

Explainable machine learning of plantar-force gait biomarkers for Hoehn–Yahr staging in Parkinson’s disease

Mei Yang

Zigong Third People’s Hospital

Xingzhong Xiong

Sichuan University of Arts and Science

Weitian Yuan

325085404331@stu.suse.edu.cn

Sichuan University of Science & Engineering

Mingju Chen

Sichuan University of Science & Engineering

Luo Zhongqiang

Sichuan University of Science & Engineering

Article

Keywords: Parkinson’s disease, Hoehn–Yahr staging, machine learning, ordinal learning, gait biomarkers, plantar force, explainable artificial intelligence

Posted Date: October 12th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10979628/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Explainable machine learning of plantar-force gait biomarkers for Hoehn–Yahr staging in Parkinson’s disease

Mei Yang^{1,2}, Xingzhong Xiong², , Weitian Yuan^{3*}, Mingju Chen³, Luo Zhongqiang³

¹ Zigong Third People’s Hospital, Zigong 643020, People’s Republic of China; yangmei829@yeah.net

² The Sichuan Province Key Laboratory of Intelligent Medicine and Health Big Data, Sichuan University of Arts and Science, Dazhou 635002, People’s Republic of China; xzxiong@suse.edu.cn

³ Intelligent Perception and Control Key Laboratory of Sichuan Province, Sichuan University of Science & Engineering, Yibin 644002, People’s Republic of China; chenmingju@suse.edu.cn; luozhongqiang@suse.edu.cn

* **Correspondence:** Intelligent Perception and Control Key Laboratory of Sichuan Province, Sichuan University of Science & Engineering, Yibin 644002, People’s Republic of China; 325085404331@stu.suse.edu.cn

Abstract

Objective assessment of Parkinson’s disease (PD) severity from gait signals remains challenging because adjacent Hoehn–Yahr (H&Y) stages overlap and machine-learning models often lack interpretability. We developed an explainable machine-learning framework with ordinal modeling to distinguish H&Y stages 2.0, 2.5, and 3.0 using plantar-force gait biomarkers. Ninety-three patients with PD from the PhysioNet Gait in Parkinson’s Disease Database were included. Clinically interpretable biomarkers extracted from multi-channel vertical ground reaction force signals were analyzed using differentiable feature selection, conditional ordinal regression, distance-aware learning, and CTGAN-based class balancing. Performance was evaluated by repeated subject-level stratified five-fold cross-validation. The model achieved 84.83% balanced accuracy, a macro F1-score of 83.96%, quadratic weighted kappa of 0.85, and mean absolute error of 0.13; recall for H&Y stage 3.0 was 80.0%. Eighteen gait biomarkers were stably selected, highlighting double-support phase, gait-cycle variability, bilateral asymmetry, and cadence. These findings support explainable ordinal machine learning of plantar-force gait biomarkers for objective PD severity staging.

Keywords: Parkinson’s disease; Hoehn–Yahr staging; machine learning; ordinal learning; gait biomarkers; plantar force; explainable artificial intelligence

1. Introduction

Parkinson’s disease (PD) is a chronic, progressive neurodegenerative disorder. Its primary clinical manifestations include bradykinesia, resting tremor, rigidity, and postural instability, often accompanied by cognitive, autonomic, and sleep abnormalities. As the disease progresses, patients experience a gradual decline in gait capacity and independence, with significantly increased risks of falls, fractures, and long-term care needs. With the aging global population and the continuous growth of PD prevalence, accurately assessing disease severity has become a crucial foundation for developing treatment plans, evaluating rehabilitation efficacy, monitoring disease progression, and conducting clinical research [1].

The Hoehn–Yahr (H&Y) scale is currently one of the most widely used clinical staging tools for PD. The modified H&Y scale categorizes patients into different stages based on symptom distribution, postural stability, and the degree of functional impairment, providing a concise and effective index for evaluating disease progression [2,3]. Unlike general multi-class classification problems, H&Y staging possesses a natural ordinal

attribute; adjacent stages reflect progressive changes in disease severity rather than mutually exclusive independent categories. Therefore, automatic staging models must not only improve classification accuracy but also fully utilize the ordinal relationship between disease stages to avoid cross-stage misclassifications that carry significant clinical implications.

Gait impairment is one of the most typical motor symptoms of PD and an important manifestation of degraded motor control. Patients typically exhibit shortened stride length, decreased walking speed, prolonged double-support phase, reduced bilateral coordination, and increased gait variability [4,5]. Consequently, gait features are considered important digital biomarkers for evaluating PD status. Compared to traditional laboratory gait analysis systems, wearable sensor technology offers advantages such as being non-invasive, portable, and capable of continuous monitoring, making it more suitable for long-term evaluation in clinical and real-world environments [6–8]. Among these, plantar pressure and vertical ground reaction force (VGRF) signals can directly reflect the interaction between the sole and the ground during walking, providing rich information on timing, load distribution, bilateral symmetry, and dynamic stability [9–14], thus serving as an important data source for automatic PD evaluation.

With the continuous enrichment of public gait databases, artificial intelligence (AI) methods have developed rapidly in PD assessment [17–22]. Early research primarily relied on manually designed gait features combined with traditional machine learning classifiers, achieving disease recognition with a certain degree of interpretability [15]. In recent years, deep learning methods such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), CNN-LSTM hybrid models, pressure map networks, and Transformers have significantly improved the automatic feature learning capability for multichannel plantar force signals [16,23–25]. Furthermore, multi-modal and multi-task learning frameworks that integrate plantar pressure, inertial sensors, and turning tasks have further enhanced model generalization across different gait tasks and external datasets [26]. These studies fully demonstrate the application potential of wearable gait analysis in automatic H&Y staging.

Despite the progress made by AI methods based on gait data in PD evaluation, automatic H&Y staging still faces several challenges. Existing studies mostly focus on PD diagnosis or binary classification between patients and healthy controls, with relatively limited research targeting different H&Y stages [22–26]. Meanwhile, the uneven distribution of samples across disease stages in public data can easily lead to insufficient recognition of minority classes by the models [29]. Some studies employ random partitioning, which may introduce subject-level data leakage, and often treat H&Y staging as a general multi-class task without fully considering its ordinal nature [27]. In addition, strong correlations exist among gait biomarkers, and the models' interpretation of key features and predictive basis remains limited, affecting the credibility of their clinical application [26,30,35].

To address these issues, we propose the Explainable Feature-Oriented Ordinal Staging Network (EFOS-Net). This method extracts clinically meaningful gait biomarkers—such as temporal, kinetic, plantar-load distribution, bilateral symmetry, and gait variability and stability features—from multichannel VGRF signals, and employs subject-level stratified five-fold cross-validation. All data preprocessing, CTGAN augmentation, and feature selection are strictly confined within the training folds. Building on this, it integrates differentiable feature selection with CORN ordinal learning to filter key gait biomarkers and utilize the ordinal structure of H&Y staging, while incorporating SHAP to analyze the model's predictive basis. Through baseline comparisons, ablation experiments, evaluation of class imbalance handling strategies, and interpretability analysis, the staging performance, ordinal consistency, minority-stage recognition ability, and interpretability of EFOS-Net are evaluated.

2. Materials and Methods

2.1. Dataset and Study Population

Our study utilized the Gait in Parkinson's Disease Database publicly available on the PhysioNet platform [22]. It combines gait data from three previous studies, comprising 166 subjects in total: 93 patients with idiopathic PD and 73 healthy controls. This dataset provides demographic information, including subject ID, study cohort, disease group, sex, age, height, and body mass, as well as disease severity indices such as H&Y staging and/or Unified Parkinson's Disease Rating Scale (UPDRS) scores.

During data collection, subjects walked continuously for approximately 2 min on level ground at a natural, self-selected speed, while their VGRF signals were acquired using the Ultraflex Computer Dyno Graphy system (Infotronic Inc.). Eight force sensors were placed on each sole, forming 16 local VGRF channels with a sampling frequency of 100 Hz and force values measured in Newtons (N). In addition to local sensor signals, the sum of forces from the 8 sensors for the left and right feet was also recorded separately. Each resulting gait record consists of one time column, 16 local VGRF channels, and 2 total-foot VGRF channels.

As our primary objective was to automatically stage disease severity in patients with diagnosed PD, healthy controls were not artificially defined as "H&Y stage 0" nor included in the primary ordinal staging analysis. A total of 93 PD patients with valid H&Y labels were finally included (58 males, 35 females; mean age: 66.30±9.50 years). The demographic characteristics across different H&Y stages are summarized in Table 1, comprising 55 patients (59.1%) at H&Y stage 2.0, 28 (30.1%) at stage 2.5, and 10 (10.8%) at stage 3.0. We defined these three stages as ordered classes with progressively increasing disease severity. To address the class imbalance caused by the relatively small number of patients at stage 3.0, we employed CTGAN for sample augmentation and balancing. Furthermore, because a small amount of height and body mass data was missing in the original database, relevant descriptive statistics were computed using available cases without imputation. Each subject was treated as an independent analysis unit, and gait cycles extracted from continuous gait records were used solely for within-subject gait biomarker computation.

Table 1. Demographic and clinical characteristics of patients with PD according to H&Y stage

Characteristic	H&Y 2.0	H&Y 2.5	H&Y 3.0	Overall PD
Participants, n (%)	55 (59.1%)	28 (30.1%)	10 (10.8%)	93 (100.0%)
Male, n (%)	36 (65.5%)	16 (57.1%)	6 (60.0%)	58 (62.4%)
Female, n (%)	19 (34.5%)	12 (42.9%)	4 (40.0%)	35 (37.6%)
Age, years	64.22±9.62	68.79±7.58	70.80±11.23	66.30±9.50
Height, cm	167.96±8.31	167.86±8.61	163.11±9.92	167.44±8.59
Body mass, kg	73.18±12.20	73.07±11.39	65.56±11.22	72.40±11.96

2.2. Signal Processing and Gait Biomarker Extraction

All gait signals were processed independently per subject. The raw data included eight local VGRF channels per foot and the total VGRF signals for both feet, sampled at 100 Hz [22]. To mitigate the influence of gait initiation and termination phases, the first and last 10 s of each record were discarded, and VGRF values below 20 N were zeroed. Records exhibiting obvious signal interruptions, critical sensor failures, or unreliable gait event identification were excluded from subsequent analyses.

Gait events were identified using the total VGRF signals of the left and right feet. The total VGRF was calculated by summing the forces from the eight ipsilateral local sensors,

which was then cross-checked for consistency against the total VGRF channels provided by the database. A 20 N threshold was applied to detect foot-contact and toe-off events; the interval between two consecutive ipsilateral foot contacts was defined as a complete gait cycle, which further partitioned the stance and swing phases [12,13,31]. To minimize false detections caused by transient noise, the minimum interval between adjacent ipsilateral foot contacts was set to 0.5 s, retaining only valid gait cycles with complete event sequences and cycle durations within a reasonable range.

Based on the valid gait cycles, candidate gait biomarkers were extracted across multiple domains, including temporal, kinetic, plantar-load distribution, relative center-of-pressure, bilateral symmetry, and gait variability and stability features (Table 2). Regional load indices were grouped according to the relative positions of local sensors and normalized as the proportion of local force or force-time integral relative to the total ipsilateral foot load. Relative center-of-pressure indices were computed via force-weighting based on relative sensor coordinates to delineate relative load positions and variations within the insole coordinate system, rather than represent spatially calibrated true center-of-pressure displacements.

All candidate features were initially calculated within individual valid gait cycles and subsequently aggregated at the subject level. For applicable continuous variables, the mean, median, standard deviation, interquartile range, and coefficient of variation were computed; bilateral averages and corresponding asymmetry indices were also derived. Because body mass information was missing for some subjects, the raw VGRF retained the unit of Newtons (N) without normalization by body mass. To prevent data leakage, the standardization parameters for the model input features were estimated solely from the training sets and fixed prior to application to the corresponding validation and test sets.

Table 2. Candidate plantar-force gait biomarkers extracted from multichannel VGRF signals

Feature domain	Candidate biomarkers	Calculation basis	Unit or representation
Temporal features	Gait-cycle duration; stance time; swing time; double-support time; stance-phase ratio; swing-phase ratio; double-support ratio; cadence	Calculated from foot-contact and toe-off events detected using the total-foot VGRF signals	s, %, steps/min
Kinetic features	Peak total VGRF; mean total VGRF; force-time integral; loading-response peak; maximum loading rate; push-off peak	Calculated from the total VGRF signal during the stance phase	N, N·s, N/s
Plantar-load distribution features	Peak force, mean force, force-time integral, and relative load proportion in the heel, midfoot, forefoot, and toe regions; forefoot-to-rearfoot load ratio; medial-to-lateral load ratio	Calculated from local sensor signals and regional load relative to the total ipsilateral foot load	N, N·s, %, or dimensionless
Relative center-of-pressure features	Mean anteroposterior position; mean mediolateral position; anteroposterior excursion range; mediolateral excursion range; within-cycle fluctuation	Force-weighted calculation based on the relative coordinates of the plantar sensors	Relative units
Bilateral symmetry features	Bilateral asymmetry indices for gait-cycle duration, stance time, swing time, peak VGRF, force-time integral, and regional load proportions	Standardized left-right difference between corresponding gait features	% or dimensionless
Variability and stability features	Standard deviation, interquartile range, and coefficient of variation of temporal, kinetic, regional-load, and relative center-of-pressure parameters across gait cycles	Within-subject statistics calculated from consecutive valid gait cycles	Original unit or %

To obtain a robust estimate of model performance while accounting for variability in subject partitioning and preventing subject-level data leakage, we employed repeated subject-level stratified five-fold cross-validation. The complete five-fold cross-validation procedure was repeated using 20 predefined random seeds. For each random seed, the 93 patients with PD were independently repartitioned into five mutually exclusive subsets, while preserving the proportions of patients at H&Y stages 2.0, 2.5, and 3.0 as closely as possible.

Within each repetition, each subset served once as the independent test set, while the remaining four subsets formed the development set and were further partitioned into training and validation subsets for model training, hyperparameter optimization, and early stopping. Consequently, every subject received exactly one out-of-fold prediction within each complete five-fold cross-validation repetition.

All gait data and derived features from a given subject were strictly maintained within the same data subset within each repetition. All data-distribution-dependent processing operations, including missing-value handling (if required), feature standardization, class-imbalance handling, CTGAN-based augmentation, and embedded feature selection, were performed exclusively using the corresponding training data. Processing parameters estimated from the training data were fixed before application to the corresponding validation and test sets; neither the validation nor the test data participated in parameter estimation, synthetic-data generation, feature selection, or model fitting.

The original PD cohort contained 55, 28, and 10 patients at H&Y stages 2.0, 2.5, and 3.0, respectively, indicating an obvious class imbalance. To enhance the representation of minority stages during model training, we employed the Conditional Tabular Generative Adversarial Network (CTGAN) for data augmentation exclusively within the training folds [29]. Using subject-level gait biomarkers and H&Y stages as input conditions, CTGAN generated synthetic samples solely for underrepresented stages until the sample size of each stage roughly equaled that of the majority stage. CTGAN was retrained independently within each training fold of every cross-validation repetition, and the synthesized samples were used exclusively within the corresponding training fold without entering the validation or test sets.

Furthermore, under this same data partitioning framework, we compared various class imbalance handling strategies, including no resampling, class weighting, random oversampling, SMOTE, and CTGAN [36]. All strategies were applied exclusively to the training data, and final performance metrics were calculated using non-augmented real test subjects to ensure a consistent comparison across different methods.

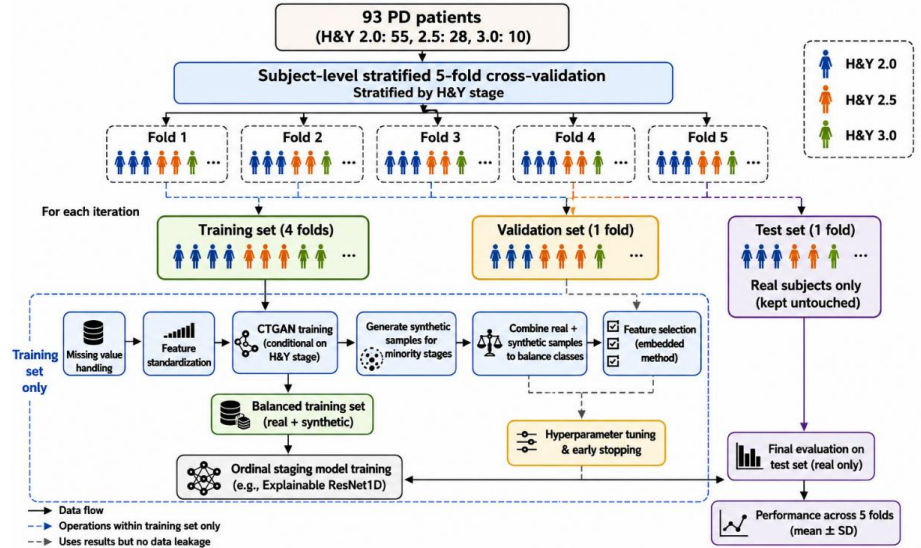


Figure 1. Workflow of subject-level data partitioning and class imbalance handling.

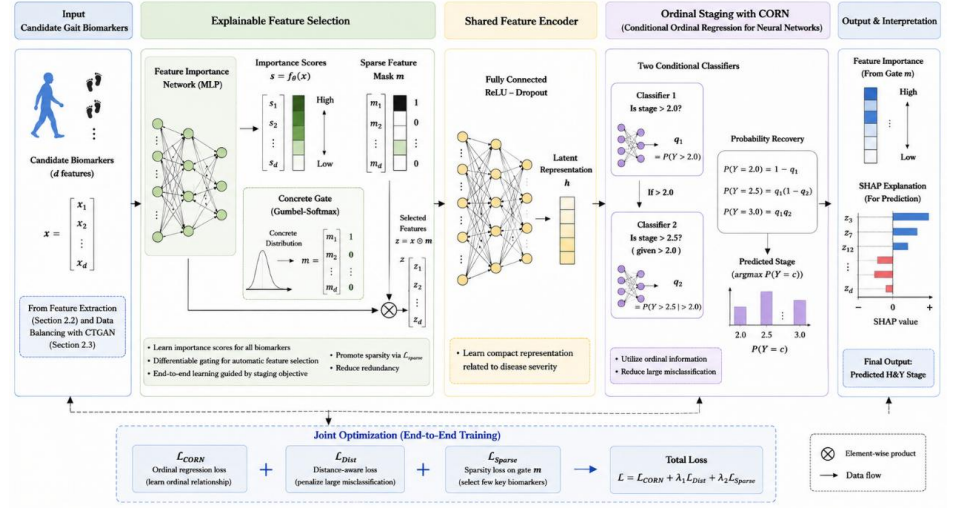


Figure 2. Workflow of the proposed EFOS-Net for explainable Parkinson's disease ordinal staging.

To filter key gait biomarkers with discriminative value for disease stages and explicitly utilize the ordinal progression of H&Y staging from mild to severe, our study constructed the Explainable Feature-Oriented Ordinal Staging Network (EFOS-Net). As illustrated in Figure 2, EFOS-Net takes subject-level gait biomarker vectors as input. After completing class imbalance handling within the corresponding training fold, the network sequentially performs feature importance estimation, differentiable feature gating, shared feature representation learning, and ordinal stage prediction. It employs end-to-end joint optimization, ensuring that the screening of gait biomarkers is directly supervised by the H&Y staging objective.

Let the candidate gait biomarker vector for the i -th subject be denoted as:

$$\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{id}]^T \in \mathbb{R}^d, \quad (1)$$

where d represents the number of candidate features, including temporal, kinetic, plantar-load distribution, relative center-of-pressure, bilateral symmetry, and gait variability and stability indices. The discriminative contribution of different gait biomarkers to disease severity may vary, and strong correlations exist among some variables. Directly inputting all candidate features into the model may increase network complexity and introduce redundant information. Therefore, EFOS-Net first employs a lightweight multi-layer perceptron to construct a Feature Importance Network, which learns a corresponding importance score for each input variable:

$$\mathbf{s}_i = f_{\theta}(\mathbf{x}_i),$$

where $f_{\theta}(\cdot)$ denotes the feature importance estimation network parameterized by θ , and $\mathbf{s}_i \in \mathbb{R}^d$ is the importance score vector for the candidate gait biomarkers. A higher score indicates that the corresponding feature has a greater potential contribution to distinguishing different H&Y stages.

Unlike traditional two-stage methods that perform offline variable selection before training a classifier, EFOS-Net embeds feature selection directly into the prediction network. The model employs a differentiable gating mechanism based on the Concrete distribution to convert continuous importance scores into learnable feature gates:

$$\mathbf{m}_i = \text{Concrete}(\mathbf{s}_i; \tau),$$

where $\mathbf{m}_i = [m_{i1}, m_{i2}, \dots, m_{id}]^T$ denotes the feature gating vector, and τ is a temperature parameter that controls the smoothness of the gating distribution. The selected feature representation is given by:

$$\tilde{\mathbf{x}}_i = \mathbf{m}_i \odot \mathbf{x}_i,$$

where $\odot$ represents the element-wise product. Gait biomarkers with larger gate weights are retained and participate in the subsequent staging prediction, while features with lower discriminative contributions or redundant information are suppressed. Because the gating process is differentiable, its parameters can receive gradients from the final staging loss, enabling feature screening and disease stage prediction to be jointly optimized under the same objective.

To encourage the model to rely on a small number of representative key gait biomarkers rather than a large set of highly correlated variables, our study further imposes a sparsity constraint on the feature gating vector:

$$\mathcal{L}_{\text{sparse}} = \frac{1}{N} \sum_{i=1}^N \|\mathbf{m}_i\|_1,$$

where N is the number of training samples. This constraint helps form a more compact feature subset and improves the interpretability of the feature selection results. After model training is completed, the gate weights across different subjects and cross-validation folds can be aggregated to evaluate the relative importance and selection stability of each gait biomarker. The gated gait biomarkers are subsequently fed into a shared feature encoder:

$$\mathbf{h}_i = f_\phi(\tilde{\mathbf{x}}_i),$$

where $f_\phi(\cdot)$ denotes a fully connected feature encoding network, and $\mathbf{h}_i$ is its output low-dimensional latent representation. The shared encoder integrates complementary information among different key gait biomarkers and enhances the separability of different disease stages within the latent feature space. Because the shared encoder is jointly trained with the frontend feature gating module using the same staging objective, the obtained latent representation can more intensely characterize the changes in PD gait function that occur with increasing disease severity.

The latent representation output by the shared encoder is further fed into a Conditional Ordinal Regression for Neural Networks (CORN) output layer. H&Y staging possesses a clear ordinal attribute, where adjacent stages represent a continuous process of progressive motor function degradation rather than mutually exclusive ordinary categories. Therefore, our study utilizes CORN instead of a traditional Softmax classifier to explicitly model the ordinal relationship between different stages. For the three-class task consisting of H&Y stages 2.0, 2.5, and 3.0, CORN transforms it into two binary classification tasks with conditional dependence:

$$q_{i1} = P(y_i > 2.0 \mid \mathbf{h}_i),$$

$$q_{i2} = P(y_i > 2.5 \mid y_i > 2.0, \mathbf{h}_i).$$

The first output determines whether the patient exceeds H&Y stage 2.0; the second output, conditioned on the patient already exceeding H&Y stage 2.0, further determines whether the patient exceeds H&Y stage 2.5. The predicted probabilities for the three disease stages are recovered via the conditional probability chain:

$$P(y_i = 2.0) = 1 - q_{i1},$$

$$P(y_i = 2.5) = q_{i1}(1 - q_{i2}),$$

$$P(y_i = 3.0) = q_{i1}q_{i2}.$$

The final prediction is determined by the stage with the maximum posterior probability. By sequentially learning the discriminative boundaries between adjacent disease stages, CORN maintains the rank consistency of the predictions, enabling the model to focus more on clinically meaningful stage transitions and reducing the risk of severe cross-stage misclassifications between H&Y stages 2.0 and 3.0.

EFOS-Net is trained end-to-end using a joint loss function that combines conditional ordinal classification, distance-aware error constraint, and feature sparsity constraint:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CORN}} + \lambda_{\text{dist}}\mathcal{L}_{\text{dist}} + \lambda_{\text{sparse}}\mathcal{L}_{\text{sparse}},$$

where $\mathcal{L}_{\text{CORN}}$ is the conditional ordinal regression loss used to learn the ordinal discriminative boundaries between different H&Y stages; $\mathcal{L}_{\text{dist}}$ is the distance-aware loss that penalizes errors based on the ordinal distance between the predicted and true stages; and $\mathcal{L}_{\text{sparse}}$ is the feature gating sparsity constraint. The hyperparameters λ_{dist} and λ_{sparse} control the contributions of the distance-aware and feature sparsity terms to the total loss, respectively, and are determined using the validation set in each cross-validation iteration. The distance-aware loss imposes a higher penalty on misclassifications spanning multiple stages compared to adjacent-stage misclassifications, thereby ensuring that model predictions align more closely with the clinical progression of PD from mild to severe.

Once model training is complete, the feature gate weights are used to illustrate which key gait biomarkers were selected by the model, and their stability is evaluated through the gate weights and selection frequencies across different cross-validation folds. Building upon this, SHapley Additive exPlanations (SHAP) is further employed to analyze the contribution directions and magnitudes of the selected features to the model's predictions. Feature gating and SHAP analysis provide interpretations at the feature selection level and the prediction contribution level, respectively, corresponding to the Output & Interpretation module in Figure 2.

In summary, EFOS-Net integrates gait biomarker screening, latent feature representation learning, and H&Y ordinal staging within a unified end-to-end framework, ensuring that the selected features directly serve the disease staging task. By providing evidence at the feature selection level via feature gating analysis and at the prediction contribution level via SHAP interpretation, the framework balances overall staging performance, ordinal consistency, feature selection stability, and clinical interpretability.

2.5. Experimental Setup and Statistical Analysis

To comprehensively evaluate the predictive performance, ordinal consistency, robustness, and interpretability of the Explainable Feature-Oriented Ordinal Staging Network (EFOS-Net), all experiments were conducted under the subject-level framework described in Section 2.3. Unless otherwise specified, all compared models utilized the same subject partitioning, data preprocessing, and test set evaluation strategies.

EFOS-Net was implemented using PyTorch and trained with the Adam optimizer. The initial learning rate was set to 1×10^{-3} , the batch size to 16, and the maximum number of training epochs to 200. The temperature parameter for the Concrete feature gating was set to 0.5. Early stopping was employed during training, which terminated the process if the validation loss did not decrease for 20 consecutive epochs, retaining the model parameters associated with the lowest validation loss. In each cross-validation iteration, hyperparameter tuning, model selection, and early stopping were performed exclusively based on the training and validation sets; the independent test set did not participate in any parameter estimation or model optimization. The weights of individual loss terms within the joint loss function were determined based on validation set performance and

remained fixed during the testing phase. A predefined set of 20 random seeds was used to repeat the complete subject-level five-fold cross-validation procedure. Each random seed generated an independent stratified five-fold subject partition, whereas the model architecture, preprocessing procedure, augmentation strategy, hyperparameter settings, and training protocol were kept unchanged across repetitions. The same random-seed-specific subject partitions were used for all compared models and experimental settings to ensure paired and reproducible comparisons.

Considering that H&Y staging exhibits both class imbalance and an ordinal structure, we evaluated model performance across two dimensions: overall classification ability and ordinal consistency. Classification metrics included accuracy (ACC), balanced accuracy (BACC), macro precision, macro recall, and macro F1-score, with BACC and macro F1-score serving as the primary metrics. Ordinal evaluation utilized the quadratic weighted kappa (QWK) and mean absolute error (MAE). In addition, we calculated the adjacent-stage accuracy (ASA) and severe misclassification rate (SMR), which represent the proportion of predictions within one stage of the true stage and the proportion of direct misclassifications between H&Y stages 2.0 and 3.0, respectively. We also reported the overall confusion matrix along with class-wise precision, recall, and F1-score.

To verify the effectiveness of the proposed method, we compared EFOS-Net against representative traditional machine learning models, standard multi-class neural networks, and ordinal learning models. All models utilized the same candidate gait biomarkers and subject-level data partitioning, with preprocessing, class imbalance handling, and model training strictly confined to the training data. Furthermore, we conducted ablation experiments by respectively removing the differentiable feature selection module, replacing the CORN ordinal output with a standard Softmax classifier, eliminating the distance-aware loss, and removing the CTGAN data augmentation, to evaluate the independent contribution of each component module. Additionally, various class imbalance handling strategies—including no resampling, class weighting, random oversampling, SMOTE, and CTGAN—were compared under the same evaluation framework. We also recorded the selection frequency of each gait biomarker across all 5×20 training folds generated by the repeated cross-validation procedure to assess feature-selection stability.

All performance metrics were calculated based on non-augmented real test subjects. For each random-seed repetition, predictions from the five mutually exclusive test folds were pooled so that each of the 93 subjects contributed exactly one out-of-fold prediction. All classification and ordinal performance metrics were calculated from the complete pooled out-of-fold predictions within each repetition. Final performance estimates were subsequently summarized as the mean±standard deviation across the 20 repeated complete five-fold cross-validation runs rather than across individual test folds. For inferential model comparisons, out-of-fold class probabilities obtained for each subject across the 20 cross-validation repetitions were first averaged at the subject level to obtain one consensus prediction per subject and model. Paired Bootstrap analysis was subsequently performed at the subject level using these consensus predictions. Subjects were resampled with replacement for 5,000 iterations, with identical resampled subject sets used for EFOS-Net and each comparator. Performance differences and their 95% confidence intervals were estimated from the paired Bootstrap distribution, and multiple comparisons were corrected using the Holm method. All statistical tests were two-sided, and a corrected $P < 0.05$ was considered statistically significant.

Depending on the data distribution, demographic and clinical variables were described using mean±standard deviation, median (interquartile range), or count (percentage). Between-group comparisons of continuous variables were performed using a one-way analysis of variance (ANOVA) or the Kruskal–Wallis test, whereas categorical variables were analyzed using the Pearson chi-square test or Fisher's exact test. When an overall

between-stage difference was identified, pairwise post-hoc comparisons were performed, and the resulting P values were adjusted for multiple comparisons using the Benjamini–Hochberg procedure.

Additionally, SHAP analysis was performed to interpret the model's predictions on EFOS-Net. Global feature importance rankings were derived via mean absolute SHAP values, and the influence direction of key feature values on staging predictions was analyzed. Concurrently, local SHAP analysis was performed on representative correctly and incorrectly classified cases to explain the contribution of key gait biomarkers to individual predictions. SHAP analysis and feature gating provided model interpretations at the prediction contribution level and the feature selection level, respectively; these results solely reflect the predictive associations within the model and do not serve as evidence of a causal relationship between gait biomarkers and PD progression.

3. Results

3.1. Study Cohort, Signal Quality, and Gait Characteristics

The primary staging analysis included 93 patients with PD, comprising 55 patients (59.1%) at H&Y stage 2.0, 28 (30.1%) at stage 2.5, and 10 (10.8%) at stage 3.0. The demographic characteristics of patients across different H&Y stages are presented in Table 1. Between-group comparisons revealed a statistically significant difference in age across the different H&Y stages ($F = 3.596$, $P = 0.031$), with the mean age increasing as the H&Y stage advanced. In contrast, no statistically significant differences were observed across the H&Y stages regarding sex composition ($\chi^2 = 0.573$, $P = 0.751$), height ($P = 0.283$), and body mass ($P = 0.197$).

All subjects included in the primary analysis passed the predefined signal quality control pipeline and successfully completed gait event identification and subject-level gait biomarker extraction. Following quality control, all 93 patients were retained for subsequent analyses; no subjects were further excluded due to obvious signal interruptions, critical sensor failures, or unreliable gait event identification. The subject-level features derived from valid gait cycles were utilized for subsequent gait feature comparisons, as well as for model training and evaluation.

The distributions of representative temporal, kinetic, plantar-load, gait-variability, and bilateral-asymmetry biomarkers across different H&Y stages are shown in Figure 3. Patients at different H&Y stages exhibited stage-related differences across multiple gait dimensions. With increasing H&Y stage, patients generally showed longer gait-cycle duration, lower cadence, altered stance- and swing-phase ratios, a higher double-support ratio, greater gait variability, and more pronounced bilateral asymmetry. Significant pairwise differences were observed for multiple temporal, kinetic, variability, and asymmetry measures after correction for multiple comparisons. In contrast, mean stance-phase VGRF, midfoot load ratio, and forefoot load ratio showed no significant pairwise differences across H&Y stages after multiple-comparison correction.

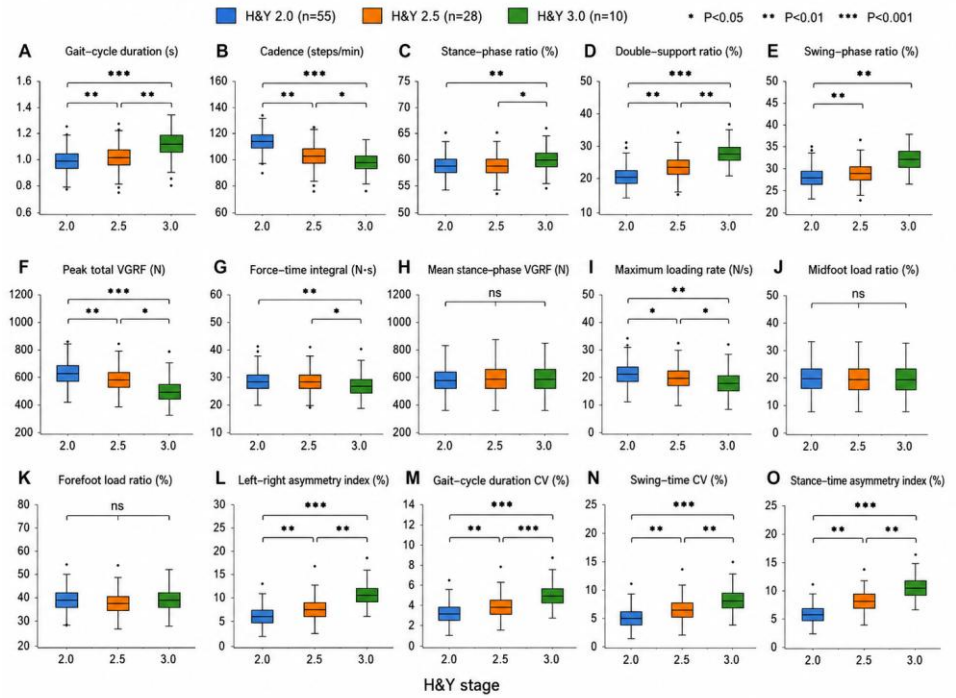


Figure 3. Distribution of gait biomarkers across H&Y stages

3.2. Primary Staging Performance of EFOS-Net

Under the repeated subject-level stratified five-fold cross-validation framework, EFOS-Net achieved favorable overall staging performance. Across the 20 complete cross-validation repetitions, the model yielded an accuracy (ACC) of $87.95 \pm 3.62\%$, a balanced accuracy (BACC) of $84.83 \pm 4.15\%$, a macro F1-score of $83.96 \pm 4.38\%$, a macro precision of 83.48% , and a macro recall of $84.83 \pm 4.15\%$. Within each repetition, predictions from the five test folds were pooled before metric calculation; therefore, the reported standard deviations represent variability across repeated complete cross-validation runs rather than variability across individual test folds.

Considering the explicit ordinal structure of H&Y staging, we further evaluated model performance using metrics related to ordinal consistency and error distance. EFOS-Net achieved a quadratic weighted kappa (QWK) of 0.85 ± 0.04 and a mean absolute error (MAE) of 0.13 ± 0.04 . Furthermore, the adjacent-stage accuracy (ASA) was $98.71 \pm 1.42\%$, and the severe misclassification rate (SMR) was strictly controlled at $1.29 \pm 1.42\%$. These results indicate that the model predictions exhibit high ordinal consistency with the clinical reference H&Y staging, and long-distance cross-stage misclassifications are relatively rare.

The class-wise recognition performance for different H&Y stages is presented in Table 3. For patients at H&Y stage 2.0, EFOS-Net achieved a precision, recall, and F1-score of $90.35 \pm 4.12\%$, $92.36 \pm 3.87\%$, and $91.28 \pm 3.54\%$, respectively. For those at stage 2.5, the corresponding metrics were $79.66 \pm 5.78\%$, $82.14 \pm 5.96\%$, and $80.76 \pm 5.31\%$. Across the repeated complete cross-validation runs, patients at H&Y stage 3.0, who constituted the smallest subgroup, achieved a precision, recall, and F1-score of $80.42 \pm 8.15\%$, $80.0 \pm 8.37\%$, and $79.83 \pm 7.26\%$, respectively. Although stage 3.0 patients accounted for only 10.8% of all PD subjects, both their recall and F1-score approached 80%, suggesting that EFOS-Net maintained good recognition performance for minority-stage patients despite the class-imbalanced cohort.

Table 3. Class-wise staging performance of EFOS-Net

H&Y stage	Precision (%)	Recall (%)	F1-score (%)
-----------	---------------	------------	--------------

H&Y 2.0	90.35±4.12	92.36±3.87	91.28±3.54
H&Y 2.5	79.66±5.78	82.14±5.96	80.76±5.31
H&Y 3.0	80.42±8.15	80.00±8.37	79.83±7.26

Note: Values are reported as mean±SD across 20 repeated complete subject-level stratified five-fold cross-validation runs. Within each repetition, predictions from the five out-of-fold test subsets were pooled before class-wise performance metrics were calculated.

To obtain a single subject-level summary of the repeated out-of-fold predictions, the predicted class probabilities for each subject were averaged across the 20 complete cross-validation repetitions. The H&Y stage with the highest mean predicted probability was then assigned as the subject-level consensus prediction, and these consensus predictions were used to construct the overall confusion matrix shown in Figure 4. Among all 93 PD patients, 82 (88.2%) were correctly staged, with 51, 23, and 8 cases accurately identified for H&Y stages 2.0, 2.5, and 3.0, respectively. Misclassifications predominantly occurred between adjacent disease stages (i.e., between stages 2.0 and 2.5, and between stages 2.5 and 3.0). Only one severe cross-stage misclassification (between stages 2.0 and 3.0) was observed among all 93 patients. Based on the aggregated out-of-fold predictions, the results for 92 out of 93 patients (98.9%) were either perfectly consistent with the clinical reference staging or deviated by only one adjacent stage. The overall confusion matrix further demonstrates that the misclassifications made by EFOS-Net were primarily concentrated in H&Y stages with similar clinical severities.

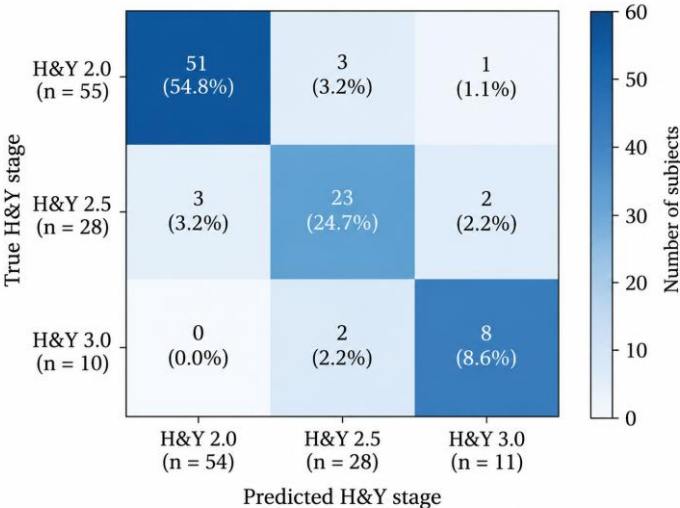


Figure 4. Confusion matrix of EFOS-Net.

3.3. Comparison with Conventional and Ordinal Benchmark Models

To evaluate the staging performance of EFOS-Net relative to different modeling strategies, we compared it with traditional machine learning classifiers, standard multi-class neural networks, and ordinal learning models under the same subject-level stratified five-fold cross-validation framework; the results are presented in Table 4. All models utilized the same subject partitioning, candidate gait biomarkers, data preprocessing pipeline, and test set evaluation strategies. We strictly confined feature standardization, class imbalance handling, and model training within the corresponding training data, while the validation and test sets consisted exclusively of non-augmented real subjects.

All model comparisons reported in this section were based on the same 20 repeated cross-validation partitions, and the reported mean±SD values represent variability across complete cross-validation repetitions.

Table 4. Comparison of staging performance between EFOS-Net and benchmark models

Model	Accuracy (%)	Balanced Accuracy (%)	Macro F1-score (%)	QWK	MAE	H&Y 3.0 Recall (%)	SMR (%)
Logistic Regression	72.04±5.76	68.31±6.12	66.85±6.28	0.61±0.07	0.31±0.06	46.00±11.40	8.6
SVM	77.42±5.02	73.18±5.47	72.84±5.31	0.69±0.06	0.27±0.05	58.00±10.44	6.45
RandomForest	75.29±5.48	71.06±5.83	70.37±5.76	0.66±0.06	0.29±0.05	54.00±10.84	7.53
XGBoost	79.35±4.63	77.11±4.94	75.26±4.87	0.73±0.05	0.23±0.05	62.00±9.75	5.38
MLP-Softmax	80.22±4.41	77.63±4.68	76.91±4.92	0.75±0.05	0.24±0.05	64.00±9.11	5.16
MLP-CORN	83.06±4.07	80.48±4.36	79.11±4.51	0.80±0.04	0.18±0.04	68.00±8.91	3.01
EFOS-Net	87.95±3.62	84.83±4.15	83.96±4.38	0.85±0.04	0.13±0.04	80.00±8.37	1.29

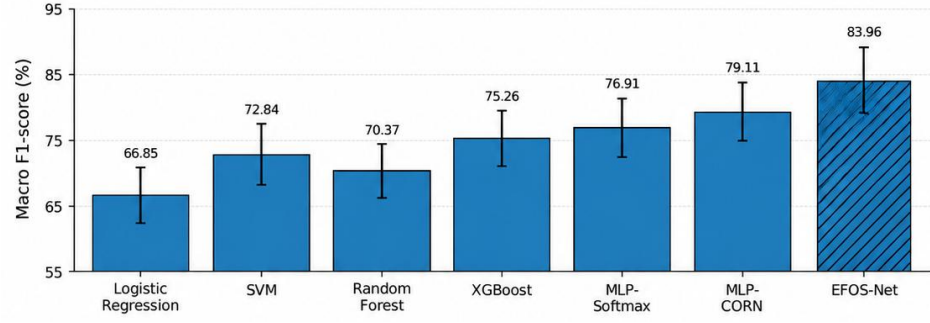
Note: Values are reported as mean±SD across 20 repeated complete subject-level stratified five-fold cross-validation runs. For each repetition, predictions from the five test folds were pooled before metric calculation. The same random-seed-specific subject partitions were used across all compared models.

Among the traditional machine learning models, XGBoost achieved the best numerical performance, yielding an ACC, BACC, and macro F1-score of 79.35±4.63%, 77.11±4.94%, and 75.26±4.87%, respectively, alongside a QWK of 0.73±0.05 and an MAE of 0.23±0.05. Among the neural network baselines, MLP-CORN with ordinal output generally outperformed MLP-Softmax, exhibiting a higher QWK and lower MAE and SMR. This suggests that ordinal modeling helps improve the consistency of H&Y staging.

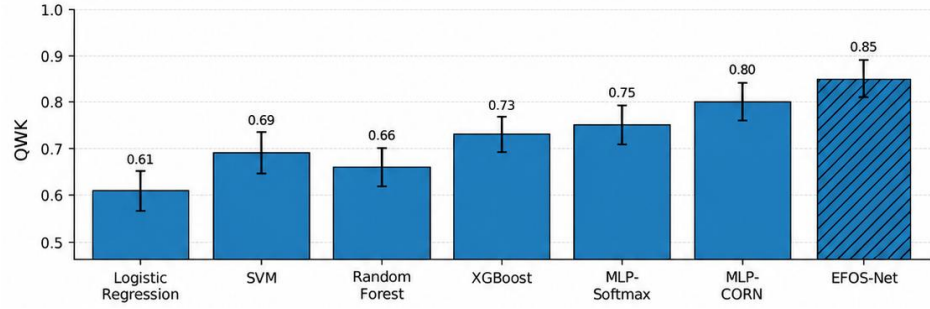
Across all compared models, EFOS-Net achieved the best overall staging performance and ordinal consistency. It reached an ACC of 87.95±3.62%, a BACC of 84.83±4.15%, and a macro F1-score of 83.96±4.38%, with a QWK of 0.85±0.04, an MAE of 0.13±0.04, and an H&Y stage 3.0 recall of 80.00±8.34%; the five-fold average SMR was only 1.29%. Compared to XGBoost (the best-performing traditional machine learning model), EFOS-Net improved the BACC and macro F1-score by 7.62 and 8.70 percentage points, respectively, and increased the stage 3.0 recall by 17.70 percentage points. Compared to MLP-CORN, EFOS-Net further improved the macro F1-score by 4.85 percentage points and the stage 3.0 recall by 11.70 percentage points, while concurrently increasing the QWK and decreasing both the MAE and SMR. These results demonstrate that EFOS-Net yields superior numerical performance in terms of overall classification, minority-stage recognition, and ordinal consistency.

Compared to the ordinal learning baseline MLP-CORN, EFOS-Net increased the ACC, BACC, and macro F1-score by 5.11, 4.25, and 4.85 percentage points, respectively; the QWK improved from 0.80 to 0.85, and the MAE decreased from 0.18 to 0.13. For patients at H&Y stage 3.0, who had the smallest sample size, the recall increased from 68.00±8.91% to 80.00±8.34% (an 12.00 percentage point gain), while the five-fold average SMR decreased from 3.01% to 1.29%. Consequently, under the current study cohort and evaluation framework, EFOS-Net achieved numerically superior results over MLP-CORN regarding overall classification performance, ordinal consistency, and H&Y stage 3.0 recognition.

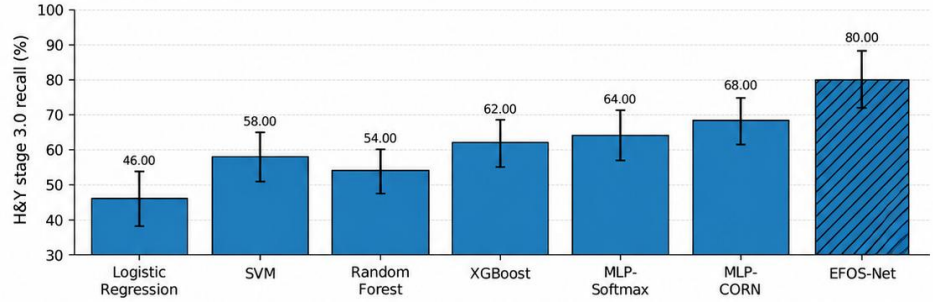
Figure 5 presents a comparison of key performance metrics across different models. Overall, EFOS-Net achieved the best numerical results among all compared models in metrics such as macro F1-score, QWK, stage 3.0 recall, and SMR. The statistical significance between models should be based on the paired Bootstrap analysis of subject-level out-of-fold predictions, with multiple comparisons corrected using the Holm method according to the predefined statistical protocol.



(A) Macro F1-score.



(B) Quadratic weighted kappa (QWK)



(C) H&Y stage 3.0 recall

Figure 5. Performance comparison of EFOS-Net and benchmark models.

3.4. Ablation and Class-Imbalance Handling Analyses

To evaluate the contributions of the primary component modules in EFOS-Net, we conducted ablation experiments under the same subject-level stratified five-fold cross-validation framework; the results are shown in Table 5 and Figure 6(A). The full EFOS-Net achieved a BACC and a macro F1-score of $84.83 \pm 4.15\%$ and $83.96 \pm 4.38\%$, respectively, with a QWK of 0.85 ± 0.04 and an MAE of 0.13 ± 0.04 .

After removing the CTGAN data augmentation, the BACC and macro F1-score dropped to $77.82 \pm 4.76\%$ and $76.98 \pm 4.91\%$, respectively, and the H&Y stage 3.0 recall decreased to $62.00 \pm 9.75\%$. When the differentiable feature selection module was removed, the macro F1-score and QWK were $79.84 \pm 4.63\%$ and 0.80 ± 0.05 , respectively. Replacing CORN with Softmax resulted in a macro F1-score of $78.92 \pm 4.71\%$, while the QWK decreased to 0.76 ± 0.05 and the MAE increased to 0.22 ± 0.05 . Removing the distance-aware loss yielded a macro F1-score and QWK of $81.47 \pm 4.52\%$ and 0.81 ± 0.04 , respectively. Overall, the removal or replacement of each primary module led to varying degrees of performance degradation.

Table 5. Ablation and class-imbalance handling analyses of EFOS-Net

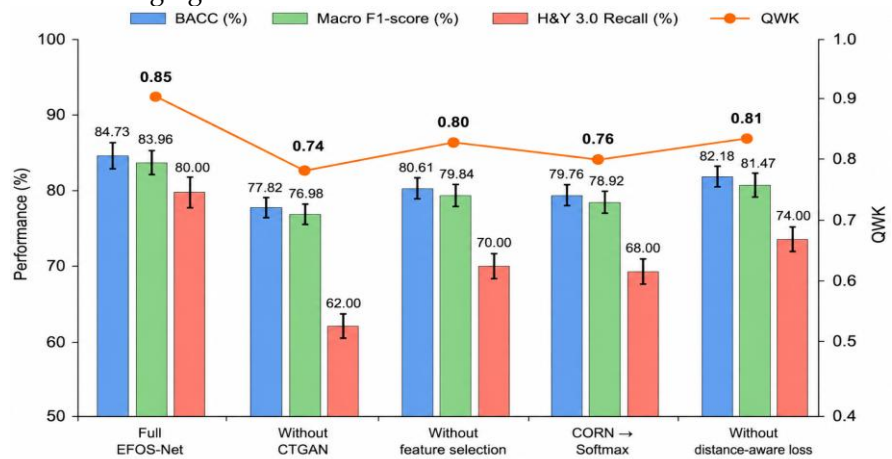
539

Analysis	Setting	BACC(%)	MacroF1-score(%)	QWK	MAE	H&Y 3.0 Recall (%)	SMR (%)
Ablation	Full EFOS-Net	84.83±4.15	83.96±4.38	0.85±0.04	0.13±0.04	80.00±8.37	1.29
	Without CTGAN	77.82±4.76	76.98±4.91	0.74±0.06	0.24±0.05	62.00±9.75	5.16
	Without feature selection	80.61±4.48	79.84±4.63	0.80±0.05	0.18±0.04	70.00±9.12	3.01
	CORN → Softmax	79.76±4.57	78.92±4.71	0.76±0.05	0.22±0.05	68.00±9.36	4.3
	Without distance-aware loss	82.18±4.31	81.47±4.52	0.81±0.04	0.18±0.04	74.00±8.77	3.01
Class imbalance	No resampling	77.82±4.76	76.98±4.91	0.74±0.06	0.24±0.05	62.00±9.75	5.16
	Class weighting	80.12±4.54	79.36±4.69	0.78±0.05	0.20±0.05	69.00±9.41	3.87
	Random oversampling	80.86±4.47	80.05±4.61	0.79±0.05	0.19±0.04	70.00±9.18	3.44
	SMOTE	82.11±4.32	81.38±4.48	0.81±0.04	0.17±0.04	72.00±8.86	2.58
	CTGAN	84.83±4.15	83.96±4.38	0.85±0.04	0.13±0.04	80.00±8.37	1.29

Note: Values are reported as mean±SD across 20 repeated complete subject-level stratified five-fold cross-validation runs. Identical random-seed-specific subject partitions were used across all ablation and class-imbalance handling settings. Within each repetition, predictions from the five out-of-fold test subsets were pooled before metric calculation.

We further compared different class imbalance handling strategies, with the results presented in Table 5 and Figure 6(B). Across the 20 repeated complete cross-validation runs, the no-resampling setting yielded an H&Y stage 3.0 recall of 62.00±9.75%; class weighting, random oversampling, and SMOTE yielded 69.00±9.41%, 70.00±9.18%, and 72.00±8.86%, respectively. CTGAN achieved a mean H&Y stage 3.0 recall of 80.0±8.37%.

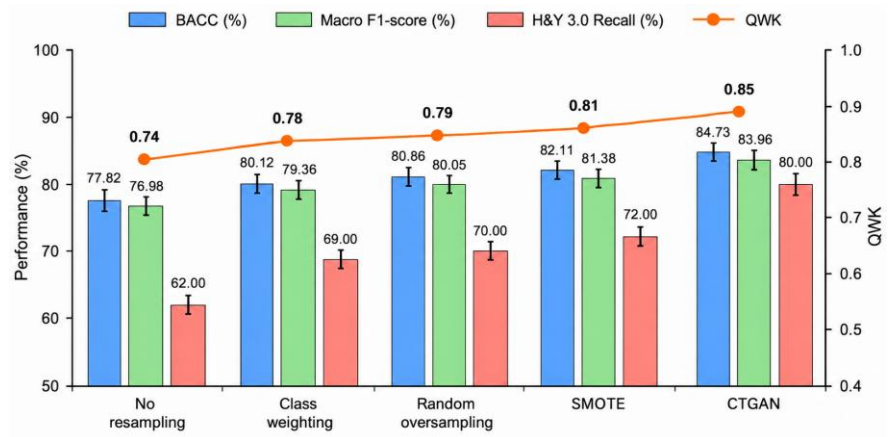
In summary, removing CTGAN caused the most noticeable decline in minority-stage recognition and overall balanced performance; eliminating feature selection primarily reduced overall classification performance; and replacing CORN with Softmax or omitting the distance-aware loss led to more prominent drops in ordinal consistency metrics. These results support the complementary roles of the primary components in EFOS-Net for automatic H&Y staging.



(A) Performance of ablation variants

555

556



(B) Effects of different class-imbalance handling strategies.

Figure 6. Ablation and class-imbalance handling analyses. (A) Performance of ablation variants.

3.5. Gait Biomarker Selection, Stability, and Model Interpretation

Under the subject-level five-fold cross-validation framework, Feature-selection stability was evaluated across all 5×20 training folds generated by the repeated cross-validation procedure. EFOS-Net identified gait biomarkers that were repeatedly retained across different subject partitions and training folds, as summarized in Table 6. The selected features cover multiple domains, including temporal, kinetic, plantar-load distribution, bilateral symmetry, and gait variability. Among them, the double-support phase, gait-cycle duration coefficient of variation, swing-time asymmetry index, stance-time asymmetry index, and cadence exhibited high repeated selection frequencies across different training folds. This suggests that these indices contribute stably to distinguishing different H&Y stages.

Table 6. Stability and gate weights of gait biomarkers selected by EFOS-Net

Rank	Selected gait biomarker	Selection frequency (%)	Mean gate weight
1	Double-support phase (%)	100	0.91
2	Gait-cycle duration coefficient of variation	100	0.88
3	Swing-time asymmetry index	100	0.84
4	Stance-time asymmetry index	100	0.81
5	Cadence	100	0.79
6	Double-support time coefficient of variation	100	0.76
7	Peak VGRF asymmetry index	100	0.73
8	Stance-time coefficient of variation	80	0.69
9	Swing-time coefficient of variation	80	0.66
10	Force-time integral asymmetry index	80	0.63
11	Forefoot-to-rearfoot load ratio	80	0.61
12	Peak total VGRF coefficient of variation	80	0.58
13	Mean stance-phase ratio (%)	80	0.55
14	Mean swing-phase ratio (%)	80	0.52
15	Relative mediolateral COP fluctuation	60	0.48
16	Push-off peak VGRF	60	0.45
17	Maximum loading rate	60	0.42
18	Regional plantar-load asymmetry index	60	0.39

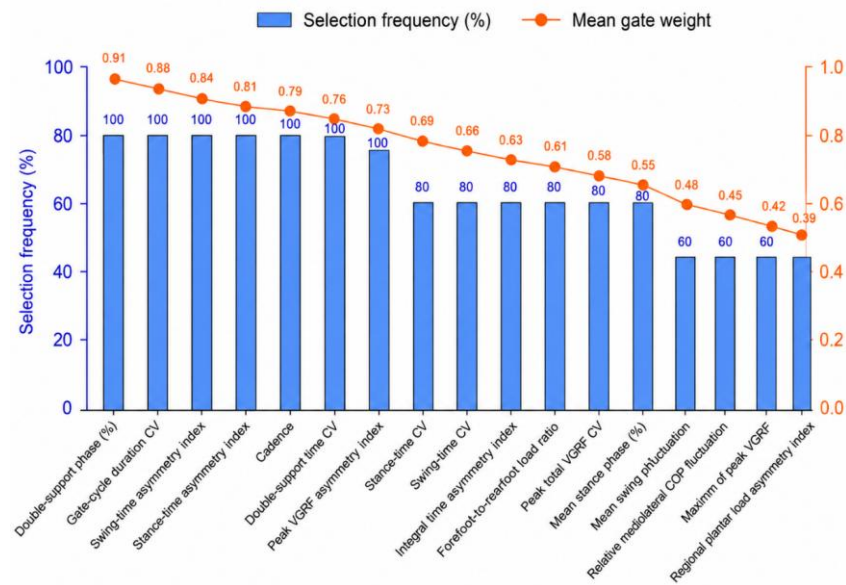
Further analysis of the feature gating results revealed that the double-support phase, gait-cycle duration coefficient of variation, swing-time asymmetry index, stance-time asymmetry index, and cadence possessed higher mean gate weights (Figure 7A). These features reflect gait phase structure, cross-cycle stability, and bilateral coordination capacity, respectively. We note that the selection frequency reflects the stability of a feature being repeatedly retained across different training folds, whereas the gate weight represents the relative selection intensity learned during model training; both metrics describe the importance of candidate gait biomarkers from different perspectives.

To evaluate the contribution of the selected features to model predictions, SHAP analysis was performed on out-of-fold subjects across the repeated five-fold cross-validation runs, and the resulting attribution values were aggregated to obtain global feature-importance estimates. The double-support phase, gait-cycle duration coefficient of variation, swing-time asymmetry index, stance-time asymmetry index, and cadence also ranked at the forefront regarding mean absolute SHAP values (Figure 7B). Among all selected features, the mean gate weight demonstrated a strong positive correlation with the mean absolute SHAP value (Spearman $\rho = 1$, $P < 0.001$). This indicates that the importance ranking learned by the feature selection module is generally consistent with the predictive contribution of the model on out-of-fold subjects.

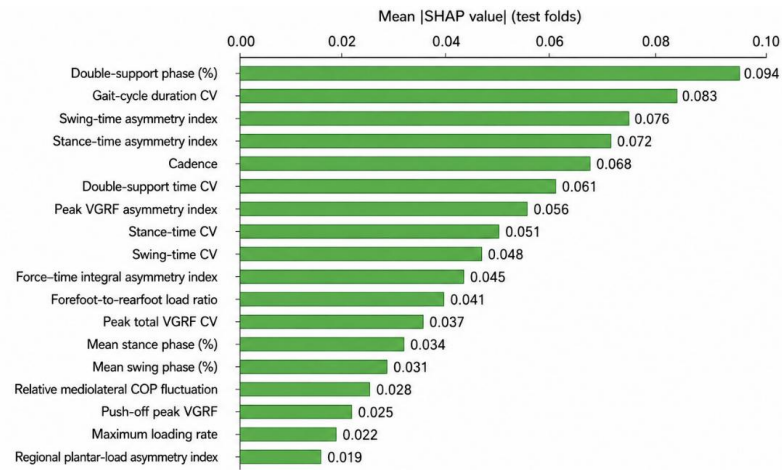
Different features also exhibited certain patterns regarding their influence on the model's prediction direction. A higher double-support phase ratio, greater gait cycle and phase variability, and more pronounced bilateral asymmetry generally tended to increase the model's contribution toward predicting a higher H&Y stage. Conversely, a higher cadence, along with lower gait variability and asymmetry, was generally associated with model outputs of lower disease stages. These results reflect the internal contribution of features to the EFOS-Net staging predictions and do not imply a causal relationship between these gait indices and PD progression.

To assess whether the model excessively relies on a few high-weight features, we performed a feature removal analysis (Figure 7C). After removing the single highest-ranking feature, the macro F1-score dropped from 83.96% to 82.71%; removing the top 3 features decreased it to 80.92%; and removing the top 5 features further reduced it to 78.64%. Although model performance gradually declined with the sequential removal of high-ranking features, removing the single highest-ranking feature only caused a 1.25 percentage point drop in the macro F1-score. This demonstrates that the model's staging performance is jointly supported by multiple complementary gait biomarkers rather than relying entirely on a single feature.

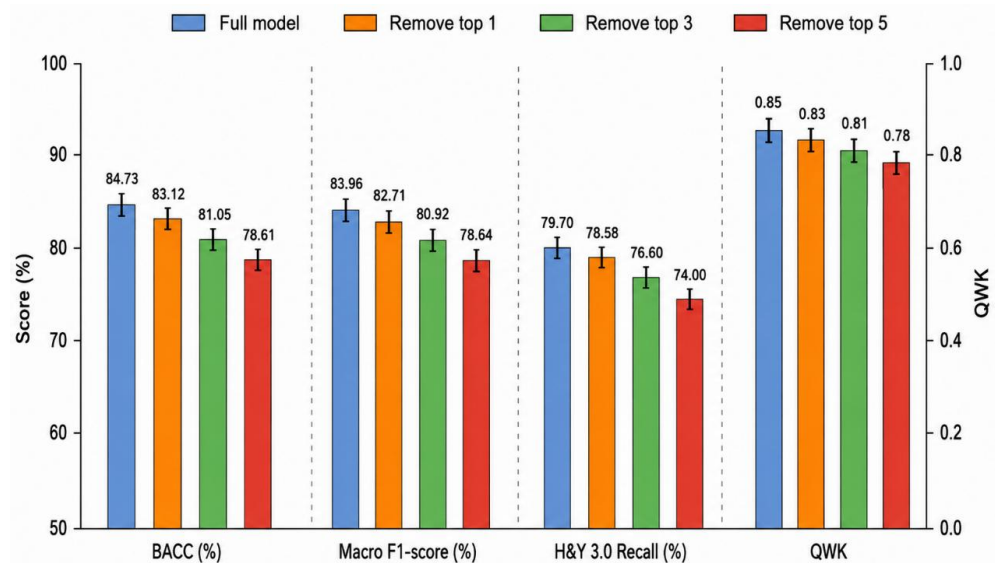
In summary, the primary gait biomarkers selected by EFOS-Net cover various domains, including gait phase structure, cross-cycle variability, kinetic load, and bilateral coordination. The feature selection frequency and gate weights show good consistency with out-of-fold SHAP importance, and the feature removal analysis further demonstrates that model predictions are not determined by a single high-weight variable. These results provide internal evidence for the stability and interpretability of the EFOS-Net staging decisions.



(A) Stability and Learned Gate Weights of Selected Gait Biomarkers.



(B) Consistency Between Gate-Weight Ranking and SHAP Importance.



(C) Sensitivity to Removal of High-Ranking Gait Biomarkers.

4. Discussion

We developed EFOS-Net for the automatic H&Y staging of PD, integrating plantar-force gait biomarker screening, intra-training class imbalance handling, and ordinal learning into a unified framework. Subject-level five-fold cross-validation results demonstrated that EFOS-Net effectively distinguished patients at H&Y stages 2.0, 2.5, and 3.0, achieving favorable results in overall classification performance, minority-stage recognition, and ordinal consistency. Compared to traditional machine learning models and neural network baselines, EFOS-Net yielded superior numerical performance across metrics such as BACC, macro F1-score, QWK, MAE, and stage 3.0 recall, alongside fewer severe cross-stage misclassifications. This indicates that under the current cohort and evaluation framework, simultaneously considering feature screening, class imbalance, and the ordinal attribute of H&Y staging helps achieve more balanced predictions that align with the hierarchical structure of disease severity.

H&Y staging possesses a clear ordinal attribute, where adjacent stages reflect progressive changes in disease severity rather than mutually independent categories. Consequently, adjacent-stage misclassifications carry different clinical implications compared to direct cross-stage misclassifications between H&Y stages 2.0 and 3.0. We utilized CORN to model the sequential relationship between stages and applied a distance-aware loss to strengthen the constraints on long-distance misclassifications. In baseline comparisons, MLP-CORN exhibited a higher QWK and lower MAE and SMR than MLP-Softmax. Furthermore, ablation experiments showed that replacing CORN with Softmax or removing the distance-aware loss led to varying degrees of decline in ordinal-related metrics. This demonstrates the necessity of evaluating rank consistency and error distance, in addition to conventional classification metrics, for the automatic H&Y staging task.

Class imbalance primarily affected the recognition of patients at H&Y stage 3.0. The current cohort contained only 10 patients at stage 3.0, notably fewer than those at stages 2.0 and 2.5. Compared to no resampling, class weighting, random oversampling, and SMOTE, CTGAN achieved better minority-stage recognition and overall balanced performance on this dataset. Conversely, removing CTGAN decreased both the stage 3.0 recall and the overall performance. This indicates that intra-training data augmentation helps mitigate the recognition bias caused by insufficient minority-stage samples. However, because CTGAN was utilized exclusively during the training phase without increasing the number of real stage 3.0 samples, it cannot substitute for real clinical cases; these results require further validation using independent data containing more advanced-stage patients.

EFOS-Net ultimately selected 18 gait biomarkers covering multiple domains, including gait phase structure, cross-cycle variability, bilateral coordination, kinetic load, and plantar-load distribution. Among these, the double-support phase, gait-cycle duration coefficient of variation, swing-time asymmetry index, stance-time asymmetry index, and cadence exhibited higher selection frequencies and gate weights, ranking highly in the SHAP analysis as well. A higher double-support phase ratio, greater gait variability, and more pronounced bilateral asymmetry generally tended to push the model toward predicting higher H&Y stages, aligning with the clinical manifestation of declining gait stability and bilateral coordination as PD severity increases. Meanwhile, the multivariate model retained certain features that lacked significant between-group differences in univariate analyses, suggesting that H&Y stage recognition likely relies on joint information across different gait dimensions rather than a single index.

The feature selection results showed good consistency with the model interpretation results. Mean gate weights demonstrated a strong positive correlation with out-of-fold

mean absolute SHAP values, and model performance gradually declined upon the removal of high-ranking features. Removing the single highest-ranking feature caused only a limited performance drop, whereas the sequential removal of multiple high-ranking features led to a more pronounced decline; this demonstrates that model predictions are jointly supported by multiple complementary gait biomarkers rather than relying on a single variable. We note that gate weights and SHAP values reflect feature contributions to model predictions and cannot infer causal relationships between the corresponding gait indices and PD progression. As a result, the features identified in our study should be interpreted as candidate digital gait biomarkers associated with disease severity.

We adopted subject-level data partitioning to reduce data leakage caused by allocating multiple gait cycles from the same patient into both the training and test sets simultaneously. Raw signals, gait cycles, and derived features from a given subject were strictly maintained within the same data subset, while operations including standardization, class augmentation, and feature selection were performed exclusively within the training data. This design helps obtain a performance evaluation that more closely approximates prediction scenarios for unseen patients. On the other hand, patient ages exhibited statistically significant differences across different H&Y stages, with the mean age increasing as the H&Y stage advanced. Consequently, age-related gait changes may still constitute a potential confounding factor. This issue should be further evaluated through age matching, covariate adjustment, or independent cohorts in future studies.

Plantar VGRF can provide objective, quantifiable gait information, while EFOS-Net outputs H&Y stage predictions alongside corresponding gait biomarker contribution insights. Accordingly, this method is better suited as an objective auxiliary tool for clinical H&Y staging rather than a replacement for clinical evaluation. It should be noted that we evaluated the cross-sectional disease stage recognition capability, which cannot be directly applied to predict individual disease progression, treatment response, or long-term functional changes. Such applications still require further validation using longitudinal repeated-measures data.

Our study has several limitations. First, the model was developed based on a single public database and evaluated primarily using internal cross-validation without independent external validation; thus, its generalization capability across different centers, devices, and patient populations remains unclear. Second, the existing data only covers H&Y stages 2.0, 2.5, and 3.0, with a limited sample size at stage 3.0, meaning the current model does not represent the complete range of H&Y staging. Third, age differences exist between different H&Y stages, and age along with other unmeasured clinical factors might exert confounding effects on gait phenotypes. Fourth, we utilized only plantar VGRF to construct gait biomarkers, without evaluating the incremental value of information from inertial sensors, kinematics, electromyography, or clinical scales. Finally, by adopting cross-sectional H&Y staging as the prediction target, we could not assess the model's sensitivity to individual disease progression. Future studies should conduct validation in cohorts containing a more complete range of disease stages, larger scale real-world patients, and multicenter independent data, while further incorporating longitudinal follow-up and multimodal data to evaluate the stability and practical application value of the model.

5. Conclusions

Our study developed EFOS-Net for the automatic Hoehn–Yahr (H&Y) staging of Parkinson's disease, integrating plantar-force gait biomarker screening, intra-training class imbalance handling, and ordinal learning into a unified framework. Based on subject-level five-fold cross-validation, EFOS-Net enabled automatic identification of H&Y stages 2.0, 2.5, and 3.0, achieving favorable performance in overall classification, minority-

stage recognition, and ordinal consistency, while reducing severe cross-stage misclassifications.

The key gait biomarkers selected by the model covered multiple dimensions, including gait phase structure, cross-cycle variability, and bilateral coordination, with the double-support phase, gait-cycle coefficient of variation, swing-time asymmetry index, stance-time asymmetry index, and cadence exhibiting high selection stability and predictive contribution. The feature gating and SHAP analyses were generally consistent, indicating that H&Y stage recognition relies on the joint information of multiple complementary gait biomarkers and providing interpretable evidence for the model's staging decisions.

Currently, our study only covered patients at H&Y stages 2.0–3.0 and lacked independent external validation; therefore, EFOS-Net is better suited as an objective auxiliary tool for H&Y staging based on plantar-force gait information, rather than a replacement for clinical H&Y evaluation. Future studies should validate the model on larger real-world cohorts encompassing a broader range of disease stages, as well as on multicenter and cross-device data, and further incorporate longitudinal follow-up to assess its generalization capability, stability, and application value in disease severity assessment and long-term monitoring.

Abbreviations

The following abbreviations are used in this manuscript:

BACC	Balanced Accuracy
CORN	Conditional Ordinal Regression for Neural Networks
CTGAN	Conditional Tabular Generative Adversarial Network
EFOS-Net	Explainable Feature-Oriented Ordinal Staging Network
H&Y	Hoehn–Yahr
MAE	Mean Absolute Error
PD	Parkinson's Disease
QWK	Quadratic Weighted Kappa
SHAP	SHapley Additive exPlanations
SMR	Severe Misclassification Rate
VGRF	Vertical Ground Reaction Force

References

1. Su D, Cui Y, He C, et al. Projections for prevalence of Parkinson's disease and its driving factors in 195 countries and territories to 2050: modelling study of Global Burden of Disease Study 2021. *BMJ*, 2025;388:e080952. DOI: 10.1136/bmj-2024-080952
2. HOEHN M M, YAHR M D. Parkinsonism: Onset, progression, and mortality[J]. *Neurology*, 1967, 17(5): 427-442. DOI: 10.1212/WNL.17.5.427.
3. GOETZ C G, POEWE W, RASCOL O, et al. Movement Disorder Society Task Force report on the Hoehn and Yahr staging scale: Status and recommendations[J]. *Movement Disorders*, 2004, 19(9): 1020-1028. DOI: 10.1002/mds.20213.
4. BONACINA D, TOSATTO D, UGOLINI A, et al. Spatiotemporal, kinematic and kinetic gait characteristics in Parkinson's disease compared to healthy individuals: A systematic review with meta-analysis[J]. *Clinical Biomechanics*, 2024, 120:106359. DOI: 10.1016/j.clinbiomech.2024.106359.
5. HAUSDORFF J M. Gait dynamics in Parkinson's disease: Common and distinct behavior among stride length, gait variability, and fractal-like scaling[J]. *Chaos*, 2009, 19(2): 026113. DOI: 10.1063/1.3147408.
6. DI BIASE L, PECORARO P M, PECORARO G, et al. Machine learning and wearable sensors for automated Parkinson's disease diagnosis aid: a systematic review[J]. *Journal of Neurology*, 2024, 271(10):6452–6470. DOI: 10.1007/s00415-024-12611-x.
7. Sapienza S, Tsurkalenko O, Giraitis M, et al. Assessing the clinical utility of inertial sensors for home monitoring in Parkinson's disease: a comprehensive review. *npj Parkinson's Disease*, 2024;10:161. DOI: 10.1038/s41531-024-00755-6

-
8. SIGCHA L, BORZÌ L, AMATO F, et al. Deep learning and wearable sensors for the diagnosis and monitoring of Parkinson's disease: A systematic review[J]. *Expert Systems with Applications*, 2023, 229: 120541. DOI: 10.1016/j.eswa.2023.120541. 756 757
 9. Sun Y, Cheng Y, You Y, et al. A novel plantar pressure analysis method to signify gait dynamics in Parkinson's disease. *Mathematical Biosciences and Engineering*, 2023;20(8):13474–13490. DOI: 10.3934/mbe.2023601. 758 759
 10. Youssef H, et al. Can wearable sensor based measures of gait accurately reflect Parkinson's disease severity? A systematic review and meta-analysis. *Parkinsonism & Related Disorders*. 2026;142:108108. DOI: 10.1016/j.parkreldis.2025.108108. 760 761
 11. GIARDINI M, TURCATO A M, ARCOLIN I, et al. Vertical ground reaction forces in Parkinson's disease: A speed-matched comparative analysis with healthy subjects[J]. *Sensors*, 2024, 24(1): 179. DOI: 10.3390/s24010179. 762 763
 12. PERRY J, BURNFIELD J M. *Gait analysis: Normal and pathological function*[M]. 2nd ed. Thorofare: SLACK Incorporated, 2010. 764
 13. WHITTLE M W. *Gait analysis: An introduction*[M]. 4th ed. Edinburgh: Butterworth-Heinemann, 2007. 765
 14. ORLIN M N, MCPOIL T G. Plantar pressure assessment[J]. *Physical Therapy*, 2000, 80(4): 399-409. DOI: 10.1093/ptj/80.4.399. 766
 15. Navita, Mittal P, Sharma YK, et al. Gait-based Parkinson's disease diagnosis and severity classification using force sensors and machine learning. *Scientific Reports*, 2025;15:328. DOI: 10.1038/s41598-024-83357-9. 767 768
 16. Franco A, Russo M, Amboni M, et al. The Role of Deep Learning and Gait Analysis in Parkinson's Disease: A Systematic Review. *Sensors*, 2024;24(18):5957. DOI: 10.3390/s24185957. 769 770
 17. HAUSDORFF J M, PURDON P, PENG C K. Gait in aging and disease database: Version 1.0.0[DB/OL]. Cambridge: PhysioNet, 2000[2026-08-01]. 771 772
 18. HAUSDORFF J M. Gait in neurodegenerative disease database: Version 1.0.0[DB/OL]. Cambridge: PhysioNet, 2000[2026-08-01]. 773 774
 19. SHIDA T K F, COSTA T M, DE OLIVEIRA C E N, et al. A public data set of walking full-body kinematics and kinetics in individuals with Parkinson's disease[J]. *Frontiers in Neuroscience*, 2023, 17: 992585. DOI: 10.3389/fnins.2023.992585. 775 776
 20. VARGHESE J, BRENNER A, PLAGWITZ L, et al. PADS—Parkinsons Disease Smartwatch dataset: Version 1.0.0[DB/OL]. PhysioNet, 2024[2026-08-01]. DOI: 10.13026/m0w9-zx22. 777 778
 21. U.S. FOOD AND DRUG ADMINISTRATION. WearGait-PD: Wearables Dataset for Gait in Parkinson's Disease and Age-Matched Controls[DB/OL].(2024-07-11)[2026-08-01]. <https://doi.org/10.7303/syn52540892>. 779 780
 22. HAUSDORFF J M. Gait in Parkinson's disease: Version 1.0.0[DB/OL]. Cambridge: PhysioNet, 2008[2026-08-01]. DOI: 10.13026/C2KW2B. 781 782
 23. EL MAACHI I, BILODEAU G A, BOUACHIR W. Deep 1D-Convnet for accurate Parkinson disease detection and severity prediction from gait[J]. *Expert Systems with Applications*, 2020, 143: 113075. DOI: 10.1016/j.eswa.2019.113075. 783 784
 24. MUÑOZ-MATA B G, DORANTES-MÉNDEZ G, PIÑA-RAMÍREZ O. Classification of Parkinson's disease severity using gait stance signals in a spatiotemporal deep learning classifier[J]. *Medical & Biological Engineering & Computing*, 2024, 62(11): 3493–3506. DOI: 10.1007/s11517-024-03148-2. 785 786 787
 25. LI S Y, CHEN M J, YANG M, et al. TDK-VAMNet: A dual-task deep learning framework for Parkinson's disease staging prediction based on 3D feature selection[J]. *Measurement Science and Technology*, 2025, 36: 065118. DOI: 10.1088/1361-6501/ade3f6. 788 789 790
 26. YU Y H, YE S, XIA Z H, et al. A dual-task gait fusion framework for classifying Parkinson's disease severity from wearable sensor data[J]. *Diagnostics*, 2026, 16(15): 2386. DOI: 10.3390/diagnostics16152386. 791 792
 27. NIU Z X, ZHOU M, WANG L, et al. Ordinal regression with multiple output CNN for age estimation[C]//*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway: IEEE, 2016: 4920-4928. DOI: 10.1109/CVPR.2016.532. 793 794
 28. CAO W Z, MIRJALILI V, RASCHKA S. Rank consistent ordinal regression for neural networks with application to age estimation[J]. *Pattern Recognition Letters*, 2020, 140: 325-331. DOI: 10.1016/j.patrec.2020.11.008. 795 796
 29. XU L, SKOULARIDOU M, CUESTA-INFANTE A, VEERAMACHANENI K. Modeling tabular data using conditional GAN[C]//*Advances in Neural Information Processing Systems*. 2019, 32. 797 798
 30. LUNDBERG S M, LEE S I. A unified approach to interpreting model predictions[C]//*Advances in Neural Information Processing Systems*. Red Hook: Curran Associates, 2017, 30: 4765-4774. 799 800
 31. ZENI J A Jr, RICHARDS J G, HIGGINSON J S. Two simple methods for determining gait events during treadmill and over-ground walking using kinematic data[J]. *Gait & Posture*, 2008, 27(4): 710–714. DOI: 10.1016/j.gaitpost.2007.07.007. 801 802
 32. FARASHI S. Analysis of the stance phase of the gait cycle in Parkinson's disease and its potency for Parkinson's disease discrimination[J]. *Journal of Biomechanics*, 2021, 129: 110818. DOI: 10.1016/j.jbiomech.2021.110818. 803 804

-
33. WANG X, XU X, ZHAO Z, et al. VGRF Signal-Based Gait Analysis for Parkinson's Disease Detection: A Multi-Scale Directed Graph Neural Network Approach[J]. IEEE Journal of Biomedical and Health Informatics, 2026, 30(1): 100–112. DOI: 10.1109/JBHI.2025.3589772.
34. LI S Y, YANG M, ZHANG X Y, et al. DM-PAFNet: A dual-stage deep learning framework for Parkinson's disease staging prediction via multi-modal MRI and clinical prior fusion[J]. Measurement Science and Technology, 2026, 37: 205701. DOI: 10.1088/1361-6501/ae5e36.
35. ROSENBLATT M, TEJAVIBULYA L, JIANG R T, et al. Data leakage inflates prediction performance in connectome-based machine learning models[J]. Nature Communications, 2024, 15(1): 1829. DOI: 10.1038/s41467-024-46150-w.
36. CHAWLA N V, BOWYER K W, HALL L O, et al. SMOTE: Synthetic minority over-sampling technique[J]. Journal of Artificial Intelligence Research, 2002, 16: 321–357. DOI: 10.1613/jair.953.

Author Contributions: Conceptualization, M.Y. and M.C.; methodology, X.X. and M.C.; software, X.X.; validation, M.Y., Z.L. and W.Y.; formal analysis, X.X.; investigation, M.Y.; data curation, M.Y. and X.X.; visualization, X.X.; writing—original draft preparation, X.X. and M.Y.; writing—review and editing, M.Y., Z.L., W.Y. and M.C.; supervision, M.C.; project administration, M.C.; funding acquisition, M.Y. and M.C. All authors have read and agreed to the published version of the manuscript.

Ethics Statement

This study was a secondary analysis of publicly available data from the Gait in Parkinson's Disease Database on PhysioNet. No new participants were recruited, and no additional experiments or clinical interventions were conducted. Ethical approval and informed consent for the original data collection were reported in the source studies. No additional ethics approval or informed consent was required for the present secondary analysis.

Funding: the open research fund of The Sichuan Province Key Laboratory of Intelligent Medicine and Health Big Data under Grant No. ZNYX2507, the Collaborative Innovation Project of Zigong Institute of Medical Technology under Grant No. 2025YJYZ0308.

Institutional Review Board Statement: Not applicable. This study was a secondary analysis of a publicly available dataset distributed without direct participant identifiers. The authors did not recruit participants, collect new human subject data, or access any direct personal identifiers. The ethical approval and informed consent procedures related to the original data collection are described by the dataset providers and the original study.

Informed Consent Statement: Not applicable. No new participant contact or data collection was performed, and no identifiable participant information was available to the authors.

Data Availability Statement: The data analyzed in this study are publicly available from the PhysioNet Gait in Parkinson's Disease database at <https://physionet.org/content/gaitpdb/1.0.0/>. This study is a secondary analysis of publicly available data, and no new datasets were generated during the current study.

Acknowledgements: The authors used ChatGPT (OpenAI) solely to assist in the design and refinement of schematic illustrations and to improve the clarity and readability of the manuscript. The authors critically reviewed and edited all AI-assisted content and take full responsibility for the accuracy, originality, and integrity of all figures and the overall content of the manuscript.

Conflicts of Interest: The authors declare no conflicts of interest. The authors alone are responsible for the study design, data analysis, interpretation of the results, manuscript preparation, and the decision to submit the manuscript for publication.