Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Interpretable Multiclass Parkinson's Disease Classification from Single-Camera Video Using Kinematic and Nonlinear Dynamic Features

Ranjan Das
University of California, Merced
Navdeep Kaur
Guru Nanak Dev Engineering College
Manish Yadav
VIT Bhopal University
Ingita Dey Munshi
Jawaharlal Nehru University
Karsimran Singh
Tatras Data Services Pvt. Ltd.
Sachin Goyal
sgoyal12@ucmerced.edu
University of California, Merced
Sukhjit Singh Sehra
Wilfrid Laurier University

Article

Keywords: Parkinson's disease classification, Nonlinear dynamic features, Pose estimation, Machine learning models, Vision-based analysis, Disease progression monitoring

Posted Date: August 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10517590/v1

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

Additional Declarations: No competing interests reported.

Interpretable Multiclass Parkinson's Disease Classification from Single-Camera Video Using Kinematic and Nonlinear Dynamic Features

Ranjan Das[a,b], Navdeep Kaur[c], Manish Yadav[d], Ingita Dey Munshi[e], Karsimran Singh[g], Sachin Goyal[a,b,*], Sukhjit Singh Sehra[f,*]

[a]*Department of Mechanical Engineering, University of California, Merced, CA, USA*
[b]*Health Sciences Research Institute, University of California, Merced, CA, USA*
[c]*Department of Information Technology, Guru Nanak Dev Engineering College, Punjab, India*
[d]*School of Computing Science Engineering and Artificial Intelligence, VIT Bhopal University, Madhya Pradesh, India*
[e]*School of Computational and Integrative Sciences, Jawaharlal Nehru University, Delhi, India*
[f]*Department of Computer Science and Physics, Wilfrid Laurier University, Ontario, Canada*
[g]*Tatras Data Services Pvt Ltd, Delhi, India*

Abstract

Parkinson's disease (PD) is a progressive neurological disorder whose motor symptoms, especially gait impairment, change across disease stages and remain difficult to measure at scale. Motion capture and force plates provide accurate gait analysis but are expensive and impractical for routine clinical use. A major challenge is the lack of large, diverse public datasets, as most clinical gait datasets are proprietary and publicly available datasets are small, imbalanced with limited diversity. To address these challenges, we propose a single-camera, markerless framework for multiclass Parkinsonian gait classification that combines kinematic and nonlinear dynamic features from video-based pose estimation. Unlike prior work focused on binary PD detection or wearable sensors, our approach enables stage-aware gait assessment from standard video. To ensure fair evaluation with limited data, we use a strictly nested cross-validation pipeline where feature selection, preprocessing, and class balancing are performed only within repeated stratified training folds, preventing information leakage. An SVM with RBF kernel achieved the highest mean cross-validation accuracy (0.8366), while a Neural Network achieved similar accuracy (0.8338) with lower variation (std = 0.017). On the held-out test set, both models reached 0.9375 accuracy and 0.8222 macro-F1. A Random Forest provided an interpretable baseline with 0.8750 accuracy. Among the 19 selected features, 16 were nonlinear-dynamic measures, suggesting gait variability in this dataset is better captured by nonlinear dynamics than by kinematics alone, though larger studies are needed. These results demonstrate accurate, interpretable multiclass classification of Parkinsonian gait and provide a foundation for scalable, AI-assisted gait assessment in neuromotor healthcare.

Keywords: Parkinson's disease classification, Nonlinear dynamic features, Pose estimation, Machine learning models, Vision-based analysis, Disease progression monitoring

1. Introduction

Human gait provides a unique window into the interplay between neural control, biomechanics, and cognition. In everyday life, walking appears effortless, yet subtle disruptions in rhythm or variability can reveal underlying pathology. Parkinson's disease (PD), a progressive neurodegenerative disorder, makes this abundantly clear: hallmark motor symptoms such as bradykinesia, rigidity, tremor, and postural instability manifest vividly in how patients move, with gait disturbances, shuffling, shortened stride length, freezing episodes, and impaired balance among the most debilitating consequences (World Health Organization, 2024; Yin et al., 2024a). Clinical scales such as the UPDRS and MDS-UPDRS remain the standard for motor assessment yet are inherently subjective and insensitive

*Corresponding author
Email addresses: rdas4@ucmerced.edu (Ranjan Das), navdeep2539007@gndec.ac.in (Navdeep Kaur), manish2022@vitbhopal.ac.in (Manish Yadav), ingita95_sit@jnu.ac.in (Ingita Dey Munshi), karsimran@tatrasdata.com (Karsimran Singh), sgoyal2@ucmerced.edu (Sachin Goyal), ssehra@wlu.ca (Sukhjit Singh Sehra)
URL: https://sites.ucmerced.edu/sgoyal (Sachin Goyal)

to subtle gait alterations, motivating the search for objective, quantitative approaches (Goetz et al., 2003; Cho et al., 2009; di Biase et al., 2020).

Early efforts centered on wearable sensors and force plates. Classifiers including SVM, decision trees, and k-nearest neighbor achieved recall rates exceeding 95% for PD detection using foot-mounted sensors (Pratihar & Sankar, 2024), while a single IMU at the lower back preserved classification accuracy above 80% with as few as seven gait traits, reducing hardware complexity without sacrificing interpretability (Sigcha et al., 2023). Parallel to these, vision-based methods emerged as a non-contact alternative: deep learning and graph-based approaches have demonstrated ROC-AUC scores up to 95% internally, though many still require specialized instrumentation or large labeled datasets to generalize (Cho et al., 2009; Kour et al., 2022, 2024; Zhong & Ng, 2023).

Markerless video-based gait analysis has since been validated across diverse clinical populations including PD, stroke, cerebral palsy, and multiple sclerosis, confirming that standard RGB cameras can reliably capture spatiotemporal gait features with clinical-grade accuracy (Olleac et al., 2025; Bogachev et al., 2023). Despite this progress, most video-based pipelines remain limited to linear spatiotemporal features or black-box deep learning outputs, with early technical efforts in silhouette matching, multimodal feature extraction, and health indicator estimation representing incremental advances rather than fully interpretable, clinically actionable systems (Chan, 2008; Ye et al., 2020; Smyrnakis et al., 2024; Liao et al., 2021).

Another practical challenge is the limited availability of suitable datasets for developing and evaluating multiclass Parkinsonian gait classifiers. Although several large gait datasets exist, many are proprietary and therefore inaccessible to the broader research community. Consequently, most published studies rely on publicly available datasets, which are typically small, imbalanced, and limited in demographic and clinical diversity. These constraints reduce statistical power, hinder model generalization, and increase the risk of optimistic performance estimates, particularly for multiclass severity classification. They also underscore the importance of evaluation protocols that maximize the reliability of results under limited-data conditions.

A complementary body of work has examined nonlinear gait dynamics. Metrics such as Hurst's rescaled range, Higuchi's fractal dimension, and approximate entropy reveal that Parkinsonian gait is slower, shorter in stride, and less predictable—particularly under dual-task conditions (Pérez-Toro et al., 2020; Lukšys & Griškevičius, 2022; Udhayakumar et al., 2025; Workman & Thrasher, 2024). Approximate entropy has emerged as a sensitive marker of altered cognitive-motor integration in PD (Sibley et al., 2021). These metrics function as validated empirical biomarkers of gait variability, not mechanistic explanations: their clinical utility lies in statistical correlation with PD severity, and their integration into an interpretable multiclass classification pipeline remains largely unexplored.

Taken together, the reviewed literature reveals three persistent gaps. First, vision-based PD gait studies have predominantly framed the problem as binary detection (PD vs. healthy), leaving severity staging a more clinically actionable task largely unaddressed. Second, while nonlinear dynamic metrics have been validated as sensitive gait biomarkers, their integration into an interpretable, evaluation-rigorous multiclass classification pipeline has not been demonstrated. Third, the limited availability of accessible PD gait datasets with large collections often remaining proprietary and public datasets typically being small, imbalanced, and limited in diversity makes reliable multiclass learning difficult and increases the risk of optimistic performance estimates if evaluation protocols are not rigorously designed. Few published studies enforce strict within-fold preprocessing to mitigate this issue. The present work addresses all three gaps simultaneously.

We address these limitations with a single-camera, markerless framework that extracts kinematic and nonlinear dynamic features from pose-estimated skeletal keypoints, without requiring specialized hardware or body-mounted sensors. Unlike prior work that treats PD detection as a binary task, the proposed approach classifies healthy, mild PD, and moderate PD within a strictly nested cross-validation pipeline where feature selection, preprocessing, and resampling are confined within folds to prevent optimistic bias common in small-sample gait studies. SHAP-based explanations link model predictions to clinically interpretable features such as stride time and joint-level variability, bridging performance with medical trust.

The specific contributions of this work are:

- **Multiclass PD staging from single-camera video.** We demonstrate that healthy, mild, and moderate PD can be distinguished using pose-estimated gait features, a more challenging and clinically meaningful task than the binary setups employed in most prior video-based studies.
- **Joint-specific nonlinear dynamics as candidate biomarkers.** Among the retained feature set, nonlinear descriptors from distal joints (ankle, elbow,

wrist) constitute the majority of selected features, suggesting meaningful predictive signal from peripheral gait irregularity. Notably, a small subset of kinematic features achieves comparable classification performance, indicating that both feature families contribute and neither is strictly dominant, a finding that motivates further investigation on larger, independent cohorts.

- **Structured feature curation under strict nested cross-validation.** From 46 candidate features, Cp, AIC, BIC, and Adjusted R^2 are used to derive three subsets (7, 9, 19 features), which are subsequently evaluated within a strictly nested cross-validation pipeline to ensure unbiased performance estimation under limited data conditions.

The objectives of this study are threefold: (1) to systematically extract and validate kinematic and nonlinear dynamic gait features from clinically motivated joint selection, establishing which descriptors carry the greatest predictive weight; (2) to develop an interpretable, strictly nested cross-validation pipeline for multiclass PD severity classification with consistent feature importance across independent models; and (3) to provide preliminary evidence that joint-specific nonlinear dynamics translate to actionable digital biomarkers, motivating future external validation of vision-based PD assessment tools. The pipeline is extensible to other gait pathologies, offering a versatile platform for vision-based clinical classification.

2. Mathematical Background

This section defines the kinematic and nonlinear dynamic features extracted from gait trajectories. These features are computed from 2D joint coordinate time series and used as inputs to the machine learning models described in later sections.

2.1. Kinematic Features

The following stride-level kinematic descriptors are computed from gait cycles:

- **Stride Time:** Duration between two successive heel strikes of the same foot.
- **Swing Time:** Time interval during which the foot is not in contact with the ground.
- **Stance Time:** Duration of ground contact within a gait cycle.
- **Stride Length:** Linear displacement between consecutive heel strikes of the same foot.
- **Stride Velocity:** Mean walking speed, defined as:

$$v = \frac{\text{stride length}}{\text{stride time}}$$

All kinematic features are computed per gait cycle and aggregated at the trial level.

2.2. Nonlinear Dynamics Features Carvajal-Castaño et al. (2022)

Nonlinear dynamic features are computed from 2D joint coordinate time series. These measures characterize temporal structure and spatial properties of the signals.

- **Hurst Exponent (H):** Estimated using rescaled range (R/S) analysis:

$$E\left[\frac{R(n)}{S(n)}\right] \propto n^H$$

where $R(n)$ is the range of cumulative deviations and $S(n)$ is the standard deviation within a window of size n. The exponent $H \in (0, 1)$ is obtained from the slope of the log-log relationship. Anis–Lloyd correction is applied for short sequences.

- **Lempel–Ziv Complexity (LZC):** Computed from a discretized version of the signal:

$$C_{\text{norm}}(S) = \frac{\log_2 n}{n \cdot C(S)}$$

where $C(S)$ is the number of distinct subsequences identified using Lempel–Ziv parsing, and n is the sequence length. Continuous signals are transformed into symbolic sequences prior to computation.

- **Detrended Fluctuation Analysis (DFA):** Defined as:

$$F(s) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[Y(i) - y_s(i)]^2}$$

where $Y(i)$ is the cumulative sum of the signal and $y_s(i)$ is the local trend at scale s. The DFA exponent α is obtained from the slope of $\log F(s)$ versus $\log s$.

- **Box-Counting Dimension (D):** Computed using:

$$D = \lim_{\epsilon \to 0} \frac{\log N(\epsilon)}{\log(1/\epsilon)}$$

where $N(\epsilon)$ is the number of boxes of size ϵ required to cover the 2D trajectory. The slope of $\log N(\epsilon)$ versus $\log(1/\epsilon)$ provides an estimate of D.

All nonlinear features are computed per joint trajectory and aggregated at the trial level prior to model training.

3. Methodology

The task of transforming raw video recordings of human gait into clinically interpretable insights requires a carefully structured pipeline. Our approach may be divided into three broad stages **preprocessing**, **training and classification**, and **analysis** each of which addresses a distinct layer of the problem as shown in Figure1. Together, these stages convert noisy and heterogeneous video data into a compact and meaningful representation of gait suitable for Parkinsonian pattern classification and disease severity assessment.

3.1. Preprocessing

Raw gait videos are rarely suitable for direct analysis: variation in quality, occlusions, and irrelevant background motion all introduce noise. Pre-processing therefore plays a critical role, standardizing the input and extracting structured motion data from each subject. This stage involves **data acquisition, pose estimation, feature extraction, and feature selection**, followed by dimensionality reduction to ensure interpretability and efficiency.

3.1.1. Data Acquisition

We used publicly available video recordings from the dataset of Kour et al. (2020, 2022), comprising neurologically healthy controls (NM) and individuals diagnosed with PD. Subjects with Knee Osteoarthritis were excluded to isolate Parkinsonian gait. All recordings were captured at a fixed distance of approximately 8 m, ensuring consistent spatial scale across subjects. Figure 2 illustrates the experimental setup. The source videos were recorded at 25 frames per second with a standard RGB camera. Pose estimation was performed using MediaPipe Holistic (v 0.8.10), Python 3.9, scikit-learn 1.2 and TensorFlow 2.11. Each video was labeled by subject ID and class (NM, PD_ML, PD_MD), yielding 60, 12, and 14 samples respectively.

Three videos were discarded due to severe occlusion, leaving 78 usable samples: 57 NM, 10 PD ML, 11 PD MD. Minor missing keypoints were interpolated linearly. Demographics covariates (age,height,gender) were available but were not included as model features. This is acknowledged as a limitation: age and height are known to influence stride length and velocity in PD (Yin et al., 2024b), and future work should examine their contributions as covariates or in subgroup analyses. The decision was made here to focus on gait kinematics and NLD features in isolation to establish a clean baseline.

Four pose estimation frameworks were evaluated. Manual annotation Cho et al. (2009) was rejected as impractical for large datasets; OpenPose offered accuracy but proved computationally expensive, while YOLOv7 produced inconsistent skeletal outputs. Ultimately, MediaPipe's holistic model was selected for its balance of computational efficiency, accuracy and open-source accessibility. The framework detects 33 anatomical landmarks. In our analysis, attention was focused on the joints of the upper and lower extremities (shoulder, elbow, wrist, knee, and ankle), as arm swing and leg dynamics are critical indicators of Parkinsonian gait.

3.1.2. Feature Extraction

Kinematic and nonlinear dynamic descriptors were derived from 2D skeletal keypoints as defined in (see Section 2). For each of the five joints and each spatial coordinate(x,y) the four NLD metrics were computed, yielding a maximum of 5x2x4 = 40 potential NLD features plus 6 kinematic features, giving 46 features. Uniform Manifold Approximation and Projection (UMAP) was applied solely for visualization of class separability on these 46 features as shown in Fig 3. Classifiers were trained on 46 features vectors, not on UMAP embeddings. The supervised UMAP projection is included solely as a high-dimensional visualization aid, providing an intuitive illustration of class structure in the 46-feature space; it does not constitute a validation of separability, as the label information is used during projection. This representation illustrates not only the discriminative power of the selected features but also the feasibility of distinguishing intra-PD severity classes—an ability not consistently achieved in earlier binary models (Hausdorff (2009)).

3.1.3. Feature Selection and Dimensionality Reduction Redundancy removal

To reduce model complexity and focus on the most informative inputs, we employed a set of well-established statistical criteria Mallows' Cp, AIC

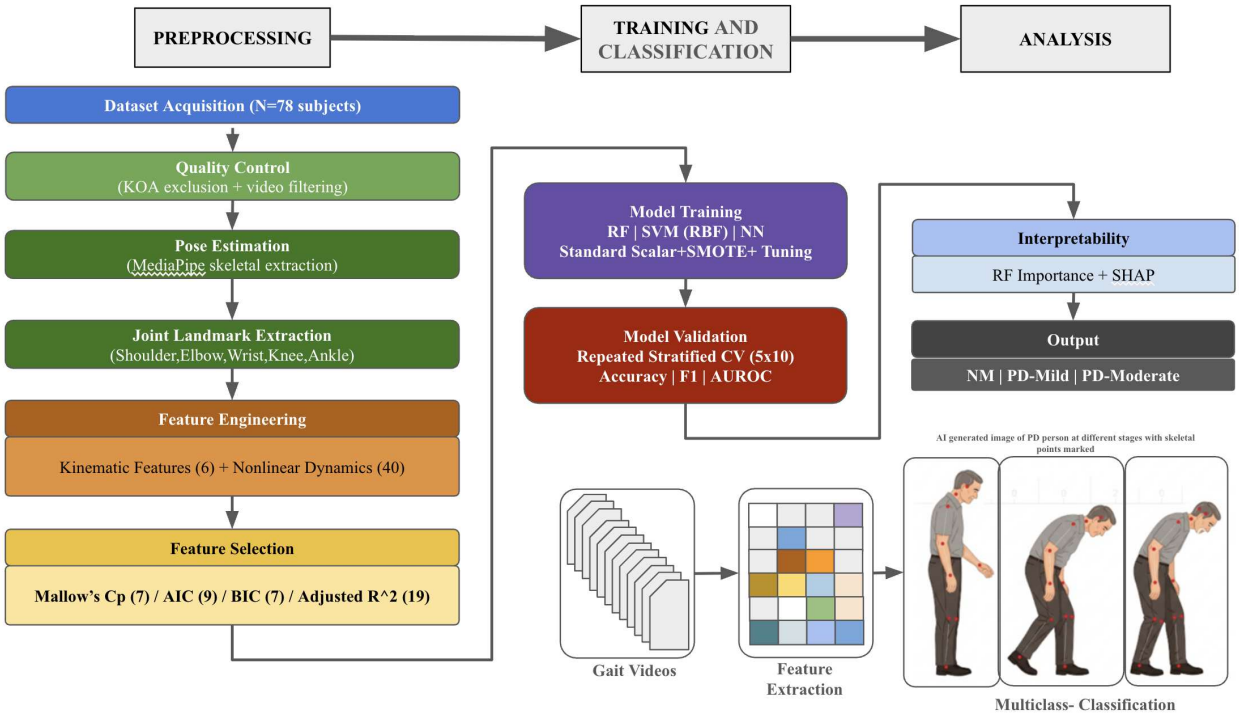


Figure 1: Workflow of the Parkinson's gait analysis from video preprocessing to Muliclass classification.

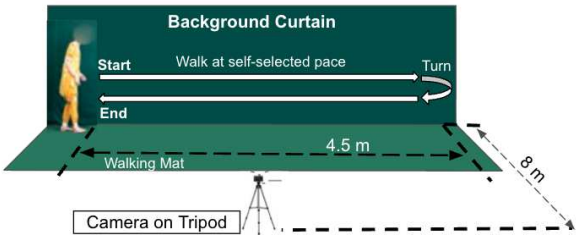


Figure 2: Experimental setup for video-based gait monitoring using a fixed camera Kour et al. (2022).

(Akaike Information Criterion), BIC (Bayesian Information Criterion), and Adjusted R^2 (James et al., 2023) as a preliminary variance-screening step to reduce the 46-dimensional feature space to compact, disease-relevant subsets. Given the limited sample size, retaining all 46 features risks overfitting, where a model performs well on available data but fails to generalise to unseen cases; structured feature selection is therefore essential for both model reliability and interpretability. Because these subset-selection criteria are defined within the ordinary least squares (OLS) framework and require a scalar response, the multiclass outcome was temporarily recoded as a binary screening target (NM vs. PD, combining PD-ML and PD-MD). This recoding is adopted solely as a computational necessity of the OLS criterion and does not imply that the classification problem is treated as binary. The selected features capture the primary disease-related variance in the data, and the multiclass distinction between NM, PD-ML, and PD-MD is handled entirely within the downstream machine learning pipeline, which receives these screened subsets and independently learns the three-way decision boundaries.

These four criteria were chosen because they each provide a different perspective on the trade-off between model accuracy and complexity. Mallows' Cp and AIC aim to identify models that achieve a good fit while avoiding unnecessary complexity, making them suitable for predictive performance. BIC applies a stronger penalty for including additional features, which is particularly advantageous for small datasets, as it favors simpler models that are less likely to overfit. Adjusted R^2, on the other hand, measures how well the model explains the variability in the data while accounting for the number of features used, allowing us to assess whether adding more variables genuinely improves explanatory power.

Because the subset selection criterion was OLS (Ordinary least square) based the multiclass outcome was temprarily recoded into a binary screening target NM

and PD(PD_MD and PD_ML). Although this feature selection was performed using a binary grouping (NM vs PD), the selected features capture the primary disease-related variance. These features were subsequently evaluated within multiclass classification models, which are capable of learning finer distinctions between PD subtypes. Applying these criteria resulted in different recommended feature subsets: Mallows' Cp and BIC suggested a compact set of 7 features, AIC selected 9 features, and adjusted R^2 indicated a larger set of 19 features as shown in Fig.4. Rather than relying on a single criterion, we constructed three fixed-feature case definitions based on these subsets to systematically compare how varying levels of model complexity affect performance. A visual representation of the distribution of subsets can be seen in the Fig.5. The smaller set of 7 and 9 features are subset of the 19 feature set among which there are 3 Kinematic features and 16 NLD features. For each case, three machine learning models were trained, and the importance of individual features was analyzed. Feature importance for Random Forest models was assessed using permutation importance, while SHAP (Shapley Additive Explanations) values were used to interpret feature contributions in Support Vector Machine (SVM) and Neural Network (NN) models. The results, presented in Table1, provide a clear comparison of how different feature selection strategies influence both predictive performance and the relative importance of the selected variables. A more detailed analysis and interpretation of these findings are provided in the Section.3.3.

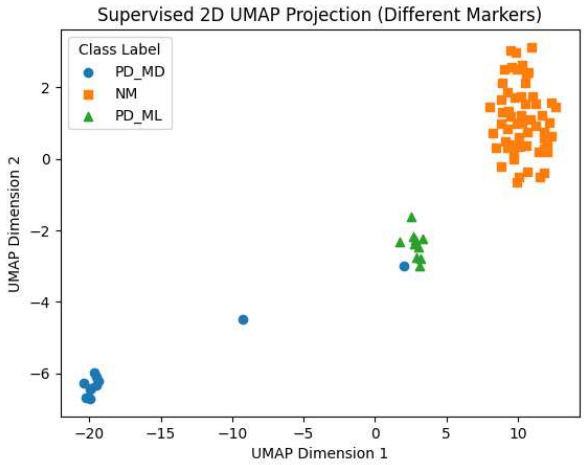


Figure 3: Supervised 2D UMAP projection for visualization of class structure across all 46 features. Class labels (NM, PD-MD, PD-ML) are used during projection to aid visual interpretation of the feature space. This figure serves as an illustration of distributional separation and is not intended as a validation of class separability.

3.2. Training and Classification

The dataset contained 78 usable videos (57 NM, 10 PD Ml, 11 PD MD). Class imbalance was addressed by applying SMOTE exclusively to the training partition (80% of data, approximately 62 samples). The test partition (20%, approximately 16 samples) was kept in its original distribution to provide an unbiased estimate of generalization. For each case, the modeling pipeline was structured to ensure fair and robust evaluation while avoiding data leakage. Specifically, the selected features were first isolated, and then model training followed a consistent sequence of preprocessing steps applied only within training data partitions. To assess model stability and generalizability which is particularly important given the small and imbalanced dataset, a repeated stratified 5-fold cross-validation was performed with 10 repeats, resulting in 50 evaluation folds per model. Stratification ensured that the class distribution was preserved in every split, preventing bias toward the majority class. Within each fold, preprocessing (such as scaling for SVM and neural networks, and class balancing using SMOTE) was applied exclusively to the training portion before being extended to validation data, thereby maintaining a strict separation between training and testing information. In addition to evaluating the full feature subset for each case, an ablation study was conducted for every model to systematically examine the contribution of different feature groups. This involved training and testing models using (i) kinematic features only, (ii) nonlinear-dynamics features only, and (iii) the combined feature set. This structured pipeline allowed for a comprehensive comparison across feature subsets, model types, and feature groups, ensuring that the reported results reflect not only predictive performance but also the robustness and interpretability of the models under different feature configurations.

We compared three machine learning models, each representing a different inductive bias:

3.2.1. Random Forest (RF)

One of the most vigorous algorithms against overfitting and with the potential to deal with the non-linear relationship.The RF model was employed as a nonlinear ensemble learning method capable of capturing complex interactions among features without requiring feature scaling. Hyperparameter tuning was performed using a randomized search strategy with 20 sampled configurations, evaluated using stratified 5-fold cross-validation on the training data. The optimization objective was the weighted F1-score to account for class imbalance.

Feature	Case 1 (BIC)			Case 2 (AIC)			Case 3 (Adjusted R^2)		
	RF	SVM	NN	RF	SVM	NN	RF	SVM	NN
stride_length	**0.321**	0.051	0.089	**0.194**	0.050	0.054	0.057	0.028	0.057
shoulder_x_lzc	0.089	**0.086**	**0.106**	0.161	**0.094**	0.078	0.097	**0.056**	**0.073**
wrist_x_hurst	0.281	0.039	0.039	0.176	0.037	0.042	0.119	0.025	0.007
ankle_y_lzc	0.224	0.026	0.015	0.176	0.013	0.017	**0.120**	0.008	0.012
stance_time	0.037	0.047	0.104	0.092	0.078	**0.100**	0.033	0.045	0.064
wrist_y_hurst	0.013	0.027	0.018	0.027	0.035	0.025	0.011	0.024	0.022
turning_angle	0.036	0.033	0.031	0.028	0.009	0.023	0.008	0.010	0.021
knee_y_dfa	-	-	-	0.084	0.025	0.033	0.072	0.006	0.008
ankle_y_dfa	-	-	-	0.063	0.012	0.012	0.068	0.003	0.015
shoulder_y_dfa	-	-	-	-	-	-	0.074	0.018	0.030
elbow_x_hurst	-	-	-	-	-	-	0.072	0.019	0.027
shoulder_x_dfa	-	-	-	-	-	-	0.079	0.020	0.015
knee_y_lzc	-	-	-	-	-	-	0.048	0.004	0.006
wrist_x_dfa	-	-	-	-	-	-	0.032	0.008	0.016
elbow_y_dfa	-	-	-	-	-	-	0.035	0.012	0.007
elbow_y_hurst	-	-	-	-	-	-	0.028	0.014	0.005
elbow_y_lzc	-	-	-	-	-	-	0.029	0.005	0.007
shoulder_y_boxcount	-	-	-	-	-	-	0.008	0.004	0.011
wrist_y_lzc	-	-	-	-	-	-	0.011	0.005	0.004

Table 1: Feature importance values across the three feature-selection cases for Random Forest (RF). SHAP values for Support Vector Machine (SVM), and Neural Network (NN). A dash indicates that the feature was not included in that case.

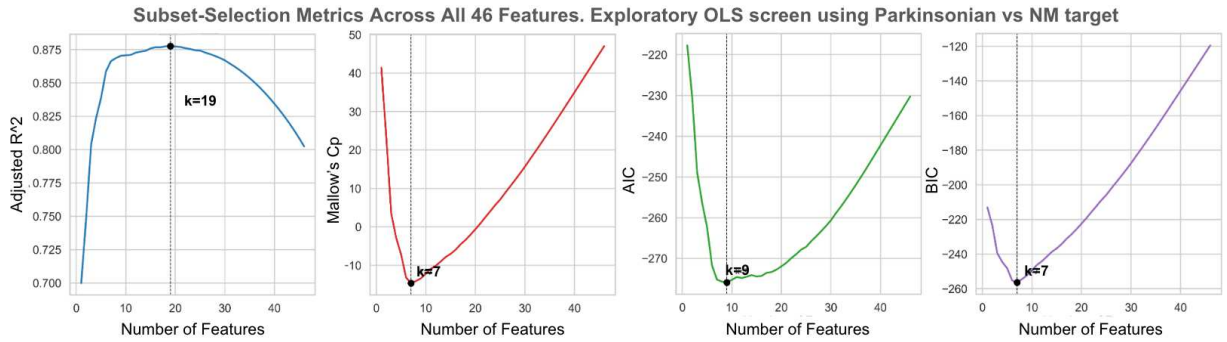


Figure 4: Visualization of OLS(Ordinary Least Square) screen using Parkinsonian(PD_MD and PD_ML) vs NM target variables. Each plot shows a subset selection metric to identify optimal number of predictors. From the left Adjusted R^2 (k=19), Mallow's C_p (k=7), AIC (Akaike Information Criterion) (k=9) and BIC (Bayesian Information Criterion) (k=7)

The hyperparameter search space included: number of trees ($n_estimators \in [100, 300]$), maximum tree depth ($max_depth \in \{\text{None}, 5, 8, 10, 15, 20\}$), minimum samples required to split a node ($min_samples_split \in [2, 10]$), minimum samples at a leaf node ($min_samples_leaf \in [1, 5]$), number of features considered at each split ($max_features \in \{sqrt, log2, \text{None}\}$), and class weighting strategy ($class_weight \in \{balanced, balanced_subsample, \text{None}\}$). SMOTE resampling was applied within each training fold to address class imbalance.

The optimal hyperparameters for each feature-selection case are summarized in Table 2. Across the cases, the tuned configurations reflect moderately constrained tree depths and controlled node-splitting criteria, which help prevent overfitting while maintaining sufficient model flexibility for the small and imbalanced dataset.

3.2.2. Support Vector Machine (SVM)

In order to learn a non-linearity relationship between the data a SVM with Radial Basis Function (RBF) was applied as a model to predict the output. RBF kernel was selected due to its good performance in terms of representing non-linear boundaries in high dimension feature space. A Support Vector Machine with a radial basis function (RBF) kernel was used to model nonlinear decision boundaries. Because SVM performance is sensitive to feature scaling, all features were standardized using a *StandardScaler* fitted exclusively on the training data within each fold.

Hyperparameter tuning was conducted using randomized search with 20 sampled configurations

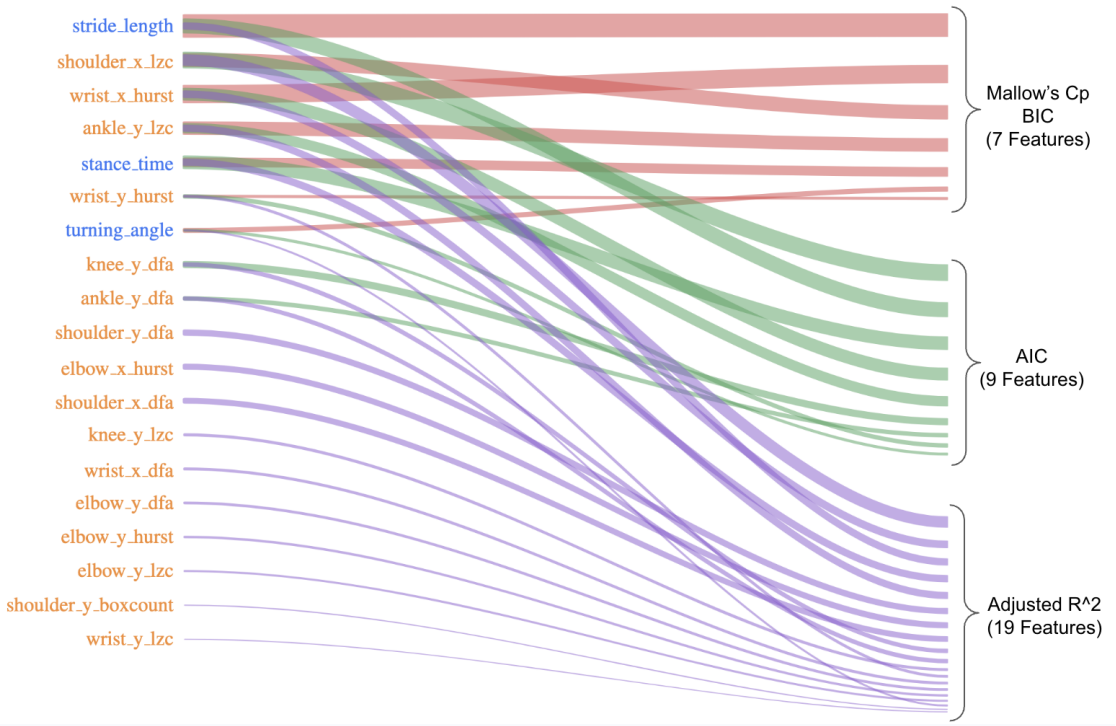


Figure 5: Visualization of the obtained optimal 19 features and their distribution to each case. The width of the ribbons represent the importance of the feature in classification as obtained from the Table 1

and stratified 5-fold cross-validation, optimizing for weighted F1-score. The regularization parameter (C) and kernel coefficient (γ) were sampled from log-uniform distributions to effectively cover multiple orders of magnitude, with $C \in [10^{-2}, 10^{3}]$ and $\gamma \in [10^{-4}, 10^{1}]$. Class imbalance was handled using *class_weight = balanced*, and SMOTE was applied after scaling within each training fold.

The final tuned hyperparameters for each case are reported in Table 2. The variation in selected C and γ values across cases reflects differences in model complexity and the degree of nonlinearity required for each feature subset.

3.2.3. Neural Network (NN)

A feed-forward neural network was implemented to capture higher-order nonlinear relationships among features. Hyperparameter tuning was performed using a predefined grid of candidate configurations, evaluated using stratified 3-fold cross-validation on the training data. Model selection was based primarily on weighted F1-score and secondarily on accuracy. A 3-fold inner cross-validation was used for neural network hyperparameter tuning, in contrast to the 5-fold protocol applied to RF and SVM. This choice was motivated by two considerations specific to the neural network setting: first, the iterative nature of backpropagation training substantially increases the computational cost of each inner fold, making a 3-fold protocol more tractable; and second, with a small and imbalanced dataset, reducing the number of inner folds helps maintain sufficient minority-class representation within each training partition during each tuning iteration, improving fold-level stability. The overall evaluation robustness is maintained by the 50-fold outer repeated stratified cross-validation applied uniformly across all models.

The tuning process explored variations in network architecture (including configurations such as $(64, 32)$, $(128, 64)$, $(128, 64, 32)$, and $(96, 48, 24)$), dropout rate (0.2–0.3), learning rate (fixed at 0.001 across candidates), batch size (e.g., 16), and number of training epochs (typically 50–60). Input features were standardized using a *StandardScaler*, and SMOTE was applied to the scaled training data to address class imbalance. An internal validation split (20%) was used during training to enable early stopping with a patience of 8 epochs.

The selected architectures and training parameters for each case are summarized in Table 2. The resulting configurations reflect a balance between network depth and regularization, where dropout and early stopping help

mitigate overfitting while maintaining sufficient representational capacity.

3.3. Analysis

Feature importance was assessed using Random Forest (RF) permutation importance and SHAP values, with the results summarized in a Table1.

3.4. Feature Importance Analysis (RF Permutation and SHAP)

Feature importance was analyzed to understand the relative contribution of each predictor across the three feature-selection cases. For the Random Forest (RF) model, permutation importance was used to quantify the decrease in model performance when a feature is randomly shuffled. For the Support Vector Machine (SVM) and Neural Network (NN), SHAP (Shapley Additive Explanations) values were computed to estimate the contribution of each feature to the model predictions. The results are summarized in Table 1, while Fig.5 illustrates the feature selection pathways across the three cases.

Case 1 (7 features: Cp/BIC). In the most compact model, feature importance is dominated by a small number of predictors. For RF, *stride_length* (0.321), *wrist_x_hurst* (0.281), and *ankle_y_lzc* (0.224) together account for approximately 82.6% of the total importance, indicating a highly concentrated importance distribution. Notably, two of these three dominant features (*wrist_x_hurst* and *ankle_y_lzc*) belong to the nonlinear-dynamics (NLD) family, demonstrating that signal-complexity descriptors are already critical even in the smallest feature subset.

Across models, SHAP values show a more distributed pattern. For NN, *stance_time* (0.104) and *shoulder_x_lzc* (0.106) are among the most influential, while SVM assigns relatively balanced contributions across features, with *shoulder_x_lzc* (0.086) being the highest. Overall, 5 out of the 7 selected features (71%) are NLD-based, indicating a strong reliance on nonlinear descriptors even in a minimal representation.

Case 2 (9 features: AIC). With the inclusion of two additional features, the importance distribution becomes more balanced. In RF, *stride_length* (0.194), *shoulder_x_lzc* (0.161), *wrist_x_hurst* (0.176), and *ankle_y_lzc* (0.176) form a core group contributing over 70% of total importance. The newly introduced features, *knee_y_dfa* (0.084) and *ankle_y_dfa* (0.063), are both detrended fluctuation analysis (DFA) features, further increasing the representation of NLD characteristics.

From a proportional perspective, 7 out of the 9 features (78%) in this case are NLD-based. Importantly, this case introduces diversity within NLD metrics, combining Lempel-Ziv complexity (LZC), Hurst exponent, and DFA features. SHAP values for SVM and NN also reflect this trend, with *shoulder_x_lzc* (0.094 for SVM) and *stance_time* (0.100 for NN) emerging as influential, but NLD features consistently maintaining comparable or higher contributions across models.

Case 3 (19 features: Adjusted R^2). In the largest feature set, importance becomes more distributed across predictors, but NLD features remain dominant. For RF, the top contributors include *ankle_y_lzc* (0.120), *wrist_x_hurst* (0.119), *shoulder_x_lzc* (0.097), and several DFA-based features such as *shoulder_y_dfa* (0.074) and *knee_y_dfa* (0.072). Although individual importance values are lower compared to smaller cases, this reflects a broader distribution rather than reduced relevance.

This case is heavily dominated by NLD features, with 16 out of 19 predictors (approximately 84%) belonging to the nonlinear-dynamics family. Among these, multiple feature types are represented:

- Hurst exponent features (e.g., wrist and elbow joints)
- DFA features (e.g., knee, shoulder, elbow)
- LZC features (e.g., ankle, knee, wrist)
- Box-counting features (e.g., shoulder_y_boxcount)

The SHAP analysis for SVM and NN mirrors this pattern, although with lower absolute magnitudes due to normalization. For example, *shoulder_x_lzc* (0.056 in SVM) and *stance_time* (0.064 in NN) remain relevant, but the importance is distributed across a larger number of NLD features.

Overall Trends and Interpretation. Fig.5 visually summarizes how features are progressively added across cases. A consistent pattern emerges in which core features such as *stride_length*, *shoulder_x_lzc*, and *wrist_x_hurst* persist across all cases, indicating their robustness. As model complexity increases, additional NLD features particularly DFA and Hurst based descriptors from upper and lower limb joints are incorporated.

Table 2: Tuned hyperparameters for all models across the three feature-selection cases.

Case	Model	Key Parameters
Case 1 (7 features)	Random Forest	$n_estimators$ = 207, max_depth = 8, $min_samples_split$ = 4, $min_samples_leaf$ = 2, $class_weight$ = *balanced*
	SVM (RBF)	C = 10.129, γ = 0.347, $class_weight$ = *balanced*
	Neural Network	Layers: [128, 64], Dropout: 0.2, LR: 0.001, Batch: 16, Epochs: 50
Case 2 (9 features)	Random Forest	$n_estimators$ = 288, max_depth = 10, $min_samples_split$ = 9, $min_samples_leaf$ = 3, $max_features$ = *sqrt*
	SVM (RBF)	C = 0.332, γ = 0.042, $class_weight$ = *balanced*
	Neural Network	Layers: [96, 48, 24], Dropout: 0.25, LR: 0.001, Batch: 16, Epochs: 50
Case 3 (19 features)	Random Forest	$n_estimators$ = 288, max_depth = 10, $min_samples_split$ = 9, $min_samples_leaf$ = 3, $max_features$ = *sqrt*
	SVM (RBF)	C = 1.445, γ = 0.00286, $class_weight$ = *balanced*
	Neural Network	Layers: [128, 64, 32], Dropout: 0.2, LR: 0.001, Batch: 16, Epochs: 60

Table 3: Ablation study across cases with different number of features. Values are mean (± std). Best-performing model per ablation is highlighted in bold. Kin (Kinematic), NLD(Non Linear Dynamics), Acc (Accuracy), F1_M (Macro F1), F1_w (Wighted F1), Rec (Recall) and AUC (AUROC). Maximum values for each categories are highlighted in bold.

Case	Abl.	Model	Acc	$F1_M$	$F1_W$	Rec	AUC
1	Kin	NN	0.894±0.059	0.729±0.158	0.883±0.068	0.894±0.059	**0.941±0.080**
		RF	0.879±0.072	**0.728±0.150**	0.873±0.077	0.879±0.072	0.934±0.081
		SVM	**0.894±0.050**	0.724±0.140	**0.881±0.059**	**0.894±0.050**	0.920±0.087
	NLD	NN	**0.790±0.086**	**0.560±0.152**	**0.793±0.077**	**0.790±0.086**	**0.965±0.022**
		RF	0.760±0.079	0.479±0.127	0.755±0.068	0.760±0.079	0.951±0.031
		SVM	0.766±0.086	0.505±0.156	0.762±0.080	0.766±0.086	0.937±0.043
	Both	NN	**0.906±0.059**	**0.761±0.147**	**0.897±0.067**	**0.906±0.059**	**0.987±0.013**
		RF	0.869±0.065	0.737±0.146	0.863±0.069	0.869±0.065	0.975±0.025
		SVM	0.898±0.059	0.749±0.156	0.883±0.072	0.898±0.059	0.973±0.028
2	Kin	NN	**0.898±0.055**	0.741±0.144	**0.888±0.062**	**0.898±0.055**	**0.942±0.076**
		RF	0.895±0.058	**0.757±0.133**	0.888±0.065	0.895±0.058	0.928±0.099
		SVM	0.897±0.058	0.755±0.142	0.888±0.064	0.897±0.058	0.942±0.090
	NLD	NN	**0.827±0.076**	**0.648±0.137**	**0.824±0.068**	**0.827±0.076**	0.966±0.028
		RF	0.815±0.078	0.605±0.139	0.808±0.071	0.815±0.078	0.960±0.022
		SVM	0.821±0.068	0.616±0.128	0.821±0.059	0.821±0.068	**0.972±0.022**
	Both	NN	0.878±0.056	0.726±0.129	0.873±0.062	0.878±0.056	0.981±0.017
		RF	0.875±0.064	**0.766±0.117**	0.875±0.060	0.875±0.064	0.976±0.021
		SVM	**0.905±0.052**	0.762±0.135	**0.896±0.059**	**0.905±0.052**	**0.988±0.015**
3	Kin	NN	0.894±0.060	0.731±0.157	0.883±0.069	0.894±0.060	0.939±0.077
		RF	0.895±0.058	0.757±0.133	0.888±0.065	0.895±0.058	0.928±0.099
		SVM	**0.896±0.074**	**0.769±0.163**	**0.888±0.079**	**0.896±0.074**	**0.933±0.084**
	NLD	NN	0.809±0.079	0.603±0.149	0.800±0.081	0.809±0.079	0.942±0.053
		RF	0.824±0.075	0.624±0.145	0.821±0.069	0.824±0.075	0.960±0.029
		SVM	**0.841±0.088**	**0.690±0.151**	**0.841±0.085**	**0.841±0.088**	**0.966±0.030**
	Both	NN	**0.880±0.086**	**0.756±0.177**	**0.876±0.089**	**0.880±0.086**	0.969±0.045
		RF	0.853±0.078	0.702±0.158	0.851±0.074	0.853±0.078	**0.971±0.026**
		SVM	0.862±0.074	0.688±0.154	0.854±0.075	0.862±0.074	0.979±0.027

Quantitatively, the proportion of NLD features increases from 71% (Case 1) to 78% (Case 2) and 89% (Case 3), demonstrating a clear shift toward nonlinear descriptors as model capacity grows. Furthermore, RF permutation importance consistently assigns higher magnitudes to NLD features compared to kinematic features, suggesting that signal complexity and scaling behavior capture critical characteristics of Parkinsonian gait.

Overall, these results indicate that while basic kinematic features such as *stride_length* and *stance_time* remain important, nonlinear-dynamics features particularly Hurst exponent, DFA, and LZC metrics derived from wrist, ankle, knee, and shoulder joints play a dominant role in classification of gait patterns. This highlights the importance of incorporating signal complexity measures when modeling movement disorders.

4. RESULTS AND DISCUSSION

The performance of three machine learning models Random Forest (RF), Support Vector Machine (SVM) with RBF kernel, and a Neural Network (NN) was evaluated for classifying Parkinsonian gait patterns. The aim was not only to assess raw predictive accuracy but also to understand how each approach balances discriminative power, sensitivity to class imbalance, and computational practicality for clinical use.

4.1. Ablation Study and Repeated Cross-Validation Analysis

To evaluate the contribution of different feature groups and assess model robustness, a repeated stratified 5-fold cross-validation (10 repeats) was performed across all three feature-selection cases. The ablation

study compared three feature configurations: kinematic features only (Kin), nonlinear-dynamics features only (NLD), and the combined feature set (Both). The results are summarized in Table 3.

Case 1 (7 features: Cp/BIC). In the most compact feature set, the combined feature configuration (Both) consistently outperforms individual feature groups across all models. The NN achieves the highest performance with an accuracy of 0.906 ± 0.059, macro F1 of 0.761, and AUC of 0.987, indicating strong predictive capability when both kinematic and nonlinear features are used together.

When considering individual feature groups, kinematic features outperform nonlinear-dynamics features in terms of accuracy and F1 scores. For example, SVM achieves 0.894 accuracy in the Kin setting compared to 0.766 in NLD. However, NLD features yield higher AUC values across all models (e.g., NN: 0.965), suggesting that nonlinear descriptors provide strong discriminative power even when classification performance is lower. This indicates that in a limited feature setting, kinematic features drive classification performance, while NLD features contribute to separability.

Case 2 (9 features: AIC). In Case 2, performance improves and becomes more balanced across feature groups. The SVM model with the combined feature set achieves the highest overall performance, with accuracy of 0.905 ± 0.052, weighted F1 of 0.896, and AUC of 0.988. This aligns with earlier results showing that the 9-feature subset provides an optimal representation.

Interestingly, the gap between Kin and NLD configurations narrows compared to Case 1. NLD-only models achieve competitive performance (e.g., NN accuracy: 0.827), and SVM attains the highest AUC (0.972) under the NLD configuration. This suggests that the inclusion of additional nonlinear features (particularly DFA-based features) enhances their standalone predictive capability.

Overall, the combined feature configuration still yields the most consistent performance across metrics, indicating complementary contributions from both feature families.

Case 3 (19 features: Adjusted R^2). In the largest feature set, performance becomes slightly more variable. While the combined feature configuration remains strong, the improvement over individual feature groups is less pronounced. The Neural Network achieves the highest accuracy (0.880 ± 0.086) and macro F1 (0.756) under the combined setting, whereas Random Forest achieves the highest AUC (0.971).

Notably, the NLD-only configuration shows improved performance compared to Case 1, with SVM achieving 0.841 accuracy and 0.966 AUC. This reflects the increased representation of nonlinear features in this case. However, the overall performance does not surpass that of Case 2, suggesting that increasing feature dimensionality beyond a certain point may lead to diminishing returns or feature redundancy.

Across all cases, several consistent patterns emerge. First, the combined feature configuration (Both) generally provides the best or most balanced performance across accuracy, F1 scores, and AUC, highlighting the complementary nature of kinematic and nonlinear-dynamics features.

Second, nonlinear-dynamics features consistently yield high AUC values across all cases, even when their standalone accuracy is lower. This indicates that these features capture important class-separating characteristics, particularly in terms of underlying signal structure.

Third, Case 2 (9 features) demonstrates the most stable and optimal performance across models and metrics, achieving the highest accuracy and balanced F1 scores with relatively low variance. In contrast, Case 1 shows limitations due to feature scarcity, while Case 3 exhibits signs of feature redundancy despite the increased number of predictors.

Overall, the ablation and repeated cross-validation results validate the effectiveness of the proposed pipeline, demonstrating that careful feature selection combined with balanced model evaluation leads to improved generalization and stable performance across different model families.

4.2. Confusion Matrix Analysis on Held-Out Test Data

The classification performance of the three models across the three feature-selection cases was further evaluated using confusion matrices computed on the held-out test set. The results provide insight into class-wise accuracy, recall, and precision, particularly under the imbalanced three-class setting as shown in Fig. 6.

Case 1 (7 features: Cp/BIC). In Case 1, both RF and SVM produced identical confusion matrices, correctly classifying all samples from the majority class (NM: 12/12) and the PD_ML class (2/2), while misclassifying one PD_MD sample as NM. This results in an overall accuracy of 15/16 = 93.75%. The recall for the NM and PD_ML classes is 100%, whereas the recall for PD_MD is reduced to 50% (1/2). Precision for NM remains high but is slightly affected by one false positive originating from PD_MD.

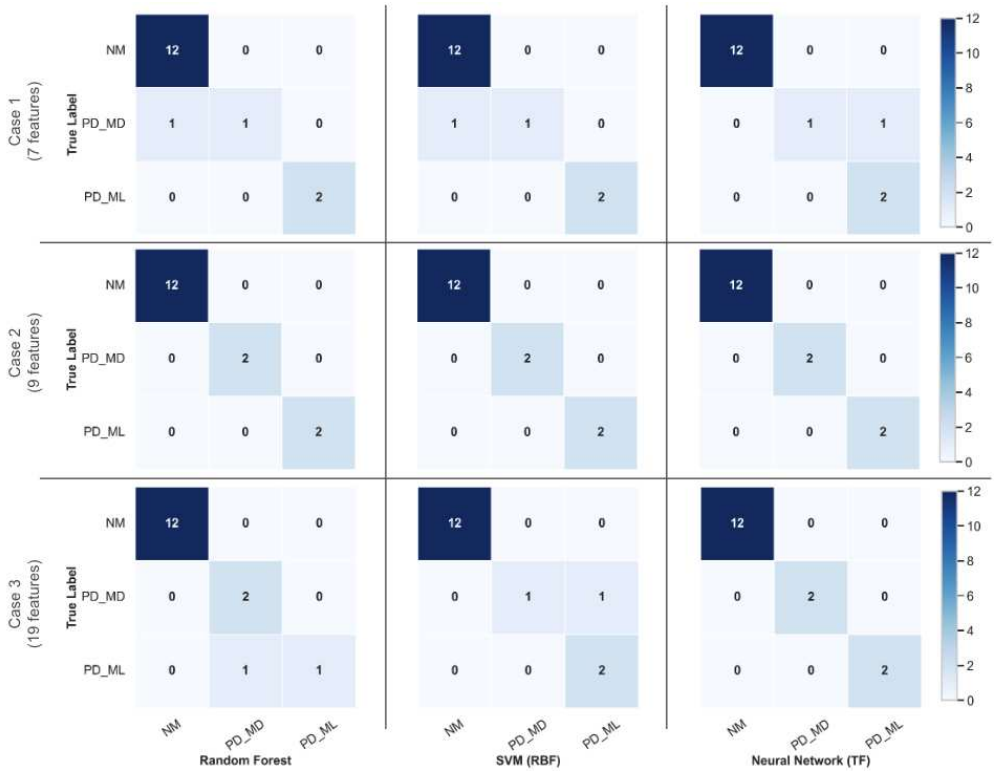


Figure 6: Confusion matrix for model RF, SVM and NN with each set of features for a Held out test.

The NN model also achieves 93.75% accuracy but exhibits a different error pattern. While NM and PD_ML are again perfectly classified, one PD_MD sample is misclassified as PD_ML, leading to a recall of 50% for PD_MD and a slight reduction in precision for the PD_ML class due to this false positive. Overall, Case 1 demonstrates strong performance for the majority and PD_ML classes, with limited sensitivity in distinguishing PD_MD.

Case 2 (9 features: AIC). In Case 2, all three models (RF, SVM, and NN) achieve perfect classification, with confusion matrices showing no misclassifications across all classes. This corresponds to an accuracy of 16/16 = 100%, with precision and recall equal to 1.00 for all three classes (NM, PD_MD, and PD_ML). It should be noted that the held-out test set comprises approximately 16 samples. On a test partition of this size, perfect classification across all three classes, while consistent with the cross-validation findings cannot be interpreted as strong evidence of generalisation. A single misclassification would reduce accuracy by approximately 6 percentage points, meaning the observed 100% result occupies a narrow and statistically sensitive margin. These findings are therefore reported as directionally supportive of the 9-feature subset's utility, with definitive performance conclusions deferred to external validation on larger cohorts.

This result indicates that the 9-feature subset selected via AIC demonstrates the most consistent performance within the constraints of the available dataset, though the small held-out set limits the strength of this claim.

Case 3 (19 features: Adjusted R^2). In Case 3, performance remains high but shows slight variability across models. The Neural Network maintains perfect classification (100% accuracy), similar to Case 2.

The Random Forest model achieves 93.75% accuracy, with one PD_ML sample misclassified as PD_MD. This results in a recall of 50% for PD_ML and 100% for NM and PD_MD. The precision for PD_MD is slightly reduced due to this misclassification.

The SVM model also achieves 93.75% accuracy but exhibits a different error pattern, where one PD_MD sample is misclassified as PD_ML. Consequently, recall for PD_MD is 50%, while NM and PD_ML retain perfect recall. Precision for PD_ML is slightly reduced due to this false positive.

Across all cases, the NM class is consistently classified with perfect recall and precision, reflecting its dom-

Case	Model	TT	HOP	HOS
C1 (7f)	RF	0.3279	0.0221	1.38
	SVM	0.0061	0.0019	0.12
	NN	20.4543	0.1212	7.57
C2 (9f)	RF	0.5021	0.0336	2.10
	SVM	0.0134	0.0021	0.13
	NN	19.0694	0.1355	8.47
C3 (19f)	RF	0.4414	0.0283	1.77
	SVM	0.0073	0.0020	0.13
	NN	22.8187	0.1194	7.46

Table 4: Performance metrics across feature subsets (C1: BIC/Cp, C2: AIC, C3: Adj. R^2; f = features), TT(Training Time) HOP (Held out prediction time) and HOS (Held out per sample)

inance in the dataset and clear separability. The PD_ML class is also generally well identified, except for minor misclassifications in Case 3. The most challenging class is PD_MD, which shows reduced recall (50%) in Cases 1 and 3, indicating difficulty in distinguishing this subgroup from either NM or PD_ML depending on the model.

From a comparative perspective, Case 2 (9 features) achieves perfect performance across all models, suggesting that this feature subset provides the most effective representation for classification. Cases 1 and 3 show slight degradation due to either underrepresentation (too few features) or increased complexity (feature dilution), highlighting the importance of selecting an optimal feature subset size.

4.3. Computational Efficiency and Practical Considerations

In addition to predictive performance, the computational efficiency of each model was evaluated in terms of training time and prediction latency. The results across all feature-selection cases are summarized in Table.4.

Training Time. The SVM consistently exhibits the lowest training time across all cases, requiring between 0.006 and 0.013 seconds. RF shows moderate training times, ranging from 0.328 to 0.502 seconds depending on the number of features. In contrast, the NN requires substantially higher training time, ranging from 19.069 to 22.819 seconds. This reflects the iterative optimization process and increased computational complexity associated with neural network training.

Prediction Latency. A similar trend is observed for prediction latency. SVM achieves the fastest inference, with held-out prediction (HOP) times between 0.0019 and 0.0021, and per-sample latency (HOS) between 0.12 and 0.13. RF shows moderate latency, with HOP ranging from 0.0221 to 0.0336 and per-sample prediction time between 1.38 and 2.10. The NN model incurs the highest inference cost, with HOP values between 0.1194 and 0.1355 and per-sample latency ranging from 7.46 to 8.47.

Effect of Feature Set Size. Across all models, increasing the number of features from 7 (Case 1) to 9 (Case 2) and 19 (Case 3) results in a modest increase in both training time and prediction latency. This trend is most evident for RF, where training time increases from 0.328 s (Case 1) to 0.502 s (Case 2), before slightly decreasing to 0.441 s (Case 3). For NN, training time varies within a relatively narrow range, increasing from 20.454 s (Case 1) to 22.819 s (Case 3), with a slight dip in Case 2. In contrast, SVM remains largely unaffected by feature set size, maintaining consistently low training and inference times across all cases.

Practical Implications. From a practical deployment perspective, SVM offers the most computationally efficient solution, combining minimal training time with extremely low prediction latency, making it well-suited for real-time or resource-constrained applications. RF provides a balanced trade-off between computational cost and model interpretability, with moderate training and inference times.

Although NN demonstrates competitive predictive capability, its substantially higher computational cost may limit its applicability in time-sensitive settings. Nevertheless, its inference latency, while higher than RF and SVM, remains within an acceptable range for offline or near-real-time clinical applications.

Overall, these results demonstrate that the proposed pipeline not only achieves strong predictive performance but also provides flexibility in model selection based on computational constraints, enabling adaptation to diverse deployment scenarios.

4.4. Broader Implications and Limitations

The results of this study demonstrate that a carefully designed, repeated stratified machine learning pipeline combined with feature selection and NLD descriptors can enable accurate and interpretable multiclass classification of Parkinsonian gait from single-camera video data. Across multiple evaluation stages including repeated cross validation, ablation analysis, and held out testing the proposed framework consistently shows that combining kinematic and nonlinear features yields the most stable and high-performing models. In particular, the predominance of nonlinear dynamic features also

has plausible physiological significance. Unlike conventional kinematic measures, which primarily quantify movement amplitude and timing, nonlinear descriptors characterize the temporal organization and complexity of movement patterns. Although these measures are not direct biomarkers of specific Parkinsonian symptoms, they may reflect underlying neuromotor impairments that influence gait and limb coordination. For example, nonlinear measures derived from wrist trajectories, such as the Hurst exponent, may capture altered movement persistence associated with reduced arm swing, rigidity, or tremor-related motor irregularities. Likewise, nonlinear descriptors extracted from ankle trajectories, including LZC, may reflect reduced movement complexity resulting from impaired lower-limb coordination, rigidity, or freezing of gait. Similarly, nonlinear features derived from knee motion may capture changes in stride dynamics and gait variability associated with disease progression. While these interpretations are physiologically plausible and consistent with known manifestations of Parkinson's disease, direct associations between individual nonlinear features and specific clinical symptoms require validation through future studies incorporating detailed clinical assessments.

A key implication of this work is that clinically meaningful gait characteristics can be captured without specialized laboratory equipment. The use of a single-camera, markerless setup significantly lowers the barrier to deployment, making the framework potentially suitable for scalable, real-world applications such as remote monitoring, early screening, and longitudinal tracking of disease progression. Furthermore, the consistent performance of models across different feature subsets and validation strategies suggests that the pipeline is not overly dependent on a specific model class, but rather benefits from the structured integration of feature engineering, selection, and evaluation.

At the same time, several limitations must be acknowledged. The dataset remains relatively small and imbalanced, with a limited number of PD subtypes, which may affect the generalizability of the results. Additionally, demographic covariates such as age, height, and gender were not included as model inputs. As noted earlier, these factors are known to influence gait characteristics and may confound certain features, particularly stride-based measures. Future work should incorporate these variables either as covariates or through stratified analysis to better understand their interaction with both kinematic and nonlinear features.

Another limitation lies in the evaluation scope. While repeated stratified cross validation and a held out test set were used to mitigate overfitting and provide robustness estimates, external validation on independent datasets is necessary to confirm the generalizability of the findings. Additionally, the nonlinear metrics used in this study function as empirical descriptors of gait variability rather than mechanistic explanations of underlying neurological processes. As such, their clinical interpretation should be considered in conjunction with broader multimodal assessments.

Finally, although the computational efficiency analysis demonstrates that models such as SVM and Random Forest are suitable for real-time or near-real-time deployment, the neural network models incur higher computational cost, which may limit their applicability in resource-constrained settings.

Overall, The class imbalance inherent to this dataset (57 NM, 10 PD-ML, 11 PD-MD) is a structural constraint of the publicly available source data and cannot be remedied by resampling alone. While SMOTE addresses the training distribution, the underlying sparsity of PD subtype samples means that per class metrics particularly macro-F1 remain sensitive to individual misclassifications on the held-out set. The notably low macro-F1 observed for Random Forest in Case 1 (0.467 against an accuracy of 0.875) directly reflects this dynamic: strong performance on the majority NM class coexists with reduced reliability on minority PD subtypes. This is an anticipated limitation of small imbalanced datasets and does not invalidate the pipeline, but it underscores the necessity of future data collection targeting PD subtype representation, including larger-scale validation, integration of demographic and clinical variables, and exploration of multimodal data sources.

5. Conclusion

This study presents a structured and interpretable machine learning framework for multiclass classification of Parkinsonian gait using features derived from single-camera video data. By combining kinematic descriptors with nonlinear dynamic features and employing a rigorous, leakage-aware evaluation pipeline, the proposed approach achieves strong and consistent performance across multiple validation settings.

A key finding is that nonlinear dynamic features particularly those derived from distal joints such as the wrist, ankle, and knee play a dominant role in classification. These features consistently contribute higher discriminative power, as evidenced by feature importance analysis, ablation studies, and their increasing proportion across feature selection cases. At the same time, kinematic features provide complementary information,

and their combination with nonlinear descriptors results in the most stable and high-performing models.

Among the evaluated feature subsets, the intermediate configuration (Case 2, 9 features) provides the most balanced trade-off between model complexity and performance, demonstrating the most consistent performance within the constraints of the available dataset. This highlights the importance of structured feature selection rather than relying solely on larger feature sets. The evaluation strategy combining repeated stratified cross validation, ablation analysis, and held-out testing further ensures that the reported results reflect robust and generalizable performance within the constraints of the dataset while accounting for the limited size and diversity of publicly available Parkinsonian gait datasets.

From a practical standpoint, the framework demonstrates that accurate and interpretable gait classification can be achieved using accessible, low-cost video data. The computational analysis further shows that efficient models such as SVM can deliver high performance with minimal latency, supporting potential real-world deployment.

In summary, this work establishes that integrating nonlinear gait dynamics with a carefully designed machine learning pipeline provides a powerful and scalable approach for Parkinson's disease classification. While further validation on larger and more diverse datasets is required, the proposed methodology offers a strong foundation for developing accessible, AI-driven tools for neuromotor assessment and disease monitoring.

6. Declaration of funding

This research received no external funding.

7. Declaration of generative AI use

During the preparation of this work, the authors used Grammarly to enhance the readability and language of the manuscript. The tool was used solely for grammar and language suggestions. All content was carefully reviewed and verified by the authors.

8. Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

9. CRediT author statement

Ranjan Das: Conceptualization, Data curation, Formal analysis, Investigation, Writing – original draft. **Navdeep Kaur**: Formal analysis, Investigation, Methodology,Validation, Visualization, Writing – review and editing. **Manish Yadav**: Formal analysis, Investigation, Methodology, Validation Visualization. **Ingita Dey Munshi**: Formal analysis, Investigation, Methodology, Validation, Visualization. **Karsimran Singh**: Supervision, Visualization. **Sachin Goyal**: Project administration, Funding acquisition, Resources, Software, Supervision, Writing – review and editing. **Sukhjit Singh Sehra**: Methodology, Data Curation, Project administration, Funding acquisition, Resources, Software, Supervision, Writing – review and editing.

10. Data Availability

The datasets generated and/or analyzed during the current study is available at Mendeley Data Repository: Kour, Navleen; Gupta, Sunanda; Arora, Sakshi(2020). "Gait dataset for knee osteoarthritis and Parkinson's disease analysis with severity levels", http://dx.doi.org/10.17632/44pfnysy89.1.

References

di Biase, L., Di Santo, A., Caminiti, M. L., De Liso, A., Shah, S. A., Ricci, L., & Di Lazzaro, V. (2020). Gait analysis in parkinson's disease: An overview of the most accurate markers for diagnosis and symptoms monitoring. *Sensors*, *20*. doi:10.3390/s20123529.

Bogachev, M., Sinitca, A., Grigarevichius, K., Pyko, N., Lyanova, A., Tsygankova, M., Davletshin, E., Petrov, K., Ageeva, T., Pyko, S., Kaplun, D., Kayumov, A., & Mukhamedshina, Y. (2023). Video-based marker-free tracking and multi-scale analysis of mouse locomotor activity and behavioral aspects in an open field arena: A perspective approach to the quantification of complex gait disturbances associated with alzheimer's disease. *Frontiers in Neuroinformatics*, *17*. doi:10.3389/fninf.2023.1101112.

Carvajal-Castaño, H., Lemos-Duque, J., & Orozco-Arroyave, J. (2022). Effective detection of abnormal gait patterns in Parkinson's disease patients using kinematics, nonlinear, and stability gait features. *Human Movement Science*, *81*, 102891. doi:10.1016/j.humov.2021.102891.

Chan, K. (2008). Video-based gait analysis by silhouette chamfer distance and kalman filter. *International Journal of Image and Graphics*, *8*, 383–418. doi:10.1142/S0219467808003155.

Cho, C.-W., Chao, W.-H., Lin, S.-H., & Chen, Y.-Y. (2009). A vision-based analysis system for gait recognition in patients with parkinson's disease. *Expert Systems with Applications*, *36*, 7033–7039. doi:10.1016/j.eswa.2008.08.076.

Goetz, C. G., Tilley, B. C., Shaftman, S. R., Stebbins, G. T., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stern, M. B., Dodel, R., Dubois, B., Holloway, R., Jankovic, J., Kulisevsky, J., Lang, A. E., Lees, A., Leurgans, S., Liechti, S., Rascol, O., Schrag, A., Teresi, J. A., van Hilten, J. J., & LaPelle, N. (2003). The unified parkinson's disease rating scale (updrs): Status and recommendations. *Movement Disorders*, *18*, 738–750. doi:10.1002/mds.10473.

Hausdorff, J. M. (2009). Gait dynamics in parkinson's disease: common and distinct behavior among stride length, gait variability, and fractal-like scaling. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *19*, 026113. doi:10.1063/1.3147408.

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An Introduction to Statistical Learning: with Applications in Python*. Springer Texts in Statistics. Springer. doi:10.1007/978-3-031-38747-0.

Kour, N., Gupta, S., & Arora, S. (2020). Gait dataset for knee osteoarthritis and parkinson's disease analysis with severity levels. Mendeley Data. doi:10.17632/44pfnysy89.1.

Kour, N., Gupta, S., & Arora, S. (2022). A vision-based clinical analysis for classification of knee osteoarthritis, parkinson's disease and normal gait with severity based on k-nearest neighbour. *Expert Systems*, *39*. doi:10.1111/exsy.12955.

Kour, N., Gupta, S., & Arora, S. (2024). A vision-based hybrid ensemble learning approach for classification of gait disorders. *Multimedia Tools and Applications*, (pp. 1–48). doi:10.1007/s11042-024-19673-z.

Liao, R., Moriwaki, K., Makihara, Y., Muramatsu, D., Takemura, N., & Yagi, Y. (2021). Health indicator estimation by video-based gait analysis. *IEICE Transactions on Information and Systems*, *E104D*, 1678–1690. doi:10.1587/transinf.2020ZDP7502.

Lukšys, D., & Griškevičius, J. (2022). Application of nonlinear analysis for the assessment of gait in patients with parkinson's disease. *Technology and Health Care*, *30*, 201–208. doi:10.3233/THC-219003.

Olleac, R., Centeno, B., Duffy, C., Nunez, M., Crespo, M., Barrios, L., Davids, J., & Kulkarni, V. (2025). Vertical video-based gait analysis for assessment of transverse plane motion: Reliability and validity in a neuromuscular population. *Journal of Pediatric Orthopaedics*, *45*, 294–299. doi:10.1097/BPO.0000000000002926.

Pratihar, R., & Sankar, R. (2024). Advancements in parkinson's disease diagnosis: A comprehensive survey on biomarker integration and machine learning. *Computers*, *13*. doi:10.3390/computers13110293.

Pérez-Toro, P. A., Vásquez-Correa, J. C., Arias-Vergara, T., Nöth, E., & Orozco-Arroyave, J. R. (2020). Nonlinear dynamics and poincaré sections to model gait impairments in different stages of parkinson's disease. *Nonlinear Dynamics*, *100*, 3253–3276. doi:10.1007/s11071-020-05691-7.

Sibley, K. G., Girges, C., Hoque, E., & Foltynie, T. (2021). Video-based analyses of parkinson's disease severity: A brief review. *Journal of Parkinson's Disease*, *11*, S83–S93. doi:10.3233/JPD-202402. PMID: 33682727.

Sigcha, L., Borzì, L., Amato, F., Rechichi, I., Ramos-Romero, C., Cárdenas, A., Gascó, L., & Olmo, G. (2023). Deep learning and wearable sensors for the diagnosis and monitoring of parkinson's disease: A systematic review. *Expert Systems with Applications*, *229*, 120541. doi:https://doi.org/10.1016/j.eswa.2023.120541.

Smyrnakis, N., Karakostas, T., & Cotton, R. (2024). Advancing monocular video-based gait analysis using motion imitation with physics-based simulation. In *2024 10th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BIOROB)* (pp. 102–108). doi:10.1109/BIOROB60516.2024.10719700.

Udhayakumar, R., Gopakumar, S., Rahman, S., & Karmakar, C. (2025). Nonlinear feature space mapping of gait dynamics for neurodegenerative disease detection. *Biomedical Signal Processing and Control*, *107*, 107800. doi:https://doi.org/10.1016/j.bspc.2025.107800.

Workman, C. D., & Thrasher, T. A. (2024). The influence of dopaminergic medication on regularity and determinism of gait and balance in parkinson's disease: A pilot analysis. *Journal of Clinical Medicine*, *13*. doi:10.3390/jcm13216485.

World Health Organization (2024). Parkinson disease. Accessed: 2025-07-19.

Ye, M., Yang, C., Stankovic, V., Stankovic, L., & Cheng, S. (2020). Distinct feature extraction for video-based gait phase classification. *IEEE Transactions on Multimedia*, *22*, 1113–1125. doi:10.1109/TMM.2019.2942479.

Yin, W., Zhu, W., Gao, H., Niu, X., Shen, C., Fan, X., & Wang, C. (2024a). Gait analysis in the early stage of parkinson's disease with a machine learning approach. *Frontiers in Neurology*, *Volume 15 - 2024*. doi:10.3389/fneur.2024.1472956.

Yin, W., Zhu, W., Gao, H., Niu, X., Shen, C., Fan, X., & Wang, C. (2024b). Gait analysis in the early stage of Parkinson's disease with a machine learning approach. *Frontiers in Neurology*, *Volume 15 - 2024*. doi:10.3389/fneur.2024.1472956.

Zhong, C., & Ng, W. W. Y. (2023). A robust frequency-domain-based graph adaptive network for parkinson's disease detection from gait data. *IEEE Transactions on Multimedia*, *25*, 7076–7088. doi:10.1109/TMM.2022.3217392.