A Multimodal In Vitro Diagnostic Method for Parkinson's Disease Combining Facial Expressions and Behavioral Gait Data

Wei Huang[1,2], Yinxuan Xu[1†], Yintao Zhou[1†], Zhengyu Li[3], Jing Huang[3], Meng Pang[1*]
mengpang@ncu.edu.cn
[1]School of Mathematics and Computer Sciences, Nanchang University, Nanchang, China
[2]Yichun University, Yichun, China
[3]Nanchang University Second Affiliated Hospital, Nanchang, China

Abstract

Parkinson's disease (PD), characterized by its incurable nature, rapid progression, and severe disability, poses significant challenges to the lives of patients and their families. Given the aging population, the need for early detection of PD is increasing. In vitro diagnosis has garnered attention due to its non-invasive nature and low cost. However, existing methods present several challenges: 1) limited training data for facial expression diagnosis; 2) specialized equipment and acquisition environments required for gait diagnosis, resulting in poor generalizability; 3) the risk of misdiagnosis or missed diagnosis when relying on a single modality. To address these issues, we propose a novel multimodal in vitro diagnostic method for PD, leveraging facial expressions and behavioral gait. Our method employs a lightweight deep learning model for feature extraction and fusion, aimed at improving diagnostic accuracy and facilitating deployment on mobile devices. Furthermore, we have established the largest multimodal PD dataset in collaboration with a hospital and conducted extensive experiments to validate the effectiveness of our proposed method.

Keywords: Parkinson's disease diagnosis; facial expression analysis; behavioral gait analysis; feature fusion

Introduction

Parkinson's disease (PD) is a prevalent neurodegenerative disorder with inconspicuous early symptoms. By the time overt motor impairments manifest, the disease often progresses to an advanced stage, severely impacting patients' daily life. Recent data indicates that PD incidence has increased by 2.7 times over the past 30 years, affecting over 11.8 million people, and PD-related deaths have risen by 1.6 times to 388,000 (Steinmetz et al., 2024). While there is currently no effective prevention or treatment, medications like levodopa can alleviate symptoms and improve patients' quality of life. Therefore, early diagnosis and prognostic management are crucial for optimizing drug efficacy, slowing disease progression, and reducing the economic burden (Buono et al., 2021; Gray et al., 2022).

Generally, PD diagnosis can be broadly categorized into two types: in vivo and in vitro diagnosis. In vivo diagnosis primarily relies on specialized medical imaging techniques, such as CT, MRI, and PET (Tolosa, Garrido, Scholz, & Poewe, 2021). While these methods offer high accuracy, they also have limitations: they require professional equipments operated by trained personnel, and certain patients, like pregnant women or those with specific medical conditions, may not be suitable for some scans. Moreover, the high costs burden ordinary families. In contrast, in vitro diagnostic methods involve collecting biomarkers from PD patients, including voice signals, gait signals, and facial expressions, offering advantages such as being non-invasive, convenient, and having fewer patient restrictions. This makes them increasingly popular among doctors and researchers. Among these in vitro biomarkers, facial expressions and behavioral gait are particularly suited for routine diagnosis due to their universal applicability across races and languages, and the rich, stable visual feature information they provide.

Figure 1: A comparison between the masked faces of PD patients and the facial expression images of normal persons under six basic emotions (i.e., anger, disgust, fear, happiness, sadness, and surprise.)

Recent studies (Ricciardi et al., 2020) have indicated that 70% of PD patients exhibit facial expression dysfunction (commonly referred to as "masked faces") when compared to non-PD patients, as illustrated in Figure. 1. While facial expression disorders have been utilized as a diagnostic criterion (Hou, Qin, & Su, 2022), a small subset of patients in the early stages of the disease can still accurately express one or several categories of basic emotions. Consequently, relying solely on facial expressions for diagnosis can possibly result in missed diagnoses. Furthermore, regarding behavioral gait diagnosis, the primary symptoms of PD include bradykinesia, myotonia, resting tremor, and abnormalities in posture and gait. Some researchers have conducted studies focusing on this biological signal (P. Liu et al., 2022). However, it is noteworthy that elderly PD patients often exhibit symptoms such as bradykinesia and postural instability, which can also be natural manifestations at this age. Therefore, relying on a single gait signal for PD diagnosis may lead to misdiagnoses.

To address the limitations of existing unimodal in vitro diagnostic methods for PD, we have collaborated with the af-

[†]These authors contributed to the work equally.
[*]Corresponding author.

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

filiated hospital of Nanchang University to create the largest known PD multimodal dataset, PDMM. This dataset encompasses video recordings of seven facial expressions (neutral, anger, disgust, fear, happiness, sadness, surprise) and gait from 95 PD patients. Based on this dataset, we innovatively propose a multimodal in vitro PD diagnostic method that integrates facial expression and behavioral gait analysis. The flowchart of the proposed method is shown in Figure 2, which includes three core stages: Firstly, we preprocess the PDMM video data to segment the patient regions, utilize HRNet (Sun, Xiao, Liu, & Wang, 2019) to obtain skeletal keypoints, and then extract patients' behavioral gait features through STGCN++ (Duan, Wang, Chen, & Lin, 2022); Secondly, with the help of StyleGAN (Karras et al., 2020), we generate facial expression images depicting other six basic emotions from a single neutral facial expression image of PD patients, thereby simulating their pre-morbid facial expression state as a reference group to train a discriminative model and extract highly discriminative facial expression features; Lastly, we propose a novel feature fusion strategy, namely hybrid fusion, aimed at effectively integrating the extracted behavioral gait and facial expression features, and based on the above process, we design an end-to-end multimodal data fusion and diagnostic model for PD prediction.

We summarize the contributions of this work as follows:

- We have created the PDMM dataset, which encompasses diverse facial expressions and gait video data from 95 PD patients. This dataset is currently the largest and most comprehensive multimodal dataset available for in vitro diagnostic research on PD.
- We have pioneered a novel in vitro diagnostic method for PD, marking the first attempt to fuse multimodal data from facial expressions and behavioral gait. Specifically, we extracted skeletal key points from gait video data to obtain deep semantic features of PD patients' behavioral gait and used StyleGAN to generate premorbid expressions for training discriminative models and extracting highly discriminative facial expression features.
- We have introduced a hybrid feature fusion strategy to effectively integrate behavior gait and facial expression features for PD diagnosis.
- Our extensive experiments demonstrate the effectiveness of StyleGAN in generating virtual facial expression data of PD patients before symptom onset, as well as the superior performance of our proposed multimodal diagnostic method in PD diagnosis.

RELATED WORKS

In vitro PD Diagnosis based on Behavioral Gait Signals. PD patients often exhibit bradykinesia, myotonia, and postural instability, which can impact gait patterns (O'Shea, Morris, & Iansek, 2002). Various studies have employed specialized equipment for quantitative gait analysis in PD diagnosis, such as piezoelectric sensors worn on the feet (Blin, Ferrandez, & Serratrice, 1990), motion sensors (Kauw-A-Tjoe, Thalen, Marin-Perianu, & Havinga, 2007), and Kinect sensors capturing 3D human skeleton motion (Li et al., 2018). However, these methods are not suitable for large-scale screening due to limitations in space, hardware, cost, and operational complexity. In contrast, gait video-based in vitro diagnosis of PD offers advantages such as easy operation, no need for wearable devices, low cost, no site limitations, and potential support from artificial intelligence. Researchers have analyzed PD patients using video data and various models, such as a 3D convolutional network model (Yin et al., 2021) and a two-stream spatio-temporal attention graph convolutional network for gait dyskinesia evaluation (Guo, Shao, Zhang, & Qian, 2021).

In vitro PD Diagnosis based on Facial Expressions. In addition to speech and gait signals, recent research has explored facial expressions as a novel in vitro biomarker for diagnosing PD. This idea is supported by studies indicating that PD patients often exhibit inconspicuous facial expressions, described as "masked faces" (Ricciardi et al., 2020) . Vinokurov *et al.* (Vinokurov, Arkadir, Linetsky, Bergman, & Weinshall, 2015) used 3D sensors for PD classification and diagnosis via linear regression, while Bo *et al.* (Jin, Qu, Zhang, & Gao, 2020) utilized the Face++ platform to extract facial expression features for diagnosis using an LSTM neural network. Huang *et al.* (W. Huang et al., 2023, 2024) achieved a good accuracy in PD diagnosis by extracting features from original and generated facial expressions of PD patients and classifying them using a deep neural network.

It is worth noting that, the aforementioned PD diagnostic methods are all based on single-modal signals, and may lack strong robustness in real and complex clinical scenarios. For example, relying solely on facial expressions may lead to miss diagnoses, especially for early-stage patients who can still express emotions accurately. Additionally, elderly PD patients often exhibit symptoms like bradykinesia and postural instability, which can be age-related manifestations. Thus, relying on a single gait signal may lead to misdiagnoses. Given the lack of cross-linguistic universality in speech signals, this paper thus proposes a multi-modal PD diagnosis solution fusing gait and facial expressions, which aims to address the deficiencies of single-modal methods, as well as improve the PD diagnosis accuracy.

THE PROPOSED METHOD

As shown in Figure 2, the proposed multimodal PD diagnostic method comprises three parts: Firstly, we employ the YOLOv8 model (Terven, Córdova-Esparza, & Romero-González, 2023) to segment patient areas from behavioral gait videos, subsequently extracting skeletal keypoints using HRNet, and further deriving behavioral gait features through STGCN++. Secondly, we use StyleGAN to generate facial expression images depicting six basic emotions from a single neutral facial image of PD patients, approximating their premorbid facial states. These images are used to train a deep learning model for extracting highly discriminative facial fea-

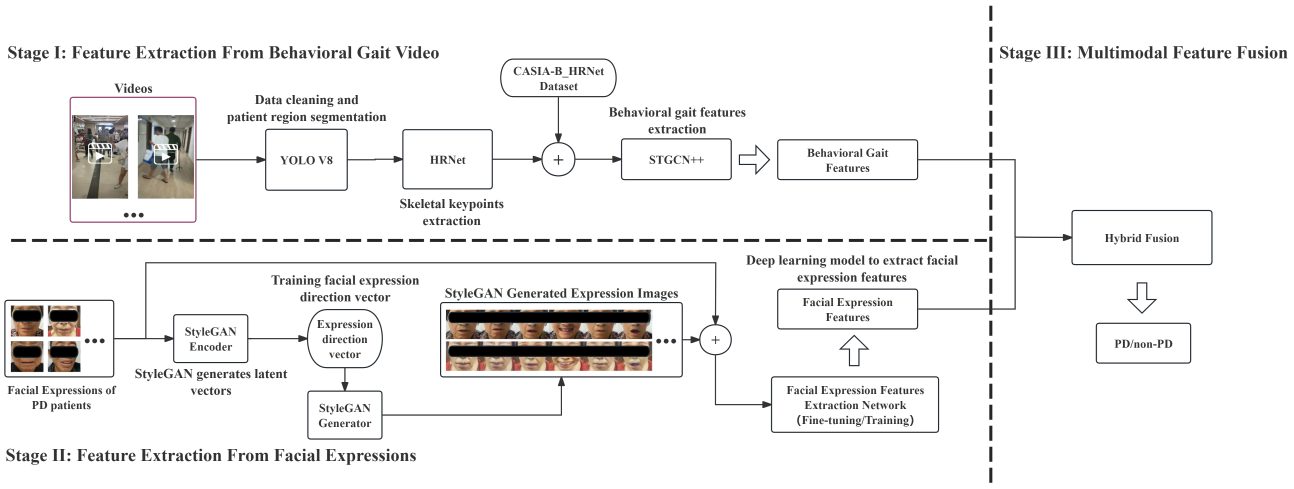

Figure 2: Overview of the proposed multimodal in vitro diagnostic method for PD.

tures. Lastly, we propose a feature fusion strategy, i.e., hybrid fusion, to effectively integrate extracted behavioral gait and facial expression features for PD diagnosis.

Feature Extraction From Behavioral Gait Video

The behavioral gait feature extraction process necessitated preprocessing due to the less rigorous recording conditions and interfering factors present in the gait video data of PD patients. The process can be distilled into three key steps:

- **Data cleaning and patient region segmentation:** We manually screened the behavioral gait videos to select unobstructed and clear clips, eliminating interfering factors such as the patients' rising and turning processes to obtain purer gait data. After cleaning, some videos contained multiple individuals. To eliminate interference from non-PD patients, we employed the YOLOv8 model for accurate tracking and segmentation, ensuring analysis focused solely on the region of interest (the PD patient).
- **Skeletal keypoints extraction:** After segmentation, we utilized HRNet to extract skeletal keypoints from the data, ensuring that these keypoints adhered to the COCO17 standard (Lin et al., 2014).
- **Behavioral gait feature extraction:** We employed STGCN++, a modified version of the original STGCN model (Yan, Xiong, & Lin, 2018), to extract effective gait features from the collected skeletal keypoints. STGCN++ incorporates multi-branch time-domain convolution, enhancing temporal modeling capability while reducing computational complexity.

Feature Extraction From Facial Expressions

Due to the lack of control samples from PD patients before they developed the disease, it is challenging to train an effective discriminative model for feature extraction from the facial expression data we collected. To address this issue, we attempt to use StyleGAN to synthesize virtual facial expression images of PD patients in their pre-morbid state. Then, we explore the use of different deep learning models for feature extraction and classification training on the augmented facial expression data. The process for facial expression feature extraction from PD patients is detailed below.

Step 1: Generating latent vectors using StyleGAN. We employ a pre-trained StyleGAN generator combined with an encoder network, inspired by (Pang et al., 2024). During training, we optimize the encoder parameters to ensure that the reconstructed images generated by the generator are as close as possible to the original input images. This approach enables us to obtain a latent vector in the latent space of StyleGAN that is similar to the identity information of the original input image. It should be noted that during training, the parameters of the encoder network are fine-tuned through a carefully designed similarity loss function, which is a weighted combination of the VGG-16 perceptual loss and the per-pixel mean squared error (MSE) loss. The definition of this loss function is as follows:

$$c^* = \min_c L_{percerp}(G(c), I_O) + \frac{\lambda_{mse}}{N} \parallel G(c) - I_O \parallel_2^2, \quad (1)$$

where c represents the target latent vector, I_O denotes the input image, G signifies the pre-trained StyleGAN generator, N denotes the total number of image pixels, and λ_{mse} is a weight hyperparameter. The perceptual loss, i.e., L_{percep}, measures the perceptual difference between the original image I_O and the synthesized image $G(c)$ using a pre-trained VGG network. This perceptual loss is defined as

$$L_{percep}(G(c), I_O) = \sum_{j=1}^{k} \frac{\lambda_j}{N_j} \parallel C_j(I_O) - C_j(G(c)) \parallel_2^2, \quad (2)$$

where $C_j(.)$ denotes the feature map output from the j-th convolutional layer of the VGG-16 network, k denotes the number of convolutional layers, N_j represents the total number of

Algorithm 1 Calculating the direction vector $\hat{n}_{AB}$

Input: Training data $(latent_x, label_x)$, $x \in \{A, B\}$; $label_A = 0$, $label_B = 1$.

Initialize model parameters $\vec{a}$ and b randomly
while not converged **do**
  $Loss = 0$
  for each training sample $(latent_x, label_x)$ **do**
    $P = \sigma((1 - 2 * label_x)\vec{a} \cdot latent_x + b)$
    $BCELoss- = label_x * log(1 - P) + (1 - label_x) * log(P)$
  end for
  Minimize the BCELoss by updating $\vec{a}$ and b using gradient descent
end while
return $\hat{n}_{AB} = \vec{a}$

Output: The direction vector $\hat{n}_{AB}$

pixels in the feature map from the j-th convolutional layer, and λ_j is a weight hyperparameter.

Step 2: Calculating facial expression direction vector. Following the method in Step 1, we can obtain latent vectors that share consistent identity information but exhibit different facial expressions A and B (e.g., neutral and happiness). Subsequently, we calculate the direction vector n_{AB} (refer to Figure 3) to achieve the transition from expression A to expression B. Specifically, we first obtain the sets of latent vectors for expressions A and B, denoted as $latent_A$ and $latent_B$, and assign them labels of $label_A = 0$ and $label_B = 1$, respectively. We then construct the mapping function $f = (1 - 2label_x)\vec{a} \cdot latent_x + b$, where $\vec{a}$ can be viewed as the normal vector from expression A to expression B, and the term $(1 - 2label_x)$ is used to control the direction*. Next, we build a logistic regression model $P = \sigma(f)$ to output the predicted class label probability, which lies between $[0, 1]$, and determine the normal vector $\vec{a}$ that manipulates the transition from A to B by optimizing a Binary Cross Entropy (BCE) loss. This $\vec{a}$ can be approximated as the required direction vector n_{AB}. For clarity, the algorithmic process for computing $\hat{n}_{AB}$ is outlined in Algorithm 1.

Similarly, by adopting the aforementioned calculation algorithm, we can obtain the direction vectors between multiple facial expression states, enabling us to synthesize different types of facial expressions while preserving identity.

Step 3: Multi facial expression synthesis and deep feature extraction.

Based on the method described in Step 2, we can successfully obtain multiple direction vectors that have the ability to control different facial expressions. For any facial image, we initially apply the method mentioned in Step 1 to extract its corresponding latent vector within the latent space of StyleGAN. Subsequently, by performing a linear calculation between the latent vector and the specific direction vector that governs facial expressions, we can utilize StyleGAN's generator G to synthesize a new image with the desired facial expression while preserving the identity information. The process of facial expression generation is formalized as follows:

$$I' = G_{dec}(G_{enc}(I) + \lambda \hat{n}_{AB}), \tag{3}$$

where G_{enc} and G_{dec} denote the encoder and the decoder of G, respectively, I represents a facial image with expression A, I' represents the same person with expression B, $\hat{n}$ is the direction vector from expression A to B, and λ controls the degree of expression change.

Using the aforementioned facial expression synthesis method, we generate images of six basic emotions (anger, disgust, fear, sadness, happiness, surprise) from a single neutral facial image of a PD patient, capturing pre-disease expressions. These images form a control group and augment our training dataset for Parkinson's disease facial expressions. We then apply deep learning models, such as ResNet and MobileNetV3, to this enriched dataset to train discriminative models for facial expression classification and feature extraction. Our objective is to find a balance between model size and classification performance, selecting one that is both effective and suitable for mobile deployment.

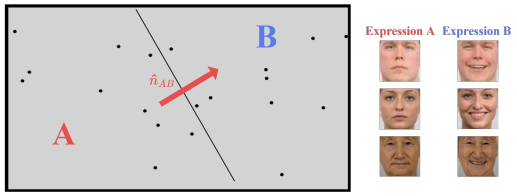

Figure 3: The direction vector $\hat{n}_{AB}$ governs the transition from expression A to expression B.

Multimodal Feature Fusion

After extracting gait features and facial expression features, we propose a multimodal feature fusion strategy, namely **Hybrid Fusion**. This strategy initially processes features from both modalities through a `FC` to obtain scores. These scores are appended to their respective features, creating new $m+1$-dimensional fused features. These two fused features are then input into another `FC` layer, producing two-dimensional outputs. These outputs are summed to obtain the final result. This approach resembles the stacking strategy in ensemble learning, where the output class probabilities (or scores) of the primary learner serve as input features for the secondary learner.

Subsequently, we designed an end-to-end multimodal feature fusion and PD diagnosis model. During the multimodal fusion training process, we utilized pre-trained models (such as STGCN++, MobileNetV3) as feature extractors and trained only the multimodal fusion layer. In this manner, we could effectively leverage existing feature extraction capabilities and reduce training costs. Furthermore, the multimodal

*When $x = A$, the term $(1 - 2label_A)$ in f evaluates to 1, and the calculated direction is $\vec{a}$. Conversely, when $x = B$, the term $(1 - 2label_B)$ in f evaluates to -1, and the calculated direction is $-\vec{a}$, which represents the opposite direction from expression B to expression A.

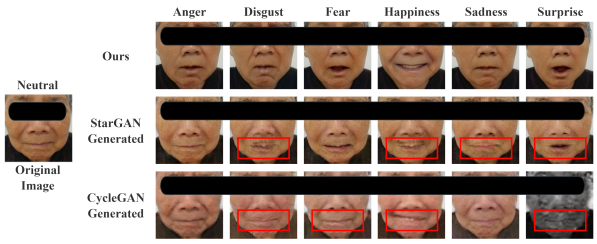

Figure 4: The neutral image of a PD patient, along with the corresponding facial images encompassing the six basic emotions generated by our method, StarGAN, and CycleGAN.

fusion layer will learn how to combine these two signals more effectively to extract more robust and highly discriminative diagnostic features.

Experimental Results

Experimental Setup

Datasets Descriptions We use three datasets to perform the evaluations, i.e., our collected PD multimodal (PDMM) dataset of PD patients, the public facial expression dataset of normal persons Tsinghua-FED (Yang et al., 2020), and the public gait dataset of normal persons CASIA-B-HRNet (Fu, Meng, Hou, Hu, & Huang, 2023). **To the best of our knowledge, PDMM is currently the largest multimodal dataset available for in vitro diagnostic research on PD.**

The PDMM dataset was collected by our group from an ongoing population-based PD study conducted at the affiliated hospital of Nanchang University. PDMM comprises 95 PD patients (55 males and 40 females), with an average age of 62.7 and a standard deviation of 9.9. For each patient, seven images were captured using a CANON EOS 5D Mark III DSLR camera equipped with an EF 24-70mm f/2.8L II USM lens. These images represent a neutral expression and six other types of basic facial expressions (i.e., anger, disgust, fear, happiness, sadness, and surprise), along with videos recording patient behavior and gait with a duration of approximately 15 seconds. *Note that the collection and use of this data have obtained informed consent from all involved PD patients. In the experiments, both the original and synthesized facial images of these PD patients were carefully censored to avoid any disclosure of personal information.*

Parameter Settings In stage 2, the values of λ_{mse} in Eq. (1), λ_j in Eq. (2), k in Eq. (2) are empirically set at 1, 1, 4, respectively, as suggested in (Abdal, Qin, & Wonka, 2019). In stage 3 of the multi-modal fusion training process, we used the pre-trained models from the first two stages (e.g., STGCN++, MobileNetV3) as feature extractors and froze their model parameters, training only the multimodal fusion layers. For this training process, we used the Adam optimizer with a learning rate of 0.001 and conducted 100 epochs of training to obtain the final PD diagnosis model. The proposed method was implemented using PyTorch on a server with Intel Xeon CPUs and NVIDIA Tesla T4 GPUs.

Table 1: Comparison of performance of different deep learning models in facial expression classification task. Ranked in descending order according to testing accuracy.

Model	Parameters	Train Acc.	**Test Acc.**
ConvNeXt	106.20 MB	0.9822	0.8889
MobileNetV3	5.93 MB	0.9872	0.8651
EfficientNetV2	77.85 MB	0.9882	0.8651
ResNet18	43.22 MB	0.9901	0.8413
MobileViT	3.76 MB	0.9803	0.8373
DenseNet121	27.13 MB	0.9842	0.8214
RegNetx_200mf	9.01 MB	0.9852	0.8095

Synthetic Facial Expression Evaluation

In this subsection, we evaluate the facial expression images synthesized in stage 2 of our method for six emotions: anger, disgust, fear, happiness, sadness, and surprise, and compare them with the synthesis results based on CycleGAN (Zhu, Park, Isola, & Efros, 2017) and StarGAN (Choi et al., 2018), as shown in Figure 4. The results indicate that the facial expression images of PD patients generated by our method can accurately depict the six basic emotions. Compared to StarGAN and CycleGAN, which exhibit local or color distortions (as highlighted by the red box in Figure 4), our method significantly outperforms in preserving facial details and maintaining image quality. These synthesis results showcase the robust image generation capability of our method using StyleGAN and also validate the effectiveness of Algorithm 1 in generating expression direction vectors.

Facial Expression Feature Extraction Evaluation

In this subsection, we utilize facial expression images of PD patients generated in Stage 2 of our method to conduct feature extraction and classification experiments on the Tsinghua-FED dataset. The objective is to evaluate the performance of various popular deep learning models, including ConvNeXt (Z. Liu et al., 2022), MobileNetV3 (Howard et al., 2019), EfficientNetV2 (Tan & Le, 2021), ResNet18 (He, Zhang, Ren, & Sun, 2016), MobileViT (Wadekar & Chaurasia, 2022), DenseNet121 (G. Huang, Liu, Van Der Maaten, & Weinberger, 2017) and RegNetx_200mf (Radosavovic, Kosaraju, Girshick, He, & Dollár, 2020), on facial expression feature extraction and classification tasks. We adopt a 4:1 partitioning strategy, dividing the facial expression dataset into training and testing sets, and report the training and testing accuracies of different deep learning models in Table 1. It is observed that all involved deep learning models achieve a testing accuracy exceeding 80%, and some lightweight models, such as MobileNetV3 and MobileViT, achieve performance comparable to large parameter models. Considering the intended application of our method in mobile device scenarios for remote healthcare, lightweight models with higher deployability are prioritized for facial feature extraction. Therefore, we choose MobileNetV3, which achieves a good bal-

Table 2: Comparison between our method and the other methods w.r.t. PD diagnosis accuracy.

Diagnosis Method	Accuracy
EfficientNetV2 (Tan & Le, 2021)	0.7125
DeiT-small (Touvron et al., 2021)	0.7193
ConvNeXt-Tiny (Z. Liu et al., 2022)	0.7336
GLCM+SVM (Hou et al., 2022)	0.4660
StarGAN+ResNet18 (W. Huang et al., 2023)	0.9543
FEPD (Zhou et al., 2024)	0.9755
Our proposed method	**1**

ance between accuracy and efficiency, as the facial expression feature extractor in Stage 2.

PD Diagnosis Evaluation

In this subsection, we compare our method using the hybrid fusion strategy with three latest representative PD diagnosis methods: co-occurrence matrix (GLCM) combined with support vector machine (SVM), dubbed GLCM+SVM (Hou et al., 2022), the data-driven unimodal facial expression PD diagnosis methods StarGAN+ResNet (W. Huang et al., 2023) and FEPD (Zhou, Pang, Huang, & Wang, 2024), and three popular deep learning models, namely DeiT (Touvron et al., 2021), ConvNeXt (Z. Liu et al., 2022), and EfficientNetV2 (Tan & Le, 2021). Specifically, we follow the evaluation protocol in (W. Huang et al., 2023; Zhou et al., 2024), and divide the PDMM dataset into 5 folds, with 4 folds (76 PD patients) utilized for training and the remaining 1 fold (19 PD patients) used for testing. To further enhance the testinig set, we incorporate facial expression images of 47 subjects aged above 60 from the Tsinghua-FED dataset, along with 47 gait video clips from the CASIA-B-HRNet dataset, forming 47 non-PD control groups. Therefore, a total of 66 subjects are used for PD diagnosis during the testing phase. All the three deep learning models mentioned above are fine-tuned using the PDMM training dataset.

The diagnostic accuracies of our method and the other comparison methods are presented in Table 2. It can be observed that our method achieves the highest diagnostic accuracy for PD, outperforming the unimodal methods StarGAN+ResNet and FEPD, and significantly surpassing the traditional GLCM+SVM PD diagnosis method as well as the other deep learning models. The superiority of our method can be attributed to its comprehensive analysis by fusing behavioral gait and facial expression features, extracting gait features with a powerful STGCN++, and proposing a multi-class facial expression generation scheme based on StyleGAN to facilitate the extraction of expressive features. The effective fusion of these two types of features mitigates the misdiagnosis issues associated with unimodal methods in PD diagnosis, thereby enhancing the accuracy of PD diagnosis. Additionally, the conventional machine learning-based GLCM+SVM performs poorly compared to the three deep learning methods and is far inferior to our method in terms of PD diagnosis accuracy, demonstrating the good representation learning capability of deep neural networks.

Table 3: PD diagnosis results using unimodal features of facial expressions or behavioral gait.

	Model	**Test Acc.**
Unimodal Facial Expression Diagnosis	MobileNetV3	0.9692
	ConvNeXt	0.9538
	EfficientNetV2	0.9538
	ResNet18	0.9230
Unimodal Behavior Gait Diagnosis	STGCN++	0.9692
	STGCN	0.9385

Ablation Study

In this subsection, we further explore the performance of our method when using only unimodal facial expression or gait features for PD diagnosis. For unimodal PD diagnosis based on facial expressions, we adopt not only the default MobileNetV3 as the feature extractor but also experiment with other deep model extractors such as ConvNeXt, EfficientNetV2, and ResNet18, aiming to assess their performance in PD diagnosis. Due to the limited availability of skeletal feature extraction models for gait-based PD diagnosis, we use only the default STGCN++ and its previous version, STGCN, for comparative analysis. The specific results are shown in Table 3. It is evident that for facial expression-based PD diagnosis, the performance of different feature extractors in PD diagnosis is closely related to their respective models' performance in classification tasks. The ranking of PD diagnosis performance is broadly consistent with the facial expression classification results shown in Table 1, indicating the importance of selecting models based on facial expression classification performance for PD diagnosis. For gait-based PD diagnosis, STGCN++ slightly outperforms STGCN. Notably, both unimodal facial expression-based PD diagnosis (using MobileNetV3) and unimodal gait-based PD diagnosis (using STGCN++) exhibit lower performance in PD diagnosis compared to our multimodal method, which combines facial expression and gait features. This not only validates the rationality of fusing these two features for PD diagnosis but also demonstrates the effectiveness of our proposed feature fusion strategy.

Conclusion

This paper presents a novel in vitro diagnostic method for Parkinson's Disease (PD), pioneering the integration of multimodal facial expression and gait behavior data. We introduce (1) a StyleGAN-based framework for synthesizing premorbid PD facial expressions to enhance expressive feature extraction, and (2) a lightweight STGCN++ model for gait analysis. A hybrid fusion strategy effectively combines facial and gait features, significantly improving PD diagnostic accuracy. Experimental results validate the superiority of our approach in PD diagnosis.

Acknowledgement

This work is supported in part by Natural Science Foundation of China (62466036, 62271239), by Natural Science Foundation of Jiangxi Province (20232BAB212025, 20232BAB206065), by High-level and Urgently Needed Overseas Talent Programs of Jiangxi Province (20232BCJ25024), and by Jiangxi Double Thousand Plan (JXSQ2023201022).

References

Abdal, R., Qin, Y., & Wonka, P. (2019). Image2stylegan: How to embed images into the stylegan latent space? In *ICCV* (pp. 4432–4441).

Blin, O., Ferrandez, A.-M., & Serratrice, G. (1990). Quantitative analysis of gait in parkinson patients: increased variability of stride length. *J. Neurol. Sci.*, *98*(1), 91–97.

Buono, V. L., Palmeri, R., De Salvo, S., Berenati, M., Greco, A., Ciurleo, R., ... others (2021). Anxiety, depression, and quality of life in parkinson's disease: the implications of multidisciplinary treatment. *Neural Regener. Res.*, *16*(3), 587–590.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., & Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR* (pp. 8789–8797).

Duan, H., Wang, J., Chen, K., & Lin, D. (2022). Pyskl: Towards good practices for skeleton action recognition. In *ACM MM* (pp. 7351–7354).

Fu, Y., Meng, S., Hou, S., Hu, X., & Huang, Y. (2023). Gpgait: Generalized pose-based gait recognition. In *ICCV* (pp. 19595–19604).

Gray, R., Patel, S., Ives, N., Rick, C., Woolley, R., Muzerengi, S., ... others (2022). Long-term effectiveness of adjuvant treatment with catechol-o-methyltransferase or monoamine oxidase b inhibitors compared with dopamine agonists among patients with parkinson disease uncontrolled by levodopa therapy: the pd med randomized clinical trial. *JAMA Neurol.*, *79*(2), 131–140.

Guo, R., Shao, X., Zhang, C., & Qian, X. (2021). Multi-scale sparse graph convolutional network for the assessment of parkinsonian gait. *IEEE Trans. Multimedia*, *24*, 1583–1594.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *CVPR* (pp. 770–778).

Hou, X., Qin, S., & Su, J. (2022). Visual detection of parkinson's disease via facial features recognition. In *CIAC* (pp. 249–257).

Howard, A., Sandler, M., Chu, G., Chen, L.-C., Chen, B., Tan, M., ... others (2019). Searching for mobilenetv3. In *ICCV* (pp. 1314–1324).

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. In *CVPR* (pp. 4700–4708).

Huang, W., Xu, W., Wan, R., Zhang, P., Zha, Y., & Pang, M. (2024). Auto diagnosis of parkinson's disease via a deep learning model based on mixed emotional facial expressions. *IEEE journal of biomedical and health informatics*.

Huang, W., Zhou, Y., Cheung, Y.-m., Zhang, P., Zha, Y., & Pang, M. (2023). Facial expression guided diagnosis of parkinson's disease via high-quality data augmentation. *IEEE Trans. Multimedia*, *25*, 7037–7050.

Jin, B., Qu, Y., Zhang, L., & Gao, Z. (2020). Diagnosing parkinson disease through facial expression recognition: video analysis. *J. Med. Internet Res.*, *22*(7), e18697.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of stylegan. In *CVPR* (pp. 8110–8119).

Kauw-A-Tjoe, R., Thalen, J., Marin-Perianu, M., & Havinga, P. (2007). Sensorshoe: Mobile gait analysis for parkinson's disease patients. In *Proc. ubicomp 2007 workshop proc.* (pp. 187–191).

Li, Q., Wang, Y., Sharf, A., Cao, Y., Tu, C., Chen, B., & Yu, S. (2018). Classification of gait anomalies from kinect. *Visual Comput.*, *34*, 229–241.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *ECCV* (pp. 740–755).

Liu, P., Yu, N., Yang, Y., Yu, Y., Sun, X., Yu, H., ... Wu, J. (2022). Quantitative assessment of gait characteristics in patients with parkinson's disease using 2d video. *PARKINSONISM RELAT D*, *101*, 49–56.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. In *CVPR* (pp. 11976–11986).

O'Shea, S., Morris, M. E., & Iansek, R. (2002). Dual task interference during gait in people with parkinson disease: effects of motor versus cognitive secondary tasks. *Phys Ther*, *82*(9), 888–897.

Pang, M., Wang, B., Ye, M., Cheung, Y.-M., Zhou, Y., Huang, W., & Wen, B. (2024). Heterogeneous prototype learning from contaminated faces across domains via disentangling latent factors. *IEEE Transactions on Neural Networks and Learning Systems*.

Radosavovic, I., Kosaraju, R. P., Girshick, R., He, K., & Dollár, P. (2020). Designing network design spaces. In *CVPR* (pp. 10428–10436).

Ricciardi, L., De Angelis, A., Marsili, L., Faiman, I., Pradhan, P., Pereira, E., ... Bologna, M. (2020). Hypomimia in parkinson's disease: an axial sign responsive to levodopa. *Eur. J. Neurol.*, *27*(12), 2422–2429.

Steinmetz, J. D., Seeher, K. M., Schiess, N., Nichols, E., Cao, B., Servili, C., ... others (2024). Global, regional, and national burden of disorders affecting the nervous system, 1990–2021: a systematic analysis for the global burden of disease study 2021. *Lancet Neurol.*, *23*(4), 344–381.

Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. In *CVPR* (pp. 5693–5703).

Tan, M., & Le, Q. (2021). Efficientnetv2: Smaller models and faster training. In *ICML* (pp. 10096–10106).

Terven, J., Córdova-Esparza, D.-M., & Romero-González, J.-A. (2023). A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas. *Mach. Learn. Knowl. Extr.*, *5*(4), 1680–1716.

Tolosa, E., Garrido, A., Scholz, S. W., & Poewe, W. (2021). Challenges in the diagnosis of parkinson's disease. *Lancet Neurol.*, *20*(5), 385–397.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In *ICML* (pp. 10347–10357).

Vinokurov, N., Arkadir, D., Linetsky, E., Bergman, H., & Weinshall, D. (2015). Quantifying hypomimia in parkinson patients using a depth camera. In *MindCare* (pp. 63–71).

Wadekar, S. N., & Chaurasia, A. (2022). Mobilevitv3: Mobile-friendly vision transformer with simple and effective fusion of local, global and input features. *arXiv preprint arXiv:2209.15159*.

Yan, S., Xiong, Y., & Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI* (Vol. 32).

Yang, T., Yang, Z., Xu, G., Gao, D., Zhang, Z., Wang, H., ... others (2020). Tsinghua facial expression database–a database of facial expressions in chinese young and older women and men: Development and validation. *PloS one*, *15*(4), e0231304.

Yin, Z., Geraedts, V. J., Wang, Z., Contarino, M. F., Dibeklioglu, H., & Van Gemert, J. (2021). Assessment of parkinson's disease severity from videos using deep architectures. *IEEE J. Biomed. Health. Inf.*, *26*(3), 1164–1176.

Zhou, Y., Pang, M., Huang, W., & Wang, B. (2024). Early diagnosing parkinson's disease via a deep learning model based on augmented facial expression data. In *ICASSP* (pp. 1621–1625).

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV* (pp. 2223–2232).