A Proxy-Based CNN-TCN Architecture for Parkinson's Disease Detection from RGB-D Gait Sequences

Jaime Peris-Tchang[1], Nélida Mirabet-Herranz[1][0000−0002−3372−1192], Nicolás Javier Salazar[2][0009−0006−8361−2839], Jorge Luis Orozco-Velez[3][0000−0001−6598−1880], Andrés Navarro-Cadavid[2][0000−0002−4154−2119], and Valery Naranjo[1][0000−0002−0181−3412]

[1] Instituto Universitario de Investigación en Tecnología Centrada en el Ser Humano, HUMAN-tech, Universitat Politècnica de València, Valencia, Spain
jpertch@alumni.upv.es, nemiher@htech.upv.es, vnaranjo@dcom.upv.es
[2] Universidad ICESI, Cali, Colombia
anavarro@icesi.edu.co, nicolas.salazar1@u.icesi.edu.co
[3] Fundación Valle del Lili, Cali, Colombia
jlorozcovelez@gmail.com

Abstract. Parkinson's disease (PD) produces motor alterations that can be reflected in gait patterns, making gait analysis a promising tool for non-invasive and objective disease detection. This work proposes a deep learning framework for PD detection from RGB-D skeletal gait sequences. The dataset collected for this study includes gait recordings from 117 participants, comprising 28 healthy controls and 89 PD patients spanning Hoehn and Yahr severity stages 1 to 4. From these recordings, individual gait-cycle samples were extracted and represented as multidimensional time series of 3D joint coordinates. The proposed architecture combines local morphological feature extraction through one-dimensional convolutional layers with long-range temporal modeling based on a Temporal Convolutional Network. Instead of using a conventional softmax classifier, the model incorporates a proxy-based deep metric learning layer, where learnable class-specific proxies define reference points in the latent space. This formulation not only improves classification performance but also provides a degree of interpretability by enabling the analysis of distances between gait embeddings and class-related proxies. The proposed model achieved an accuracy of 95.45% and an F1-score of 89.41% and showed strong performance across disease severity groups.

Keywords: Gait analysis · Parkinson's disease · RGB-D video · Deep Learning.

1 Introduction

Parkinson's disease (PD) represents the second most prevalent neurodegenerative disorder globally, surpassed only by Alzheimer's disease, and by 2030, the

prevalence of PD is expected to double, particularly in developing countries [6]. It is a progressive disease that presents significant clinical heterogeneity, with a high degree of variability in terms of symptom burden and rate of disease progression, mainly affecting the non-motor and dopaminergic circuits of the brain which is an essential neurotransmitter for the control of movement. Currently, there are no objective tests for PD diagnosis, which still relies mainly on visual inspection of the specialist, on clinical criteria and supportive clinical criteria such as asymmetric motor impairment and response to levodopa [14]. As a result, diagnostic accuracy remains limited, with an estimated error rate of over 20%, a figure that has not significantly improved over the last few decades [2]. Besides, a survey conducted in 2019 revealed that 26% of individuals with PD were initially misdiagnosed, often receiving inappropriate treatments that resulted in worsening health outcomes for more than one-third of those affected [13].

In addition, by the time PD is clinically diagnosed, it is estimated that around 60% of dopaminergic neurons have already degenerated, meaning diagnosis often occurs at a relatively advanced stage of the disease [1]. While no cure for PD currently exists, accurate early diagnosis and intervention are increasingly recognized as essential for preserving neuronal function, slowing disease progression, and improving patients' quality of life [17]. As the disease progresses, symptoms often become less responsive to pharmacological interventions [7], making early accurate diagnosis vital for initiating timely management strategies that could slow disease progression. Delayed and incorrect diagnosis has critical consequences for patients and caregivers alike. It leads to faster deterioration of motor and non-motor functions, a significant decline in quality of life, and increased psychological and physical strain on caregivers [4].

One of the earliest observable symptoms of PD is gait disturbance, reflecting subtle motor impairments in walking patterns [5]. Gait analysis has shown promise as a potential tool for early-stage PD detection [2] especially given that current diagnosis relies primarily on clinical observation of motor and non-motor symptoms [15]. Although academic efforts have explored more objective automated methods such as Inertial Measurement Units (IMUs) or commercial gait mats, each of these alternatives presents significant limitations. IMUs provide valuable spatiotemporal data covering relevant body regions such as the arms and legs, where asymmetry and balance issues can be detected. However, sensors may miss other symptoms such as postural shifts and tremor in areas not covered by the devices. On the other hand, commercial systems like gait mats offer only predefined parameters, require expensive hardware, and may alter natural movement due to their obtrusiveness [5, 16]. Other alternatives, such as high-accuracy marker-based motion capture systems, remain impractical outside of controlled environments due to their cost and complexity [15].

Video-based methods, including recent computer vision approaches using standard cameras [5, 12], offer a low-cost and unobtrusive alternative. However, because these sensors only capture 2D image projections, models must either classify PD directly from 2D data or infer the missing depth information, which may introduce spatial inaccuracies and increase the complexity of the classifi-

cation task. In this work, we propose an end-to-end deep learning framework for PD detection from RGB-D skeletal gait sequences. By directly providing depth information, RGB-D cameras enable markerless acquisition of 3D body-joint trajectories during walking, offering a practical alternative to both standard 2D video and marker-based motion capture systems. Moreover, RGB-D sensing reduces the practical limitations of marker-based approaches, which may introduce physical discomfort, alter natural walking patterns, or increase participant awareness during the recording process.

The main contributions of this work are threefold. First, we propose a deep learning architecture for gait-based PD detection that combines local morphological feature extraction, long-range temporal modeling, and proxy-based deep metric learning. The proposed model achieved an accuracy of 95.45% and an F1-score of 89.41% under a strict inter-subject validation protocol. Second, this study evaluates model performance across different levels of disease severity, including early-stage PD patients. To the best of our knowledge, this is the first study that analyses RGB-D skeletal gait-based PD detection across Hoehn and Yahr stages ranging from 1 to 4. In particular, the proposed model achieved an accuracy of 83.81% in early-stage patients, corresponding to H&Y stages 1-1.5, where motor alterations are often subtle and more difficult to detect. Third, the proposed proxy-based classification layer provides a degree of interpretability by design, which is particularly relevant in clinical applications, where model decisions should not rely solely on opaque classification scores.

This paper is organized as follows: Section 2 provides a review of the state-of-the-art and related work. Section 3 describes the dataset, the preprocessing pipeline, the proposed architecture and the implementation details. In Section 4, we present the experimental setup and the results. Finally, section 5 draws the conclusions.

2 Related work

This section provides a comprehensive overview of current state-of-the-art methodologies engineered to differentiate individuals with PD from healthy control subjects using gait sequences. It is organized into two parts according to the modality of the data used.

2.1 IMU-based detection methods

IMU-based methods represent one of the main approaches for gait-based Parkinson's disease detection, relying on wearable inertial sensors to capture movement-related signals during walking. These methods characterize walking dynamics from body-mounted inertial measurements and enable the identification of disease-related motor patterns.

In 2022, Carvajal-Castaño et al. [3] introduced one of the first end-to-end deep learning approaches for gait-based PD detection using foot-mounted IMUs. The authors proposed three architectures to analyze gait data acquired from two

IMUs placed on the participants' feet: a convolutional neural network (CNN), a gated recurrent unit (GRU) network, and a hybrid CNN-GRU model. Overall, the hybrid CNN-GRU architecture achieved the best performance when signals from both feet were processed jointly. Upper-body inertial sensing has also been explored as an alternative configuration for PD gait analysis. Tumbaco-Sellan et al. [18] used three IMUs positioned on the lower spine and both wrists of the participants. The collected signals were modeled as temporal sequences and used to train two architectures: a long short-term memory (LSTM) network and a transformer-based model. Although both strategies achieved high performance, the transformer-based model obtained the best results, reaching an accuracy of 99.82% under a 5-fold cross-validation scheme.

Although IMU-based approaches have demonstrated strong potential for gait-based Parkinson's disease detection, their practical implementation presents several limitations. These systems require wearable sensors to be attached to specific body locations, which may affect user comfort, introduce variability due to sensor placement, and reduce their suitability for unsupervised or large-scale assessments.

2.2 Video-analysis-based detection method

Recent studies have increasingly explored video-analysis-based methods as a non-invasive alternative for gait-based Parkinson's disease detection. Unlike wearable-sensor approaches, these methods do not require physical devices to be attached to the body; instead, they rely on skeletal data extracted from RGB and RGB-D videos to analyze gait cycles, offering a more accessible and natural framework for clinical and real-world assessment.

To our knowledge, one of the first works that identified the potential of gait videos for PD detection was conducted by He et al. [8] in 2022. They introduced the PD-Walk dataset, which comprises RGB video recordings of 96 PD patients and 95 healthy controls performing a 5-meter walking task at a natural pace. The videos were processed using HRNet to extract 2D skeletal representations consisting of 13 key joints. The authors proposed the ADGCN model, characterized by an architecture featuring two parallel ST-GCN-based branches. With this approach, they reported an accuracy of 84.1% via 5-fold cross-validation, establishing the potential of this modality. Following their promising results, in 2024 Ma et al. moved away from the graph-based skeletal representation towards a more abstract one [10]. The authors used the PD-Walk dataset to demonstrate their approach based on representations derived from the Euclidean distances between pre-selected pairs of joints in each frame. The proposed architecture comprises two parallel blocks based on transformer methodologies incorporating self-attention layers. Their model achieved an average performance of 86.8% via 5-fold cross-validation. From another perspective, Zhang et al. [19] represented skeletal data as graphs defined by adjacency matrices and proposed an ST-GCN-based architecture to extract spatial and temporal features. The model's performance was evaluated using a single fold validation strategy comprising 32 healthy and 30 diseased samples, achieving an accuracy of 87.10%. However,

the authors reported that the control group comprised samples from 50 healthy participants recorded in a controlled environment, while the PD samples were derived from six videos sourced from YouTube. This disparity in data acquisition suggests that the reported performance may not generalize to clinical settings as the model may have learned dataset-specific recording conditions rather than gait patterns directly associated with Parkinson's disease.

In the most recent approach, Kumawat et al. explored the use of 3D skeletons by introducing TrPrNet, a transformer-based architecture designed to detect short and long-term temporal dependencies [9]. These 3D sequences were reconstructed from 2D RGB video recordings via MediaPipe BlazePose. While the authors reported a 99.38% test accuracy rate, this evaluation should be interpreted with caution as it was derived from an exceptionally small test set consisting of only 2 PD patients and 2 control subjects selected through non-random sampling. Nevertheless, despite the limitations of the evaluation protocol, this work suggests that the use of three-dimensional inferred visual information may hold relevant potential for gait-based Parkinson's disease detection. This motivates our research into native depth-based gait analysis, where RGB-D data are explored as a non-invasive source of direct 3D motion information for Parkinsonian gait assessment.

Overall, although previous video-based studies have demonstrated the potential of visual gait analysis for PD detection, several of them present limitations related to dataset composition, acquisition conditions, and evaluation protocols. To address these issues, our work emphasizes a more rigorous and reproducible experimental design, based on a clinically collected RGB-D gait dataset and a clearly defined validation protocol.

3 Methodology

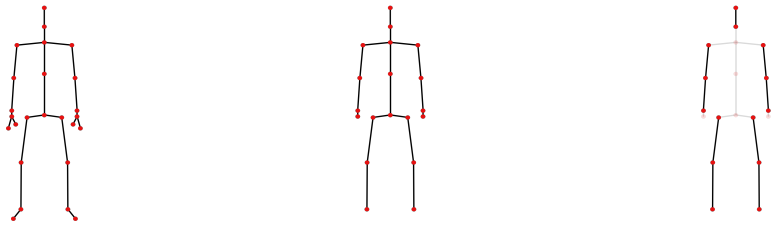
This section details the dataset collected for this study, the implemented preprocessing pipeline, and the proposed neural network architecture to classify PD from the preprocessed data.

3.1 Dataset description

The dataset collected for this study was compiled by researchers from *Universidad ICESI* in collaboration with the departament of neurology of the univeristy hospital *Fundación Valle del Lili*. Data collection was approved by the Institutional Review Board on August 10, 2022 (Approval No. 2022.1916). This dataset includes gait recordings from 117 participants: 28 control subjects and 89 patients diagnosed with Parkinson's disease. In addition to the diagnostic classification, the dataset provides pathological severity labels for the PD samples based on the Hoehn and Yahr (H&Y) scale. The PD samples are distributed across H&Y stages as follows: 30 samples in stage 1, 61 in stage 1.5, 168 in stage 2, 16 in stage 2.5, 72 in stage 3, and 24 in stage 4. Notably, this distribution represents a key strength of the dataset, as most existing databases include few

or no patients in the early stages of the disease. Labels were assigned through clinical assessments conducted by a neurologist specialized in movement disorders. The dataset showed a comparable age distribution between groups, with a mean age of 64.48 years for PD patients and 59.91 years for controls. This balance helps reduce the risk of models learning age-related gait differences.

Data acquisition was performed under standardized conditions. Participants were instructed to walk naturally toward the imaging system over a distance of three meters. Recordings were acquired in RGB-D format using either a Kinect v2 or an Orbbec Astra depth sensor. From these RGB-D recordings, skeletal time series were extracted using the native pose estimation frameworks provided by each camera. The Kinect v2 recordings provided 25 joints per participant (Fig.1a), whereas the Orbbec Astra recordings provided 19 joints (Fig.1b). In Section 4, we also evaluate a simplified version of the skeleton (Fig.1c). For each gait cycle sample, the corresponding joint trajectories were stored in JSON format.



(a) **(b)** **(c)**

Fig. 1: Comparison of skeleton configurations. (a) Standard 25-joint Kinect model. (b) 19-joint shared set. (c) Reduced 14-joint configuration; red markers indicate selected joints, while semi-transparent represent discarded ones.

3.2 Data pre-process

Once the gait sequences were recorded, a structured preprocessing pipeline was applied to improve data quality, reduce acquisition-related artifacts, and ensure consistency across recordings.

First, all gait sequences were evaluated in terms of joint availability. To guarantee temporal continuity and reliability, only recordings with a joint availability rate above 96% were retained. Missing values were then handled according to their location and duration. Missing samples located at the beginning or end of a sequence were removed by truncating the corresponding frames, thereby preserving temporal synchronization across all joint signals. For internal gaps of 10 or fewer consecutive missing samples, cubic spline interpolation was used to estimate the missing joint coordinates. Conversely, sequences containing gaps longer than 10 consecutive samples in one or more joint trajectories were discarded to avoid interpolation-induced artifacts.

To reduce high-frequency noise, including jitter introduced by the pose estimation algorithms, all joint trajectories were filtered using a fourth-order low-pass Butterworth filter with a cutoff frequency of 6 Hz. In addition, some recordings contained a stationary period before gait initiation. Therefore, the beginning of the walking phase was automatically identified using a segmentation procedure based on the moving variance of the depth coordinate (z) of the SpineBase joint. This joint was selected because it provides a stable approximation of the body center of mass. The onset of gait was defined as the point at which the moving variance (calculated using a window size of 10 frames) exceeded a threshold of 0.0005, and the preceding static frames were removed from the sequence.

Although the depth cameras nominally operate at 30 frames per second, the effective sampling rate may fluctuate due to acquisition conditions, such as lighting and sensor-related variability. To mitigate these temporal inconsistencies, all sequences were uniformly resampled to a constant sampling frequency. Joint positions were recalculated at equally spaced time intervals through cubic spline interpolation, ensuring temporal consistency across recordings and enabling reliable time-series analysis. After resampling, recordings shorter than one second, corresponding to fewer than 30 samples, were excluded from the dataset. This minimum-duration criterion was established to ensure that each retained sequence contained sufficient temporal information to represent a clinically meaningful gait pattern.

A perspective-related artifact was also identified in recordings acquired with the Orbbec Astra camera. This artifact produced an apparent increase in body scale as participants approached the sensor and was particularly evident in several control recordings. To mitigate this effect, anthropometric normalization was applied by scaling the skeletal segments with respect to torso length, defined as the distance between the SpineBase and SpineMid joints. Furthermore, to remove the influence of global body translation and focus the analysis on intrinsic gait dynamics, spatial normalization was performed by anchoring the SpineBase joint at the origin of the coordinate system.

Because the Kinect v2 and Orbbec Astra systems provide different skeletal models, the joint representation was harmonized by retaining only the 19 anatomical joints common to both pose estimation frameworks. This step ensured that all samples shared the same feature structure before subsequent analysis. Finally, each joint coordinate signal was standardized independently to place all features on a comparable numerical scale and facilitate model optimization. The standardization transformed each signal to have zero mean and unit variance, according to $\text{StandardizedSignal} = \frac{\text{signal}-\text{mean}}{\text{std}}$.

After preprocessing, the dataset comprised 424 recordings, including 65 control recordings and 359 PD recordings. To increase the number of training samples and improve the robustness of the learning process, a sliding-window strategy was applied to the preprocessed sequences. Specifically, one-second temporal windows with 50% overlap, equivalent to a 0.5-second shift, were extracted from each recording. This procedure yielded 147 samples for the control class and 1261 samples for the PD class.

3.3 Model architecture

The proposed architecture combines local morphological feature extraction, temporal modeling, and metric-based classification. The first block extracts local gait patterns using one-dimensional convolutional layers, whereas the second block models longer-range temporal dependencies through a Temporal Convolutional Network (TCN) with dilated convolutions. Finally, the learned representation is passed to a proxy-based metric classification layer.

The local feature extraction block consists of three sequential 1D convolutional units, each followed by a ReLU activation function. The first two units use convolutional layers with a kernel size of 5 and expand the feature representation to $(n_{sensors} \times 3) \times 64$ channels. The third unit uses a kernel size of 3 and reduces the channel dimensionality to $(n_{sensors} \times 3) \times 32$, producing a compact representation for the temporal modeling stage.

The temporal reasoning block is implemented as a TCN composed of four cascaded dilated convolutional layers, each followed by a ReLU activation function and a residual connection. The first TCN layer uses $(n_{sensors} \times 3) \times 32$ filters, whereas the remaining three layers use 128 filters. The dilation factor increases exponentially across layers, following $d = 2^i$, which enlarges the receptive field without increasing the number of trainable parameters. To preserve the temporal length of the feature maps, the padding of each layer is computed as:

$$padding = \frac{(kernel_{size} - 1) \cdot 2^i}{2}$$

This padding strategy maintains the output dimensionality with respect to the TCN input and reduces information loss at the temporal boundaries. Residual connections are included in each TCN layer to facilitate gradient flow and stabilize the learning process. The temporal embeddings produced by the TCN are then aggregated into a single 128-dimensional feature vector using global average pooling across the temporal dimension. This vector is projected into a lower-dimensional latent space through a two-stage compression module. First, a fully connected layer maps the representation to 64 dimensions and applies an ELU activation function. A second dense layer then produces the final 32-dimensional embedding, which is used by the proxy-based metric classification layer.

Proxy-based classification layer: For the final classification stage, the proposed architecture replaces the conventional softmax classifier with a proxy-based deep metric learning layer. This strategy follows the approach introduced by Movshovitz-Attias et al. [11], in which classification is performed by comparing each learned embedding with a set of trainable class representatives, referred to as proxies. In contrast to pair- or triplet-based metric learning methods, which require explicit sample-to-sample comparisons, proxy-based learning reduces the computational cost by comparing each sample directly with a limited number of learnable proxies. Each proxy acts as a representative point in the latent space

and is optimized jointly with the rest of the network. During training, embeddings are encouraged to lie close to proxies associated with their corresponding class and far from proxies associated with other classes.

In this work, the final embedding produced by the feature extraction module is projected into a 32-dimensional latent space. The classification layer contains 11 learnable proxies in this space: 2 proxies are assigned to the control class and 9 proxies to the PD class. This asymmetric allocation was introduced to account for the higher intra-class variability and larger sample density of the PD group. In addition to supporting classification, the proxy-based formulation provides a degree of interpretability by defining explicit class-related reference points in the latent space. The distance between each sample embedding and the learned proxies can therefore be analyzed to assess whether the model associates a given gait pattern more strongly with control-like or PD-like representations.

The similarity between embeddings and proxies is computed using the squared Euclidean distance, which provides a stable distance-based objective during optimization. For each class, the minimum distance between the embedding and the proxies assigned to that class is computed. The resulting distances are multiplied by -1, so that larger output values correspond to higher class likelihoods. These scores are then used as logits for the cross-entropy loss.

3.4 Implementation details

The models were implemented and trained using Python 3.11, PyTorch 2.10.0, and CUDA 12.6. All models were trained on a computer equipped with an NVIDIA GeForce RTX 2080 GPU. In the local morphological feature extraction block, each convolutional unit included batch normalization and a 20% dropout rate to mitigate overfitting and improve generalization. In the subsequent stages, each TCN unit included batch normalization, while both the TCN units and the fully connected layers used a 50% dropout rate.

The cross-entropy loss function was minimized using the Adam optimizer, configured with an initial learning rate of 10^{-5} and a weight decay of $4{*}10^{-2}$. The weight decay was introduced to reduce the risk of memorization and overfitting. A Reduce-on-Plateau strategy was also employed, reducing the learning rate by a factor of 0.5 if no improvement in the validation F1-score was observed for 10 consecutive epochs. In addition, early stopping was applied with a patience of 20 epochs based on the validation F1-score. Thus, training was stopped if the validation F1-score did not improve for 20 consecutive epochs.

4 Experiments

This section presents the experimental protocol designed to assess the proposed architecture and its evaluation. It is organized into five experiments. First, an empirical comparison of two class-imbalance mitigation strategies is performed. Second, an architectural ablation is conducted to contrast the proposed model backbone with alternative designs. Third, the contribution of individual skeletal

joints to classification performance is analyzed. Fourth, a head-to-head comparison is drawn between the proposed proxy-based classification head and a conventional softmax decision boundary. Fifth, the model's viability is evaluated across the isolated H&Y stages.

4.1 Experimental setup

To guarantee the learning of robust and generalizable features, the dataset was partitioned using a stratified 5-fold cross-validation. For each fold, the data were split into three subsets dedicated to training, validation and test. The validation subset was used exclusively to monitor the early-stopping criterion and to trigger the Reduce-on-plateau learning rate schedule described in Section 3.4.

A significant methodological challenge in gait-based analysis is the risk that the model may memorize subject-specific biometric traits instead of learning pathological patterns associated with PD. Without appropriate safeguards, gait sequences belonging to the same individual may be disturbed across training and test partitions, leading to inflated performance metrics that fail to generalize to new subjects. To address this concern, an inter-subject data splitting strategy was implemented. This approach ensures that all data gait sequences from a given individual are restricted to a single partition within each fold.

All experiments used the same hyperparameter configuration and preprocessing pipeline to isolate the effect of the specific variable analyzed in each experiment. Model performance was assessed using four standard metrics: accuracy, precision, recall, and F1-score, computed with the Scikit-Learn library. Precision, recall, and F1-score were computed first independently for each class and then averaged across classes, assigning equal importance to both control and PD samples regardless of class imbalance between folds. The values reported in Tables 1 to 5 correspond to the average across the five folds of the test partition.

4.2 Experiments

Evaluation of Bias Mitigation Strategies To address class imbalance, two mitigation strategies were evaluated under the inter-subject validation protocol described above: class-weighted loss and weighted oversampling. In the class-weighted approach, a penalty factor derived from the class distribution was incorporated into the cross-entropy loss function. Instead of using the raw inverse class frequencies, a smoothed weighting scheme was applied to reduce the risk of unstable gradient updates during training: $\text{PenaltyFactor} = \sqrt{\frac{\text{TotalDatasetSize}}{2.0 \times \text{ClassSize}}}$. In the weighted oversampling approach, a weighted random sampler was used to increase the sampling probability of the minority class during training, allowing both classes to be observed with comparable frequency in each epoch.

As shown in Table 1, weighted oversampling outperformed class weighting across all evaluation metrics, achieving an F1-score of 89.41% compared with 87.95% for the class-weighted approach. Therefore, weighted oversampling was selected as the imbalance mitigation strategy for the remaining experiments.

Table 1: Comparison between class weights and weighted oversampling approach.

	Accuracy (%)	**Precision (%)**	**Recall (%)**	**F1-score (%)**
Class weights	94.75 (±1.70)	84.08	94.11	87.95 (±3.47)
Weighted oversampling	**95.45 (±1.13)**	**85.38**	**95.39**	**89.41 (±2.58)**

Comparison between different architectures To isolate the contribution of the different backbone components to the overall performance of the proposed model, three ablated variants were trained and evaluated. The first variant consisted of only the local morphological feature extractor, described in Section 3.3, followed by the proxy-based classification head. The remaining variants incorporated an additional temporal modeling block after the CNN structure, while preserving the same proxy-based classification head. Specifically, one variant used an LSTM block, whereas the other used a GRU block.

Table 2: Comparison between different architectures.

	Accuracy (%)	**Precision (%)**	**Recall (%)**	**F1-score (%)**
ProxyConv model	95.17	84.89	94.31	88.68
ProxyConvLSTM model	95.17	**85.83**	90.40	87.81
ProxyConvGRU model	94.74	83.54	95.29	88.11
ProxyConvTCN (prop.)	**95.45**	85.38	**95.39**	**89.41**

Table 2 shows that the proposed model achieved the best overall performance, obtaining the highest accuracy, recall, and F1-score. Among the ablated variants, the ProxyConv model, which relies only on the local morphological feature extractor, achieved a competitive F1-score. Adding recurrent temporal modeling did not consistently improve performance: the ProxyConvLSTM model obtained the highest precision but showed a reduction in recall, whereas the ProxyConvGRU model improved recall at the expense of precision. In this context, the precision-recall trade-off is particularly relevant, as both false positives and false negatives have important implications in medical applications. A model with low recall may fail to detect PD cases, whereas a model with low precision may incorrectly classify control gait patterns as pathological. Therefore, achieving a balance between these two metrics is essential. The proposed architecture provided the best balance between precision and recall, while also achieving the highest recall among all evaluated models. This result is particularly relevant in the context of PD detection, where minimizing missed PD cases is clinically important.

Assessment of Joint-Specific Signal Contributions to Model Performance To demonstrate that providing the model with the largest possible set of joints does not necessarily yield the best discriminative performance, an empirical evaluation was conducted to determine the optimal input configuration

for the network by varying the set of joints provided to the model. In this experiment, two configurations were contrasted: a baseline using all 19 joints common to the Kinect and Orbbec Astra skeletons (see Fig. 1b), and a reduced configuration retaining 14 joints (shown in Fig. 1c), obtained by discarding 3 trunk joints and both hands. The results in Table 3 show that the 14-joint configuration outperforms the full 19-joint baseline across all four metrics.

Table 3: Comparison between different set of joints.

	Accuracy (%)	**Precision (%)**	**Recall (%)**	**F1-score (%)**
19 joint model	94.25	82.88	93.21	86.88
14 joint model (prop.)	**95.45**	**85.38**	**95.39**	**89.41**

Evaluation of softmax decision boundaries vs proxy-based classification The proposed model replaces the conventional softmax classification head with the proxy-based approach defined in Section 3.3. To quantify the performance gain associated with this design choice, a comparative model was trained using the same backbone but replacing the proxy-based head with a conventional fully connected softmax layer. As reported in Table 4, the proxy-based model outperformed the softmax baseline across all evaluation metrics. This behavior is consistent with the motivation behind the proxy-based formulation: by assigning multiple learnable proxies to the heterogeneous PD class, the model can better represent the variability of Parkinsonian gait patterns within the latent space.

To further analyze the interpretability provided by the proxy-based formulation, Fig. 2 presents a two-dimensional UMAP projection of the latent embeddings obtained from the test samples of one validation fold, together with the learned class proxies represented as crosses. Each point corresponds to the feature representation of an individual gait-cycle sample in the learned embedding space. The visualization reveals a clear organization of the samples into compact clusters, where the proxies naturally emerge as representative reference points for different regions of the latent space. Notably, the multiple proxies associated with the PD class are distributed across several separated clusters, highlighting the heterogeneity of Parkinsonian gait patterns captured by the model. In contrast, the control samples form a more compact distribution around their corresponding proxies. This behavior supports learned proxies as an interpretable representation of the decision space while simultaneously modeling the intra-class variability of PD gait patterns.

Table 4: Comparison between a softmax classification head and a proxy-based approach.

	Accuracy (%)	**Precision (%)**	**Recall (%)**	**F1-score (%)**
Softmax	94.89 ($\pm$1.41)	84.00	95.36	88.45 ($\pm$2.81)
Proxy (proposed)	**95.45 ($\pm$1.13)**	**85.38**	**95.39**	**89.41 ($\pm$2.58)**

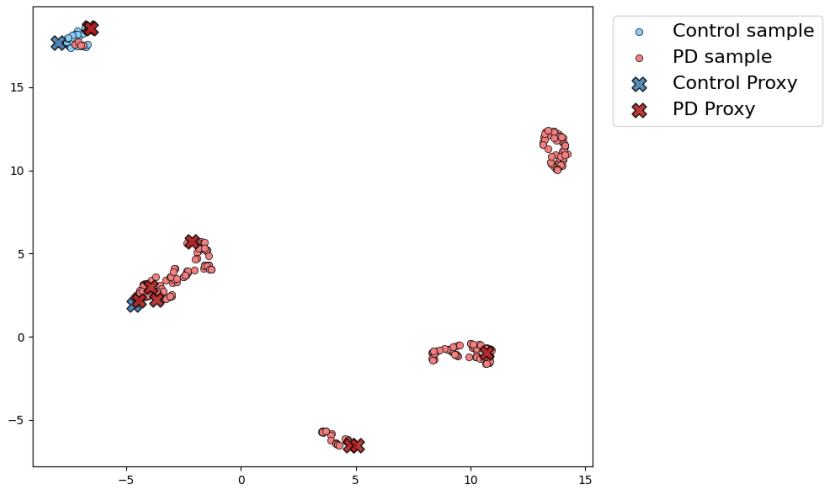


Fig. 2: 2D projections of embeddings and proxies.

Performance over different H&Y stages This experiment evaluates the performance of the model across four independent test sub-cohorts: healthy controls and patients stratified according to H&Y stage. The results reported in Table 5 show that the proposed architecture exhibits a conservative classification profile, characterized by a low false-positive rate, as reflected by the high accuracy obtained for the control group, 95.31%. The lowest accuracy was observed for the H&Y 1-1.5 subgroup, 83.81%, reflecting the difficulty of identifying early-stage PD, where motor alterations may be subtle and predominantly unilateral. This subgroup also showed the highest standard deviation, 12.32%, suggesting greater variability across folds and indicating that early-stage cases are more challenging and less consistently detected by the model. In contrast, once the disease reaches stages associated with bilateral motor involvement, namely H&Y 2-2.5 and H&Y 3-4, the model achieves higher accuracy, with 97.03% and 96.87%, respectively. These groups also presented lower standard deviations, 3.32% and 3.84%, indicating more stable performance across folds for moderate and advanced stages.

Table 5: Model performance over different H&Y stages

Group	**Accuracy (%)**	**Std. Dev. (%)**
Control	95.31	± 7.77
H&Y 1-1.5	83.81	± 12.32
H&Y 2-2.5	97.03	± 3.32
H&Y 3-4	96.87	± 3.84

5 Conclusion and future work

This work presented a subject-independent deep learning framework for Parkinson's disease detection based on RGB-D skeletal gait sequences. The proposed approach combines local morphological feature extraction, temporal modeling through a TCN, and a proxy-based deep metric learning classification layer. This framework enables non-invasive and markerless gait analysis without requiring wearable sensors or physical markers. The experimental results showed that the proposed architecture achieved strong classification performance under a strict inter-subject validation protocol, reaching an accuracy of 95.45% and an F1-score of 89.41%. Moreover, the learned proxies provide class-related reference points in the latent space, offering a degree of interpretability that is particularly relevant for clinical applications. The stratified analysis across Hoehn and Yahr stages showed that the model performed particularly well in moderate and advanced PD stages, while early-stage cases remained more challenging. This result is clinically meaningful, as motor alterations in early PD are often subtle and may overlap with healthy gait variability. Nevertheless, the ability to evaluate model behavior across severity stages represents an important step toward more transparent and clinically informative gait-based PD detection systems. Future work will focus on further strengthening the clinical robustness and interpretability of the proposed framework. First, medication state information should be incorporated during data acquisition, since dopaminergic treatment may modulate motor symptoms and gait patterns. In addition, by integrating different fluid or tissue biomarkers, genetic markers, diagnostic imaging, or digital markers we aim to achieve more accurate diagnoses. Finally, further analysis of proxy behavior and latent-space organization will be explored to enhance the explainability of the model and provide clinicians with more transparent decision-support information.

Acknowledgements This work was supported in part by the Generalitat Valenciana under Grant CIPROM/2022/20 (COMTACTS2)

References

1. Becker, G., Müller, A., Braune, S., Büttner, T., Benecke, R., Greulich, W., Klein, W., Mark, G., Rieke, J., Thümler, R.: Early diagnosis of parkinson's disease. Journal of neurology **249**(Suppl 3), iii40–iii48 (2002)
2. Burtscher, J., Moraud, E.M., Malatesta, D., Millet, G.P., Bally, J.F., Patoz, A.: Exercise and gait/movement analyses in treatment and diagnosis of parkinson's disease. Ageing research reviews **93**, 102147 (2024)
3. Carvajal-Castaño, H.A., Pérez-Toro, P.A., Orozco-Arroyave, J.R.: Classification of parkinson's disease patients—a deep learning strategy. Electronics **11**(17), 2684 (2022)
4. Chaudhuri, K.R., Azulay, J.P., Odin, P., Lindvall, S., Domingos, J., Alobaidi, A., Kandukuri, P.L., Chaudhari, V.S., Parra, J.C., Yamazaki, T., et al.: Economic burden of parkinson's disease: a multinational, real-world, cost-of-illness study. Drugs-Real World Outcomes **11**(1), 1 (2024)

5. Connie, T., Aderinola, T.B., Ong, T.S., Goh, M.K.O., Erfianto, B., Purnama, B.: Pose-based gait analysis for diagnosis of parkinson's disease. Algorithms **15**(12), 474 (2022)
6. Erkkinen, M.G., Kim, M.O., Geschwind, M.D.: Clinical neurology and epidemiology of the major neurodegenerative diseases. Cold Spring Harbor perspectives in biology **10**(4), a033118 (2018)
7. Gallagher, D.A., Schrag, A.: Impact of newer pharmacological treatments on quality of life in patients with parkinson's disease. CNS drugs **22**(7), 563–586 (2008)
8. He, Y., Yang, T., Yang, C., Zhou, H.: Integrated equipment for parkinson's disease early detection using graph convolution network. Electronics **11**(7), 1154 (2022)
9. Kumawat, V., Sharma, Y., Bhatia, A., Tiwari, K.: Trprnet: Early parkinson detection network using marker-less gait analysis. In: International Conference on Advanced Information Networking and Applications. pp. 415–427. Springer (2025)
10. Ma, L., Huo, H., Liu, W., Zhao, C., Wang, J., Xu, N.: Twin-tower transformer network for skeleton-based parkinson's disease early detection. Complex & Intelligent Systems **10**(5), 6745–6765 (2024)
11. Movshovitz-Attias, Y., Toshev, A., Leung, T.K., Ioffe, S., Singh, S.: No fuss distance metric learning using proxies. In: Proceedings of the IEEE international conference on computer vision. pp. 360–368 (2017)
12. Muñoz-Ospina, B., Alvarez-Garcia, D., Clavijo-Moran, H.J.C., Valderrama-Chaparro, J.A., García-Peña, M., Herrán, C.A., Urcuqui, C.C., Navarro-Cadavid, A., Orozco, J.: Machine learning classifiers to evaluate data from gait analysis with depth cameras in patients with parkinson's disease. Frontiers in human neuroscience **16**, 826376 (2022)
13. Parkinson's UK: Poll finds a quarter of people with Parkinson's are wrongly diagnosed. https://www.parkinsons.org.uk/news/poll-finds-quarter-people-parkinsons-are-wrongly-diagnose (2020), last accessed 5 May 2026
14. Postuma, R.B., Berg, D., Stern, M., Poewe, W., Olanow, C.W., Oertel, W., Obeso, J., Marek, K., Litvan, I., Lang, A.E., et al.: Mds clinical diagnostic criteria for parkinson's disease. Movement disorders **30**(12), 1591–1601 (2015)
15. Ricciardi, C., Amboni, M., De Santis, C., Improta, G., Volpe, G., Iuppariello, L., Ricciardelli, G., D'Addio, G., Vitale, C., Barone, P., et al.: Using gait analysis' parameters to classify parkinsonism: A data mining approach. Computer methods and programs in biomedicine **180**, 105033 (2019)
16. Stenum, J., Hsu, M.M., Pantelyat, A.Y., Roemmich, R.T.: Clinical gait analysis using video-based pose estimation: Multiple perspectives, clinical populations, and measuring change. PLOS Digital Health **3**(3), e0000467 (2024)
17. Tinelli, M., Kanavos, P., Grimaccia, F.: The value of early diagnosis and treatment in parkinson's disease: A literature review of the potential clinical and socioeconomic impact of targeting unmet needs in parkinson's disease (2016)
18. Tumbaco-Sellan, K., Tapia-Rosero, A., Loayza, F.R., Pelaez, E., Navarro-Cadavid, A., Orozco-Velez, J.L., Salazar, N., Muñoz, B., Agila, W., Rubio-Roldán, A., et al.: Wearable gait analysis using imu sensors and deep learning for parkinson's disease detection. In: 2025 IEEE Latin American Conference on Computational Intelligence (LA-CCI). pp. 1–6. IEEE (2025)
19. Zhang, J., Lim, J., Kim, M.H., Hur, S., Chung, T.M.: Wm–stgcn: A novel spatiotemporal modeling method for parkinsonian gait recognition. Sensors **23**(10), 4980 (2023)