TOPOLOGICAL GAIT ANALYSIS: A NEW FRAMEWORK AND ITS APPLICATION TO THE STUDY OF HUMAN GAIT - ADDITIONAL INFORMATION

SHREYAM MISHRA, DEBASISH CHATTERJEE *SENIOR MEMBER, IEEE*, AND NEETA KANEKAR

1. INTRODUCTION

This material accompanies the main manuscript titled *"Topological Gait Analysis: A New Framework and its Application to the Study of Human Gait"* by Shreyam Mishra, Debasish Chatterjee, and Neeta Kanekar. This document provides a comprehensive theoretical overview of cubical homology, clinical information for each subject from [1, 2] and conventional analysis of gait, a detailed explanation of the method used for tuning the hyperparameter σ, and an exploration of an additional dataset [3] comprising subjects with neurodegenerative conditions as well as aging to underscore the generalizability of the proposed topological gait analysis (TGA) framework. We also discuss the minimum amount of gait data (in terms of number of strides) required for the proposed TGA framework to produce accurate results and how its applicability can be extended to subjects with severe conditions who face significant challenges in walking for extended periods.

2. TOPOLOGICAL DATA ANALYSIS USING CUBICAL HOMOLOGY

Homology is a tool used to study the structure and connectivity of multi-dimensional data. 'Persistence' homology offers a method to inspect the data across various scales and identify 'significant' features that persist over a vast range. We review some basic concepts of cubical homology; for more formal treatment, readers are guided to [4, pp 57–59].

Definition 2.1 (Abstract grid cubes)**.** Consider $p \in \mathbb{Z}^d$ and $\ell \in \{0,1\}^d$. The set $c(p,\ell) \coloneqq [p_1, p_1+\ell_1] \times \cdots \times [p_d, p_d+\ell_d]$ is called the *abstract grid cube* in $\mathbb{R}^d$ associated with vectors p, ℓ with p being the base point of $c(p,\ell)$. For the case of $d = 2$, we call the elementary cube $c(p,\ell)$ as a *pixel* in X.

Definition 2.2 (Cubical complex)**.** A collection of grid cubes X is called a *grid cubical complex* if whenever $c \in X$, all the grid cubes contained as subsets in c also belong to X.

Definition 2.3 (Boundary operator)**.** To determine homologies of the grid complex, we define a linear map called as the nth boundary operator on every elementary cube of X as

$$\begin{aligned}\partial_n\big(c(p,\ell)\big) \coloneqq \sum_{m=1}^{d} (-1)^{\ell_1+\cdots+\ell_m} \big([p_1, p_1+\ell_1] \times \cdots \times \{p_m\} \times \cdots \times [p_d, p_d+\ell_d]\big) \\ - \big([p_1, p_1+\ell_1] \times \cdots \times \{p_m+\ell_m\} \times \cdots \times [p_d, p_d+\ell_d]\big),\end{aligned} \tag{1}$$

where $n = \dim\big(c(p,\ell)\big) = \ell_1 + \cdots + \ell_d$.

Definition 2.4 (Homology group)**.** For every $n \in \mathbb{N}$, let $Q_n(X)$ denote the *free abelian group* generated by all the n dimensional grid cubes in X. The homology groups $QH_n(X)$ are then defined as

$$QH_n(X) = \mathrm{Ker}(\partial_n) \big/ \mathrm{Im}(\partial_{n+1}). \tag{2}$$

The work was supported, in part, by IoE Cell, IIT Bombay. *(Corresponding author: Neeta Kanekar.)*

Shreyam Mishra and Debasish Chatterjee are with the Systems and Control Engineering, Indian Institute of Technology Bombay, Mumbai 400076, India (e-mail: mishra.shreyam@iitb.ac.in; dchatter@iitb.ac.in).

Neeta Kanekar is with the Department of Biosciences and Bioengineering, Indian Institute of Technology Bombay, Mumbai 400076, India (e-mail: nkanekar@iitb.ac.in).

Digital Object Identifier of main manuscript 10.1109/JBHI.2024.3427700 .

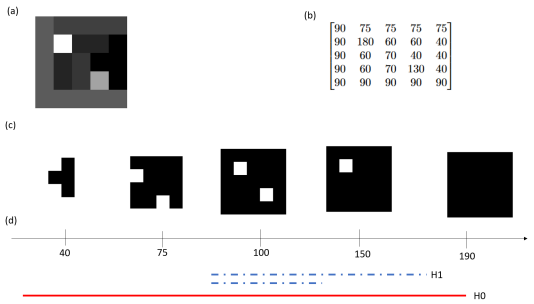


FIGURE 1. Application of cubical homology: (a) Shows a gray scale image (b) A matrix representing the pixel values of the image (c) The sequence of filtered cubical complexes (d) Persistence barcode associated with the filtration.

Intuitively the 0th homology group H_0 captures the number of connected components, the 1st homology group H_1 counts cycles, H_2 identifies voids, and so forth.

Consider the 5×5 gray-scale image in Figure 1(a). The collection of pixels $\left(q_{ij}\right)_{i,j=1}^{5}$ constitute the grid complex X, while the gray-scale values of each pixel shown in the matrix of Figure 1(b) define the function $X \ni q_{ij} \mapsto f(q_{ij})$. The sub-level sets of f are then defined by

$$L_t(f) \coloneqq \left\{q_{ij} \in X \;\middle|\; f(q_{ij}) \leqslant t\right\}.$$

Definition 2.5 (Filtration)**.** The sequence

$$L_{t_1}(f) \subset L_{t_2}(f) \subset \cdots \subset L_{t_n}(f),$$

for an increasing sequence of values $t_1 < t_2 \cdots < t_n$ is called a *filtration of* f. Figure 1(c) displays the sub-level sets of the image for $t \in \{40, 75, 100, 150, 190\}$.

The sub-level sets in Figure 1(c) capture the image's topological features effectively. The two 'holes' (relatively whiter pixels) appear to be born together at the filtration value $t = 90$. Formally, this is evaluated using (1) and (2), and is visually represented by the onset of the blue bar, denoted as H_1, at $t = 90$ in Figure 1(d). The 'gray-er' pixel (corresponding to pixel value 130 as shown in Figure 1(b)) disappears at $t = 130$, which is seen by the termination of one of the blue lines at $t = 130$ while the blue line corresponding to the second hole continues to grow until $t = 180$. The second 'hole' dies and the blue line tracking its evolution is also terminated. Since the image exists as a single connected component for all values of t, the red H_0 line never terminates. Such a pictorial representation of recording the appearance (birth) and disappearance (death) of topological features as the filtration parameter t is varied is called a *persistence barcode*. This information can also be captured using a single graph known as the *persistence diagram* $D \coloneqq \{(b_1, d_1), \ldots, (b_m, d_m)\}$, which is a collection of points in $\mathbb{R}^2$ with (b_i, d_i) representing the birth and death values of the ith feature. The persistence diagram for the given example is shown in Figure 2.

3. CLINICAL INFORMATION

The main manuscript covered the analysis of the Gait Dynamics in Neuro-Degenerative Disease Dataset from [1, 2]. Details regarding the data collection protocol and experimental conditions have been mentioned in §II. of the main manuscript. Clinical aspects such as subject age, disease severity (Sev.) along with essential statistics related to the left and right limb stride, stance and swing intervals (henceforth referred as LStr, RStr, LStn, RStn, LSw and RSw respectively) have been summarized in Table 1–Table 4. These statistics include the average (μ), standard deviation ($\hat{\sigma}$) and coefficient of variation (CV). The coefficient of variation was calculated as CV = $100(\frac{\hat{\sigma}}{\mu})$. Please note that in the Table 1–Table 4 mean and standard deviation have been rounded off for representational purposes. The CV values were calculated prior to the rounding-off. Please also note that for stride interval (s), the differences between right and left are at the level of the 4th decimal position. Any other information can be found in the Physionet database [1].

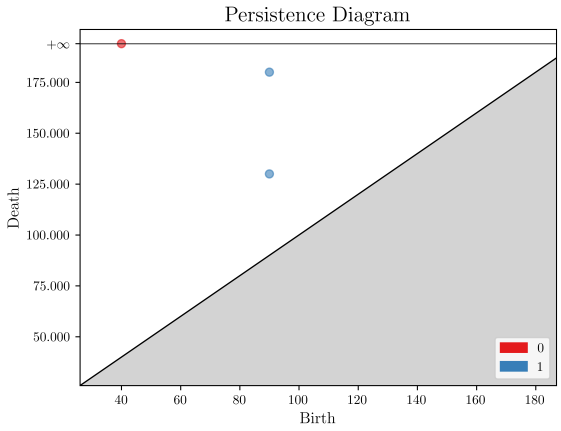


FIGURE 2. Persistence diagram for the gray-scale image in Figure 1(a). The two blue dots in the persistence diagram indicate the presence of two local local maximas.

TABLE 1. Clinical information and gait parameters of healthy controls.

Sub.	Age	Sev.	LStr (s)			RStr (s)			LStn (s)			RStn (s)			LSw (% GC)			RSw (% GC)		
			Mean	Std-Dev	CV	Mean	Std-Dev	CV	Mean	Std-Dev	CV	Mean	Std-Dev	CV	Mean	Std-Dev	CV	Mean	Std-Dev	CV
HC-1	57	0	1.07	0.04	3.81	1.07	0.04	3.52	0.73	0.04	5.57	0.69	0.03	4.63	32.39	2.07	6.39	35.55	1.60	4.51
HC-2	22	0	1.16	0.11	9.48	1.15	0.05	4.60	0.71	0.10	14.08	0.71	0.06	7.99	38.88	1.84	4.73	38.06	2.36	6.19
HC-3	23	0	1.09	0.03	3.02	1.09	0.04	3.36	0.70	0.02	3.56	0.75	0.03	3.78	35.71	1.21	3.40	31.32	1.14	3.63
HC-4	52	0	1.04	0.02	1.91	1.04	0.02	1.90	0.65	0.02	2.33	0.65	0.02	2.63	37.28	0.87	2.34	37.07	0.90	2.43
HC-5	47	0	1.11	0.05	4.88	1.11	0.05	4.83	0.70	0.05	6.86	0.72	0.05	6.80	37.13	1.99	5.36	34.84	2.29	6.57
HC-6	30	0	1.03	0.03	2.91	1.03	0.03	2.92	0.65	0.02	3.46	0.71	0.03	3.74	36.75	1.07	2.92	31.34	1.35	4.31
HC-7	22	0	1.07	0.03	2.85	1.07	0.03	2.88	0.67	0.02	3.63	0.67	0.02	2.73	37.45	0.74	1.99	36.81	0.89	2.42
HC-8	22	0	1.06	0.04	3.78	1.06	0.04	3.68	0.68	0.03	4.30	0.68	0.03	3.96	35.72	1.01	2.83	36.36	0.95	2.61
HC-9	32	0	1.01	0.04	3.69	1.01	0.04	3.81	0.64	0.03	4.75	0.65	0.03	4.08	36.67	1.33	3.62	36.02	0.98	2.72
HC-10	38	0	1.00	0.04	4.18	1.00	0.04	3.94	0.63	0.04	6.38	0.64	0.03	4.57	37.25	1.69	4.52	36.24	1.00	2.76
HC-11	69	0	1.04	0.04	3.56	1.04	0.04	3.56	0.69	0.03	4.33	0.67	0.03	5.15	33.16	1.07	3.23	35.34	1.45	4.09
HC-12	74	0	1.14	0.07	6.50	1.14	0.06	5.10	0.73	0.07	9.59	0.72	0.05	6.66	36.65	1.89	5.16	36.73	1.88	5.12
HC-13	61	0	1.11	0.04	3.52	1.11	0.04	3.78	0.70	0.03	4.56	0.70	0.04	5.35	37.08	1.28	3.45	36.52	2.18	5.96
HC-14	20	0	1.12	0.05	4.47	1.12	0.05	4.79	0.71	0.04	5.50	0.71	0.04	5.19	36.65	1.55	4.22	36.05	1.36	3.76
HC-15	20	0	1.40	0.07	5.00	1.40	0.07	5.07	0.89	0.04	4.81	0.89	0.05	5.91	36.17	1.24	3.44	36.14	1.07	2.95
HC-16	40	0	1.11	0.08	7.45	1.11	0.08	6.91	0.74	0.07	9.87	0.72	0.06	8.29	33.82	1.88	5.56	35.30	1.66	4.71
Group Mean			1.10	0.05	4.44	1.10	0.04	4.04	0.70	0.04	5.85	0.71	0.04	5.09	36.17	1.42	3.95	35.61	1.44	4.05

TABLE 2. Clinical information and gait parameters of people with PD.

Sub.	Age	Sev.	LStr (s)			RStr (s)			LStn (s)			RStn (s)			LSw (% GC)			RSw (% GC)		
			Mean	Std-Dev	CV	Mean	Std-Dev	CV	Mean	Std-Dev	CV	Mean	Std-Dev	CV	Mean	Std-Dev	CV	Mean	Std-Dev	CV
PD-1	77	4	1.13	0.04	3.69	1.13	0.05	4.26	0.74	0.05	6.32	0.78	0.05	7.01	34.98	3.06	8.74	31.56	3.59	11.38
PD-2	44	1.5	1.01	0.06	5.54	1.01	0.06	5.56	0.63	0.04	6.55	0.64	0.04	6.97	37.81	1.07	2.83	36.07	2.08	5.76
PD-3	80	2	1.21	0.11	9.36	1.21	0.11	9.30	0.81	0.10	11.94	0.86	0.10	11.74	32.97	2.08	6.31	28.64	2.57	8.99
PD-4	74	3.5	1.25	0.08	6.38	1.25	0.08	6.64	0.86	0.05	6.34	0.77	0.06	7.60	30.85	3.71	12.02	38.29	2.23	5.82
PD-5	75	2	1.06	0.05	4.77	1.06	0.05	4.94	0.71	0.04	6.15	0.72	0.04	5.92	32.92	1.95	5.92	31.85	2.30	7.24
PD-6	53	2	1.04	0.05	4.59	1.04	0.05	4.90	0.66	0.05	7.39	0.67	0.04	6.03	35.91	3.63	10.11	35.56	2.39	6.73
PD-7	64	4	1.23	0.59	47.64	1.22	0.34	28.11	0.84	0.60	70.99	0.82	0.23	28.57	32.98	3.76	11.39	32.75	3.21	9.80
PD-8	64	4	1.37	0.14	10.26	1.36	0.15	10.67	0.97	0.10	10.54	0.95	0.11	11.48	28.80	4.30	14.94	30.42	3.51	11.52
PD-9	68	1.5	1.25	0.11	8.81	1.25	0.11	8.72	0.80	0.09	11.28	0.81	0.10	12.79	36.27	1.45	4.00	35.35	2.15	6.08
PD-10	60	3	0.97	0.08	8.21	0.97	0.07	7.57	0.64	0.06	9.97	0.67	0.05	7.71	33.51	2.39	7.14	30.52	3.11	10.19
PD-11	74	3	1.21	1.73	142.98	1.15	1.42	123.23	0.90	1.71	189.80	0.90	1.42	158.72	29.98	5.13	17.12	24.97	4.69	18.79
PD-12	57	3	1.13	0.05	4.64	1.13	0.05	4.70	0.73	0.04	6.05	0.77	0.05	6.87	35.33	1.74	4.93	32.24	2.25	6.97
PD-13	79	3	1.11	0.11	9.99	1.10	0.11	9.63	0.74	0.12	15.97	0.75	0.10	13.28	33.17	4.00	12.07	31.88	2.54	7.96
PD-14	57	3	1.00	0.13	12.78	1.01	0.20	19.46	0.69	0.09	12.46	0.60	0.06	10.35	30.98	2.58	8.33	39.63	6.12	15.43
PD-15	76	2.5	1.17	0.05	4.46	1.17	0.05	3.88	0.79	0.05	5.93	0.81	0.05	6.45	32.79	3.21	9.78	30.63	3.19	10.41
Group Mean			1.14	0.23	18.94	1.14	0.19	16.77	0.77	0.21	25.18	0.77	0.17	20.10	33.28	2.94	9.04	32.69	3.06	9.54

4. HYPER-PARAMETER TUNING OF σ

To determine an optimal value for σ, the dataset was systematically divided into two sets: (1) training, and (2) validation, each containing a random sample of 50% of the subjects. The purpose of this splitting is to ensure that the hyper-parameter tuning occurs systematically and is not prone to overfitting.

The training set serves the purpose of tuning the hyper-parameter σ through visual inspection of the CDF plots of persistence entropy. The σ value that provided the clearest separation between healthy and

TABLE 3. Clinical information and gait parameters of people with HD.

Sub.	Age	Sev.	LStr (s)			RStr (s)			LStn (s)			RStn (s)			LSw (% GC)			RSw (% GC)		
			Mean	Std-Dev	CV	Mean	Std-Dev	CV	Mean	Std-Dev	CV	Mean	Std-Dev	CV	Mean	Std-Dev	CV	Mean	Std-Dev	CV
HD-1	42	8	0.90	0.05	5.74	0.90	0.05	5.69	0.56	0.04	6.67	0.54	0.03	6.28	38.19	2.87	7.52	39.49	2.32	5.89
HD-2	41	11	1.23	0.10	8.43	1.23	0.10	8.19	0.81	0.08	9.50	0.81	0.07	8.55	34.17	2.17	6.35	34.58	2.96	8.56
HD-3	66	4	1.20	0.12	9.80	1.20	0.13	10.55	0.79	0.09	11.27	0.77	0.07	9.64	34.16	3.92	11.48	35.40	3.95	11.15
HD-4	47	2	1.04	0.09	8.74	1.04	0.08	7.22	0.67	0.08	11.56	0.78	0.07	9.41	35.91	4.47	12.43	25.37	5.45	21.46
HD-5	36	10	1.06	0.12	11.42	1.05	0.09	8.50	0.65	0.09	13.10	0.66	0.08	12.41	38.48	2.49	6.48	37.08	2.70	7.28
HD-6	41	8	1.06	0.08	7.20	1.06	0.09	8.39	0.67	0.07	10.64	0.66	0.06	9.25	36.16	3.43	9.48	37.20	2.53	6.81
HD-7	71	2	1.20	0.10	7.98	1.20	0.09	7.82	0.81	0.07	8.15	0.79	0.06	7.26	32.27	3.85	11.94	34.33	3.14	9.14
HD-8	53	9	1.09	0.07	6.62	1.08	0.05	4.93	0.68	0.07	9.60	0.69	0.05	6.60	37.72	1.62	4.30	36.44	1.35	3.71
HD-9	54	12	1.03	0.07	6.86	1.03	0.08	7.41	0.65	0.07	10.25	0.70	0.07	10.36	36.86	1.82	4.93	31.98	2.23	6.99
HD-10	47	4	1.26	0.08	6.40	1.26	0.08	6.23	0.83	0.06	6.89	0.85	0.06	7.32	34.12	2.45	7.19	32.73	2.36	7.19
HD-11	33	11	1.16	0.07	6.21	1.16	0.07	6.23	0.74	0.06	8.13	0.87	0.09	10.29	36.04	2.78	7.73	25.46	6.09	23.91
HD-12	47	8	1.08	0.10	8.88	1.08	0.10	9.55	0.70	0.07	10.05	0.72	0.08	11.22	35.31	3.59	10.16	33.40	4.38	13.12
HD-13	40	5	1.67	0.39	23.27	1.67	0.38	22.65	1.07	0.28	26.08	1.12	0.31	27.74	35.69	8.39	23.52	33.30	8.86	26.60
HD-14	36	12	1.09	0.56	51.18	1.09	0.57	52.10	0.70	0.50	71.48	0.70	0.57	81.46	36.44	2.38	6.52	37.25	3.63	9.75
HD-15	34	3	1.14	0.30	25.94	1.13	0.26	22.80	0.89	0.27	30.11	0.81	0.25	30.69	22.66	6.29	27.75	28.97	7.30	25.19
HD-16	70	5	1.46	0.31	20.97	1.50	0.30	20.20	0.95	0.20	20.91	1.07	0.27	24.89	33.88	8.81	25.99	28.52	7.32	25.67
HD-17	29	12	1.13	0.08	7.19	1.12	0.07	5.84	0.74	0.07	9.52	0.75	0.06	7.91	34.45	1.96	5.69	33.52	2.21	6.59
HD-18	54	2	1.11	0.34	30.70	1.11	0.35	31.94	0.76	0.32	42.86	0.71	0.32	45.47	32.44	9.29	28.65	36.39	9.64	26.49
HD-19	59	1	1.15	0.09	8.13	1.15	0.09	8.11	0.76	0.07	9.46	0.80	0.07	8.22	33.45	3.81	11.40	30.09	4.27	14.18
Group Mean			1.16	0.16	13.77	1.16	0.16	13.39	0.76	0.13	17.17	0.78	0.14	17.63	34.65	4.02	12.08	33.24	4.35	13.67

TABLE 4. Clinical information and gait parameters of people with ALS.

Sub.	Age	Sev.	LStr (s)			RStr (s)			LStn (s)			RStn (s)			LSw (% GC)			RSw (% GC)		
			Mean	Std-Dev	CV	Mean	Std-Dev	CV	Mean	Std-Dev	CV	Mean	Std-Dev	CV	Mean	Std-Dev	CV	Mean	Std-Dev	CV
ALS-1	68	1	1.30	0.33	25.74	1.30	0.34	25.92	0.87	0.33	38.10	0.91	0.32	34.91	33.32	3.81	11.42	30.59	3.01	9.83
ALS-2	63	14	1.15	0.02	1.84	1.15	0.02	1.77	0.76	0.02	2.30	0.75	0.02	3.11	33.66	0.95	2.81	34.78	1.65	4.74
ALS-3	70	13	1.29	0.48	37.13	1.30	0.52	40.24	0.85	0.47	55.57	0.91	0.52	57.44	35.14	4.50	12.80	30.81	4.20	13.64
ALS-4	70	54	2.06	3.23	156.62	2.05	3.10	151.06	1.64	3.25	198.42	1.58	3.13	197.38	25.98	5.42	20.84	28.66	5.94	20.72
ALS-5	36	5.5	1.31	0.56	42.92	1.27	0.09	7.02	0.86	0.55	64.34	0.85	0.09	10.11	35.14	3.43	9.76	33.43	2.03	6.08
ALS-6	43	17	1.58	0.10	6.49	1.57	0.10	6.65	1.12	0.10	8.87	1.15	0.11	9.40	28.88	2.57	8.92	27.07	2.77	10.24
ALS-7	65	9	1.75	0.13	7.42	1.75	0.12	6.93	1.28	0.11	8.87	1.22	0.11	9.31	27.07	2.00	7.38	30.46	2.58	8.46
ALS-8	51	3	1.20	0.07	6.14	1.20	0.07	6.03	0.77	0.05	6.80	0.79	0.06	7.65	35.19	2.36	6.71	34.06	1.97	5.77
ALS-9	50	54	1.31	0.08	6.49	1.32	0.16	12.41	0.84	0.06	7.36	0.85	0.15	17.46	35.67	1.24	3.47	35.78	2.28	6.38
ALS-10	40	14.5	1.13	0.03	3.03	1.13	0.03	2.90	0.77	0.03	4.16	0.76	0.02	3.14	32.50	1.19	3.66	32.56	1.19	3.65
ALS-11	39	7	1.22	0.06	5.10	1.22	0.06	4.66	0.82	0.05	6.27	0.80	0.05	6.55	32.68	1.71	5.22	33.88	1.63	4.82
ALS-12	62	12	2.28	5.82	254.89	2.33	5.95	255.51	1.62	5.39	333.68	1.65	5.33	322.61	33.34	8.06	24.17	31.47	7.35	23.36
ALS-13	66	34	1.52	0.09	6.10	1.52	0.10	6.29	1.04	0.09	8.85	1.02	0.08	7.69	31.47	2.79	8.88	32.70	2.59	7.92
Group Mean			1.47	0.85	43.07	1.47	0.82	40.57	1.02	0.81	57.20	1.02	0.77	52.83	32.31	3.08	9.70	32.02	3.01	9.66

neurodegenerative gait was selected. This method is akin to the training-validation-test split commonly used in machine learning experiments.

The selected σ was then tested on the validation set to verify that the trend identified in the training set generalizes to the validation set. This step ensured that the chosen σ is not a consequence of overfitting to the dataset, but rather identifies the underlying patterns in the dataset effectively.

As illustrated in Figures 3-4, $\sigma = 0.0015$ was selected for stride analysis and $\sigma = 0.0020$ for stance analysis. Once an appropriate σ is fixed for each gait parameter, the CDF plots of persistence entropy are plotted to check if the trend identified based on training samples generalizes to a disjoint set of subjects, i.e., the validation set. Indeed, based on Figures 5a, 5b, 5c and 5d, the CDF plots continue to exhibit a clear trend with disease severity, indicating that these trends are not a consequence of overfitting. Although this analysis was not conducted for swing interval (% GC), similar results are anticipated with a high degree of confidence based on the consistency of findings observed in stride and stance intervals (s). The systematic tuning and validation process enhances the credibility of the findings and demonstrates that the identified trends are meaningful and reliable.

5. GENERALIZABILITY OF TGA FRAMEWORK

To verify the generalizability of the proposed TGA framework, we applied it to an additional, distinct dataset from PhysioBank [3].

5.1. **Procedure.** This database comprises of stride interval time series for 15 subjects, with 5 subjects per group:

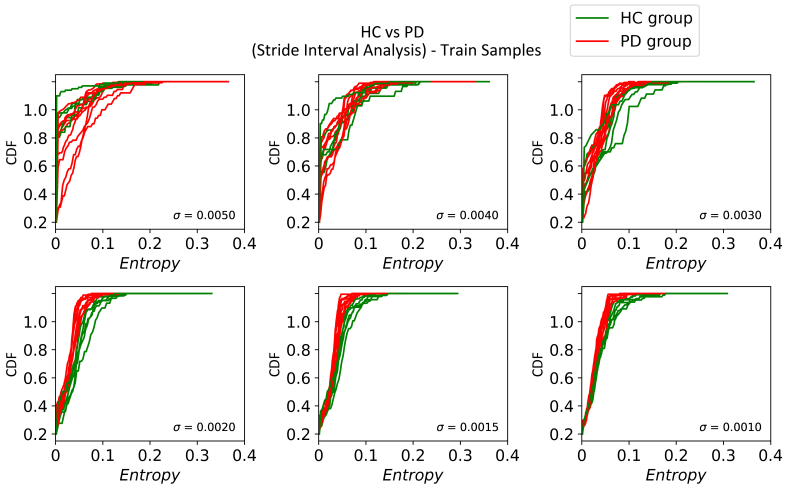


FIGURE 3. Persistence entropy plots for stride interval analysis corresponding to the training set. $\sigma = 0.0020$, 0.0015, and 0.0010 offer separation between HC and PD groups. $\sigma = 0.0015$ was chosen for validation.

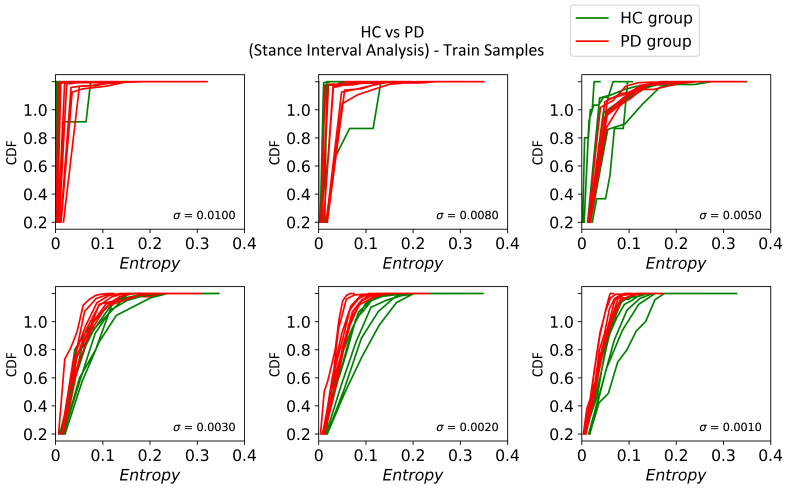
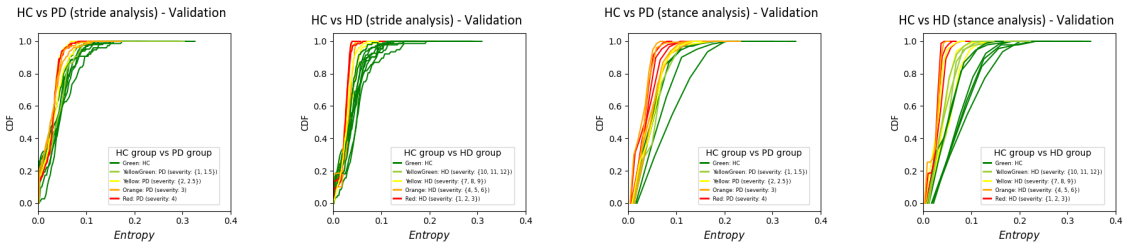


FIGURE 4. Persistence entropy plots for stance interval analysis corresponding to the training set. $\sigma = 0.0030$, 0.0020, and 0.0010 offer separation between HC and PD groups. $\sigma = 0.0020$ was chosen for validation.



(a) (b) (c) (d)

FIGURE 5. Validation comparison for stride and stance interval analysis between PD, HD, and HC groups using $\sigma = 0.015$ for stride analysis and $\sigma = 0.0020$ for stance analysis. The chosen σ values effectively identify trends based on disease severity in the validation set, which is entirely disjoint from the training set. This demonstrates that the identified patterns are true trends and not a result of parameter overfitting.

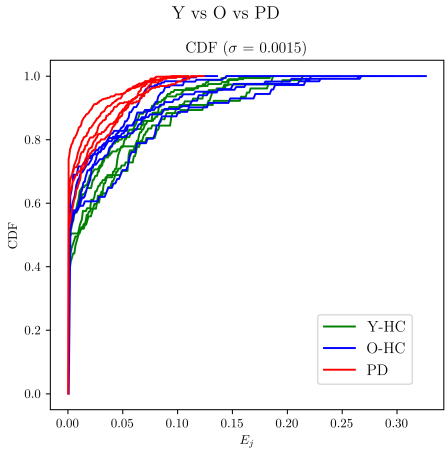


FIGURE 6. Persistence entropy plots clearly distinguish the healthy group from PD subjects. Additionally, the entropy plots for some older adults are closer to those of PD subjects in comparison to healthy, younger adults. This indicates that the method effectively captures gait deterioration due to neurodegeneration as well as due to aging. *Y-HC: healthy young adults*, *O-HC: healthy old adults*, *PD: subjects with Parkinson's disease.*

(1) Healthy young adults (Y-HC) (23-29 years old)
(2) Healthy old adults (O-HC) (71-77 years old)
(3) Older adults (60-77 years old) with PD

The experimental protocol involved subjects walking an obstacle free path on level ground. The stride interval (s) was measured using insole-based force sensitive resistors. The force signal was sampled at 300 Hz, converted to a digital signal via a 12 bit A/D converter using an ambulatory, ankle-worn micro-computer that also recorded the data. Subsequently, the time between foot-strikes was automatically computed. Data from the healthy subjects (young and old) was collected as subjects walked in a roughly circular path for 15 minutes. Data from subjects with PD was collected for 6 minutes along a long hallway. For the purpose of this analysis, the first 5 minutes of data from each subject was used.

5.2. **Analysis.** The reconstruction of the configuration space from the raw time-series involves the use of invoking Taken's theorem [5] which requires evalutation of the time-delay parameter τ and embedding dimension d as described in Section II of the main manuscript. In the main manuscript, we assume $d = 2$ and use the left and right limb time-series ($\eta_L(\cdot), \eta_R(\cdot)$ respectively) to construct the configuration space as $M := \big((\eta_L(n), \eta_R(n)\big)_{n=1}^{N}$ which subsumes and exploits the inherent time-delay between the limb movements. For this dataset, however, the time-series data for only a single limb was available. Therefore, we fixed $d = 2$ and determined an optimal time-delay τ for the dataset.

To determine the optimal τ for configuration space reconstruction, an empirical strategy was adopted. An optimal τ_s for each subject s was determined by minimizing the mutual information [6]. Once the optimal τ_s was determined for each subject, the average of all τ_s values was used as the optimal time-delay τ for the entire dataset.

Thus, if the raw time-series of a subject s is denoted by $\mathbb{N}^* \ni n \mapsto y_s(n) \in \mathbb{R}$, then the configuration space $M := \big((y_s(n), y_s(\tau + n)\big)_{n=1}^{N}\big)$, where N is the number of strides used for analysis.

5.3. **Results.** The results post implementation of the TGA framework on this new dataset have been encapsulated in Figure 6. CDF plots of persistence entropy revealed stark differences between PD and HC gait. The entropy plots of some older healthy adults (O-HC) appear to be closer to PD subjects in comparison to young healthy adults (Y-HC), indicating deterioration in gait with aging. Clearly this shows that the framework captures the deterioration in gait that occurs due to neurodegenerative conditions as well as due to aging.

Tables 5-7 demonstrate that the scalars α_s and E_s not only distinguish between healthy and neurodegenerative gait but also reveal significant differences between young and old healthy adults. This suggests

TABLE 5. Best-fit α_s and E_s values for healthy young adults.

Subject	Age	α_s (stride)	E_s (stride)
Y-1	23	20.9	0.039
Y-2	29	30.2	0.038
Y-3	23	28.8	0.045
Y-4	21	34.8	0.030
Y-5	26	40.6	0.035
Mean ± SD		31.06 ± 7.31	0.037 ± 0.006

TABLE 6. Best-fit α_s and E_s values for healthy old adults.

Subject	Age	α_s (stride)	E_s (stride)
O-1	76	42.7	0.026
O-2	74	39.5	0.018
O-3	75	26.4	0.022
O-4	77	36.6	0.025
O-5	71	33.6	0.040
Mean ± SD		35.76 ± 6.11	0.026 ± 0.008

TABLE 7. Best-fit α_s and E_s values for PD subjects.

Subject	Age*	α_s (stride)	E_s (stride)
PD-1	-	43.1	0.020
PD-2	-	41.6	0.017
PD-3	-	53.7	0.012
PD-4	-	48.1	0.022
PD-5	-	40.4	0.025
Mean ± SD		45.38 ± 5.10	0.019 ± 0.005

*The exact age of subjects with PD was not available. However, all subjects with PD were old adults in the age range of 60 - 77 years [3].

that the TGA framework generalizes well across different gait parameters and datasets and can effectively identify gait deterioration due to aging and neurodegenerative conditions.

6. DATA LENGTH AND STRIDE REQUIREMENTS FOR TGA FRAMEWORK

Section 5 shows that the TGA framework generalizes to different datasets and distinct gait parameters. However, both the datasets [1] and [3] analysed had a key advantage of having trials with a sufficiently large number of strides for the TGA framework to discern between healthy and neurodegenerative gait and identify clear trends with disease severity.

The practical challenges faced by patients in advanced stages of diseases, such as PD, where long trial durations may not be feasible, are recognized. Constructing a density estimate from time-series data typically benefits from longer trials, as they provide more data points, resulting in a more accurate density estimate and better overall analysis. The framework naturally extends to accommodate longer trials.

To determine the minimum trial duration in minutes or number of strides required for reasonable analysis, a heuristic strategy was adopted. The fraction of the total number of strides used for analysis was defined as ρ. For example, $\rho = 0.4$ indicates that only the first 40% of the total strides were used for analysis across all subjects. The final output for various values of ρ was then compared with the baseline where $\rho = 1$ (i.e. all available strides in the dataset were used). The value of ρ at which the results began to closely resemble

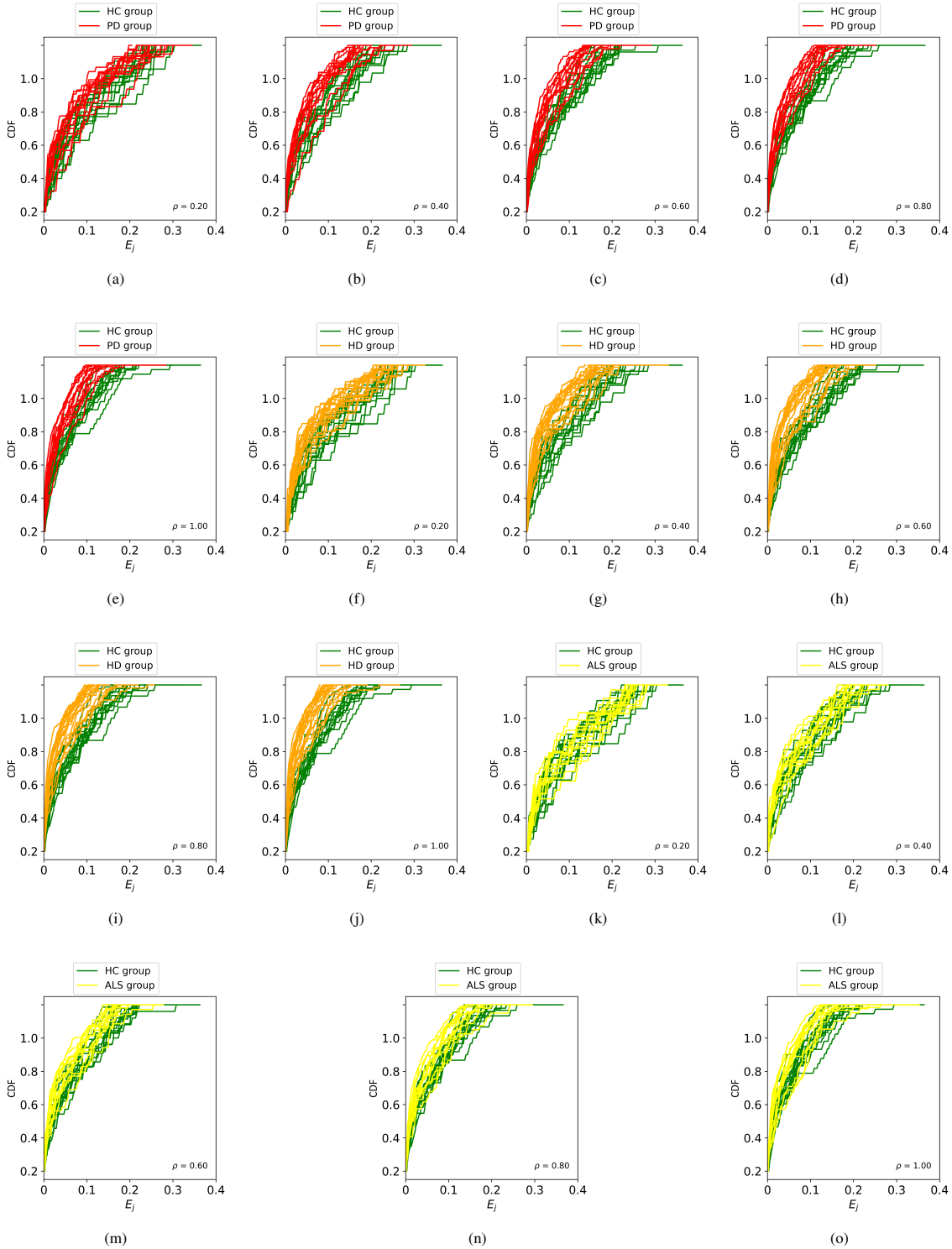


FIGURE 7. Persistence entropy plots of swing interval (in % gait cycle) from the TGA framework using various dataset fractions. Clear trends emerge for $\rho = 0.6$, but not for $\rho = 0.2$ or 0.4. Thus, at least 60% of a 5-minute walking trial (≈3 minutes or ≈180 strides) is required for accurate TGA results.

those obtained with $\rho = 1$ served as an estimate of the minimum fraction of the trial length required for the framework to produce reliable results.

As shown in Figure 7, for $\rho = 0.2$ and $\rho = 0.4$, there was no clear separation between the CDF of persistence entropy for healthy subjects and those with neurodegenerative conditions across all three groups. For $\rho = 0.6$, the plots begin to resemble those obtained from the entire dataset (i.e $\rho = 1$) for most subjects, while the results for $\rho = 0.8$ are nearly identical to the analysis of the complete dataset. Hence, we contend that a minimum of 60% of the 5-minute walking trial (typically comprising of nearly 300 strides), i.e. a 3 minute walking trial comprising of nearly 180 strides is required to make reasonable claims using the proposed TGA framework. For subjects unable to walk for this duration in a single trial, this framework may not be suitable and could lead to erroneous assessments. However, it is important to note that the recorded data used for analysis need not be part of a single trial comprising of consecutive strides. Instead, the strides can be collected across several trials with appropriately spaced breaks to aid patients with severe gait impairment. This could enable in extending the utility of this framework to severely affected subjects as well.

References

[1] A. L. Goldberger et al., "Physiobank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," pp. 215–220, 2000. [Online]. Available: https://physionet.org/content/gaitndd/1.0.0/

[2] J. M. Hausdorff et al., "Dynamic markers of altered gait rhythm in Amyotrophic Lateral Sclerosis," *Journal of Applied Physiology*, vol. 88, no. 6, pp. 2045–2053, 2000.

[3] A. L. Goldberger et al., "Physiobank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," pp. 215–220, 2000. [Online]. Available: https://physionet.org/content/gaitdb/1.0.0/

[4] D. N. Kozlov, *Organized Collapse: An Introduction to Discrete Morse Theory*, ser. Graduate Studies in Mathematics. American Mathematical Society, 2021, vol. 207.

[5] E. Bradley and H. Kantz, "Nonlinear time-series analysis revisited," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 25, no. 9, 2015, paper no. 097610.

[6] A. M. Fraser and H. L. Swinney, "Independent coordinates for strange attractors from mutual information," *Physical review A*, vol. 33, no. 2, p. 1134, 1986.