Evaluating the Impact of Dataset Size on Classical and Topological Feature Extraction for Gait Classification

JHONATHAN BARRIOS*
Centre of Mathematics, School of Sciences, University of Minho
Braga, Portugal
id10605@uminho.pt

MATHEUS CARVALHO
School of Sciences, University of Minho
Braga, Portugal
pg52254@uminho.pt

WOLFRAM ERLHAGEN
Centre of Mathematics, School of Sciences, University of Minho
Braga, Portugal
wolfram.erlhagen@math.uminho.pt

MIGUEL F. GAGO
Hospital da Senhora da Oliveira de Guimarães and ICVS, School of Medicine, University of Minho
Braga, Portugal
miguelgago@hospitaldeguimaraes.min-saude.pt

ESTELA BICHO
Algoritmi Centre, School of Engineering, University of Minho
Guimarães, Portugal
estela.bicho@dei.uminho.pt

FLORA FERREIRA
Centre of Mathematics, School of Sciences, University of Minho
Braga, Portugal
fjferreira@math.uminho.pt

ABSTRACT

This study investigates the impact of dataset size on the performance of classical and topological feature extraction strategies in gait classification tasks. Using a ground reaction force dataset (GaitRec), we compare two feature extraction strategies applied to machine learning models: the classical approach, where models are trained directly with preprocessed data, and the topological data analysis approach, where time series are transformed using phase-space reconstruction followed by persistent homology to generate topological descriptors. Results show that topological descriptors, particularly Persistence Landscapes (PL), outperform the classical approach in small data scenarios. CatBoost with PL achieved an AUC of 0.943 in the Healthy Control versus Calcaneus Fractures task using the entire subject pool, while Random Forest maintained >86% accuracy even with 50% of the data. These findings highlight the potential of topological data analysis to improve classification performance in data-constrained clinical contexts.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning**; Feature selection; Machine learning approaches; • **Mathematics of computing** → *Topological data analysis.*

KEYWORDS

Gait Analysis; Machine Learning; Topological Data Analysis; Dataset Size Impact; GaitRec Dataset

AMMS 2025, September 24–26, 2024, Athens, Greece

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
JHONATHAN BARRIOS, MATHEUS CARVALHO, WOLFRAM ERLHAGEN, MIGUEL F. GAGO, ESTELA BICHO, and FLORA FERREIRA. 2025. Evaluating the Impact of Dataset Size on Classical and Topological Feature Extraction for Gait Classification . In *Proceedings of 7th International Applied Mathematics, Modelling and Simulation Conference (AMMS 2025).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Gait analysis is gaining an increasingly important role in both clinical and biomechanical fields, becoming a valuable tool for the evaluation of human locomotion [1]. The inherent characteristics of gait data such as high dimensionality, inter- and intra-subject variability, nonlinear relationships, temporal dependencies, and multiple correlations, generate large amounts of information that are often difficult to analyze and interpret [1, 2]. This complexity poses significant challenges for data analysis, requiring highly skilled clinical professionals to draw valid inferences. In response to these challenges, various machine learning (ML) approaches have been proposed in recent years, aimed at assisting clinicians in identifying and classifying gait patterns into clinically relevant categories, thus supporting the medical decision-making process [3–6].

Typically, ML algorithms become more effective as the size of the dataset increases. However, gait datasets derived from clinical studies, particularly those focused on neurodegenerative diseases such as Parkinson's disease, normal pressure hydrocephalus, or other conditions, are often small [4, 6, 7]. The challenge of applying ML to small datasets is a topic that is being widely explored and has recently been addressed through different strategies such as the use of normalization techniques [6, 8], the use of convolutional neural networks [4], or the more recent and growing field of computational topology, the topological data analysis (TDA), which has shown significant results in gait analysis [9, 10]. In this context, the main objective of this paper was to evaluate and compare the performance of ML techniques, to compare two feature extraction strategies applied to ML models: the classical approach, in which models are trained directly on preprocessed (yet relatively raw)

data, and the TDA approach, where the data are transformed using persistent homology to generate topological descriptors. For this purpose, the GaitRec database [11], released in 2020, was used. This dataset provides a large collection of ground reaction force (GRF) gait analysis data from more than 2,000 individuals, including healthy participants and individuals with four different types of gait disorders. The data offer an excellent opportunity to explore how different ML techniques perform on varying dataset sizes.

The paper is structured as follows. Section 2 provides a description of the GRF data and the proposed methodological approach. Section 3 presents the results and discussion, including the evaluation of ML models with classical and topological approaches. Section 4 outlines the conclusions and future directions for this research, emphasizing the potential of topological descriptors to improve the performance of ML models when applied to small database contexts.

2 MATERIALS AND METHODS

2.1 Database

GaitRec is a public database [11], which represents one of the most extensive and comprehensive datasets currently available for clinical gait analysis. The dataset consists of a total of 2,295 subjects, including 2,084 patients with various musculoskeletal impairments (*Gait Disorders*) and 211 healthy controls (HC) [11]. The data were collected between 2007 and 2018 at the Rehabilitation Center Weißer Hof, Austrian Workers' Compensation Board (AUVA), located in Klosterneuburg, Austria, following a standardized clinical protocol and with ethical approval [11].

Ground Reaction Forces were recorded bilaterally using embedded force plates along a 10-meter walkway. Each subject performed multiple walking tests at self-selected walking speeds, with valid recordings determined by clinical assessors [11]. GRF signals included three components: vertical, anterior-posterior, and mediolateral, initially sampled at 2000 Hz and subsequently downsampled to 250 Hz for storage. In addition to raw data, the dataset also provides a preprocessed and normalized version, where the signals were filtered (second-order Butterworth filter with a 20 Hz cutoff), temporally normalized to 100% of the stance phase (resulting in vectors of 101 points per cycle), and amplitude-normalized relative to body weight. This preprocessing minimizes the influence of anthropometric factors on force signals [11].

Patients were clinically classified according to the anatomical region affected, following a hierarchical taxonomy that includes impairments at the hip (H), knee (K), ankle (A), and calcaneus (C), with additional subcategories based on the type and combination of injuries [11]. Clinical annotations were performed manually by experienced therapists based on each patient's medical records. In total, the dataset comprises 75,732 bilateral walking trials, enabling both inter-subject studies (e.g., diagnostic classification) and intra-subject analyses (e.g., rehabilitation progress monitoring) [11].

2.2 Experimental Design

A specific selection and preparation of the GaitRec database was carried out to facilitate comparative analysis and control for clinical and methodological variability. First, only preprocessed data were used, performing a selection of diagnostic classes, focusing the analysis on three specific subgroups of musculoskeletal disorders with localized impairments: Calcaneus fractures (380 patients), Ankle fractures (156 patients), and Knee fractures (157 patients).

To simplify classification and reduce missing data, the analysis was limited to signals from the left leg. This ensured standardized comparisons, improved data consistency, and minimized biases from bilateral asymmetries or incomplete recordings. Only patients with left-side impairments were included. Additional filtering removed records with incomplete or invalid data, ensuring dataset integrity. Of the original 211 healthy controls, 209 complete records were retained. The final cohort distribution is shown in Table 1.

Table 1: Final cohort distribution.

Group	Number of Subjects
Calcaneus fractures (C_F)	178
Ankle fractures (A_F)	156
Knee fractures (K_F)	157
Healthy Controls (HC)	209

Note: All fracture cases involve the left side.

2.3 Classification Methodology

In the classical approach, machine learning models were trained directly on the preprocessed time series of 101 points. Each time series was concatenated into a one-dimensional fixed-length feature vector without additional feature extraction, allowing a direct comparison with the topological representations.

In the TDA approach, topological features were extracted from the time series following the standard persistent homology pipeline applied to temporal data. Phase space reconstruction was first performed using Takens embedding [12], with parameters set to $d = 2$ and $\tau = 1$, transforming each series into a two-dimensional point cloud that preserves the underlying dynamics of the gait cycle. Persistence diagrams for homology dimensions H_0 and H_1 were then computed from the Vietoris-Rips complexes built on these embedded point clouds (for more details see [13]). Finally, topological descriptors were extracted from the persistence diagrams following the methodology proposed by Yan et al. [10]. The Betti Curves (BC), Persistence Landscapes (PL), and Silhouette Landscapes (SL) were computed and vectorized using 25 bins, generating fixed-length feature vectors compatible with ML models. All topological processing was performed using the Giotto-TDA library [14].

For both feature extraction strategies, the same supervised classifiers were applied to ensure a fair and consistent comparison. The models included Random Forest (RF), Support Vector Machine (SVM), Multiple Linear Perceptron (MLP), and CatBoost Classifier (CatBoost). For RF, SVM, and MLP, hyperparameters optimized in previous studies [6] were adopted, ensuring consistency with previous experiments in small data sets.ets. In the case of CatBoost, hyperparameters were chosen following the work of Darshan et al. [5], where this algorithm demonstrated strong performance for gait disorder classification using the GaitRec dataset. Specifically, Random Forest was configured with 100 trees, square root feature

selection, and class balancing; SVM employed a linear kernel, class balancing, and probability estimation enabled; MLP used a single hidden layer with 50 neurons and a maximum of 300 iterations; and CatBoost was trained with 100 iterations, a learning rate of 0.1, tree depth of 6, and L2 leaf regularization set to 3. All models were implemented using `scikit-learn` (for SVM, RF, and MLP) and the official `CatBoost` library in Python, integrated into a unified training and evaluation framework. For each experimental scenario, k-fold cross-validation was used to assess model stability, and performance metrics included accuracy, area under the ROC curve (AUC), as well as sensitivity and specificity.

In addition, to evaluate the impact of the size of the dataset on model performance, a series of experimental scenarios were constructed by progressively reducing the number of subjects available. Specifically, models were evaluated using 100%, 90%, 80%, 70%, 60%, 50%, 40%, 30%, 20%, and 10% of the total selected subjects for each class. In each scenario, subjects were randomly sampled while preserving class proportions, and cross-validation techniques were applied to ensure robust performance estimation. This procedure allows for the investigation of model sensitivity to data availability and evaluates their generalization capacity under limited sample conditions.

3 RESULTS AND DISCUSSION

Table 2 displays the results obtained using the classical approach. Mean values and standard deviations for the classification metrics across all classifiers, considering the different proportions of the total subject.

It can be observed that RF and CatBoost consistently achieve the best results across all levels of subject reduction. In particular, when using 100% of the available data, CatBoost reaches an AUC of up to 0.965 for the classification task distinguishing HC vs C_F, with a sensitivity of 87.4% and a specificity of 93.8%. Similarly, RF and MLP show strong and stable performance, especially in the C_F and A_F tasks, maintaining accuracy rates above 85% even when only 60% of the total subjects are used. In contrast, SVM exhibits a more pronounced downward trend as the training set size decreases, with a notable drop in sensitivity in the 20% and 10% scenarios, particularly in the HC vs A_F classification task. However, its performance remains relatively competitive in intermediate scenarios (70%–90%), with AUC scores exceeding 0.86 in most cases.

An important observation is that the HC vs C_F classification task tends to be the most accurate among the three, consistently producing higher AUC values and greater stability in all metrics, even with small datasets. This suggests that alterations in gait pattern caused by calcaneus fractures generate GRF signals that are more distinguishable compared to those associated with knee (K_F) or ankle (A_F) fractures, at least when using raw signals as input.

As the percentage of subjects used decreases, all models show a progressive decline in performance, particularly in sensitivity. This degradation becomes more pronounced below the 30% level and is especially evident at 10%, where most classifiers lose generalization capacity, with AUC values falling below 0.8 in several cases. The results indicate that the classical approach achieves strong performance in scenarios with medium to large datasets, but suffers significant degradation as the number of available subjects decreases, especially in the absence of additional feature extraction or enrichment mechanisms.

In general, the RF classifier is the best performing model for the classification task across the different experimental scenarios. Although the CatBoost model occasionally achieves higher accuracy or AUC values in specific tasks, RF displays more stable and consistent performance as the number of subjects decreases. Furthermore, its lower computational cost makes it a preferable alternative when seeking a balance between performance and efficiency.

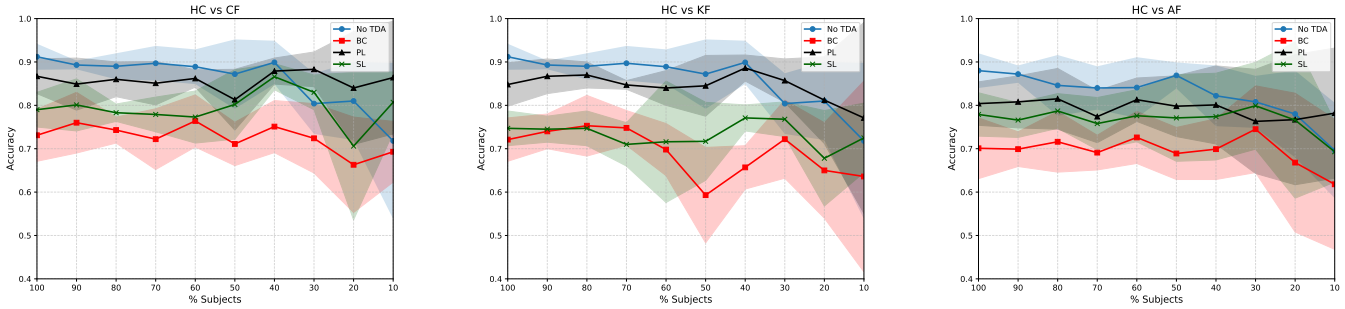
Figure 1 shows the accuracy curves of the RF model in the three classification tasks considered, comparing the classical approach (red line) with the three topological descriptors (BC, PL, and SL). In all cases, the topological approaches consistently outperform the classical approach when the number of available subjects is reduced, confirming their ability to extract more discriminative representations from complex time series. Furthermore, it can be seen that PL (black line) is the topological descriptor that best enriches the classification task. This may be due to the fact that, unlike BC, both PL and SL encode not only the number of topological features across scales, but also their persistence and prominence over time. This richer and more robust representation improves class separability and tends to be more resilient to noise, which could explain the superior performance of PL and SL, particularly in complex or highly variable data scenarios.

Table 3 presents the results obtained using the PL topological descriptor. As in Table 2, average values and standard deviations for all metrics are reported in different proportions of the subject pool. The results clearly show that PL significantly improves classifier performance in reduced data scenarios, especially when compared to the classical approach. In particular, RF and CatBoost continue to stand out as the most robust models when using PL, achieving accuracies close to or above 86% in multiple tasks even when the training set is reduced to 50% or less.

In the HC vs. C_F task, all classifiers exhibit stable and high results, with AUCs ranging from 0.92 and 0.95 down to 40% of the subjects. RF and CatBoost maintain sensitivities above 75% in most scenarios, and CatBoost achieves the highest overall performance with an AUC of 0.943 when 100% of the data are used. This behavior suggests that PL is particularly effective in capturing alterations in gait cycle associated with calcaneus fractures. For HC vs. K_F task, the overall performance of the models clearly improves compared to the classical approach. CatBoost and RF prove especially competitive, consistently maintaining AUCs above 0.87 even with 50% of the subjects. Although sensitivity shows some variability, PL enables the models to retain acceptable discriminative capacity, reinforcing its utility in more challenging clinical cases, such as knee-related gait disorders. Similarly, in the HC vs A_F task, both accuracy and AUC values exceed those obtained with the classical approach. RF and CatBoost maintain stable performance around 80% accuracy down to 30% of the dataset, and the SVM model reaches an AUC of 0.931 using only 30% of subjects, which is remarkable given the difficulty of this task. Although MLP results are more variable, PL generally helps preserve model generalization even under data scarcity. The results support the hypothesis that Persistence Landscape is a robust and effective topological descriptor, enriching classification models by capturing nonlinear and complex temporal patterns in gait signals. Its ability to maintain high levels of performance with

Table 2: Classifier Performance for the Classical Approach

% Subjects	Model	HC vs C_F				HC vs K_F				HC vs A_F			
		Acc	AUC	Spec	Sens	Acc	AUC	Spec	Sens	Acc	AUC	Spec	Sens
100	RF	**0.912±0.03**	**0.968±0.02**	**0.961±0.05**	**0.850±0.04**	**0.849±0.04**	**0.923±0.03**	**0.914±0.04**	**0.752±0.07**	**0.840±0.03**	**0.925±0.04**	**0.904±0.07**	**0.751±0.09**
	SVM	0.819±0.04	0.897±0.05	0.837±0.06	0.797±0.05	0.809±0.03	0.872±0.04	0.828±0.07	0.780±0.11	0.832±0.04	0.907±0.02	0.833±0.06	0.832±0.09
	MLP	0.885±0.04	0.955±0.02	0.904±0.04	0.862±0.05	0.857±0.02	0.913±0.02	0.895±0.03	0.801±0.05	0.851±0.03	0.933±0.03	0.847±0.08	0.859±0.05
	CatBoost	0.909±0.03	0.965±0.02	0.938±0.03	0.874±0.04	0.863±0.03	0.927±0.04	0.890±0.04	0.823±0.03	0.863±0.04	0.930±0.02	0.899±0.04	0.811±0.08
90	RF	**0.893±0.01**	**0.964±0.02**	**0.952±0.03**	**0.820±0.06**	**0.853±0.02**	**0.915±0.02**	**0.936±0.01**	**0.730±0.05**	**0.885±0.03**	**0.918±0.04**	**0.931±0.01**	**0.819±0.07**
	SVM	0.816±0.04	0.894±0.04	0.845±0.07	0.780±0.04	0.790±0.01	0.868±0.03	0.793±0.02	0.786±0.06	0.813±0.05	0.881±0.06	0.814±0.03	0.811±0.12
	MLP	0.890±0.05	0.956±0.03	0.931±0.05	0.840±0.05	0.850±0.05	0.919±0.04	0.920±0.04	0.747±0.10	0.841±0.07	0.924±0.07	0.851±0.05	0.826±0.15
	CatBoost	0.911±0.03	0.966±0.01	0.936±0.06	0.880±0.04	0.818±0.02	0.914±0.04	0.861±0.06	0.754±0.03	0.872±0.04	0.925±0.04	0.888±0.04	0.849±0.10
80	RF	**0.890±0.03**	**0.963±0.02**	**0.934±0.05**	**0.835±0.04**	**0.824±0.04**	**0.913±0.04**	**0.917±0.06**	**0.687±0.05**	**0.853±0.03**	**0.927±0.03**	**0.904±0.03**	**0.780±0.05**
	SVM	0.813±0.06	0.908±0.02	0.827±0.11	0.798±0.06	0.778±0.04	0.854±0.02	0.790±0.06	0.758±0.07	0.800±0.03	0.873±0.05	0.766±0.04	0.848±0.05
	MLP	0.860±0.04	0.955±0.01	0.905±0.08	0.806±0.10	0.824±0.04	0.890±0.03	0.875±0.06	0.749±0.14	0.828±0.06	0.918±0.06	0.820±0.05	0.837±0.18
	CatBoost	0.880±0.02	0.955±0.02	0.917±0.05	0.836±0.08	0.832±0.04	0.914±0.04	0.857±0.08	0.795±0.08	0.856±0.01	0.920±0.03	0.880±0.04	0.822±0.04
70	RF	**0.897±0.04**	**0.955±0.03**	**0.945±0.07**	**0.836±0.04**	**0.856±0.08**	**0.921±0.07**	**0.931±0.05**	**0.744±0.19**	**0.860±0.08**	**0.924±0.05**	**0.904±0.07**	**0.798±0.12**
	SVM	0.813±0.05	0.882±0.04	0.843±0.05	0.776±0.08	0.762±0.05	0.849±0.05	0.787±0.08	0.723±0.13	0.811±0.05	0.867±0.03	0.808±0.07	0.815±0.07
	MLP	0.790±0.05	0.881±0.05	0.849±0.10	0.714±0.13	0.811±0.07	0.887±0.05	0.876±0.06	0.714±0.15	0.835±0.07	0.907±0.05	0.863±0.06	0.796±0.13
	CatBoost	0.885±0.05	0.955±0.03	0.911±0.06	0.853±0.07	0.852±0.08	0.927±0.07	0.891±0.06	0.795±0.17	0.855±0.07	0.919±0.05	0.876±0.06	0.826±0.11
60	RF	**0.889±0.04**	**0.967±0.02**	**0.944±0.05**	**0.820±0.10**	**0.871±0.03**	**0.926±0.04**	**0.944±0.04**	**0.762±0.09**	**0.836±0.04**	**0.924±0.02**	**0.888±0.04**	**0.763±0.12**
	SVM	0.787±0.04	0.890±0.04	0.800±0.09	0.770±0.04	0.842±0.07	0.899±0.05	0.856±0.10	0.821±0.08	0.742±0.01	0.842±0.03	0.760±0.12	0.718±0.16
	MLP	0.902±0.05	0.960±0.03	0.928±0.05	0.870±0.08	0.875±0.05	0.936±0.03	0.920±0.06	0.810±0.08	0.794±0.02	0.882±0.04	0.808±0.08	0.775±0.13
	CatBoost	0.880±0.04	0.950±0.04	0.920±0.06	0.830±0.10	0.856±0.02	0.922±0.03	0.880±0.03	0.821±0.06	0.836±0.03	0.904±0.02	0.880±0.04	0.774±0.09
50	RF	**0.872±0.08**	**0.953±0.04**	**0.943±0.05**	**0.784±0.14**	**0.833±0.05**	**0.907±0.03**	**0.932±0.03**	**0.686±0.10**	**0.831±0.07**	**0.906±0.08**	**0.885±0.12**	**0.756±0.12**
	SVM	0.812±0.09	0.890±0.07	0.846±0.06	0.772±0.18	0.782±0.08	0.854±0.05	0.809±0.11	0.743±0.11	0.770±0.11	0.878±0.11	0.760±0.12	0.784±0.14
	MLP	0.866±0.08	0.922±0.07	0.923±0.04	0.796±0.17	0.833±0.07	0.875±0.07	0.885±0.05	0.757±0.16	0.787±0.10	0.879±0.09	0.788±0.12	0.782±0.15
	CatBoost	0.877±0.06	0.957±0.04	0.914±0.06	0.832±0.12	0.822±0.08	0.900±0.04	0.876±0.06	0.743±0.13	0.837±0.07	0.918±0.05	0.875±0.09	0.783±0.14
40	RF	**0.899±0.05**	**0.951±0.04**	**0.927±0.03**	**0.865±0.10**	**0.798±0.02**	**0.890±0.04**	**0.881±0.07**	**0.679±0.10**	**0.816±0.07**	**0.895±0.05**	**0.891±0.12**	**0.717±0.19**
	SVM	0.778±0.06	0.923±0.03	0.774±0.15	0.786±0.17	0.755±0.10	0.815±0.15	0.771±0.13	0.738±0.21	0.725±0.12	0.792±0.13	0.710±0.14	0.750±0.19
	MLP	0.866±0.03	0.954±0.03	0.868±0.05	0.863±0.07	0.827±0.09	0.882±0.05	0.854±0.09	0.789±0.14	0.803±0.07	0.859±0.05	0.795±0.12	0.817±0.14
	CatBoost	0.865±0.07	0.914±0.04	0.867±0.10	0.864±0.06	0.755±0.08	0.862±0.04	0.807±0.02	0.683±0.20	0.831±0.03	0.889±0.03	0.892±0.11	0.748±0.10
30	RF	**0.804±0.07**	**0.920±0.05**	**0.819±0.15**	**0.780±0.13**	**0.807±0.14**	**0.831±0.13**	**0.856±0.13**	**0.728±0.24**	**0.698±0.08**	**0.816±0.08**	**0.824±0.06**	**0.528±0.19**
	SVM	0.741±0.11	0.811±0.12	0.742±0.25	0.740±0.18	0.769±0.09	0.826±0.05	0.790±0.14	0.739±0.15	0.631±0.07	0.771±0.12	0.662±0.07	0.592±0.12
	MLP	0.839±0.05	0.897±0.11	0.886±0.14	0.780±0.08	0.797±0.09	0.839±0.07	0.854±0.10	0.714±0.16	0.688±0.07	0.751±0.10	0.728±0.10	0.636±0.14
	CatBoost	0.875±0.07	0.942±0.03	0.871±0.11	0.880±0.04	0.779±0.14	0.792±0.14	0.808±0.13	0.728±0.23	0.726±0.02	0.776±0.03	0.810±0.11	0.617±0.14
20	RF	**0.810±0.09**	**0.885±0.13**	**0.825±0.21**	**0.786±0.19**	**0.654±0.11**	**0.744±0.21**	**0.783±0.15**	**0.473±0.26**	**0.771±0.08**	**0.848±0.18**	**0.853±0.14**	**0.653±0.27**
	SVM	0.837±0.04	0.901±0.06	0.803±0.11	0.876±0.14	0.780±0.15	0.822±0.18	0.803±0.19	0.760±0.19	0.757±0.04	0.837±0.10	0.758±0.07	0.767±0.15
	MLP	0.810±0.08	0.858±0.13	0.800±0.19	0.824±0.12	0.811±0.12	0.866±0.14	0.853±0.06	0.753±0.23	0.757±0.04	0.817±0.03	0.831±0.06	0.660±0.16
	CatBoost	0.823±0.09	0.895±0.05	0.800±0.19	0.852±0.10	0.695±0.04	0.724±0.18	0.783±0.09	0.573±0.16	0.786±0.10	0.831±0.18	0.831±0.11	0.720±0.20
10	RF	**0.718±0.18**	**0.700±0.17**	**0.800±0.33**	**0.617±0.16**	**0.729±0.15**	**0.817±0.18**	**0.900±0.22**	**0.500±0.17**	**0.586±0.19**	**0.675±0.20**	**0.700±0.45**	**0.433±0.15**
	SVM	0.804±0.08	0.863±0.10	0.850±0.22	0.750±0.28	0.738±0.11	0.892±0.11	0.850±0.22	0.600±0.28	0.705±0.14	0.775±0.23	0.700±0.21	0.733±0.28
	MLP	0.779±0.12	0.775±0.15	0.850±0.22	0.700±0.18	0.757±0.16	0.787±0.23	0.850±0.22	0.633±0.25	0.676±0.12	0.704±0.07	0.750±0.25	0.600±0.28
	CatBoost	0.661±0.17	0.667±0.10	0.750±0.31	0.550±0.20	0.790±0.14	0.758±0.15	0.900±0.22	0.667±0.33	0.586±0.13	0.667±0.20	0.700±0.33	0.433±0.15



(a) HC vs C_F **(b) HC vs K_F** **(c) HC vs A_F**

Figure 1: Accuracy curves for Random Forest model

limited data makes it a valuable tool in clinical contexts where data availability is often restricted.

These results suggest that topological analysis offers greater stability and generalizability in limited data scenarios.

4 CONCLUSIONS

This work investigated the impact of dataset size on the performance of machine learning models applied to gait pattern classification, comparing a classical approach based on raw force signals with an approach based on TDA. The results showed that topological descriptors, particularly Persistence Landscape, significantly improve the generalization ability of the models in low-data contexts, outperforming the classical approach in most scenarios. These findings reinforce the utility of TDA as an effective tool for extracting nonlinear structures in complex time series, offering both practical and theoretical value in the field of gait analysis. Furthermore, they highlight that TDA-based strategies may be especially useful in clinical settings, where data availability is often limited. However, the study presents some limitations, such as the fixed choice of topological parameters, which can affect the generalization of the

Table 3: Classifier Performance for Persistence Landscape descriptor from the topological Approach

% Subjects	Model	HC vs C_F				HC vs K_F				HC vs A_F			
		Acc	AUC	Spec	Sens	Acc	AUC	Spec	Sens	Acc	AUC	Spec	Sens
100	RF	0.867±0.04	0.939±0.03	0.923±0.07	0.795±0.09	0.806±0.03	0.887±0.04	0.866±0.03	0.717±0.05	0.804±0.05	0.896±0.04	0.863±0.06	0.724±0.07
	SVM	0.848±0.05	0.911±0.03	0.885±0.02	0.801±0.12	0.786±0.04	0.861±0.03	0.804±0.05	0.759±0.07	0.738±0.02	0.818±0.03	0.792±0.05	0.667±0.06
	MLP	0.864±0.04	0.928±0.03	0.881±0.04	0.843±0.10	0.774±0.04	0.849±0.03	0.770±0.05	0.780±0.06	0.771±0.05	0.848±0.04	0.796±0.08	0.737±0.10
	CatBoost	0.862±0.05	0.943±0.03	0.904±0.06	0.807±0.10	0.834±0.04	0.912±0.03	0.876±0.05	0.773±0.05	0.834±0.05	0.916±0.03	0.886±0.04	0.764±0.07
90	RF	0.849±0.06	0.916±0.05	0.904±0.03	0.780±0.11	0.796±0.04	0.874±0.05	0.851±0.06	0.715±0.09	0.808±0.06	0.888±0.05	0.862±0.07	0.736±0.13
	SVM	0.867±0.04	0.921±0.05	0.883±0.07	0.847±0.03	0.815±0.02	0.854±0.04	0.830±0.04	0.793±0.08	0.781±0.04	0.853±0.05	0.804±0.04	0.750±0.06
	MLP	0.861±0.05	0.926±0.05	0.883±0.05	0.833±0.07	0.761±0.03	0.844±0.03	0.771±0.02	0.746±0.05	0.757±0.04	0.841±0.04	0.788±0.07	0.714±0.06
	CatBoost	0.846±0.06	0.927±0.05	0.899±0.03	0.780±0.11	0.835±0.06	0.901±0.03	0.883±0.02	0.763±0.16	0.812±0.05	0.900±0.05	0.857±0.05	0.750±0.10
80	RF	0.860±0.04	0.943±0.03	0.929±0.07	0.775±0.12	0.778±0.07	0.869±0.05	0.857±0.07	0.660±0.10	0.815±0.07	0.894±0.05	0.857±0.07	0.759±0.13
	SVM	0.870±0.03	0.922±0.03	0.886±0.06	0.850±0.04	0.810±0.03	0.892±0.03	0.827±0.07	0.786±0.08	0.754±0.07	0.818±0.07	0.774±0.09	0.727±0.14
	MLP	0.847±0.05	0.911±0.03	0.844±0.03	0.848±0.11	0.767±0.06	0.862±0.04	0.779±0.07	0.750±0.09	0.750±0.04	0.840±0.02	0.762±0.09	0.735±0.14
	CatBoost	0.860±0.03	0.929±0.04	0.922±0.03	0.782±0.10	0.821±0.05	0.907±0.04	0.887±0.07	0.725±0.12	0.822±0.08	0.898±0.04	0.857±0.12	0.775±0.11
70	RF	0.851±0.05	0.927±0.04	0.890±0.08	0.801±0.04	0.819±0.05	0.894±0.05	0.863±0.03	0.754±0.13	0.774±0.06	0.849±0.05	0.823±0.05	0.707±0.08
	SVM	0.847±0.01	0.917±0.01	0.842±0.04	0.853±0.05	0.787±0.05	0.881±0.04	0.802±0.08	0.765±0.08	0.762±0.07	0.846±0.04	0.774±0.09	0.743±0.13
	MLP	0.851±0.03	0.916±0.02	0.822±0.08	0.888±0.02	0.778±0.08	0.882±0.06	0.794±0.08	0.755±0.12	0.754±0.06	0.846±0.06	0.720±0.07	0.799±0.07
	CatBoost	0.866±0.06	0.945±0.02	0.924±0.06	0.792±0.10	0.803±0.05	0.901±0.04	0.849±0.03	0.734±0.10	0.773±0.06	0.863±0.04	0.795±0.06	0.744±0.12
60	RF	0.862±0.02	0.936±0.02	0.920±0.04	0.790±0.04	0.818±0.04	0.896±0.05	0.912±0.03	0.679±0.06	0.813±0.05	0.919±0.05	0.857±0.09	0.750±0.14
	SVM	0.840±0.04	0.900±0.07	0.872±0.08	0.800±0.05	0.809±0.03	0.865±0.03	0.864±0.05	0.726±0.07	0.754±0.07	0.820±0.05	0.754±0.03	0.753±0.13
	MLP	0.840±0.07	0.898±0.07	0.856±0.12	0.820±0.08	0.799±0.02	0.863±0.03	0.808±0.06	0.787±0.09	0.758±0.06	0.813±0.04	0.698±0.08	0.837±0.10
	CatBoost	0.849±0.06	0.942±0.03	0.872±0.12	0.820±0.06	0.847±0.03	0.915±0.03	0.920±0.05	0.739±0.10	0.849±0.05	0.924±0.04	0.873±0.09	0.816±0.08
50	RF	0.813±0.07	0.899±0.07	0.857±0.07	0.760±0.19	0.798±0.05	0.870±0.08	0.884±0.09	0.671±0.06	0.798±0.07	0.900±0.05	0.838±0.09	0.744±0.08
	SVM	0.845±0.07	0.896±0.08	0.856±0.11	0.832±0.12	0.764±0.02	0.837±0.05	0.780±0.07	0.743±0.13	0.699±0.03	0.755±0.05	0.724±0.06	0.666±0.06
	MLP	0.840±0.06	0.903±0.08	0.827±0.10	0.854±0.10	0.770±0.02	0.804±0.05	0.790±0.08	0.743±0.12	0.776±0.04	0.833±0.04	0.781±0.04	0.769±0.03
	CatBoost	0.839±0.08	0.921±0.06	0.885±0.05	0.784±0.17	0.822±0.10	0.904±0.07	0.894±0.13	0.714±0.16	0.814±0.06	0.893±0.05	0.876±0.09	0.731±0.05
40	RF	0.879±0.03	0.933±0.04	0.927±0.05	0.819±0.07	0.791±0.08	0.840±0.06	0.880±0.09	0.661±0.08	0.801±0.09	0.892±0.05	0.833±0.09	0.756±0.11
	SVM	0.886±0.03	0.937±0.04	0.878±0.10	0.895±0.09	0.784±0.04	0.845±0.03	0.820±0.04	0.732±0.11	0.726±0.03	0.814±0.04	0.726±0.04	0.727±0.07
	MLP	**0.913±0.03**	**0.949±0.01**	**0.927±0.05**	**0.893±0.12**	0.734±0.14	0.801±0.07	0.738±0.15	0.732±0.17	0.794±0.02	0.873±0.04	0.821±0.07	0.758±0.08
	CatBoost	**0.906±0.04**	**0.925±0.04**	**0.952±0.03**	**0.847±0.08**	0.806±0.06	0.870±0.09	0.893±0.08	0.679±0.05	0.808±0.07	0.895±0.04	0.857±0.07	0.741±0.09
30	RF	**0.883±0.04**	**0.940±0.02**	**0.933±0.09**	**0.820±0.15**	0.721±0.05	0.816±0.11	0.787±0.12	0.611±0.19	0.763±0.12	0.893±0.06	0.842±0.09	0.660±0.25
	SVM	0.857±0.05	0.931±0.06	0.869±0.09	0.840±0.05	0.653±0.07	0.722±0.10	0.694±0.03	0.600±0.17	0.706±0.09	0.801±0.12	0.745±0.10	0.660±0.22
	MLP	0.866±0.09	0.916±0.06	0.901±0.14	0.820±0.11	0.712±0.07	0.782±0.09	0.726±0.15	0.692±0.16	0.744±0.09	0.817±0.08	0.778±0.06	0.702±0.19
	CatBoost	**0.883±0.04**	**0.936±0.01**	**0.900±0.11**	**0.860±0.09**	**0.801±0.10**	**0.743±0.12**	**0.790±0.11**	**0.658±0.21**	0.745±0.12	0.855±0.08	0.827±0.12	0.638±0.23
20	RF	**0.840±0.13**	**0.928±0.06**	**0.850±0.14**	**0.824±0.16**	0.798±0.12	0.813±0.14	0.850±0.10	0.713±0.19	0.767±0.15	0.855±0.10	0.858±0.15	0.652±0.28
	SVM	0.812±0.10	0.849±0.09	0.806±0.14	0.824±0.12	**0.826±0.11**	**0.889±0.15**	**0.881±0.08**	**0.747±0.28**	0.727±0.08	0.779±0.13	0.761±0.12	0.681±0.15
	MLP	0.771±0.16	0.865±0.11	0.706±0.26	0.852±0.10	0.796±0.17	0.867±0.15	0.828±0.14	0.747±0.20	0.726±0.09	0.792±0.09	0.664±0.11	0.810±0.13
	CatBoost	0.813±0.14	0.907±0.12	0.828±0.11	0.795±0.22	0.769±0.08	0.821±0.13	0.856±0.10	0.640±0.12	0.710±0.18	0.804±0.13	0.783±0.15	0.619±0.28
10	RF	**0.864±0.13**	**0.925±0.10**	**0.950±0.11**	**0.767±0.22**	**0.762±0.09**	**0.833±0.08**	**0.900±0.14**	**0.567±0.15**	**0.782±0.15**	**0.837±0.17**	**0.900±0.14**	**0.600±0.43**
	SVM	0.771±0.22	0.858±0.15	0.750±0.31	0.800±0.30	0.762±0.09	0.733±0.25	0.700±0.11	0.833±0.24	0.700±0.13	0.787±0.19	0.720±0.19	0.667±0.33
	MLP	0.718±0.11	0.817±0.12	0.650±0.29	0.817±0.17	0.705±0.10	0.738±0.05	0.650±0.14	0.767±0.22	0.475±0.17	0.558±0.15	0.480±0.31	0.467±0.30
	CatBoost	**0.839±0.13**	**0.904±0.13**	**0.950±0.11**	**0.717±0.31**	**0.824±0.12**	**0.883±0.10**	**0.900±0.14**	**0.733±0.28**	**0.725±0.13**	**0.807±0.18**	**0.900±0.14**	**0.467±0.30**

results to other domains or biomedical signals. To mitigate this, future work could explore new topological descriptors and evaluate different combinations of topological parameters.

ACKNOWLEDGMENTS

This work has been supported by national funds through FCT - Fundação para a Ciência e Tecnologia through the doctoral scholarship https://doi.org/10.54499/2023.02242.BDANA and UID/00013: Centro de Matemática da Universidade do Minho (CMAT/UM).

REFERENCES

[1] Angkoon Phinyomark, Giovanni Petri, Esther Ibáñez-Marcelo, Sean T Osis, and Reed Ferber. Analysis of big data in gait biomechanics: Current trends and future directions. *Journal of Medical and Biological Engineering*, 38(2):244–260, 2018.

[2] Tom Chau. A review of analytical techniques for gait data. part 1: fuzzy, statistical and fractal methods. *Gait and Posture*, 13(1):49–66, 2001.

[3] Jianning Wu, Jue Wang, and Li Liu. Feature extraction via kpca for classification of gait patterns. *Human Movement Science*, 26(3):393–411, 2007.

[4] Carlos Fernandes, Flora Ferreira, Rui L. Lopes, Estela Bicho, Wolfram Erlhagen, Nuno Sousa, and Miguel F. Gago. Discrimination of idiopathic parkinson's disease and vascular parkinsonism based on gait time series and the levodopa effect. *Journal of Biomechanics*, 125(1):110214, aug 2021.

[5] Darshan Jani, Vijayakumar Varadarajan, Rushirajsinh Parmar, Mohammed Husain Bohara, Dweepna Garg, Amit Ganatra, and Ketan Kotecha. An efficient gait abnormality detection method based on classification. *Journal of Sensor and Actuator Networks*, 11(3), 2022.

[6] Jhonathan Barrios, Bárbara Araújo, Miguel Gago, Wolfram Erlhagen, Estela Bicho, and Flora Ferreira. Neurological disease classification based on gait analysis through transformation-based multiple linear regression normalization. In *New Frontiers in Statistics and Data Science*, pages 381–392, Cham, 2025. Springer Nature Switzerland.

[7] Çiğdem B. Erdaş, Eda Sümer, and Süleyman Kibaroğlu. Neurodegenerative diseases detection and grading using gait dynamics. *Multimedia Tools and Applications*, 82:22925–22942, 2023. Issue Date: June 2023.

[8] Flora Ferreira, Jhonathan Barrios, Paulo Barbosa, Miguel F Gago, Estela Bicho, and Wolfram Erlhagen. Impact of variable transformations on multiple regression models for enhancing gait normalization. In *Proceedings of the 2023 6th International Conference on Mathematics and Statistics*, ICoMS '23, page 103–107, New York, NY, USA, 2023. Association for Computing Machinery.

[9] Yan Yan, Olatunji Mumini Omisore, Yu-Cheng Xue, Hui-Hui Li, Qiu-Hua Liu, Ze-Dong Nie, Jianping Fan, and Lei Wang. Classification of neurodegenerative diseases via topological motion analysis—a comparison study for multiple gait fluctuations. *IEEE Access*, 8:96363–96377, 2020.

[10] Yan Yan, Yu-Shi Liu, Cheng-Dong Li, Jia-Hong Wang, Liang Ma, Jing Xiong, Xiu-Xu Zhao, and Lei Wang. Topological descriptors of gait nonlinear dynamics toward freezing-of-gait episodes recognition in parkinson's disease. *IEEE Sensors Journal*, 22(5):4294–4304, 2022.

[11] Brian Horsak, Djordje Slijepcevic, Anna-Maria Raberger, Caterine Schwab, Marianne Worisch, and Matthias Zeppelzauer. Gaitrec: A large-scale ground reaction force dataset of healthy and impaired gait, Apr 2020.

[12] Floris Takens. Detecting strange attractors in turbulence. In David Rand and Lai-Sang Young, editors, *Dynamical Systems and Turbulence, Warwick 1980*, pages 366–381, Berlin, Heidelberg, 1981. Springer Berlin Heidelberg.

[13] Frédéric Chazal and Bertrand Michel. An introduction to topological data analysis: Fundamental and practical aspects for data scientists. *Frontiers in Artificial Intelligence*, 4:667963, Sep 2021.

[14] G. Tauzin, U. Lupo, L. Tunstall, J. Burella Pérez, M. Caorsi, Anibal M. Medina-Mardones, A. Dassatti, and K. Hess. giotto-tda: : A topological data analysis toolkit for machine learning and data exploration. *Journal of Machine Learning Research*, 22(39):1–6, 2021.

Received ; revised ; accepted