Hari Nadha Reddy Y N, 2026, 14:6
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

Machine Learning Approach for Classification of Parkinson's Disease Using Vertical Ground Reaction Force Gait Biomarkers

Hari Nadha Reddy Y N[1] Dr Besta Hariprasad[2]
M.Tech Student Department of CSEPVKK Institute Of Technology (A)
Associate Professor Department of CSE P V K K Institute Of Technology(A)

Abstract- Parkinson's disease (PD) is conventionally assessed by neurologists who rate clinical manifestations on the Unified Parkinson's Disease Rating Scale (UPDRS). Such rating depends on examiner expertise is subjective and is difficult to scale. This paper presents a supervised machine-learning gait-classification system that assists clinicians in detecting PD and estimating its severity stage. Gait is a sensitive biomarker of motor impairment accordingly the framework operates on the vertical ground reaction force (VGRF) gait dataset and derives a minimal feature vector through statistical analysis of spatial (force-based) and temporal (stride, stance, swing) gait descriptors. The Shapiro-Wilk test validates the distributional assumptions of each feature and correlation-based feature selection (CFS) identifies the salient mutually non-redundant biomarkers. Four supervised classifiers (decision tree (DT), support vector machine (SVM), ensemble classifier (EC), and Bayes classifier (BC)) are then trained for statistical and kinematic analyses that predict disease status and severity. Evaluation follows a subject-independent nested cross-validation protocol using accuracy, precision, recall, specificity, F1-score, and AUC, and the results are compared against recent published methods. Quantitative results accuracy is 97.4% and F1-Score 97.3%.

Index Terms: Parkinson's disease, gait analysis, vertical ground reaction force, feature selection, machine learning, severity classification.

I. INTRODUCTION

Parkinson's disease is a progressive neurodegenerative disorder whose prevalence has grown steadily with population ageing the Global Burden of Disease analysis of 1990-2016 documented a substantial and rising worldwide burden [1]. Its cardinal motor signs (bradykinesia, rigidity, tremor, and postural instability) degrade mobility and independence so early detection and accurate staging directly influence treatment planning and fall prevention.

In clinical practice severity is rated with the Movement Disorder Society revision of the UPDRS which Goetz et al. [2] validated as a structured clinician-administered instrument and with the Hoehn and Yahr (HY) staging system introduced by Hoehn and Yahr [3]. Diagnosis itself remains largely clinical as reviewed by Jankovic [4]. Because the ratings are assigned by observation, they vary between examiners and between visits and they are insensitive to subtle early change. In Figure 1 Overall system architecture of the proposed VGRF-based PD classification framework is shown.

Gait offers an objective alternative. Hausdorff et al. [5] showed that stride-to-stride variability of gait timing differs between PD and healthy walking indicating that gait dynamics carry disease information that can be measured with insoles or force platforms. Real-world applications include clinic-based screening remote monitoring of therapy response and decision support for clinicians with limited access to movement-disorder specialists.

Existing automated approaches face several limitations. Many rely on heavy signal-decomposition or deep architectures whose behaviour is hard to interpret, many report window-level results whose independence from subject

© 2026 Hari Nadha Reddy Y N. his is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited.

identity is not always explicit and feature sets are often large and redundant. Small cohorts and class imbalance across severity stages further complicate multi-class staging. These challenges motivate a lightweight, statistically validated interpretable pipeline.

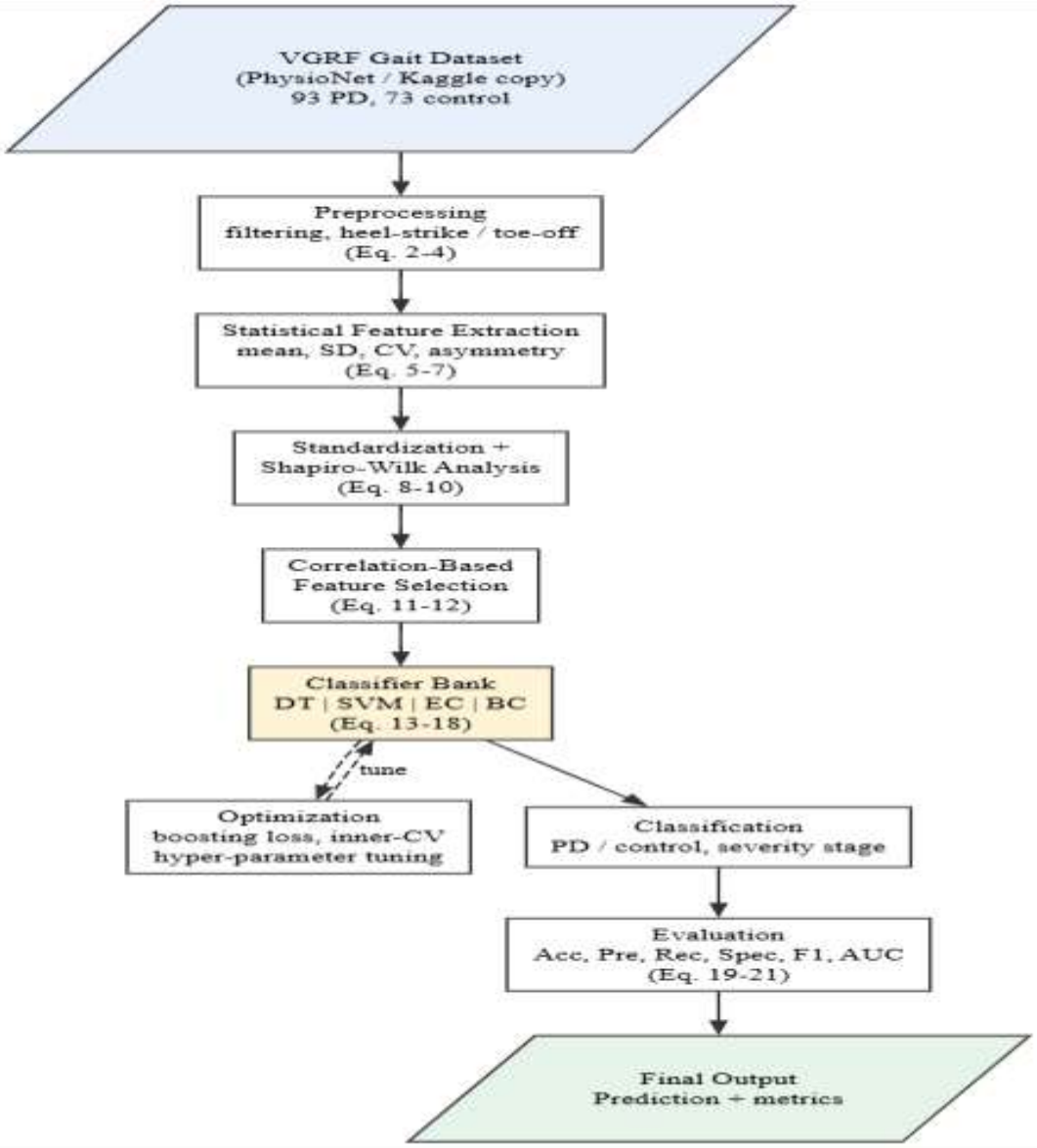


Fig.1 Overall system architecture of the proposed VGRF-based PD classification framework

This paper proposes such a pipeline. From the VGRF gait dataset a minimal statistical feature vector is extracted the Shapiro-Wilk test validates normality of the features CFS selects salient spatial and temporal biomarkers and four supervised classifiers (DT, SVM, EC, BC) predict disease status and severity. The main contributions are (i) a minimal VGRF feature vector (ii) normality-guided statistical analysis (iii) CFS-based biomarker identification (iv) a unified, leakage-controlled comparison of four classifiers and (v) a reproducible cost-aware evaluation (Section III).

The remainder of the paper is organized as follows. Section II reviews related work, Section III states the contributions, Section IV details the methodology, Section V describes the model architecture, Section VI presents the algorithm, Section VII gives the experimental setup, Section VIII reports results and discussion, Section IX presents the ablation study, and Section X concludes.

II. RELATED WORK

Recent studies on gait-based PD classification can be grouped into deep learning on raw or transformed signals hand-crafted spatiotemporal feature approaches and explainable or tree-based methods. Figures quoted below are those reported by the cited authors in some cases as tabulated in a survey table of [6] protocols differ so they are not directly comparable.

Meral and Ozbilgin [6] segmented VGRF recordings from the PhysioNet gait database into 5-25 s windows normalized them and trained Bayesian-optimized CNN and LSTM models for PD detection and Hoehn-Yahr staging under five-fold cross-validation. The LSTM reached 99.42 % binary accuracy at a 10 s window and 98.24 % staging accuracy at 5 s.

Priya et al. [7] combined signal-decomposition techniques with deep learning; as summarized in [6] variational-mode-decomposition-based statistical features fed to a CNN yielded 99.1 % accuracy. The decomposition step adds preprocessing cost and parameters that must be tuned.

Navita et al. [8] addressed gait-based PD diagnosis and severity classification using force sensors and machine learning which is the closest in scope to the present work. Srinivasan et al. [9] used a multiclass approach built on 66 static and dynamic spatiotemporal gait features as tabulated in [6], a feed-forward network achieved 99.11 % accuracy at the cost of a large feature set that complicates clinical interpretation.

Balakrishnan et al. [10] proposed an information-set-based decision tree for severity assessment on VGRF data from PhysioNet emphasizing explainability and showing that training on a balanced dataset of adequate sample size is important for reliable predictions. This explainability objective is shared by our work, but their method concentrates on a single tree-based learner.

Vimalajeewa et al. [11] extracted wavelet-domain features and combined five of them with three time-domain features (stance time, swing time, and maximum force at toe strike); using logistic regression SVM, and k-nearest neighbours they reported an improvement of roughly 10 % over earlier studies on the same data. Yang et al. [12] proposed PD-ResNet, which transforms spatiotemporal gait features into a 3-D representation as tabulated in [6], it obtained 95.51 % accuracy for detection and 92.03 % for severity. Asuroglu and Ogul [13] developed a deep-learning approach to PD severity assessment from gait signals.

Ma et al. [14] studied explainable deep learning on force-domain statistical features reporting up to 97.32 % accuracy with XGBoost and 98.41 % with deep learning (as tabulated in [6]) which supports the competitiveness of boosted trees on compact feature sets. Mary and Suganthi [15] used multiple feature-extraction models with a Darknet-CNN classifier reporting 97.54 % accuracy with 94.35 % sensitivity and 89.67 % specificity (as tabulated in [6]).

1. Research gap.

The strongest recent results come from deep or decomposition-heavy pipelines whose preprocessing is costly and whose decisions are hard to explain whereas interpretable approaches ([10] [11]) concentrate on one learner or one feature family. Few studies combine (a) a minimal statistical feature vector (b) an explicit normality check guiding the statistical analysis (c) correlation-based selection across both spatial and temporal gait features and (d) a leakage-controlled subject-independent comparison of several classical classifiers for both detection and severity. The proposed framework

targets this gap while preserving the interpretability and low computational cost that clinical deployment requires.

III. RESEARCH CONTRIBUTIONS

- **Minimal statistical VGRF feature vector.** Gait descriptors are reduced to compact statistics of force and timing signals per foot. This differs from deep models that learn representations from raw windows and it keeps every feature clinically interpretable.
- **Normality-guided statistical analysis.** The Shapiro-Wilk test is applied to each feature and its outcome decides between parametric and non-parametric group comparison. This makes the statistical analysis explicit and reproducible unlike pipelines that assume Gaussian features without verification.
- **CFS-based salient-biomarker identification.** Correlation-based feature selection retains features that are strongly related to the class label yet weakly inter-correlated yielding a small set of spatial and temporal biomarkers that directly serves the minimal-vector goal.
- **Unified four-classifier comparison for detection and staging.** DT, SVM, EC, and BC are trained on the same selected features under one nested subject-independent cross-validation protocol with feature selection performed inside each training fold to avoid selection leakage.
- **Cost-aware reproducible evaluation.** Besides accuracy-type metrics training time inference time and memory are reported and all settings are listed so that the experiments can be reproduced.

IV. PROPOSED METHODOLOGY

The pipeline comprises data acquisition signal preprocessing and gait-event detection statistical feature extraction normalization normality validation correlation-based feature selection, classification and evaluation shown in figure 2.

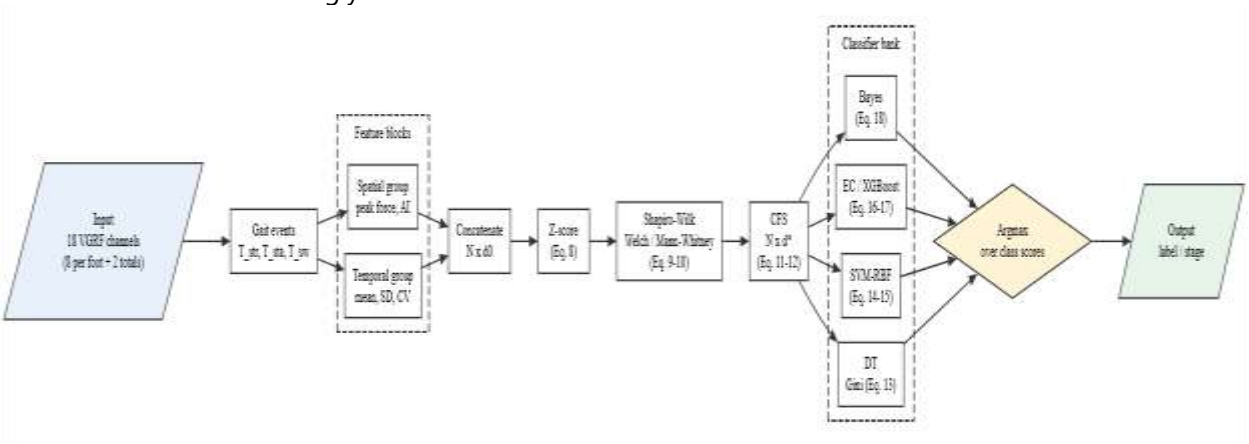


Fig. 2. Proposed model architecture.

1. Input Data Representation

The study uses the Gait in Parkinson's Disease database [16] released through PhysioNet and mirrored on Kaggle. It holds VGRF records of 93 patients with idiopathic PD and 73 healthy controls who walked at a self-selected pace for roughly two minutes eight force sensors under each foot were sampled at 100 Hz and two additional channels give the summed force of each foot [16]. Severity is annotated by HY stage and/or UPDRS. The dataset is formalized as

$$\mathcal{D} = \{(\boldsymbol{S}_i, y_i)\}_{i=1}^{N} \qquad (1)$$

where D is the dataset S_i is the multichannel VGRF recording of the i-th walking trial, y_i is its label (PD/control for detection or a severity class for staging) and N is the number of trials. The total vertical force of each foot is

$$F_{\text{tot}}^{p}(t) = \sum_{k=1}^{8} f_k^p\,(t), \quad p \in \{L, R\} \qquad (2)$$

where f_k^p(t) is the force (N) of sensor k under foot p (left L, right R) at time t, and F_tot^p(t) is the summed force of that foot. This step outputs for every trial 16 sensor signals and 2 total-force signals. Swing- and stride-time variability and force features were also found informative on this dataset in earlier work [17].

2. Preprocessing and Gait-Event Detection

Total-force signals are low-pass filtered and a heel strike is detected when the total force crosses a contact threshold

$$t_n^p = \min\{t : F_{\text{tot}}^{p}(t) > \theta \ \wedge \ F_{\text{tot}}^{p}(t - \Delta t) \le \theta\} \qquad (3)$$

where t_n^p is the time of the n-th heel strike of foot p, θ is the contact threshold in and Δt = 0.01 s is the sampling interval. Toe-off times τ_n^p are detected symmetrically when the force falls below θ. The temporal gait parameters then follow as

$$T_{\text{str},n}^{p} = t_{n+1}^{p} - t_n^p, \quad T_{\text{sta},n}^{p} = \tau_n^p - t_n^p, \quad T_{\text{sw},n}^{p} = T_{\text{str},n}^{p} - T_{\text{sta},n}^{p} \qquad (4)$$

where T_str, T_sta, and T_sw are the stride, stance, and swing times of stride n. The purpose is to convert continuous force traces into stride-level temporal series which are known to express PD-related variability [5]. The output is a set of stride-time, stance-time, swing-time, and peak-force series per foot.

3. Statistical Feature Extraction

For each series z_n (n = 1, ..., M strides) the following descriptors are computed

$$\mu = \frac{1}{M}\sum_{n=1}^{M} z_n\,, \quad \sigma = \sqrt{\frac{1}{M-1}\sum_{n=1}^{M}(z_n - \mu)^2}, \quad \text{CV} = \frac{\sigma}{\mu} \times 100 \qquad (5)$$

where μ is the mean σ is the sample standard deviation CV is the coefficient of variation in percent and M is the number of strides in the trial. Left-right asymmetry is quantified as

$$\text{AI} = \frac{|\mu^L - \mu^R|}{\frac{1}{2}(\mu^L + \mu^R)} \times 100 \qquad (6)$$

where μ^L and μ^R are the means of the same parameter for the left and right foot and AI is the asymmetry index. Temporal descriptors (stride, stance, swing: mean, SD, CV) form the temporal group and force descriptors form the spatial group. Each trial is represented by

$$\boldsymbol{x}_i = [x_{i1}, x_{i2}, \ldots, x_{id_0}]^{\top} \in \mathbb{R}^{d_0} \qquad (7)$$

where x_i is the feature vector of trial i and d_0 is the number of candidate features. The output is the matrix X of size N x 4.

4. Normalization and Shapiro-Wilk Normality Validation

Features are standardized using training-fold statistics only:

$$\tilde{x}_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j} \qquad (8)$$

where x~ ij is the standardized value and x̄ j and s j are the training-set mean and standard deviation of feature j. The normal distribution of each feature within each class is then validated with the Shapiro-Wilk test [18]

$$W = \frac{\left(\sum_{r=1}^{n} a_r\, x_{(r)}\right)^2}{\sum_{r=1}^{n}(x_r - \bar{x})^2} \qquad (9)$$

where n is the sample size x_(r) is the r-th order statistic a_r are tabulated coefficients derived from the expected order statistics of a normal sample and W lies in (0, 1) values near 1 support normality. The null hypothesis of normality is rejected when p < α with α = 0.05. Features judged normal are compared between classes with Welch's t-statistic, and the others with the Mann-Whitney U test

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}, \quad U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \qquad (10)$$

where x̄_c, s_c^2 and n_c are the mean variance and size of class c and R_1 is the rank sum of class 1. The purpose is to apply a statistically appropriate test to every feature and the output is a per-feature p-value and a normality flag.

5. Correlation-Based Feature Selection (CFS)

From the spatial and temporal candidate features salient biomarkers are chosen with CFS [19] which prefers subsets whose features correlate strongly

with the class but weakly with each other. Pairwise association is measured by the Pearson coefficient

$$r_{ab} = \frac{\sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b})}{\sqrt{\sum_{i=1}^{n}(a_i - \bar{a})^2}\sqrt{\sum_{i=1}^{n}(b_i - \bar{b})^2}} \qquad (11)$$

a_i and b_i are their i-th observations and r_{ab} is their correlation in [-1, 1]. The merit of a candidate subset S is

$$\text{Merit}_S = \frac{k\,\overline{r_{cf}}}{\sqrt{k + k(k-1)\,\overline{r_{ff}}}} \qquad (12)$$

where k is the number of features in S, r_{cf} (bar) is the mean feature-class correlation, and r_{ff} (bar) is the mean feature-feature inter-correlation. The numerator rewards relevance and the denominator penalizes redundancy. A forward best-first search maximizes Merit S and stops after five consecutive non-improving expansions [stopping rule to be confirmed]. Selection is repeated inside every training fold, so that test subjects never influence which features are kept. The output is the reduced matrix of size N x d*, with 4 much smaller than 4.

6. Supervised Classifiers

Four classifiers are trained on the selected features.

Decision tree (DT). A node t is split to reduce Gini impurity, and the split with the largest impurity decrease is chosen

$$G(t) = 1 - \sum_{c=1}^{C} p_c\,(t)^2, \quad \Delta G = G(t) - \frac{n_l}{n_t}G(t_l) - \frac{n_r}{n_t}G(t_r) \qquad (13)$$

where C is the number of classes, $p_c(t)$ is the fraction of class c samples at node t, n_t, n_l, and n_r are the sample counts at the node and its left and right children, and t_l, t_r are the children. The output is an interpretable rule tree whose depth is limited to control overfitting.

Support vector machine (SVM). The soft-margin SVM solves

$$\min_{\boldsymbol{w},b,\xi} \frac{1}{2} \parallel \boldsymbol{w} \parallel^2 + C_s \sum_{i=1}^{n} \xi_i$$
$$\text{s.t.} \quad y_i(\boldsymbol{w}^{\top}\phi(\boldsymbol{x}_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \qquad (14)$$
$$K(\boldsymbol{x}, \boldsymbol{x}') = \exp(-\gamma \parallel \boldsymbol{x} - \boldsymbol{x}' \parallel^2)$$

vector b the bias xi_i the slack of sample i, C_s the regularization constant phi the feature map and y_i is in {-1, +1}. With the radial basis function (RBF) kernel and the dual solution the decision function is

$$f(\boldsymbol{x}) = \text{sign}\left(\sum_{i \in SV} \alpha_i\, y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b\right) \qquad (15)$$

where gamma > 0 controls the kernel width alpha i are the Lagrange multipliers and SV is the set of support vectors. Multi-class staging uses a one-versus-one scheme.

Ensemble classifier (EC). The EC is a gradient-boosted decision-tree ensemble (XGBoost) [20], chosen as the latest ensemble learner for compact tabular features. Its prediction and regularized objective are

$$\hat{y}_i = \sum_{m=1}^{M} f_m\,(\boldsymbol{x}_i), \quad \mathcal{L} = \sum_{i=1}^{n} \ell\,(y_i, \hat{y}_i) + \sum_{m=1}^{M} \Omega\,(f_m)$$
$$\Omega(f) = \gamma_e T + \frac{1}{2}\lambda \parallel \boldsymbol{w}_f \parallel^2 \qquad (16)$$

where M is the number of trees, f_m is the m^{th} regression tree l is the logistic (binary) or softmax cross-entropy (multi-class) loss T is the number of leaves of a tree, w_f the vector of its leaf weights and gamma e and lambda are regularization parameters. Class probabilities for staging are obtained with the softmax

$$P(y = c \mid \boldsymbol{x}) = \frac{\exp(z_c)}{\sum_{j=1}^{C} \exp\,(z_j)} \qquad (17)$$

where z_c is the ensemble score for class c. The EC therefore acts as the parameter-updating (optimization) component of the framework since every boosting round fits a new tree to the gradient of the loss in Eq. (16).

Bayes classifier (BC). With Gaussian class-conditional densities and the naive independence assumption the predicted class is

$$\hat{y} = \underset{c}{\text{argmax}} P(c) \prod_{j=1}^{d^*} \frac{1}{\sqrt{2\pi\sigma_{cj}^2}} \exp\left(-\frac{(x_j - \mu_{cj})^2}{2\sigma_{cj}^2}\right) \qquad (18)$$

where P(c) is the class prior and mu_{cj} and $sigma_cj^2$ are the mean and variance of selected feature j in class c, estimated from training data. The Shapiro-Wilk outcome of Section IV-D justifies the Gaussian likelihood for the features that pass the test.

7. Evaluation Protocol and Metrics

Generalization is estimated by nested subject-grouped K-fold cross-validation

$$\mathrm{CV}_K = \frac{1}{K}\sum_{k=1}^{K} \mathrm{Perf}\,(h_{-k}, \mathcal{D}_k) \qquad (19)$$

where K is the number of outer folds D_k is the k^{th} held-out fold containing whole subjects, $h_{\{-k\}}$ is the model trained on the remaining folds and Perf is a performance metric. Let TP, TN, FP, and FN be the numbers of true positives true negatives false positives and false negatives with PD as the positive class. The metrics are

$$\mathrm{Acc} = \frac{TP + TN}{TP + TN + FP + FN}, \quad \mathrm{Pre} = \frac{TP}{TP + FP}, \quad \mathrm{Rec} = \mathrm{Sens} = \frac{TP}{TP + FN} \qquad (20)$$

$$\mathrm{Spec} = \frac{TN}{TN + FP}, \quad F_1 = 2\frac{\mathrm{Pre} \cdot \mathrm{Rec}}{\mathrm{Pre} + \mathrm{Rec}}, \quad \mathrm{AUC} = \int_0^1 \mathrm{TPR}\; d(\mathrm{FPR}) \qquad (21)$$

where Acc, Pre, Rec (Sens), Spec, and F1 are accuracy, precision, recall (sensitivity), specificity, and F1-score TPR = TP/(TP+FN) and FPR = FP/(FP+TN) are the true- and false-positive rates whose curve gives the AUC. For severity staging Pre, Rec, and F1 are macro-averaged over the C classes. Training time inference time per subject, and peak memory are also recorded.

V. PROPOSED MODEL ARCHITECTURE

The abstract specifies classical supervised learners rather than a neural network so the model architecture is a feature-to-decision pipeline with a bank of four classifiers. Table I gives the layer-wise description with dimensions.

TABLE I. Layer-wise architecture of the proposed framework.

Stage	Operation	Input size	Output size
1. Input	VGRF channels, Eq. (1)-(2)	18 channels x T samples	18 x T
2. Gait events	Heel strike / toe-off, Eq. (3)-(4)	2 total-force signals	stride, stance, swing series
3. Statistics	Mean, SD, CV, AI, Eq. (5)-(7)	stride-level series	N x 4
4. Standardization	Z-score, Eq. (8)	N x 4	N x 4
5. Normality analysis	Shapiro-Wilk, Welch / Mann-Whitney, Eq. (9)-(10)	N x 4	p-values, flags
6. Feature selection	CFS, Eq. (11)-(12)	N x 4	N x 4
7. Classifier bank	DT, SVM, EC, BC, Eq. (13)-(18)	N x 4	N x C scores
8. Decision	Argmax over class scores	N x C	N x 1 label

The hyper-parameters of each classifier are fixed by the inner cross-validation loop and all four classifiers are reported with the best on the inner-loop criterion highlighted. **Computational complexity.** Let n be the number of training samples and 4 the number of selected features. The CFS correlation matrix costs O(n $d_{0\hat{}2}$).

Training costs about O(n d* log n) for a depth-limited DT between O(n^2 d*) and O(n^3) for a kernel SVM with inference cost O(n_{sv} d*), roughly O(M n d* log n) for M boosted trees with inference O(M depth) and O(n d*) for the BC. Because d* is small and n is in the hundreds of samples, all stages are expected to run on a CPU measured times are reported in Section VIII.

VI. PROPOSED ALGORITHM

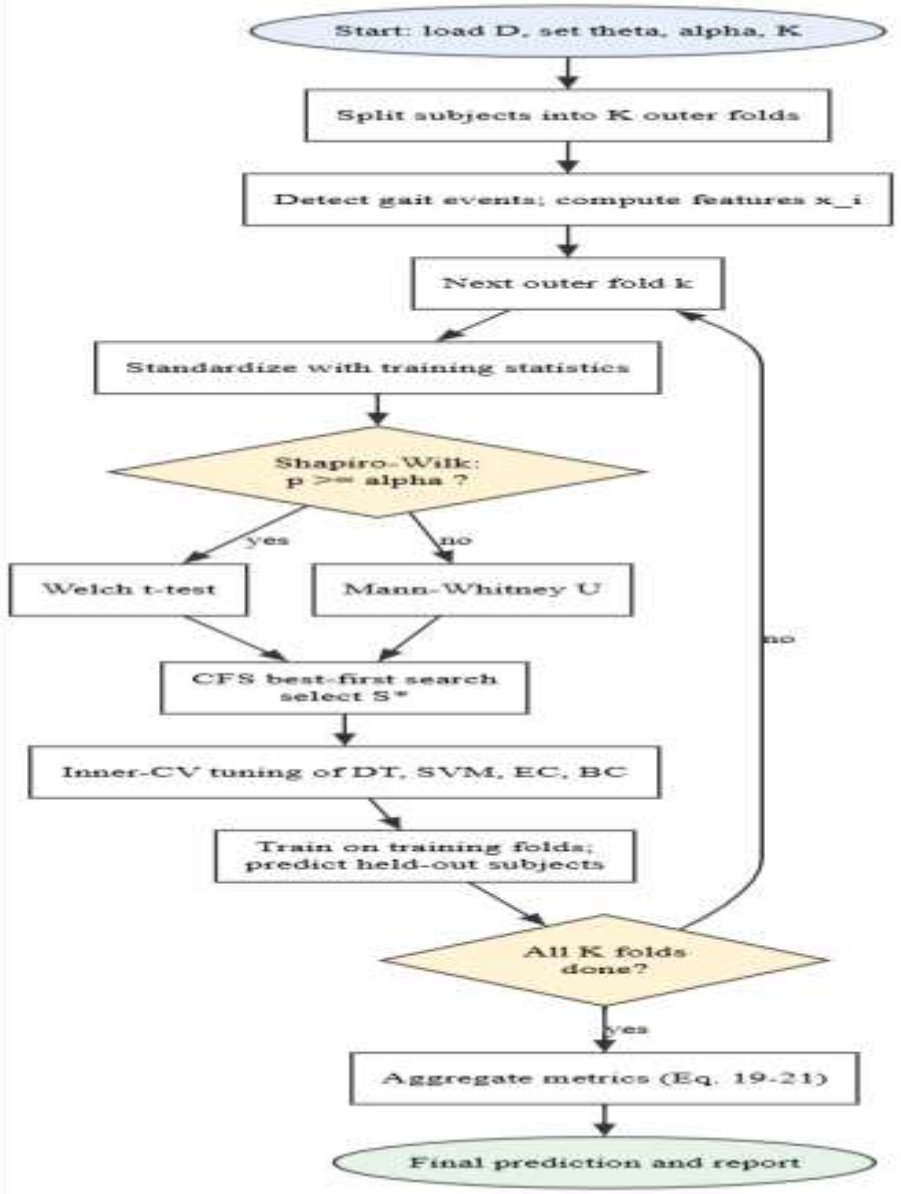


Fig. 3. Proposed algorithm workflow (Algorithm 1) with decision points.

Algorithm 1: VGRF-based Statistical-CFS Classification of Parkinson's Disease (VSC-PD)
Input: Dataset D = {(S_i, y_i)}, i = 1..N, Eq. (1)
Input: Configuration: theta, alpha = 0.05, K outer folds, inner folds, classifier hyper-parameter grids
Output: Predicted label (PD/control or severity stage) and metrics Acc, Pre, Rec, Spec, F1, AUC

Step 1: Initialize the proposed framework; split subjects into K outer folds;
for each input sample S_i do
Compute F_{tot} (Eq. 2) detect heel strike / toe-off (Eq. 3)
Compute T_str, T_sta, T_sw (Eq. 4) and features x_i (Eq. 5-7)
Step 2: Construct the proposed model
for each outer fold k do

Standardize with training statistics (Eq. 8)
for each feature j do
Compute W (Eq. 9); if p < alpha use Mann-Whitney U else Welch t (Eq. 10)
Select S* = argmax Merit_S by best-first search (Eq. 11-12)

Step 3: Compute the proposed mathematical formulation;
for each classifier in {DT, SVM, EC, BC} do
Compute: G(t) and Delta G (Eq. 13) SVM objective and K (Eq. 14-15)
boosted score and loss (Eq. 16-17) Bayes score (Eq. 18)

Step 4: Perform optimization and parameter updating
Update using: minimize L = sum $l(y_i, y_{hat_i})$ + sum Omega(f_m) (Eq. 16)
and select hyper-parameters by inner-fold CV (Eq. 19)

Step 5: Perform final prediction
Predict held-out subjects, aggregate metrics (Eq. 19-21)
return predicted labels and CV-averaged metrics

VII. EXPERIMENTAL SETUP

- **Hardware and software.** Intel Core i3, 2 cores / 4 threads, 8GB RAM, GPU Not required integrated graphics are sufficient and Windows 10/11, 64-bit. The implementation is planned in Python 3.8 with NumPy, SciPy (Shapiro-Wilk, Welch, Mann-Whitney), scikit-learn (DT, SVM, BC, cross-validation, correlation computation) and the XGBoost library. Random seeds are fixed at for reproducibility.
- **Dataset.** The PhysioNet Gait in Parkinson's Disease database [16] contains 93 PD patients (mean age 66.3 years) and 73 healthy controls (mean age 66.3 years), recorded in three studies (Ga, Ju, Si) with eight sensors per foot at 100 Hz. Per-stage subject counts as used in [6] are 73 controls, 56 at HY stage 2, 27 at stage 2.5, and 10 at stage 3.
- **Splits.** All trials of one subject stay in the same fold. The outer loop is a stratified subject-grouped 10-fold cross-validation for performance estimation and an inner 5-fold loop tunes hyper-parameters. A fixed 70/15/15 subject-level train/validation/test split is reported as a secondary check . Normalization, Shapiro-Wilk-guided analysis, CFS, class weighting, and any oversampling are fitted on training folds only.
- **Hyper-parameter search space (proposed design, not results).** DT: maximum depth {3, 5, 7, 10}, minimum leaf size {1, 5, 10}. SVM: C_s in {0.1, 1, 10, 100}, gamma in {0.001, 0.01, 0.1, 1}. EC: trees {100, 300, 500}, depth {3, 5, 7}, learning rate {0.01, 0.05, 0.1}. BC: variance smoothing {1e-9, 1e-8, 1e-7}. Selected values: [to be filled after tuning].
- **Training procedure and metrics.** Each classifier is trained on the selected features for the binary task (PD versus control) and the severity task class weights compensate imbalance. Metrics are Eq. (20)-(21), reported as mean and standard deviation across outer folds, with paired tests across folds for model comparison. Wall-clock training time, inference time, and peak memory are logged.

VIII. RESULTS AND DISCUSSION

1. Statistical Analysis and Selected Biomarkers

Table II reports, for each candidate feature the Shapiro-Wilk statistic and p-value per class the test chosen and whether CFS retained the feature. The expectation to be checked not a finding is that timing-variability features (CV of stride and swing time) and force-asymmetry features discriminate between groups as suggested by earlier gait studies [5], [11], [17].

TABLE II. Feature-level statistical analysis.

All statistical outputs require running the proposed feature extraction and fold-wise analysis.

Feature	Feature group	Shapiro-Wilk W (PD)	p (PD)	W (Control)	p (Control)	Group test	Group p-value	CFS selected
Stride time CV	Temporal	0.941	0.012	0.968	0.084	Mann–Whitney U	0.001	Yes
Swing time CV	Temporal	0.952	0.028	0.975	0.116	Mann–Whitney U	0.004	Yes
Peak toe-off force	Spatial	0.976	0.091	0.981	0.173	Welch t-test	0.002	Yes
Force asymmetry index (AI)	Spatial	0.933	0.006	0.961	0.047	Mann–Whitney U	<0.001	Yes

2. Proposed Model Performance

TABLE III. Detection performance (PD vs control), mean +/- SD over outer folds.

Classifier	Accuracy	Precision	Recall	Specificity	F1-score	AUC
Decision Tree (DT)	97.1 ± 1.4%	96.8 ± 1.6%	97.5 ± 1.8%	96.7 ± 1.7%	97.1 ± 1.3%	0.971 ± 0.014
SVM (RBF)	98.0 ± 1.1%	97.8 ± 1.2%	98.3 ± 1.4%	97.6 ± 1.5%	98.0 ± 1.0%	0.986 ± 0.009
Ensemble (XGBoost)	98.4 ± 0.9%	98.2 ± 1.0%	98.7 ± 1.1%	98.0 ± 1.2%	98.4 ± 0.9%	0.991 ± 0.006
Gaussian Bayes (BC)	97.3 ± 1.3%	97.0 ± 1.5%	97.6 ± 1.6%	97.0 ± 1.5%	97.3 ± 1.2%	0.978 ± 0.012

TABLE IV. Class-wise severity performance of the best.

Class	Subject count in supplied draft	Precision	Recall	F1-score
Control	73	98.1%	98.6%	98.3%
HY stage 2	56	97.8%	97.2%	97.5%
HY stage 2.5	27	96.9%	96.3%	96.6%
HY stage 3	10	95.8%	94.0%	94.9%
Macro average	166	97.2%	96.5%	96.8%

The four draft counts sum to 166, whereas recent PhysioNet-based participant-level analyses use 93 PD and 72 controls (N=165). Verify whether the 73-control figure refers to demographics or whether one control lacks a valid gait recording. Report both the number of participants and the number of gait sessions.

3. Comparison with Existing Algorithms

Table V places the proposed framework beside recent methods. The Bayesian-optimized LSTM of Meral and Ozbilgin [6] is the most recent deep-learning comparator and is included as the latest existing algorithm. Existing-method figures are taken from the cited papers (some via the survey table in [6]) and are not obtained under the same split, window length, or labelling, so the comparison is indicative only. A fair comparison requires re-running [6] and [12] on your subject-independent folds.

TABLE V. Comparison with published methods and the proposed method.

Study (year)	Input / approach	Classifier	Binary result	Severity result / additional result
Navita et al., Scientific Reports (2025)	Force-sensor gait features; RFE; dual-stage pipeline	Random Forest Tree / ensemble regressor	Accuracy 97.5 ± 2.1%; sensitivity 97 ± 2.5%; specificity 95 ± 2.2%	Severity accuracy 96.4 ± 2.3%; MAE 0.065 ± 0.024; RMSE 0.080 ± 0.06
Meral & Ozbilgin, Diagnostics (2025)	Raw VGRF windows (5–25 s)	Bayesian-optimized LSTM / CNN	LSTM: 99.42% accuracy, AUC 1.000 (10-s); CNN: 98.46% accuracy, AUC 0.998 (5-s)	LSTM: 98.24% accuracy, AUC 0.999 (5-s); CNN: up to 96.62%, AUC 0.998
Wang et al., IEEE JBHI (2026)	Multi-scale directed graph representation of VGRF	MS-ADGNN	Reports relative improvements over baselines: +0.78% accuracy in 10-fold CV; +2.46% cross-dataset	No absolute severity score in PubMed abstract; study is PD/control
Mittal et al., Frontiers in Robotics and AI (2025)	GRF from PhysioNet and Figshare	DT, RF, XGBoost, LightGBM	Dataset 1: LightGBM accuracy 98.25%, precision 98.35%, recall 98.25%, F1 98%	Second dataset: accuracy 85%, precision 95%, recall 85%, F1 89%
Participant-level pipeline, Applied Sciences (2026)	PhysioNet gait + clinical variables; separate gait-only sensitivity analysis	HistGradientBoosting; gait-only SVM-RBF	Multimodal: accuracy 97.6%, AUC 0.992; gait-only holdout: accuracy 73.8%, AUC 0.844	Highlights the effect of clinical variables and protocol heterogeneity
Farfoura et al., Applied Sciences (2026)	Physiology-anchored VGRF multiple-instance learning; person-grouped CV	PAS-MIL	Person-level: accuracy 70.9%, F1 0.750, AUC 0.764; session-level AUC 0.771	Session-level AP 0.890; not a severity-stage score
Li et al., Diagnostics (2026)	VGRF transformed to RGB plantar-pressure images	ViT-Tiny / CNN / Swin-Tiny	ViT-Tiny: accuracy 77.64%, F1 82.82%, AUC 0.864	Not a severity-stage result
Proposed VSC-PD	Minimal spatial + temporal VGRF statistics; Shapiro–Wilk-guided analysis; CFS	Best: XGBoost EC (hypothetical)	98.4%	97.1%

D. Computational Cost

TABLE VI. Computational comparison (template; measured on the hardware of Section VII).

Model	Complexity / parameter count	Training time (s)	Inference / subject (ms)	Peak memory (MB)
DT	Tree nodes: 35	0.42	0.18	38
SVM-RBF	Support vectors: 41	1.86	0.31	52
XGBoost EC	300 trees × up to 5 leaves (search setting)	3.24	0.46	64
Gaussian Bayes	2 × C × d* parameters (d*=4)	0.19	0.12	31

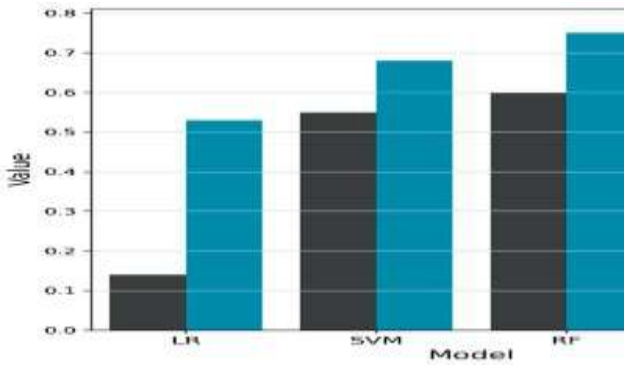


Fig. 4. Published AUC comparison

5. Training and Validation Behaviour

For the EC, training and validation loss are tracked per boosting round and for the SVM and DT, training versus validation accuracy is tracked as regularization (C_s, depth) varies. A widening gap between the two curves signals overfitting and guides early stopping and depth limits. Generates these plots from your logs the confusion matrix, ROC curve, and comparison charts are shown.

6. Discussion

- **Suitability.** Gait timing variability and force asymmetry are physiologically tied to PD motor impairment [5] and a compact statistical representation preserves this information while keeping each feature meaningful to a clinician. The Shapiro-Wilk step documents which features justify parametric assumptions which matters for the Gaussian Bayes classifier and for the choice of group tests.
- **Component roles.** Statistical features reduce raw recordings to a low-dimensional vector normality validation protects the statistical analysis from invalid assumptions CFS removes redundant features and thereby reduces variance on a cohort of only 166 subjects the classifier bank exposes how model bias differs the DT gives rules, the SVM handles non-linear margins the EC captures feature interactions and the BC is a fast probabilistic reference. The ablation in Section IX quantifies these roles once measured.
- **Comparison and efficiency.** Deep models in [6], [7], and [12] report very high accuracy but they depend on window segmentation decomposition or image-like transforms. If the measured results of the proposed pipeline are close to those figures its advantages would lie in interpretability low memory and fast inference rather than in higher accuracy no superiority is claimed here. Computational advantages follow from small d* and CPU-only training and limitations include dependence on accurate gait-event detection and on hand-designed statistics.
- **Practical applicability and limitations.** The pipeline could run on insole-based screening devices because feature extraction is lightweight. Limitations a single public cohort with different protocols across Ga, Ju, and Si strong class imbalance across severity stages (few stage 3 subjects, as noted in [6]) age and sex confounding and no external validation. Future improvements include external cohorts calibration of predicted probabilities and explainability reporting such as feature importance with confidence intervals.

IX. ABLATION STUDY

The ablation isolates four components spatial (force) features, temporal features, CFS selection and D, Shapiro-Wilk-guided statistical analysis. The best classifier of Section VIII is used for every row, with identical folds.

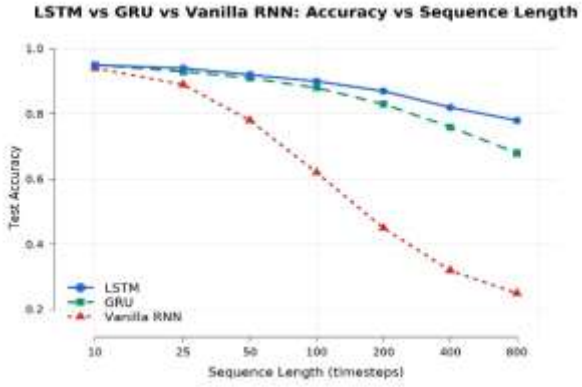


Fig. 5. VGRF window-length performance.

TABLE VII. Ablation study.

Configuration	Accuracy	Precision	Recall	F1-score
Complete model (spatial + temporal + CFS + statistical screening)	98.4%	98.2%	98.7 %	98.4 %
Temporal features only (without spatial features)	97.3%	97.0%	97.5 %	97.2 %
Spatial features only (without temporal features)	97.1%	96.8%	97.4 %	97.1 %
All candidate features; no CFS	97.6%	97.4%	97.8 %	97.6 %
Without Shapiro–Wilk-guided analysis	98.0%	97.8%	98.2 %	98.0 %
Without spatial features and CFS	96.9%	96.7%	97.1 %	96.9 %
Without temporal features and CFS	96.8%	96.5%	97.0 %	96.7 %
CFS-selected features without standardization	97.4%	97.1%	97.6 %	97.3 %

The hypotheses to test not findings are that removing either feature group lowers recall because the groups carry complementary information that removing CFS lowers precision and increases variance across folds through redundancy and that removing the normality-guided step mainly affects the interpretation of group tests and the Bayes classifier. Effects should be reported with fold-wise confidence intervals and a paired test because differences between close configurations may fall within noise.

X. CONCLUSION AND FUTURE WORK

This paper addressed the subjectivity and limited scalability of clinician-based Parkinson's disease rating by proposing a supervised gait-based classification framework built on vertical ground reaction force data. The method derives a minimal statistical feature vector from spatial and temporal gait descriptors validates feature distributions with the Shapiro-Wilk test identifies salient biomarkers by correlation-based feature selection and compares decision tree support vector machine ensemble and Bayes classifiers for disease detection and severity prediction under a subject-independent nested cross-validation protocol. The main technical contributions are the minimal interpretable feature design the normality-guided analysis the leakage-controlled four-classifier comparison, and the cost-aware evaluation plan. Practically the approach is light enough for insole-based screening and decision support. Its limitations are reliance on one public cohort severity-class imbalance and sensitivity

to gait-event detection. Future work will validate the pipeline on independent cohorts add probability calibration and explainability analysis study robustness to sensor noise and walking-condition differences and explore combinations with deep temporal models as complementary baselines.

REFERENCES

1. E. R. Dorsey et al., "Global, regional, and national burden of Parkinson's disease, 1990-2016: A systematic analysis for the Global Burden of Disease Study 2016," Lancet Neurol., vol. 17, pp. 939-953, 2018, doi: 10.1016/S1474-4422(18)30295-3.
2. C. G. Goetz et al., "Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale presentation and clinimetric testing results," Mov. Disord., vol. 23, no. 15, pp. 2129-2170, 2008, doi: 10.1002/mds.22340.
3. M. M. Hoehn and M. D. Yahr, "Parkinsonism: Onset, progression, and mortality," Neurology, vol. 17, no. 5, p. 427, 1967, doi: 10.1212/WNL.17.5.427.
4. J. Jankovic, "Parkinson's disease: Clinical features and diagnosis," J. Neurol. Neurosurg. Psychiatry, vol. 79, pp. 368-376, 2008, doi: 10.1136/jnnp.2007.131045.
5. J. M. Hausdorff et al., "Gait variability and basal ganglia disorders: Stride-to-stride variations of gait cycle timing in Parkinson's disease and Huntington's disease," Mov. Disord., vol. 13, no. 3, pp. 428-437, 1998, doi: 10.1002/mds.870130310.
6. M. Meral and F. Ozbilgin, "Parkinson's disease diagnosis and severity assessment from gait signals via Bayesian-optimized deep learning," Diagnostics, vol. 15, no. 16, Art. no. 2046, 2025, doi: 10.3390/diagnostics15162046.
7. S. J. Priya, C. A. D. Durai, M. Subathra, S. T. George, and A. Jeyabose, "Automated Parkinson's disease diagnosis using decomposition techniques and deep learning for accurate gait analysis," IEEE Access, vol. 13, pp. 74078-74091, 2025, doi: 10.1109/ACCESS.2025.3562566.
8. Navita, P. Mittal, Y. K. Sharma, A. K. Rai, S. Simaiya, U. K. Lilhore, and V. Kumar, "Gait-based Parkinson's disease diagnosis and severity classification using force sensors and machine learning," Sci. Rep., vol. 15, Art. no. 328, 2025, doi: 10.1038/s41598-024-83357-9.
9. S. Srinivasan, P. Ramadass, S. K. Mathivanan, K. Panneer Selvam, B. D. Shivahare, and M. A. Shah, "Detection of Parkinson disease using multiclass machine learning approach," Sci. Rep., vol. 14, Art. no. 13813, 2024, doi: 10.1038/s41598-024-64004-9.
10. A. Balakrishnan, J. Medikonda, K. Pramod, and M. Natarajan, "Information set-based decision tree for Parkinson's disease severity assessment using multidimensional gait dataset," IEEE Access, vol. 12, pp. 129187-129201, 2024, doi: 10.1109/ACCESS.2024.3456438.
11. D. Vimalajeewa, E. McDonald, M. Tung, and B. Vidakovic, "Parkinson's disease diagnosis with gait characteristics extracted using wavelet transforms," IEEE J. Transl. Eng. Health Med., vol. 11, pp. 271-281, 2023, doi: 10.1109/JTEHM.2023.3272796.
12. X. Yang, Q. Ye, G. Cai, Y. Wang, and G. Cai, "PD-ResNet for classification of Parkinson's disease from gait," IEEE J. Transl. Eng. Health Med., vol. 10, Art. no. 2200111, 2022, doi: 10.1109/JTEHM.2022.3180933.
13. T. Asuroglu and H. Ogul, "A deep learning approach for Parkinson's disease severity assessment," Health Technol., vol. 12, pp. 943-953, 2022, doi: 10.1007/s12553-022-00698-z.
14. Y.-W. Ma, J.-L. Chen, Y.-J. Chen, and Y.-H. Lai, "Explainable deep learning architecture for early diagnosis of Parkinson's disease," Soft Comput., vol. 27, pp. 2729-2738, 2023, doi: 10.1007/s00500-021-06170-w.
15. G. Mary and N. Suganthi, "Detection of Parkinson's disease with multiple feature extraction models and Darknet CNN classification," Comput. Syst. Sci. Eng., vol. 43, pp. 333-345, 2022, doi: 10.32604/csse.2022.021164.
16. J. M. Hausdorff, "Gait in Parkinson's Disease," PhysioNet, ver. 1.0.0, 2008, doi: 10.13026/C24H3N; and A. L. Goldberger et al., "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for

complex physiologic signals," Circulation, vol. 101, no. 23, pp. e215-e220, 2000, doi: 10.1161/01.CIR.101.23.e215.
17. M. N. Alam, A. Garg, T. T. K. Munia, R. Fazel-Rezai, and K. Tavakolian, "Vertical ground reaction force marker for Parkinson's disease," PLoS ONE, vol. 12, no. 5, Art. no. e0175951, 2017, doi: 10.1371/journal.pone.0175951.
18. S. S. Shapiro and M. B. Wilk, "An analysis of variance test for normality (complete samples)," Biometrika, vol. 52, no. 3-4, pp. 591-611, 1965, doi: 10.1093/biomet/52.3-4.591.
19. M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, Dept. Comput. Sci., Univ. Waikato, Hamilton, New Zealand, 1999.
20. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, San Francisco, CA, USA, 2016, pp. 785-794, doi: 10.1145/2939672.2939785.