POLITECNICO DI TORINO
Repository ISTITUZIONALE

EDA-Based Arousal Analysis During Gamified Leg Agility in Parkinson's Disease: A Cross-Subject XAI Approach

Original
EDA-Based Arousal Analysis During Gamified Leg Agility in Parkinson's Disease: A Cross-Subject XAI Approach / Masi, G., Amprimo, G., Ferraris, C., Olmo, G.. - (2026), pp. 1108-1113. (IEEE International Symposium on Computer-Based Medical Systems CBMS Lymassol (CY) 3-05 June 2026) [10.1109/CBMS69103.2026.00225].

Availability:
This version is available at: 11583/3016461 since: 2026-10-21T14:45:37Z

Publisher:
IEEE

Published
DOI:10.1109/CBMS69103.2026.00225

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
IEEE postprint/Author's Accepted Manuscript

©2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collecting works, for resale or lists, or reuse of any copyrighted component of this work in other works.

(Article begins on next page)

19 November 2026

EDA-Based Arousal Analysis During Gamified Leg Agility in Parkinson's Disease: A Cross-Subject XAI Approach

1st Giulia Masi
Politecnico di Torino
Turin, Italy
giulia.masi@polito.it

2nd Gianluca Amprimo
Politecnico di Torino
Turin, Italy
gianluca.amprimo@polito.it

4th Claudia Ferraris
CNR-IEIIT Italy
Turin, Italy
claudia.ferraris@cnr.it

5th Gabriella Olmo
Politecnico di Torino
Turin, Italy
gabriella.olmo@polito.it

***Abstract*—Leg Agility (LA) is a standard motor task for motor assessment in Parkinson's disease (PD), and recent work has explored its gamification to sustain patient engagement in unsupervised rehabilitation scenarios. However, the effect of this gamification on autonomic arousal during task execution has not been objectively characterised. This paper investigates electrodermal activity (EDA) arousal patterns in 33 PD patients performing traditional LA and two difficulty levels of a gamified Bouncing Ball (BB) exercise. A JMIM-based feature selection pipeline fed a Leave-One-Subject-Out XGBoost classifier with SHAP explainability, complemented by Wicoxon testing with Benjamini-Hochberg FDR correction and per-leg-session monotonic trend analysis. The transition from LA to the game (BB1) is robustly detectable, with several spectral features at $p < .001$ and showing a monotonically increasing EDA trend in 70–80% of sessions. In contrast, discriminating between difficulty levels (BB1 vs BB2) yields only 61% accuracy, despite significant group-level statistical differences, indicating that difficulty-related arousal changes are present at the group level but heterogeneous across individuals. Inter-subject SHAP variance analysis identifies a rank divergence between feature importance and patient-to-patient variability in the BB1 vs BB2 comparison, a direct XAI signature of idiosyncratic responses, and pinpoints cvxPr$_{f_{max}}$ as the most population-consistent biomarker. These findings suggest that gamification reliably increases autonomic arousal in PD patients, while difficulty modulation requires personalised rather than universal calibration, with implications for real-time adaptive game design in motor rehabilitation.**

***Index Terms*—SHAP, XAI, gamification, PD, EDA, Emotional Arousal, Serious Game**

I. Introduction

Parkinson's disease (PD) is the fastest-growing neurological disorder worldwide, with global prevalence having doubled in the last 25 years and disability-adjusted life years increasing by over 80% since 2000 [1]. Beyond the canonical motor triad of bradykinesia, rigidity, and tremor, PD involves pervasive autonomic dysfunction [2]: sympathetic pathways are altered early in the disease, with sympathetic skin response abnormalities correlating with motor subtype severity and UPDRS-III scores [3]. This dual nature of PD, simultaneously a motor and an autonomic disorder, motivates the use of electrodermal activity (EDA) as a physiological window into the functional state of patients during structured tasks.

Gamification has emerged as a viable strategy to increase engagement and adherence in PD motor rehabilitation. A growing body of evidence supports VR-based exergames and Kinect-driven systems as tools that match, and in some cases exceed, conventional physiotherapy outcomes for balance, gait, and upper-limb motor function [4]–[6]. Despite this progress, the monitoring of patient engagement and arousal during gameplay remains largely absent from the clinical loop. Questionnaires are retrospective and prone to recall bias; kinematic performance scores reflect motor output but not emotional involvement. Real-time physiological feedback, specifically from wearable EDA sensors, represents an unexplored but potentially powerful mechanism for adaptive game control in PD [7], [8].

EDA, measured as the change in skin electrical conductance driven by eccrine sweat gland activity under sympathetic innervation, has been validated as a biomarker of arousal in healthy populations across cognitive load, stress, and affective stimuli [9]. Tonic EDA reflects slow changes in baseline arousal, while the phasic component captures the rapid, stimulus-locked sympathetic response. In PD specifically, wearable EDA has been explored for detecting medication ON/OFF states, with SVM classifiers achieving around 72% accuracy on tonic amplitude features [10]. However, the use of EDA to characterise autonomic arousal during structured motor tasks, and in particular during gamified variants of clinical assessments, has not been systematically studied.

Explainable AI (XAI) is increasingly recognised as a prerequisite for clinical adoption of machine learning in neurodegenerative disease monitoring [11]. SHAP (Shapley additive explanations) has become the dominant post-hoc explainability framework in this domain, applied to wearable motor features for PD ON/OFF classification [12], to clinical biomarkers for severity prediction [13], and to transcriptomic features for motor progression forecasting [14]. These prior studies typically report SHAP contributions as population-level averages. Here, we additionally examine the inter-subject distribution of SHAP values as an indicator of feature-specific heterogeneity across patients. This distinction is particularly relevant in PD, where motor and autonomic phenotypes vary substantially.

In this work, we present an analysis of EDA collected from 33 PD patients (H&Y $\leq$ 3) during three sequential conditions: a standard Leg Agility (LA) task from the UPDRS-III protocol, and two difficulty levels of a gamified Bouncing Ball exergame (BB1, BB2) implemented in Unity with Kinect body tracking. We address three questions: (1) Does gamification produce a detectable, cross-subject EDA arousal signature compared to the standard clinical task? (2) Does in-game difficulty modulation produce an additional, individually consistent arousal shift? (3) Which EDA features act as population-level biomarkers, and which are better suited for personalised modelling due to high inter-subject variability?

Our contributions are: (i) a cross-validated pipeline combining JMIM feature selection and Leave-One-Subject-Out XGBoost achieving 87.1% balanced accuracy for LA vs. BB1 discrimination; (ii) a within-subject monotonic trend analysis showing that 70-80% of experimental session exhibit individually consistent EDA escalation across conditions for spectral phasic features, despite population-level classification of BB1 vs. BB2 being near chance (61%); and (iii) an inter-subject SHAP variance metric that operationalises the distinction between robust digital biomarkers and individually heterogeneous features.

II. Materials&Methods

For this study, 33 PD subjects (H-Y: max 3; average age: 73.48 ± 7.16; gender: 21M/12F) were recruited for the experimental test at the Associazione Amici Parkinsoniani Piemonte Onlus. The ethics committee approved the study in accordance with the Declaration of Helsinki (1964) and its amendments. All participants were volunteers and provided written informed consent before testing. The data acquisition protocol consisted of: illustrative explanation; EDA sensor positioning and calibration; Leg Agility (LA) execution for 15 seconds (for each leg); play of the Bouncing ball (BB) game, calibrated to be similar to the subject's LA performance (for each leg), referred as BB1; play of the BB game calibrated to be challenging with respect to the LA executed (for each leg), referred as BB2. Between each task some rest was assured. The study follows a within-subject repeated-measures design with three fixed-order conditions (LA, BB1, BB2). The order was not randomized; therefore, condition and temporal effects cannot be fully disentangled.

A. Assessment Tasks: Leg Agility and Bouncing Ball

LA is a standard Unified Parkinson's Disease Rating Scale (UPDRS) assessment task in which the subject is asked to raise and lower the knee repeatedly. The clinician usually score the task considering the speed, amplitude and regularity of the movement to evaluate disease severity. In this study, subjects were asked to perform LA during video recording and providing body tracking with the setup soon to be illustrated.

The custom BB game was originally developed in [15], [16], where kinematic characteristics of LA and BB were explored, showing good descriptive results for PD condition and highlighting similarities and differences between the two tasks. The gaming and acquisition platform exploits the integration of: video-based body tracking (from which kinematic features are retrieved); Kinect; and Unity, where the game is developed. In the game the subject's avatar is showed reproducing the real movements in the game scene. Hence, the system is completely non invasive. To summarize, BB game require the same knee motion as LA, but embeds it in a goal orienting task involving a virtual ball. The subject has to hit the ball that appears on screen at a certain height from the knee (this set the amplitude of the knee motion), when asked by a halo (this set the speed of the movement). For this experimentation, BB was minimally modified adding an automatic parameter setting tool. Indeed, the game difficulty can be modified setting the ball initial height (bH) and the halo turn on time frequency (hF), that determines the speed of the movement. The automatic bH and hF setting is based on the previously performed LA kinematic features (average amplitude and speed of the movement). In the first BB execution (BB1), the parameters were set as much as possible similar to those found in LA. On the contrary, in the second execution (BB2), the weakest feature was stressed. For instance, if LA was run slow, in BB2 hF was set higher, while if LA showed small amplitude, in BB2 bH was set greater.

B. EDA Features Extraction

The wearable Empatica E4 device was employed for recording 4Hz-sampled Electrodermal activity (EDA) during the whole experimental session. The EDA signal feature extraction was performed in MATLAB R2024a, while feature analysis in Python 3.14. For each subject, 6 EDA windows were selected in correspondence to the execution of the LA and BB for each leg, synchronizing the device with the game platform. For each EDA window, a total of 50 features were computed. Two simple custom functions were created to extract two sets of features: *StatDesc* that computes descriptive statistic, such mean, std, median, kurtosis, skewness, range, and max; and *EDAfeat2*, computing features from frequency, derivative and non linear domains, popular for EDA analysis [17], illustrated in Table I.

TABLE I
Description of the features obtained by *EDAfeat2*

Feature	Description
Der_{mean}	Mean of I derivative
Der_{max}	Mean of I derivative
Negder_{mean}	Mean of I derivative < 0
Activity	(Hjorth) Variance
Mobility	(Hjorth) Square root of the variance of the first derivative, divided by activity
ApEnt	Approximate Entropy
Pwr_{mean}	Mean of spectral power (Lomb-Scargle periodogram)
Pwr_{std}	Std of the spectral power
Pwr_{B5}	Spectral power in band [0.05-0.5] Hz
f_{max}	Frequency corresponding to the max of spectral power
f_{90}	Frequency corresponding to the 90% of spectral power

These two functions were applied to the filtered (low pass 1°order Butterworth filter with a cut-off of 1.5 Hz) and z-scored EDA signal, *EDAfz*. In addition, EDA was decom-

posed in tonic ($cvxT$), sparse sudomotor nerve activity coefficients ($cvxP$), and reconstructed phasic ($cvxPr$) components throughout convex optimisation [18]. *StatDesc* was used on the three obtained components ($cvxT$, $cvxP$, and $cvxPr$). Finally, *EDAfeat2* was applied to $cvxPr$, obtaining the full feature dataset. Hence, the final dataset consisted of 200 windows (33 subjects x 2 legs x 3 conditions), with 50 EDA-derived features per window.

C. Feature Relevance and Statistics

The features were explored in order to observe differences in the emotional response between LA, BB1, and BB2. To progressively study the effect, first of gamification, then of setting the game difficulty, feature relevance was computed for targeting the comparison between: LA vs BB1 vs BB2, LA vs BB1, and BB1 vs BB2.

To obtain a compact and non-redundant subset of physiologically meaningful features for interpretability and statistical analysis, Joint Mutual Information Maximization (JMIM) from parallelized mutual information based feature selection (mifs) [19] was employed. JMIM ranks features based on mutual information with the condition labels while penalizing redundancy among selected features [20]. Indeed, the structuring of the feature dataset is inherently prone to redundancy.

For the obtained ranked feature sets, in each condition separately, cross-correlation (Spearman's coefficent ρ) was explored and cross-correlated features ($\rho > 0.8$) were identified and removed.

The chosen features were statistically tested by means of Wilcoxon signed-rank test (non-normal distribution found, i.e. Shapiro Wilk $p < 0.05$), confronting: LA vs BB1, LA vs BB2, and BB1 vs BB2. False Discovery Rate (FDR) correction (Benjamini–Hochberg procedure) was applied across features within each comparison to control for multiple testing. Effect size was quantified using rank-biserial correlation (r). Box plots were observed for further meaning exploration.

To evaluate individual-level monotonic trends across conditions (LA, BB1, BB2), Spearman's ρ was computed between the ordinal condition vector [0,1,2] and the feature value per experimental session. The two legs were kept separate to preserve the asymmetric motor profile inherent to PD. Given the three-level design, ρ assumes discrete values and is interpreted descriptively as an indicator of trend consistency rather than magnitude.

D. Cross-Subject Classification and Explainability

Cross subject generalization was evaluated using Leave One Subject Out (LOSO) validation. Three classification problems were considered: LA versus BB1, to quantify the autonomic effect of gamification; BB1 versus BB2, to evaluate the effect of difficulty modulation; LA versus BB1 versus BB2, to assess full three condition separability. An XGBoost classifier was employed. Accuracy was computed as the ratio of correctly classified windows over total test windows for each fold and averaged across folds. Classification was performed using the JMIM derived non redundant feature subsets. Because feature selection was label-aware, results may be affected by data leakage. Therefore, LOSO validation was repeated using the full set of 50 EDA features as a reference.

Model explainability was assessed using SHAP values computed with TreeExplainer. SHAP assigns to each feature a contribution value for each prediction. Positive SHAP values increase the probability of the predicted class. Negative values decrease it. Two complementary explainability metrics were extracted:

1) Global feature importance - Mean absolute SHAP value was computed across all subjects and folds. This metric quantifies the overall contribution of each feature to the classification task, independently of direction.
2) Inter subject heterogeneity index - For each subject, mean SHAP values were first averaged across that subject test windows. The standard deviation of these subject level mean SHAP values was then computed across subjects. A low standard deviation indicates consistent feature contribution across individuals. A high standard deviation indicates that the feature contributes differently across subjects.

Features showing high global importance and low heterogeneity were interpreted as robust population level biomarkers. Features showing high heterogeneity relative to their mean importance were interpreted as individually variable and potentially suitable for personalized modeling. This framework allows discrimination between globally stable autonomic markers and subject specific response patterns.

III. Results

A. Statistical Analysis

Feature sets obtained by JMIM (default options) had size 17, 13, and 14 in LA vs BB1 vs BB2, LA vs BB1, and BB1 vs BB2, respectively. Exclusion of cross-correlated features resulted in the final size of the ranked sets of 12, 10 and 8. Even though each subset of features was tested for redundancy, when considering all the sets together is important to report for interpretation that EDAfz_{mean} and EDAfz_{median} are strongly correlated between each other ($\rho > 0.9$). Same for EDAfz_{range} with $\text{EDAfz}_{mobility}$, and also for cvxP_{range} with $\text{cvxPr}_{der_{mean}}$. The selected features were statistically tested to evaluate the difference in autonomic arousal between LA, BB1, and BB2. The results of the paired test are shown in Table II.

The statistical analysis showed a significant difference between LA, BB1 and BB2 for the majority of the identified features. The frequency domain provides the strongest features, especially in terms of f_{max} and Pwr_{B5}. These differences indicate a systematic modulation of autonomic features between LA and BB1. In particular, $\text{cvxPr}_{f_{max}}$, is strongly indicative that the gamified environment, introduced with BB1, provoked an augmented phasic activation. Indeed, phasic activity isolates the fast changes in the signal, hence, an augmented frequency in cvxPr, corresponds to an augmented arousal.

Moreover, many features provide significant p-values in BB1 vs BB2 pointing the varied solicitation provoked by the

TABLE II

WILCOXON SIGNED RANK TEST P VALUES COMPARING LA, BB1 AND BB2. P = RAW MANN-WHITNEY P-VALUE; P_FDR = BENJAMINI-HOCHBERG FDR-CORRECTED (PER CONDITION PAIR, $\alpha = 0.05$); R = RANK-BISERIAL CORRELATION (EFFECT SIZE).

Feature	LA vs BB1			LA vs BB2			BB1 vs BB2		
	p	**p_fdr**	***r***	**p**	**p_fdr**	***r***	**p**	**p_fdr**	***r***
EDAfz$_{f90}$	**<.001**	***<.001***	0.703	**<.001**	***<.001***	0.450	**0.006**	***0.014***	-0.280
cvxPr$_{Pwr_{B5}}$	**<.001**	***<.001***	0.789	**<.001**	***0.002***	0.343	**<.001**	***<.001***	-0.429
EDAfz$_{Pwr_{B5}}$	**<.001**	***<.001***	0.714	**0.006**	***0.011***	0.278	**<.001**	***<.001***	-0.426
EDAfz$_{f_{max}}$	**<.001**	***<.001***	0.673	**0.002**	***0.004***	0.314	**0.015**	***0.029***	-0.245
cvxPr$_{f_{max}}$	**<.001**	***<.001***	0.648	**<.001**	***0.002***	0.343	**0.001**	***0.006***	-0.324
EDAfz$_{Mobility}$	**<.001**	***<.001***	0.518	**0.008**	***0.014***	0.267	**0.045**	*0.055*	-0.203
cvxPr$_{f_{90}}$	**<.001**	***<.001***	0.553	**<.001**	***<.001***	0.438	0.28	*0.30*	-0.109
cvxT$_{range}$	**<.001**	***<.001***	-0.437	0.62	*0.70*	-0.051	**0.002**	***0.008***	0.307
EDAfz$_{range}$	**<.001**	***<.001***	-0.372	0.36	*0.44*	-0.092	**0.015**	***0.029***	0.246
cvxPr$_{Mobility}$	**0.004**	***0.005***	0.295	0.76	*0.80*	0.031	**0.020**	***0.034***	-0.235
cvxPr$_{skew}$	**0.002**	***0.003***	-0.319	0.80	*0.80*	-0.025	**0.023**	***0.036***	0.230
EDAfz$_{kur}$	**0.020**	***0.028***	0.235	**0.004**	***0.008***	0.292	0.71	*0.71*	0.038
cvxPr$_{kur}$	0.13	*0.15*	-0.152	**0.022**	***0.033***	0.231	**<.001**	***<.001***	0.383
cvxP$_{range}$	0.29	*0.30*	-0.108	0.054	*0.071*	0.194	**0.004**	***0.012***	0.289
EDAfz$_{mean}$	0.052	*0.063*	0.197	**<.001**	***<.001***	0.396	**0.031**	***0.044***	0.218
EDAfz$_{median}$	0.052	*0.063*	0.197	**<.001**	***<.001***	0.394	**0.039**	*0.052*	0.208
cvxT$_{median}$	0.33	*0.33*	0.098	**0.024**	***0.033***	0.229	0.22	*0.25*	0.124

game setting. The setting of bH and hF in BB2 is meant to require additional effort with respect to the natural motor gesture seen in LA. An easily interpretable feature is EDAfz$_{mean}$, indicative of the slow signal trend, usually referred to as tonic activity (extracted as an isolated component in cvxT). The corresponding box plots (omitted for brevity) suggest a progressive increase in tonic activity across conditions, with significant pairwise differences.

In the monotonic trend analysis, across the 66 sessions, most frequency and tonic features showed a monotonic increase from LA to BB1 to BB2, but with variable consistency, as shown in Table III. Only EDAfz$_{f90}$ showed a strong directional consistency with 80.3 percent of leg sessions showing an increasing trend. In contrast, 27 to 40 percent of recordings displayed flat or decreasing trajectories, and features such as EDAfz$_{range}$ and cvxPr$_{skew}$ showed no consistent direction. Overall, the majority of recordings support a progressive arousal pattern across conditions, but the substantial minority with non increasing trends suggest a subject-specific response.

TABLE III

MONOTONIC TREND ANALYSIS. SPEARMAN ρ WAS CALCULATED BETWEEN CONDITION ORDER [0,1,2] AND EACH INDIVIDUAL RECORDING. ρ = +1: PERFECT MONOTONIC INCREASE; ρ= -1: PERFECT DECREASE. WITH 3 CONDITIONS, ρ CAN ONLY TAKE VALUES IN -1, -0.5, 0.5, 1.

Feature	mean ρ	% $\rho > 0$	% $\rho > 0.5$
EDAfz$_{f90}$	0.386	80.3	19.7
EDAfz$_{Mobility}$	0.326	72.7	27.3
EDAfz$_{f_{max}}$	0.331	69.7	28.8
cvxPr$_{Pwr_{B5}}$	0.280	69.7	19.7
cvxPr$_{f90}$	0.324	69.7	33.3
EDAfz$_{Mean}$	0.311	68.2	45.5
cvxPr$_{f_{max}}$	0.205	65.2	19.7
EDAfz$_{median}$	0.242	63.6	43.9
EDAfz$_{Pwr_{B5}}$	0.182	62.1	22.7
cvxT$_{median}$	0.205	62.1	39.4
EDAfz$_{kur}$	0.182	62.1	21.2
cvxP$_{range}$	0.205	60.6	30.3
cvxPr$_{kur}$	0.152	60.6	18.2
cvxPr$_{Mobility}$	0.068	50.0	25.8
cvxPr$_{skew}$	-0.045	47.0	18.2
cvxT$_{range}$	-0.121	43.9	12.1
EDAfz$_{range}$	-0.189	37.9	7.6

B. Condition Classification

The classification performance were: accuracy 87% (88% with full set of features) in LAvsBB1, 61% (68% with full set) in BB1vsBB2, 60% (59% with full set) in the three class problem. These results indicate poor performance for difficulty classification, but no remarkable inflation due to the leakage from the feature selection process.

Looking at the SHAP values obtained in the LAvsBB1 classification, the inter-subject Std ranking is nearly identical to the inter-subject Mean ranking as shown in Table IV. EDAfz$_{f90}$ leads both, cvxPr$_{f_{max}}$ is lowest in both. This coherence is consistent with the high classification accuracy observed in this task: the features that drive the model most are also the ones with the largest absolute spread, but proportionally (CV 0.20–0.53) no feature shows disproportionate patient-to-patient variability. cvxPr$_{f_{max}}$ (CV = 0.20, Std = 0.153) is the most stable population-level biomarker: its contribution is both the strongest relative to its importance level, and the most consistent across patients. .

In contrast, BB1 vs BB2 accuracy remained close to binary chance despite significant statistical differences, suggesting that the underlying patterns do not generalise well across subjects. In particular, the observation of the SHAP values in this task, reported in Table V, shows the higher rank divergence for EDAfz$_{Pwr_{B5}}$: it ranks 1st by inter-subject Mean (0.752) but its inter-subject Std (0.392) is also the highest in absolute terms. Critically, compare with the same feature in LA vs BB1: there, EDAfz$_{Pwr_{B5}}$ had Mean = 0.895 and

TABLE IV

GLOBAL | SHAP|: MEAN AND STD ACROSS ALL HELD-OUT SAMPLES (ALL FOLDS). INTER-SUBJECT: PER-SUBJECT MEAN | SHAP| AGGREGATED ACROSS SUBJECTS. STD REFLECTS HOW MUCH A FEATURE'S CONTRIBUTION VARIES FROM PATIENT TO PATIENT. CV = STD/MEAN (NORMALISED). A HIGHER RANK SHIFT MEANS THAT THE FEATURE RANKS HIGHER IN INTER-SUBJECT STD THAN IN MEAN, INDICATING DISPROPORTIONATE INDIVIDUAL VARIABILITY.

LA vs BB1	***Global* | *SHAP*|**		***Inter-subject* | *SHAP*| *variability***	
	Mean	**Std**	**Std**	**CV**
EDAfz$_{f90}$	1.140	0.504	0.315	0.28
cvxPr$_{Pwr_{B5}}$	0.913	0.431	0.254	0.28
EDAfz$_{Pwr_{B5}}$	0.895	0.406	0.271	0.30
cvxPr$_{f_{max}}$	0.775	0.221	0.153	0.20
EDAfz$_{f_{max}}$	0.672	0.311	0.206	0.31
EDAfz$_{median}$	0.368	0.230	0.179	0.49
cvxP$_{range}$	0.263	0.185	0.139	0.53
EDAfz$_{Mobility}$	0.251	0.214	0.107	0.43
EDAfz$_{kur}$	0.224	0.123	0.087	0.39
cvxPr$_{Mobility}$	0.074	0.053	0.039	0.52

TABLE V

GLOBAL |SHAP|: MEAN AND STD ACROSS ALL HELD-OUT SAMPLES (ALL FOLDS). INTER-SUBJECT: PER-SUBJECT MEAN | SHAP| AGGREGATED ACROSS SUBJECTS. STD REFLECTS HOW MUCH A FEATURE'S CONTRIBUTION VARIES FROM PATIENT TO PATIENT. CV = STD/MEAN (NORMALISED). A HIGHER RANK SHIFT MEANS THAT THE FEATURE RANKS HIGHER IN INTER-SUBJECT STD THAN IN MEAN, INDICATING DISPROPORTIONATE INDIVIDUAL VARIABILITY.

BB1 vs BB2	***Global* | *SHAP*|**		***Inter-subject* | *SHAP*| *variability***	
	Mean	**Std**	**Std**	**CV**
EDAfz$_{Pwr_{B5}}$	0.752	0.596	0.392	0.52
cvxPr$_{kur}$	0.610	0.302	0.186	0.31
EDAfz$_{f90}$	0.533	0.307	0.196	0.37
cvxT$_{median}$	0.485	0.271	0.194	0.40
cvxP$_{range}$	0.446	0.283	0.175	0.39
cvxPr$_{Pwr_{B5}}$	0.367	0.278	0.181	0.49
cvxPr$_{f_{max}}$	0.323	0.265	0.154	0.48
cvxT$_{range}$	0.299	0.255	0.138	0.46

Std = 0.271. In BB1 vs BB2 the mean dropped moderately (0.752) but the std increased substantially (0.392): the absolute spread of contributions across patients grew even as average importance declined. This pattern indicates increased inter-subject variability in feature contribution for BB1 vs BB2. No other feature showed comparable inter subject inconsistency relative to its importance. Finally, the global Std (sample-level) for EDAfz$_{Pwr_{B5}}$ is 0.596, larger than the mean itself (0.752) only by a factor of 0.8, and much of this spread is within-subject recording noise rather than between-subject heterogeneity. The inter-subject Std (0.392) is the cleaner estimate of the patient-level variability.

IV. DISCUSSION

This study investigated whether gamification and difficulty modulation induce measurable changes in autonomic arousal in Parkinson's disease, and whether such changes generalise across subjects. Two main findings emerge: gamification induces a robust and generalisable autonomic modulation, difficulty modulation produces a real but heterogeneous effect.

In more details, the transition from LA to BB1 produced strong and consistent effects across statistical, trend, and classification analyses. Most of the frequency-domain features (EDAfz$_{f90}$, EDAfz$_{Pwr_{B5}}$, cvxPr$_{Pwr_{B5}}$, EDAfz$_{f_{max}}$, cvxPr$_{f_{max}}$) showed highly significant differences after FDR correction. The LOSO classification achieved 87% balanced accuracy, indicating that the autonomic pattern associated with gamification generalises to unseen subjects. The SHAP analysis further supports this stability. In LA vs BB1, the inter-subject standard deviation ranking closely follows the mean importance ranking. The most important features, such as EDAfz$_{f90}$ (mean |SHAP| = 1.140) and cvxPr$_{Pwr_{B5}}$ (0.913), also show proportional variability across subjects. Notably, cvxPr$_{f_{max}}$ exhibits the lowest coefficient of variation (CV = 0.20), indicating high consistency relative to its importance level. This coherence between importance and variability aligns with the high classification performance. The monotonic trend analysis reinforces this result at the individual recording level. EDAfz$_{f90}$ showed an increasing trajectory in 80.3% of leg sessions, the highest among all features. A second group of spectral and tonic features showed increasing trends in approximately 60–73% of sessions. Together, these findings indicate that the introduction of the game context induces a consistent autonomic modulation measurable across individuals. Although the motor gesture is similar between LA and BB1, the game context imposes timing constraints and attentional demands, plausibly contributing to the observed autonomic engagement.

In contrast, the transition from BB1 to BB2 yielded weaker cross-subject generalisation. Although several features remained statistically significant after FDR correction, the LOSO accuracy dropped to 61%, close to binary chance. This discrepancy indicates that statistical separability does not necessarily imply a stable population-level decision boundary. The SHAP analysis clarifies this behaviour. In BB1 vs BB2, EDAfz$_{Pwr_{B5}}$ ranks first in mean importance but also exhibits the highest inter-subject standard deviation. Compared to LA vs BB1, its mean importance decreases (from 0.895 to 0.752) while its inter-subject variability increases (from 0.271 to 0.392). This pattern indicates that the feature strongly contributes to classification for some patients but much less for others. No feature in this comparison demonstrates both high importance and low relative variability. The monotonic trend analysis is consistent with this interpretation. While several features show increasing trajectories in approximately 60–70% of sessions, a substantial proportion (30–40%) present flat or decreasing trends. This level of heterogeneity is sufficient to impair cross-subject classification performance. Difficulty modulation therefore appears to induce genuine autonomic changes, but these changes are not uniformly expressed across individuals.

From these results, it is possible to draw some implications for digital biomarkers and personalized adaptation. In fact, the results support EDA frequency-domain features as promising digital biomarkers of game-induced autonomic

engagement in Parkinson's disease. In particular, EDAfz$_{f90}$ consistently emerged as the feature with the strongest monotonic consistency and high classification relevance. However, the difficulty-related modulation highlights a limitation of population-level models. Beyond performance metrics, the explicit modelling of feature contributions through SHAP provides a transparency layer that is central to XAI. The comparison between mean importance and inter-subject variability makes it possible to distinguish features that generalise reliably from those that behave inconsistently across patients. The increased inter-subject SHAP variability observed in BB1 vs BB2 suggests that adaptive difficulty calibration may require subject-specific modelling. A fixed difficulty increment does not elicit uniform autonomic responses, which may reflect differences in disease severity, medication timing, cognitive load tolerance, or motivational factors. From a clinical perspective, this distinction is important. Gamification itself appears to reliably enhance engagement, whereas progressive difficulty may need dynamic adjustment to avoid under-stimulation in some patients and excessive load in others.

A. Limitations and future directions

The sample size limits subgroup analysis. The fixed protocol order (LA, then BB1, then BB2) introduces potential order or habituation effects that cannot be disentangled from condition effects. Clinical covariates such as disease stage or medication state were not available for stratified analysis.

Future work should investigate adaptive difficulty strategies and incorporate clinical metadata to determine whether the observed heterogeneity can be explained by identifiable patient characteristics. Longitudinal designs may also clarify whether autonomic engagement predicts training responsiveness over time.

V. Conclusion

This study aimed to investigate the impact of gamification on autonomic arousal during motor tasks in PD patients, utilizing wearable EDA recordings. Gamification induces a robust and generalisable autonomic modulation detectable through EDA frequency features. Difficulty increments produce measurable but individually variable responses that limit cross-subject generalisation. Together, these findings support the use of explainable machine learning to distinguish stable population-level biomarkers from features requiring personalised calibration in digital neurorehabilitation.

Acknowledgment

The authors would like to thank the volunteers that took part in this experiment at Associazione Amici Parkinsoniani Onlus. A special mention goes to the valuable Antonella T.

References

[1] D. Su, Y. Cui, C. He, P. Yin, R. Bai, J. Zhu, J. S. Lam, J. Zhang, R. Yan, X. Zheng, *et al.*, "Projections for prevalence of parkinson's disease and its driving factors in 195 countries and territories to 2050: modelling study of global burden of disease study 2021," *bmj*, vol. 388, 2025.

[2] S. Jain, "Multi-organ autonomic dysfunction in parkinson disease," *Parkinsonism & related disorders*, vol. 17, no. 2, pp. 77–83, 2011.

[3] J.-Y. Wang, M.-Y. Wang, R.-P. Liu, Y. Li, W.-Y. Zhang, B. Ovlyakulov, X. Zhang, and J.-H. Zhu, "Association analyses of autonomic dysfunction and sympathetic skin response in motor subtypes of parkinson's disease," *Frontiers in neurology*, vol. 11, p. 577128, 2020.

[4] P. Bosch-Barceló, O. Martínez-Navarro, M. Masbernat-Almenara, C. Tersa-Miralles, A. Pakarinen, and H. Fernández-Lago, "Gamification integration in technological devices for motor rehabilitation in parkinson disease: scoping review," *JMIR Serious Games*, vol. 13, no. 1, p. e69433, 2025.

[5] I. Quirosa-Galán, S. García-Bravo, A. C. Fabara-Rodríguez, M. P. Rodríguez-Pérez, E. Huertas-Hoyas, J. Pérez-Corrales, G. Fernández-Gómez, M. Donovan, and C. García-Bravo, "Video games in rehabilitation programs for people with parkinson's disease: A systematic review," *Applied Sciences*, vol. 15, no. 1, p. 311, 2024.

[6] G. Abbruzzese, R. Marchese, L. Avanzino, and E. Pelosin, "Rehabilitation for parkinson's disease: current outlook and future challenges," *Parkinsonism & related disorders*, vol. 22, pp. S60–S64, 2016.

[7] A. Calvo-Morata, M. Freire, I. Martínez-Ortiz, and B. Fernández-Manjón, "Scoping review of bioelectrical signals uses in videogames for evaluation purposes.," *IEEE Access*, 2022.

[8] R. Robinson, K. Wiley, A. Rezaeivahdati, M. Klarkowski, and R. L. Mandryk, "" let's get physiological, physiological!" a systematic review of affective gaming," in *Proceedings of the annual symposium on computer-human interaction in play*, pp. 132–147, 2020.

[9] R. Sánchez-Reolid, F. López de la Rosa, D. Sánchez-Reolid, M. T. López, and A. Fernández-Caballero, "Machine learning techniques for arousal classification from electrodermal activity: A systematic review," *Sensors*, vol. 22, no. 22, p. 8886, 2022.

[10] M. e. A. Barrachina-Fernández, "Advanced analysis of electrodermal activity measures to detect the onset of on state in parkinson's disease," *Mathematics*, vol. 11, no. 23, p. 4822, 2023.

[11] B. Ndlovu, K. Maguraushe, and O. Mabikwa, "Machine learning and explainable ai for parkinson's disease prediction: a systematic review," *The Indonesian Journal of Computer Science*, vol. 14, no. 2, 2025.

[12] X. Wu, L. Ma, P. Wei, Y. Shan, P. Chan, K. Wang, and G. Zhao, "Wearable sensor devices can automatically identify the on-off status of patients with parkinson's disease through an interpretable machine learning model," *Frontiers in Neurology*, vol. 15, p. 1387477, 2024.

[13] J.-Z. Xiang, Q.-Y. Wang, Z.-B. Fang, J. A. Esquivel, X.-Y. Li, and X.-Q. Xu, "Advancing parkinson's disease detection through multi-dimensional machine learning: a comprehensive framework using wearable movement sensor analytics," *Frontiers in Physiology*, vol. 16, p. 1737585, 2025.

[14] Y. Fırat, "An explainable ensemble machine learning model using baseline blood transcriptomics to predict parkinson's disease motor progression," *Frontiers in Digital Health*, vol. 8, p. 1774436.

[15] G. Amprimo, G. Masi, C. Ferraris, G. Olmo, and L. Priano, "A preliminary comparison between traditional and gamified leg agility assessment in parkinsonian subjects," in *2022 IEEE 10th International Conference on Serious Games and Applications for Health (SeGAH)*, pp. 1–8, IEEE, 2022.

[16] G. e. A. Amprimo, "Assessment tasks and virtual exergames for remote monitoring of parkinson's disease: An integrated approach based on azure kinect," *Sensors*, vol. 22, no. 21, p. 8173, 2022.

[17] J. Shukla, M. Barreda-Angeles, J. Oliver, G. C. Nandi, and D. Puig, "Feature extraction and selection for emotion recognition from electrodermal activity," *IEEE Transactions on Affective Computing*, vol. 12, no. 4, pp. 857–869, 2019.

[18] A. Greco *et al.*, "CvxEDA: A convex optimization approach to electrodermal activity processing," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 4, pp. 797–804, 2016.

[19] D. Homola, "mifs." https://github.com/danielhomola/mifs. Accessed: 2023-8-6.

[20] M. Bennasar, Y. Hicks, and R. Setchi, "Feature selection using joint mutual information maximisation," *Expert Systems with Applications*, vol. 42, no. 22, pp. 8520–8532, 2015.