 35–41 (2020)
7. Ghai, S., Ghai, I., Schmitz, G., Effenberg, A.O.: Effect of rhythmic auditory cueing on parkinsonian gait: a systematic review and meta-analysis. Sci. Rep. 8(1), 1–19 (2018)
8. Sveinbjornsdottir, S.: The clinical symptoms of parkinson's disease. J. Neurochem. 139, 318–324 (2016)
9. Impedovo, D., Pirlo, G., Vessio, G.: Dynamic handwriting analysis for supporting earlier parkinson's disease diagnosis. Information 9(10), 247 (2018)
10. Cheng, Z., Jian, S., Rashidi, T.H., Maghrebi, M., Waller, S.T.: Integrating household travel survey and social media data to improve the quality of od matrix: a comparative case study. IEEE Trans. Intell. Transp. Syst. 21(6), 2628–2636 (2020)
11. Thomas, M., Lenka, A., Kumar Pal, P.: Handwriting analysis in parkinson's disease: current status and future directions. Mov. Disord. Clin. Pract. 4(6), 806–818 (2017)
12. Zhao, A., Li, J.: Two-channel lstm for severity rating of parkinson's disease using 3d trajectory of hand motion. Multimed. Tools Appl. 1–16 (2022)
13. Cheng, Z., Rashidi, T.H., Jian, S., Maghrebi, M., Waller, S.T., Dixit, V.: A spatio-temporal autocorrelation model for designing a carshare system using historical heterogeneous data: Policy suggestion. Transp. Res. Part C Emerg. Technol. 141, 103758 (2022)
14. Zhao, A., Dong, J., Li, J., Qi, L., Zhou, H.: Associated spatio-temporal capsule network for gait recognition. IEEE Trans. Multimedia. 24, 846–860 (2021)
15. Sahu, B., Mohanty, S.N.: Cmba-svm: a clinical approach for parkinson disease diagnosis. Int. J. Inf. Technol. 13(2), 647–655 (2021)
16. Xu, S., Pan, Z.: A novel ensemble of random forest for assisting diagnosis of parkinson's disease on small handwritten dynamics dataset. Int. J. Med. Inform. 144, 104283 (2020)
17. Li, Y., Yang, L., Wang, P., Zhang, C., Xiao, J., Zhang, Y., et al.: Classification of parkinson's disease by decision tree based instance selection and ensemble learning algorithms. J. Med. Imaging & Health Infor. 7(2), 444–452 (2017)
18. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232. IEEE, Piscataway (2017)
19. Hou, Q., Zhou, D., Feng, J.: Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13713–13722. IEEE, Piscataway (2021)
20. Khachnaoui, H., Mabrouk, R., Khlifa, N.: Machine learning and deep learning for clinical data and pet/spect imaging in parkinson's disease: a review. IET Image Process. 14(16), 4013–4026 (2020)
21. Kamran, I., Naz, S., Razzak, I., Imran, M.: Handwriting dynamics assessment using deep neural network for early identification of parkinson's disease. Future Gener. Comput. Syst. 117, 234–244 (2021)
22. Taleb, C., Khachab, M., Mokbel, C., Likforman Sulem, L.: Feature selection for an improved parkinson's disease identification based on handwriting. In: 2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR), pp. 52–56. IEEE, Piscataway (2017)
23. Hamid, N.A., Sjarif, N.N.A.: Handwritten recognition using svm, knn and neural network. arXiv preprint, arXiv:170200723 (2017)
24. Drotár, P., Mekyska, J., Rektorová, I., Masarová, L., Smékal, Z., Faundez Zanuy, M.: Analysis of in-air movement in handwriting: A novel marker for parkinson's disease. Comp. Meth. Prog. Biomed. 117(3), 405–411 (2014)
25. Drotár, P., Mekyska, J., Rektorová, I., Masarová, L., Smékal, Z., Faundez Zanuy, M.: Decision support framework for parkinson's disease based on novel handwriting markers. IEEE Trans. Neural Syst. Rehabilitation Eng. 23(3), 508–516 (2014)
26. Drotár, P., Mekyska, J., Rektorová, I., Masarová, L., Smékal, Z., Faundez Zanuy, M.: Evaluation of handwriting kinematics and pressure for differential diagnosis of parkinson's disease. Artif. Intell. Med. 67, 39–46 (2016)
27. Pereira, C.R., Weber, S.A., Hook, C., Rosa, G.H., Papa, J.P.: Deep learning-aided parkinson's disease diagnosis from handwritten dynamics. In: 2016 29th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), pp. 340–346. IEEE, Piscataway (2016)
28. Sivaranjini, S., Sujatha, C.: Deep learning based diagnosis of parkinson's disease using convolutional neural network. Multimed. Tools Appl. 79(21), 15467–15479 (2020)
29. Diaz, M., Moetesum, M., Siddiqi, I., Vessio, G.: Sequence-based dynamic handwriting analysis for parkinson's disease detection with one-dimensional convolutions and bigrus. Expert Syst. Appl. 168, 114405 (2021)
30. Nolazco Flores, J.A., Faundez Zanuy, M., De La Cueva, V., Mekyska, J.: Exploiting spectral and cepstral handwriting features on diagnosing parkinson's disease. IEEE Access 9, 141599–141610 (2021)
31. Afonso, L.C., Rosa, G.H., Pereira, C.R., Weber, S.A., Hook, C., Albuquerque, V.H.C., et al.: A recurrence plot-based approach for parkinson's disease identification. Future Gener. Comput. Syst. 94, 282–292 (2019)
32. Naseer, A., Rani, M., Naz, S., Razzak, M.I., Imran, M., Xu, G.: Refining parkinson's neurological disorder identification through deep transfer learning. Neural Comput. Appl. 32(3), 839–854 (2020)
33. Moetesum, M., Siddiqi, I., Vincent, N., Cloppet, F.: Assessing visual attributes of handwriting for prediction of neurological disorders'a case study on parkinson's disease. Pattern Recognit. Lett. 121, 19–27 (2019)
34. Gazda, M., Hireš, M., Drotár, P.: Multiple-fine-tuned convolutional neural networks for parkinson's disease diagnosis from offline handwriting. IEEE Trans. Syst. Man Cybern. 52(1), 78–89 (2021)
35. Fei, Y., Li, Z., Zhu, T., Ni, C.: A lightweight attention-based convolutional neural networks for fresh-cut flower classification. IEEE Access, (2023)
36. Park, J., Woo, S., Lee, J.Y., Kweon, I.S.: Bam: Bottleneck attention module. arXiv preprint arXiv:180706514 (2018)
37. Zhang, Q.L., Yang, Y.B.: Sa-net: Shuffle attention for deep convolutional neural networks. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2235–2239. IEEE, Piscataway (2021)
38. Zhang, Y., Lei, H., Huang, Z., Li, Z., Liu, C.M., Lei, B.: Parkinson's disease classification with self-supervised learning and attention mechanism. In: 2022 26th International Conference on Pattern Recognition (ICPR), pp. 4601–4607. IEEE, Piscataway (2022)
39. Duong, L.T., Le, N.H., Tran, T.B., Ngo, V.M., Nguyen, P.T.: Detection of tuberculosis from chest x-ray images: Boosting the performance with vision transformer and transfer learning. Expert Syst. Appl. 184, 115519 (2021)
40. Pereira, C.R., Weber, S.A., Hook, C., Rosa, G.H., Papa, J.P.: Deep learning-aided parkinson's disease diagnosis from handwritten dynamics. http://wwwp.fc.unesp.br/~papa/pub/datasets/Handpd/. Accessed May 2023
41. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6848–6856. IEEE, Piscataway (2018)
42. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR (2019)
43. Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., et al.: Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:170404861 (2017)
44. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4510–4520. IEEE, Piscataway (2018)
45. Mehta, S., Rastegari, M.: Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. arXiv preprint arXiv:211002178 (2021)
46. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11976–11986. IEEE, Piscataway (2022)
47. Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., Hu, Q.: Eca-net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11534–11542. IEEE, Piscataway (2020)

48. Chen, C.F.R., Fan, Q., Panda, R.: Crossvit: Cross-attention multi-scale vision transformer for image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 357–366. IEEE, Piscataway (2021)
49. Dai, Z., Liu, H., Le, Q.V., Tan, M.: Coatnet: Marrying convolution and attention for all data sizes. NIPS 34, 3965–3977 (2021)
50. Chu, X., Tian, Z., Wang, Y., Zhang, B., Ren, H., Wei, X., et al.: Twins: Revisiting the design of spatial attention in vision transformers. NIPS 34, 9355–9366 (2021)

How to cite this article: Wang, N., Niu, X., Yuan, Y., Sun, Y., Li, R., You, G., Zhao, A.: A coordinate attention enhanced swin transformer for handwriting recognition of Parkinson's disease. IET Image Process. 17, 2686–2697 (2023). https://doi.org/10.1049/ipr2.12820