MaTiLDA: An Integrated Machine Learning and Topological Data Analysis Platform for Brain Network Dynamics

Katrina Prantzalos, Dipak Upadhyaya
Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH 44106, USA
Email: Katrina.prantzalos@case.edu; Dipak.upadhyaya@case.edu

Nassim Shafiabadi, Guadalupe Fernandez-BacaVaca
Department of Neurology, University Hospitals Cleveland Medical Center, Cleveland, OH 44106, USA
Email: Nassim.Shafiabadi@uhhospitals.org; Guadalupe.Fernandez-BacaVaca@uhhospitals.org

Nick Gurski
Department of Mathematics, Case Western Reserve University, Cleveland, OH 44106, USA
Email: Nick.gurski@case.edu

Kenneth Yoshimoto, Subhashini Sivagnanam, Amitava Majumdar
San Diego Supercomputer Center, University of California, San Diego, CA, USA
Email: kenneth@sdsc.edu; sivagnan@sdsc.edu; majumdar@sdsc.edu

Satya S. Sahoo
Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH 44106, USA
Email: Satya.sahoo@case.edu

Topological data analysis (TDA) combined with machine learning (ML) algorithms is a powerful approach for investigating complex brain interaction patterns in neurological disorders such as epilepsy. However, the use of ML algorithms and TDA for analysis of aberrant brain interactions requires substantial domain knowledge in computing as well as pure mathematics. To lower the threshold for clinical and computational neuroscience researchers to effectively use ML algorithms together with TDA to study neurological disorders, we introduce an integrated web platform called MaTiLDA. MaTiLDA is the first tool that enables users to intuitively use TDA methods together with ML models to characterize interaction patterns derived from neurophysiological signal data such as electroencephalogram (EEG) recorded during routine clinical practice. MaTiLDA features support for TDA methods, such as persistent homology, that enable classification of signal data using ML models to provide insights into complex brain interaction patterns in neurological disorders. We demonstrate the practical use of MaTiLDA by analyzing high-resolution intracranial EEG from refractory epilepsy patients to characterize the distinct phases of seizure propagation to different brain regions. The MaTiLDA platform is available at: https://bmhinformatics.case.edu/nicworkflow/MaTiLDA

Keywords: Epilepsy Seizure Network; Topological Data Analysis (TDA); Machine Learning; Neurological Disorders

1. Introduction

The increasing availability of multimodal brain activity recordings highlights an emergent demand for accurate and reliable analytical methods to characterize brain interaction dynamics to meet clinical research goals and to improve patient care[1]. The analysis of brain recordings provide insights into the dynamics of interaction patterns involving specialized brain regions that may be responsible for higher-order brain functions[2]. Understanding disruptions in brain interaction patterns is crucial to characterizing neurological disorders, revealing pathophysiological mechanisms, and defining biomarkers for clinical diagnoses[1–3]. These research goals are particularly important in epilepsy, which is a complex neurological disorder affecting over 50 million individuals worldwide[4]. Epilepsy is characterized by recurrent seizures stemming from abnormal electrical discharges that spread throughout the brain[4]. Similar to other disease domains, there has been a rapid increase in the use of machine learning (ML) algorithms to study brain interaction dynamics in epilepsy patients[5,6]. ML algorithms such as support vector machines (SVM) have used features extracted from neurophysiological signal data, such as electroencephalogram (EEG), to lateralize seizure onset zone for subsequent surgical intervention [5,6].

Graph-based models of networks are commonly applied to characterize interaction patterns in the brain; however, recent studies have used rigorous algebraic topology methods to analyze brain recordings to address several limitations of graph-based models[5,7–10]. Topological data analysis (TDA) is a quantitative framework that can be used to characterize higher-dimensional interaction patterns by using robust, scale-invariant methods, such as persistent homology[11]. Specifically, quantitative measures generated from persistent homology values, such as persistence landscapes, persistence images, and persistent entropy, have highlighted the promise of applying TDA methods to analyze EEG data with respect to seizure (ictal) activity[5,7,9,10] and to distinguish seizure onset from preictal activity[5,7]. Moreover, TDA methods have been integrated with ML algorithms for several applications[12], including characterizing brain interaction dynamics[5].

The development and use of an integrated ML and TDA tool to characterize brain interaction dynamics is a resource-intensive endeavor that demands expertise in domains such as mathematics, neurology, and computing. Therefore, there is a high entry barrier for the wider neuroscience community to use TDA methods and ML algorithms together for research studies[13,14]. To address this critical barrier, we introduce MaTiLDA as the first integrated web platform for TDA methods and ML algorithms to analyze neurophysiological recordings. We demonstrate the practical utility of MaTiLDA by characterizing brain interaction dynamics in refractory epilepsy patients using high resolution intracranial EEG (iEEG) recordings.

2 Background

2.1 The Neuro-Integrative Connectivity platform

Over the past decade we have developed an integrated neuroinformatics workflow tool called the Neuro-Integrative Connectivity (NIC) platform to automate the multi-step methods used to characterize brain interaction dynamics using signal data[15–17]. The NIC platform is a modular tool that supports addition of new modules in a flexible manner as support for new functionalities, including ML, are added. One module transforms neurophysiological signal recording stored in European Data Format (EDF) into a JSON- based human-readable format with semantic annotations using an epilepsy domain ontology that is more suitable for storage and analysis[15]. A second module computes signal coupling measures using both frequency and amplitude features of the signal data[16]. A third module computes a variety of graph model-based metrics[17]. A fourth module supports persistent homology functions using open source libraries such as GUDHI[18]. We refer to our previous work for additional details of the NIC tool[15–17]. MaTiLDA is an extension of the NIC tool to enable users to use TDA with ML algorithms for integrated analysis of signal data.

2.2 Topological data analysis of EEG

Brain functions are often characterized by interaction between multiple brain regions[2]; therefore, TDA is well-suited to characterizing these interaction patterns with high dimensionality, which cannot be easily represented using graph models[14]. Persistent homology is a TDA method that has been successfully used to identify brain states by analyzing multi-dimensional interactions across brain regions[5,7,9,14]. Specifically, studies applying persistent homology to neurophysiological signal data have shown the promise of TDA in characterizing aberrant brain interaction dynamics in neurological disorders[5,7,9,14]. In this section, we briefly describe the terminology associated with TDA methods to facilitate understanding of the subsequent sections of the paper.

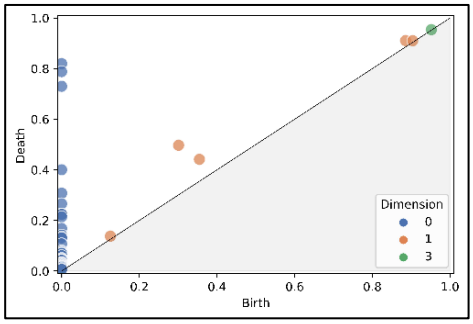

Figure 1: A persistence diagram from our analysis (section 2.6). A persistence diagram is a visualization of the results from persistent homology, where each point represents one homology class.

Persistent homology is a TDA method used to quantify the presence of topological structures, called homology classes, across various thresholds, or filtration values[14,19,20]. A homology class is a boundary composed of simplices, defined as the convex hull of a set of p+1 vertices[20]. A simplex has dimension p, and is referred to as a p-simplex, if it has a cardinality of p+1[13]. Persistent homology tracks the filtration at which each homology class is created (birth), the filtration at which it is terminated (death), and dimension of each homology class. These values can be visualized with a persistence diagram (Figure 1), a plot representing birth along the x axis and death along the y axis[11,13,19]. The lifespan, (death minus birth) of homology classes, as displayed in the persistence diagram, can be analyzed across various periods of neurophysiological signal recording to identify changes in topological structures and gain insights into the topology of brain networks[11,13,14]. We refer interested readers to Edelsbrunner and Harer[11] for further descriptions of persistent homology.

3. Methods

The computation and analysis of topological features from neurophysiological signal data entails multiple stages of processing, which include extraction of signal data, computation of signal coupling measures, TDA of signal coupling, data cleaning, and comparative analysis of topological features (Figure 2). Scientific workflow systems like the NIC platform have been used to automate these multi-step processes[17]. In this paper, we describe MaTiLDA as an extension of the NIC platform to implement integrated support for TDA and ML algorithms for brain interaction studies.

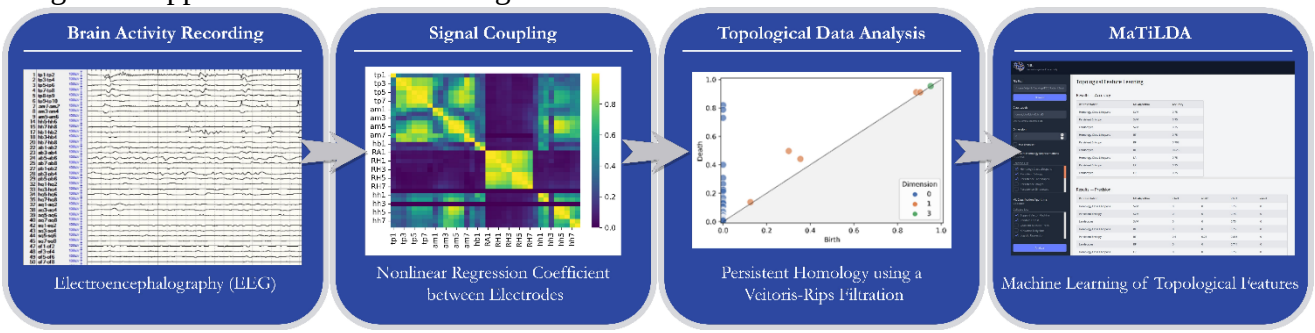

Figure 2: *Our framework for computing and comparing topological features from neurophysiological recordings. EEG from intracranial electrodes is used to extract signal data during epileptic seizures. Signal coupling is calculated using the nonlinear regression coefficient developed by Pijn et al.[21]. Persistent homology is applied to the signal coupling values using a Vietoris-Rips filtration as implemented in GUDHI[18]. MaTiLDA then allows users to select specialized data structures such as persistence landscapes or persistence images to use as input for user-selected machine learning classification such as SVM.*

3.1 MaTiLDA architecture and development

The MaTiLDA platform was built using the Django web application framework, which uses the Python programming language and features several libraries and modules that support a variety of data processing and analysis tasks including libraries for ML and TDA. MaTiLDA adopts the Model View Template (MVT) approach, with user inputs managed by an object relational data component (Model), the user interface handled by the View component, and user interaction mediated by the Template component.

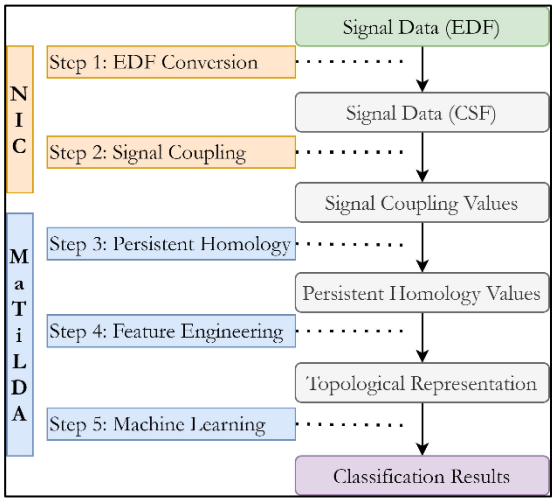

Figure 3: The MaTiLDA workflow leverages the NIC workflow to compute signal coupling. MaTiLDA applies persistent homology to the coupling values and allow users to select representations of the resulting persistent homology values for input into machine learning classifications of their choice.

3.2 A framework for classifying brain states

MaTiLDA leverages modules from the NIC tool and maintains a modular analysis process (Figure 3). Before analysis with MaTiLDA, neurophysiological recordings such as those from EEG are processed with the NIC tool

to convert from EDF to CSF and to compute signal coupling measures that can be used as input into MaTiLDA for a desired ML classification task. Users are required to provide a set of folders each containing a set of coupling measure values (Figure 4). Users can subsequently apply MaTiLDA's persistent homology function, using a Vietoris-Rips filtration, to each input using the GUDHI[18] library. The persistent homology values are transformed into a specialized data structure as requested; these data structures are used as input values for ML models selected by the user. A ML model is trained using an 80% data partition. Labels are predicted for the remaining 20% data partition as a test set. The test set accuracy score is reported alongside the precision, recall, and the area under the receiver operating characteristic (ROC) curve. Accuracy scores are calculated as the number of correctly identified predictions out of total predictions[22]. Precision is calculated as the number of true positive predictions divided by the number of positive predictions[22,23]. Recall, or true positive rate, is calculated as the number of true positive predictions divided by the number of positive samples[22,23]. The ROC curve is a plot of the true positive rate along the y-axis against the false positive rate along the x-axis for varying vales of a threshold used to classify samples[23].

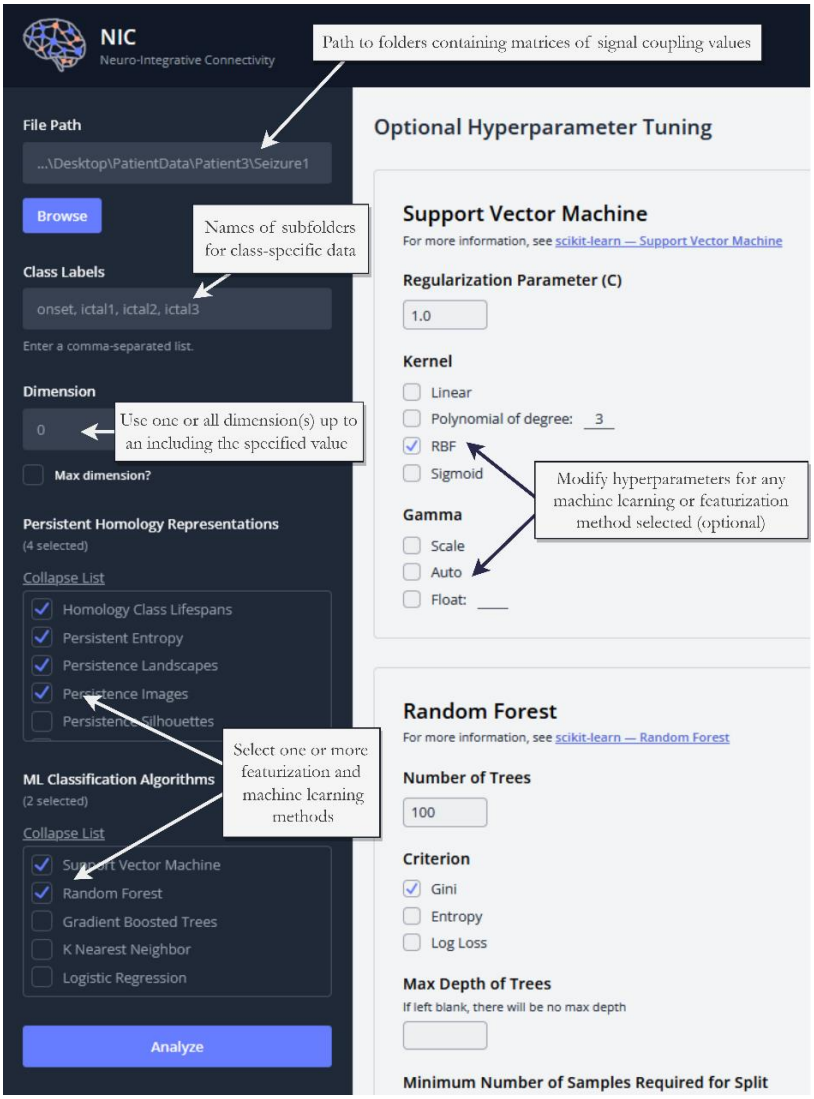

Figure 4: MaTiLDA supports various representations of persistent homology values in specialized data structures and ML algorithms with optional hyperparameter inputs. Users provide a folder including subfolders of outputs from the NIC correlator module, a list of all class labels (subfolder names), and a dimension for analysis. Users may select multiple data structures and multiple machine learning classification algorithms for their analysis using the checkboxes. For any selected representation or machine learning algorithm, a set of hyperparameters will appear in the left of the screen. The user may refine these parameters or use the preselected defaults. MaTiLDA will run each combination of representation-algorithm pairs selected for analysis. In the example provided above, the results from 8 analyses will be given.

The area under the ROC curve (AUC) measures the average classification accuracy across all thresholds[23]. A separate ML model is run for each combination of selected data structures and ML algorithms. By default, all ML models are implemented using default model parameters from Scikit-learn and GUDHI; however, users have the option to modify these parameters.

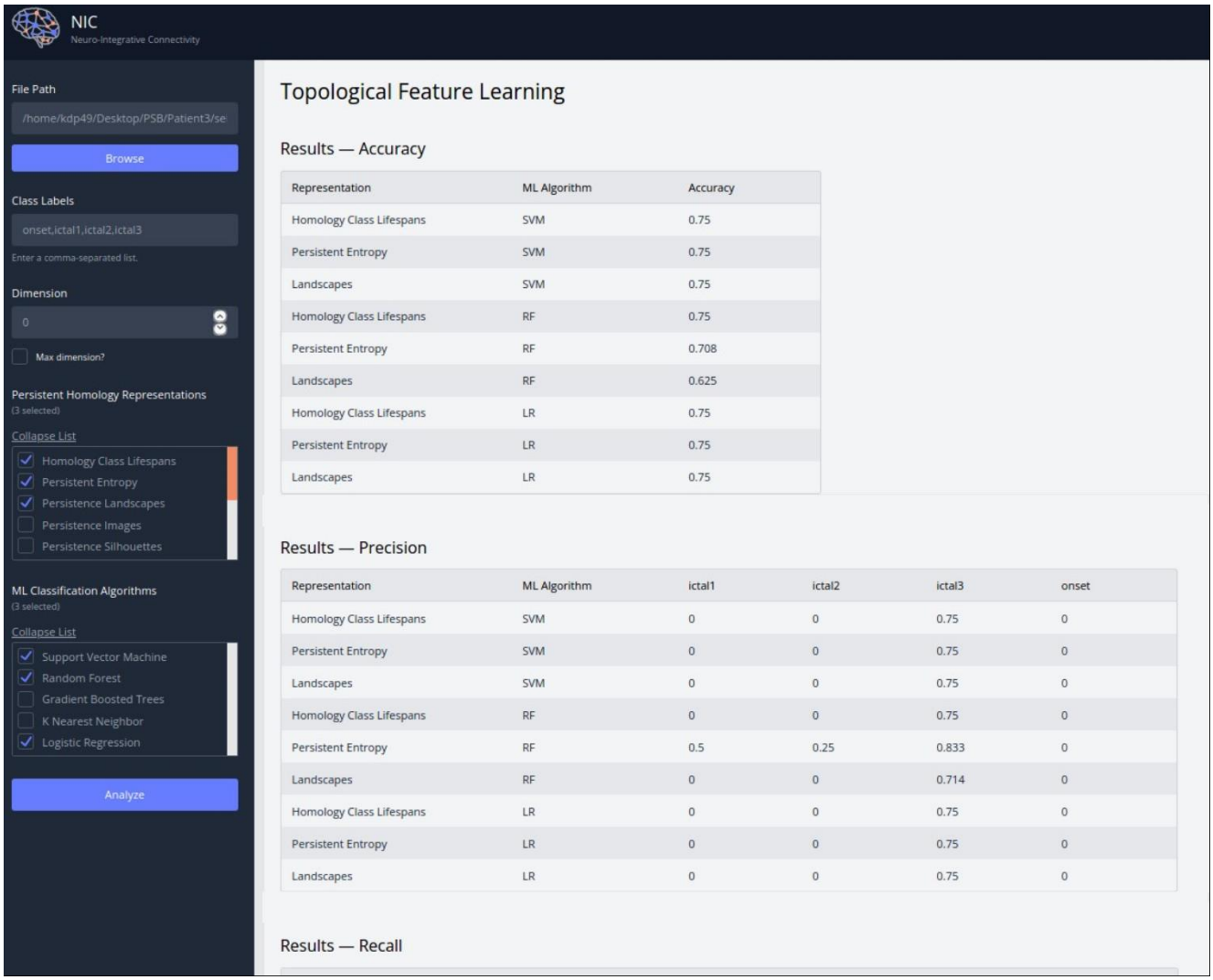

Figure 5: *Results for one seizure from a multiclass classification of ictal phases for patient one using homology class lifespans, persistent entropy, persistence landscapes, or persistence images as input to SVM, random forest, and logistic regression models.*

3.3 MaTiLDA user interface

The MaTiLDA user interface (Figure 4) consists of an intuitive data entry module and a minimal results table (Figure 5). MaTiLDA requires users to specify a directory containing several subdirectories, each of which should contain signal coupling values derived from neurophysiological signal data. MaTiLDA

internally manages all data preprocessing, expecting signal coupling values to be in the format produced by the NIC tools. A list of labels must be specified by the user; these labels will be matched to the subdirectory names to select and label signal coupling data from the main directory provided. Users must select a dimension for analysis; they may limit analysis to homology classes of that dimension, or they may analyze homology classes of dimension 0 through that dimension. Users may select several specialized data structures as representations for persistent homology values as well as several ML algorithms from a set of available options and may refine parameters for each selection using simple radio buttons and numeric input fields. Results are generated for all representation-algorithm pairs selected. The results table displays the representation chosen, the ML algorithm used, the model's accuracy in testing data, the true positive rate, the false negative rate, and the AUC.

3.4 Topological feature representation for machine learning

A key challenge for applying persistent homology lies in the difficulty of statistical interpretation of results[13]. Persistent homology values lack geometric properties that would allow for the definition of basic statistical concepts such as mean or median[13]. While persistence diagrams are an intuitive visualization method for representing the attributes of topological structures, the visual component of persistence diagrams makes it challenging to use statistical methods to quantitatively analyze them[12,13,19]. Additionally, persistence diagrams are not vectors in a Hilbert or Banach space and thus a unique mean cannot be established to define statistical measures[12,13]. Moreover, persistent homology values, and the persistence diagrams representing them, do not maintain a consistent number of homology classes, which creates a challenge for conducting balanced comparisons[12]. Consequently, a range of quantitative methods have been devised to facilitate the integration of persistence diagrams and persistent homology values into ML classifications. These methods for feature engineering can be used to represent persistent homology values as specialized data structures that can be used as input to ML models[12,19]. We provide the necessary background for the five quantitative methods for persistent homology value representation that have been implemented in the initial version of MaTiLDA: homology class lifespans, persistence

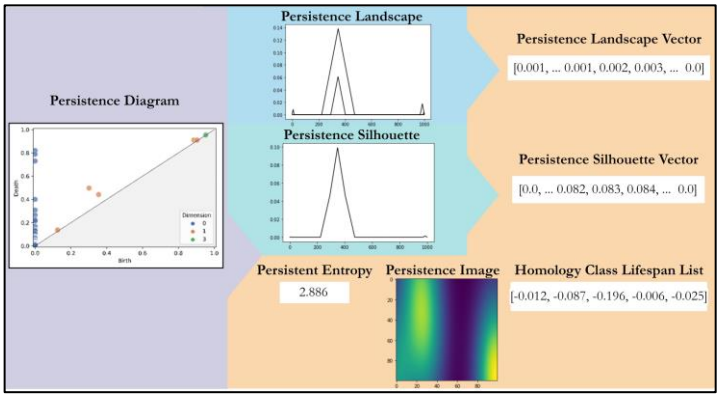

Figure 6: MaTiLDA offers several options for representing persistent homology values as vectors in Euclidean space, including persistence landscapes, persistence silhouettes, persistence images, persistent entropy, and homology class lifespans. Homology class lifespans create a list of values from the lifespans of all homology classes in a persistence diagram. Persistence landscapes and silhouettes transform persistence diagrams and apply a tent function before sampling uniformly across the transformed axis to create a list of values. Persistence images convert a persistence diagram into a two-dimensional image where each pixel represents a rectangular area of the diagram, and the intensity of the image represents the frequency of occurrence of homology classes. Persistent entropy is the Shannon

landscapes, persistence silhouettes, persistence images, and persistent entropy (Figure 6). In this work, we show how MaTiLDA can be used to intuitively conduct analyses by using these quantitative methods to represent persistent homology values derived from coupling measures computed from neurophysiological recordings and using the resulting features as input into ML algorithms.

3.4.1 Homology class lifespan

We calculate the lifespan for each homology class resulting from persistent homology and store the values in a list. Lifespan lists are ordered based on the lifespan values such that the first value in the lifespan list is the longest lifespan within that list. The lifespan list has a length equivalent to the sum of the Betti numbers (the number of homology classes) from all dimensions included in analysis. We create the input features for ML algorithms using tensor data structures that are padded with zero values to account for varying length of the tensors corresponding to different homology class lifespan values. Our methods are similar to the work described in the study by Bendich et al.[24]; however, unlike Bendich et al., we do not limit the number of lifespan values included in a list.

3.4.2 Persistence landscapes & silhouettes

The persistence landscape is a sequence of piecewise-linear functions, $\lambda_1, \lambda_2, \ldots : \mathbb{R} \rightarrow \mathbb{R}$, that map persistent homology values to a vector space, where λ_k refers to the kth persistence landscape function[25]. The persistence landscape can be calculated using Eq 1, where t denotes the filtration value, *kmax* denotes the kth largest element in the set of persistent homology values, *I*, and each homology class in *I* has a birth b_i and a death d_i[25].

$$\lambda(k,t) = kmax\{\max(0, \min(birth_i + t, death_i - t)\}_{i \in I} \quad (1)$$

The persistence landscape is plotted with the filtration along the x axis and the persistence landscape value $\lambda(k,t)$ along the y axis (Figure 6). A vector is created by uniformly sampling points along the x-axis and calculating the maximum of the persistence landscape functions at that point[19]. A persistence silhouette is a variation of the persistence landscape in which a vector is created by taking the weighted average of the functions, rather than the maximum[19,26]. The advantages of persistence landscapes and silhouettes are that they are invertible, parameter-free, nonlinear, and have desirable properties for statistical modeling including a unique mean[19,25].

3.4.3 Persistence images

To create a persistence image, a Gaussian function is applied to each homology class resulting from persistent homology[27]. The weighted sum of Gaussian functions are discretized to define a grid, and a matrix of pixel values is created by taking the integral of this grid on each grid box[27]. Consequently, each pixel value in the persistence image represents a rectangular area of the persistence diagram, and the intensity of the image represents the frequency of occurrence of homology classes[19,27]. Persistence images require a distribution, a resolution, and a weighting function to calculate[19]. The advantages of

persistence images are that they are stable, interpretable, and computationally efficient representations in $\mathbb{R}^{n}$[19,27].

3.4.4 Persistent entropy

Persistent entropy is a single value representing the Shannon entropy of a probability distribution obtained from persistent homology[28]. The persistent entropy of a set of persistent homology values can be calculated using Eq (2), where l_i is the lifespan of a topological structure[28].

$$\sum -\frac{l_i}{\sum -l_i} \log \left(\frac{l_i}{\sum -l_i}\right) \quad (2)$$

3.5 Machine Learning of Topological Features

In the MaTiLDA pipeline (Figure 4), persistent homology is applied to signal coupling values derived from neurophysiological signal recordings. Based on user specification (section 2.3), feature engineering is applied to the resulting persistent homology values to create specialized data structures (section 3.4) to be used as input features for ML models. Five common algorithms for ML classification were selected to be implemented in the initial version of MaTiLDA: support vector machines, random forest, gradient boosted trees, K-nearest neighbor, and logistic regression. In this section, we provide a brief introduction to each of these algorithms.

3.5.1 Support vector machine

Support vector machine (SVM) is a supervised learning algorithm that aims to find the best-separating function, called a kernel, to classify data into different categories[22]. While kernels do not naturally distinguish between more than two classes, SVM can be extended to multi-class classification problems using approaches such as the one-vs-one and one-versus-rest approaches[22]. For MaTiLDA, multi-class classifications using SVM are handled using the one-versus-rest approach. In the one-versus-rest approach, for a classification of K classes, SVM will fit K kernels where each kernel will compare one of the K classes to the remaining K-1 classes[22].

3.5.2 Random forest and gradient boosted trees

Random forest (RF) is a form of decision tree bagging (generating several training sets by sampling from the original training set with replacement) that focuses on making the ensemble of decision trees more diverse[29]. As in bagging, an ensemble of trees is built based on bootstrapped training samples[22]. However, rather than varying the training sets, a random sampling of attributes is selected at each split point in the tree; of this sample, the attribute with the highest information gain is selected as the split[29]. A majority vote from the tree-specific predictions is used to classify each example[29].

Gradient boosted trees (GBT), like random forest, is a powerful learning algorithm that can learn complex, non-linear relationships[29]. GBT is a boosting algorithm using gradient descent[29]. While

bagging builds trees on bootstrapped data independently of other trees, boosting uses a modified version of the original dataset to sequentially grow trees such that each tree is grown using information from previously grown trees[22].

3.5.3 K-nearest neighbor

K-nearest neighbor (KNN) is a non-parametric, supervised learning classifier that facilitates classification for observations by leveraging their proximity to the K nearest datapoints, or neighbors, in the training data[22,29]. The classification decision is made through a majority voting scheme among the K nearest neighbors[29]. KNN has a high computational cost due to performing distance calculations for each observation[29].

3.5.4 Logistic regression

Logistic regression (LR) models the probability that an observation belongs to a particular class[22]. By employing a logistic function, a linear combination of predictors is mapped to the range [0, 1], allowing LR to estimate the probability of class membership using maximum likelihood estimation[22].

3.6 Validation of MaTiLDA

Epilepsy is the second most common neurological disorder[4] and presents a unique opportunity for the application of TDA to study aberrant brain interaction dynamics. Epilepsy is characterized by recurrent seizures stemming from abnormal electrical discharges that spread throughout the brain and disrupt normal functioning[4,30]. Most significant changes to brain interactions during seizures occur during the spread of aberrant activity to new brain regions (referred to as ictal phases such as ictal 1 phase, ictal 2 phase, etc.) and the termination of a seizure[30]. One approach to understanding these changes in brain interaction dynamics is the classification of these ictal phases. To validate the use of the MaTiLDA interface for characterizing aberrant brain interaction dynamics using TDA and ML, we apply the MaTiLDA pipeline to analyze neurophysiological signal data from a cohort of four refractory epilepsy patients undergoing pre-surgical evaluation in the epilepsy monitoring unit (EMU) at University Hospitals Cleveland Medical Center's level 4 epilepsy facility that regularly performs epilepsy surgery. All patients were between the ages of 25 and 50 and had refractory epilepsy; 75% of the patients were women. ***Table 1*** shows the characteristics of these patients. Using MaTiLDA, we applied TDA and ML to analyze iEEG recordings from two seizures from each of these patients to classify ictal phases including seizure onset and propagation to different brain regions.

3.6.1 Study Data

We selected iEEG recordings from two seizures each from four refractory epilepsy patients undergoing pre-surgical evaluation. Intracranial electrodes are implanted based on a presurgical protocol described in work by Wu et al.[31]. Retrospective visual analyses of EEG recordings were conducted using a Nihon-Kohden Neurofax system (Nihon Kohden America, Foothill Ranch, CA, U.S.A.) with AC amplifiers,

a high sampling rate of 2,000 Hz, and an acquisition rate spanning 0.016-300 Hz[31,32]. The EEG was filtered at 600 Hz with a 0.03s time constant and sensitivity ranging from 30-100 μV based on optimal seizure visibility for each implant[31,32]. A 60 Hz notch filter was applied to all EEG recordings[31]. Clinicians defined seizure onset as the earliest distinctive occurrence of rhythmic sinusoidal activity or repetitive spikes; the region of activity was noted as the seizure onset zone[31]. Ictal phases were defined as the subsequent spread of seizure activity to new brain regions. EEG sequences were broken down into one second epochs and features were computed for each epoch.

Table 1: Characteristics of two seizures from four randomly selected refractory epilepsy patients.

Patient	Age Range	Sex	Epileptogenic Zone	Medication	Seizure Duration (s)	Active Electrodes	Ictal Phases	Seizure Semiology
1	25-30	F	Left Hemisphere	Trileptal, Keppra	48	IM1, IM8-9, SM1-3, IL6-8, ML1-8, SP2-5, IP1-3, MP1-3, HH1-10	2	Aura → mouth and hand automatisms → mild combativeness & amnesia
					43	IM1, IM8-9, SM1-3, IL6-8, ML1-8, SP2-5, IP1-3, MP1-3, HH1-10	2	Aura
2	45-50	M	Bitemporal	Lamotrigine, Phenytoin, Valproic Acid	90	TP1-8, AM1-8, HB1-2, RA1-2 RH1-8, HH1-8	2	Aura → postictal aphasia
					120	TP1-8, AM1-4, HB1-2	2	Aura → postictal aphasia
3	20-25	F	Left Mesial Temporal	Trileptal, Vimpat	120	HH1-3, HB1-3, AM1-3, MI1-12, PI1-12, IA1-12, IM1-12, SA1-12, MA1-12	4	Abdominal aura.
					120	HH1-3, HB1-3, AM1-3, MI1-12, PI1-12, IA1-12, IM1-12, SA1-12, MA1-12	4	Abdominal & gustatory aura
4	30-35	F	Right Mesial Temporal	Keppra, Lacosamide	60	HH2-3, EM8-9, HH1-12, HB1-12, TT1-12, OF1-12	4	After stimulating AM3 with 50Hz, 4.6mA, 3s, patient felt "oozy"
					60	AM1-2, EM9-10, HH1-12, HB1-12, TT1-12, OF1-12	4	After stimulating AM4 with 5Hz, 7mA, 3 seconds, patient felt funny

3.6.2 Study Design

All seizure data was preprocessed using the NIC tools. For each seizure, we used MaTiLDA to apply persistent homology to signal coupling values from one-second epochs of iEEG data and to create data structures representing the resulting persistent homology values that were used as input into ML models to classify epochs as belonging to an ictal phase. Of the eight seizures selected, four seizures were analyzed in binary classification tasks to classify seizure onset from ictal 1 phase, and the remaining four seizures were analyzed in multiclass classification tasks to

Table 2: The sample size of each class is equal to the duration of the associated ictal phase.

Patient	Seizure	Duration of Ictal Phase			
		Onset	Ictal 1	Ictal 2	Ictal 3
1	1	15	33	-	-
	2	15	28	-	-
2	1	10	80	-	-
	2	5	115	-	-
3	1	10	15	5	90
	2	10	15	5	90
4	1	10	15	5	30
	2	10	15	5	30

classify ictal phases (seizure onset, ictal 1 phase, ictal 2 phase, ictal 3 phase, and ictal 4 phase). Each seizure was analyzed separately. The number of one-second epochs in each ictal phase of each seizure, equivalent to the sample sizes of each class label in each seizure-specific analysis, is provided in Table 2. Default parameters were used for all representations of persistent homology values and for all ML algorithms in the analysis of each of the eight seizures to show the baseline capabilities of MaTiLDA.

4. Results

To validate the use of the MaTiLDA interface, we aimed to classify ictal phases within a seizure for eight seizures from four refractory epilepsy patients, as described in section 2.6. For brevity, we present only the results from the analysis of persistent homology values in dimension 0.

Binary classifications were used to compare seizure onset and ictal 1 phase for the four seizures from patient one and patient two, as these seizures were limited to these two ictal phases. Due to space constraints, we review only the results for RF, SVM, and LR models using either the lifespan or persistence landscape methods. ROC curves can be seen for each of these models for all four seizures in Figure 7. Model performance varied across all seizures, and no ML algorithm or representation of persistent homology values outperformed others to consistently distinguish seizure onset and ictal 1 phase (Figure 8). This may be due to imbalanced class sizes (Table 2). For example, the 20% test partition of patient two's second seizure contained only one epoch from seizure onset, and only four epochs from seizure onset were included in the 80% training partition. For all combinations of ML algorithms and representations of persistent homology values, this one epoch was misclassified as belonging to ictal 1 phase, resulting in precision and recall values of 0 and an AUC of 0.50 but an accuracy

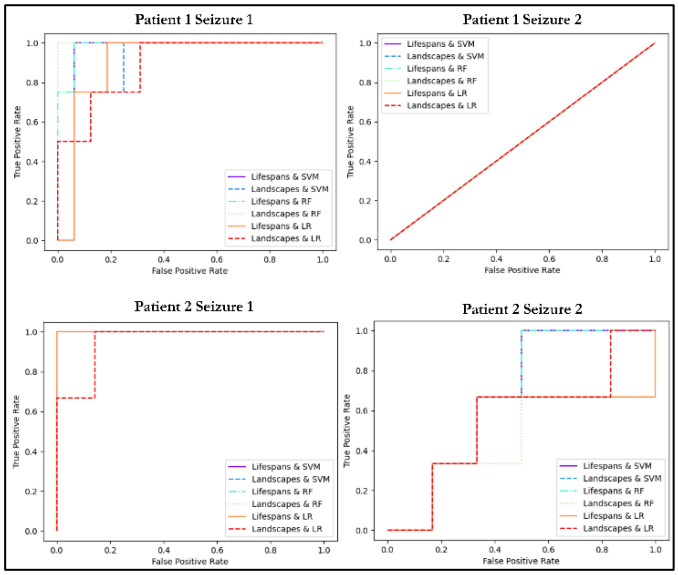

Figure 7: *ROC curves for each seizure from the binary classifications for seizures from patients one and two using lifespans or persistence landscapes in SVM, RF, or LR.*

Patient	Seizure	Algorithm	Featurization	Accuracy	Precision	Recall	AUC
1	1	RF	Lifespan	1	1	1	1
			Landscape	0.9	1	0.67	0.83
		SVM	Lifespan	0.9	1	0.67	0.83
			Landscape	0.8	0.67	0.67	0.76
		LR	Lifespan	0.8	1	0.33	0.67
			Landscape	0.8	1	0.33	0.67
	2	RF	Lifespan	0.67	0.5	0.33	0.58
			Landscape	0.67	0.5	0.33	0.58
		SVM	Lifespan	0.67	0.5	0.33	0.58
			Landscape	0.67	0.5	0.33	0.58
		LR	Lifespan	0.56	0	0	0.42
			Landscape	0.56	0.5	0.33	0.58
2	1	RF	Lifespan	0.95	0.8	1	0.97
			Landscape	1	1	1	1
		SVM	Lifespan	0.8	0	0	0.5
			Landscape	0.9	1	0.5	0.75
		LR	Lifespan	0.8	0	0	0.5
			Landscape	0.85	0.67	0.5	0.72
	2	RF	Lifespan	0.96	0	0	0.5
			Landscape	0.96	0	0	0.5
		SVM	Lifespan	0.96	0	0	0.5
			Landscape	0.96	0	0	0.5
		LR	Lifespan	0.96	0	0	0.5
			Landscape	0.96	0	0	0.5

Figure 8: *MaTiLDA's model performance for RF, SVM, and LR using lifespans or persistence landscapes for the four seizures from patients one and two.*

of 0.96. Increasing the number of samples from seizure onset may improve the ML models (as seen for patient two's first seizure). MaTiLDA's implementation of data augmentation, however, is still under development.

Multiclass classifications were used to classify seizure phases for each of the remaining four seizures from patients three and four which included multiple ictal phases (seizure onset, ictal 1 phase, ictal 2 phase, ictal 3 phase, and ictal 4 phase). Due to space constraints, we limit our results to the RF models using the lifespans and persistence landscapes (Figure 9). No algorithm or representation of persistent homology values consistently outperformed others to classify ictal phases, and there was high variation in model performance within and across seizures (Figure 9).

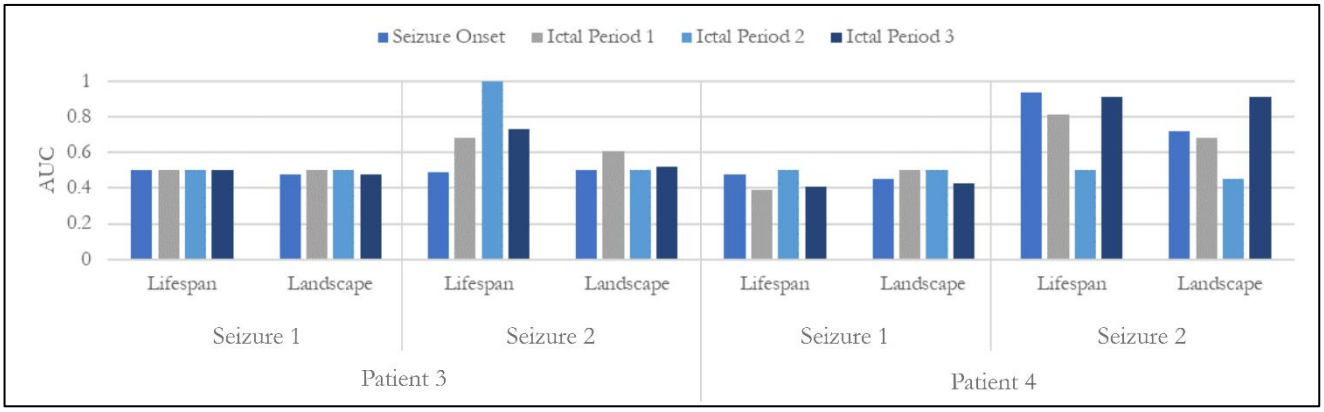

Figure 9*: MaTiLDA's One-vs-Rest AUC values for RF classification of ictal phases using lifespans or persistent landscapes for each of the four seizures from patients three and four show high variation in model performance within and across seizures.*

5. Discussion & Conclusion

The results of this evaluation demonstrate that MaTiLDA is an effective tool for analyzing complex topological features, enabling the detection of changes in brain interactions during seizures. We have developed a novel pipeline that can classify brain states, such as the ictal phases of several seizures in this study, using various common TDA methods and ML algorithms. The MaTiLDA platform provides a robust, accessible, and reliable framework for applying TDA and ML algorithms to datasets from neurophysiological recordings to characterize brain interaction dynamics in neurological disorders. MaTiLDA enables the wider neuroscience research community, who have limited experience in both TDA and ML algorithm implementation to use ML and TDA algorithms to analyze the increasingly large volumes of brain activity recordings and characterize brain interaction dynamics. We believe that the MaTiLDA tool can be used in future research to investigate complex brain interaction patterns in neurological disorders such as epilepsy, and allow clinicians and researchers to characterize neurological disorders, understand pathophysiological mechanisms, and identify biomarkers for clinical diagnoses.

References

1. Bassett, D. S. & Sporns, O. Network neuroscience. Nat Neurosci 20, 353–364 (2017).
2. Menon, V. Large-scale brain networks and psychopathology: a unifying triple network model. Trends in Cognitive Sciences 15, 483–506 (2011).
3. Bullmore, E. T. & Bassett, D. S. Brain Graphs: Graphical Models of the Human Brain Connectome. Annual Review of Clinical Psychology 7, 113–140 (2011).
4. World Health Organization. Epilepsy. World Health Organization Epilepsy Fact Sheet https://www.who.int/news-room/fact-sheets/detail/epilepsy (2023).
5. Caputi, L., Pidnebesna, A. & Hlinka, J. Promises and pitfalls of topological data analysis for brain connectivity analysis. NeuroImage 238, 118245 (2021).
6. Grinenko, O. et al. A fingerprint of the epileptogenic zone in human epilepsies. Brain 141, 117–131 (2018).
7. Merelli, E., Piangerelli, M., Rucco, M. & Toller, D. A topological approach for multivariate time series characterization: the epileptic brain. in Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS) (ACM, 2016). doi:10.4108/eai.3-12-2015.2262525.
8. Zhang, J. et al. Characterizing Brain Network Dynamics using Persistent Homology in Patients with Refractory Epilepsy. AMIA Annu Symp Proc 2021, 1244–1253 (2022).
9. Wang, Y., Ombao, H. & Chung, M. K. Topological Data Analysis of Single-Trial Electroencephalographic Signals. Ann Appl Stat 12, 1506–1534 (2018).
10. Piangerelli, M., Rucco, M., Tesei, L. & Merelli, E. Topolnogical classifier for detecting the emergence of epileptic seizures. BMC Res Notes 11, 392 (2018).
11. Edelsbrunner, H. & Harer, J. Computational Topology: An Introduction. (American Mathematical Society, 2009).
12. Pun, C. S., Lee, S. X. & Xia, K. Persistent-homology-based machine learning: a survey and a comparative study. Artif Intell Rev 55, 5169–5213 (2022).
13. Otter, N., Porter, M. A., Tillmann, U., Grindrod, P. & Harrington, H. A. A roadmap for the computation of persistent homology. EPJ Data Sci. 6, 1–38 (2017).
14. Sizemore, A. E., Phillips-Cremins, J. E., Ghrist, R. & Bassett, D. S. The importance of the whole: Topological data analysis for the network neuroscientist. Netw Neurosci 3, 656–673 (2019).

15. Jayapandian, C. et al. A scalable neuroinformatics data flow for electrophysiological signals using MapReduce. Frontiers in Neuroinformatics 9, (2015).

16. Gershon, A. et al. Computing Functional Brain Connectivity in Neurological Disorders: Efficient Processing and Retrieval of Electrophysiological Signal Data. AMIA Jt Summits Transl Sci Proc 2019, 107–116 (2019).

17. Sahoo, S. S. et al. NeuroIntegrative Connectivity (NIC) Informatics Tool for Brain Functional Connectivity Network Analysis in Cohort Studies. AMIA Annu Symp Proc 2020, 1090–1099 (2021).

18. Maria, C., Boissonnat, J.-D., Glisse, M. & Yvinec, M. GUDHI library. GUDHI library https://project.inria.fr/gudhi/software/ (2014).

19. Barnes, D., Polanco, L. & Perea, J. A. A Comparative Study of Machine Learning Methods for Persistence Diagrams. Front Artif Intell 4, 681174 (2021).

20. Giusti, C., Ghrist, R. & Bassett, D. S. Two's company, three (or more) is a simplex: Algebraic-topological tools for understanding higher-order structure in neural data. J Comput Neurosci 41, 1–14 (2016).

21. Pijn, J. P. & Lopes da Silva, F. Propagation of Electrical Activity: Nonlinear Associations and Time Delays between EEG Signals. in Basic Mechanisms of the EEG (eds. Zschocke, S. & Speckmann, E.-J.) 41–61 (Birkhäuser Boston, 1993). doi:10.1007/978-1-4612-0341-4_4.

22. James, G., Witten, D., Hastie, T., Tibshirani, R. & Taylor, J. An Introduction to Statistical Learning: with Applications in Python. (Springer International Publishing, 2023). doi:10.1007/978-3-031-38747-0.

23. Zou, K. H., O'Malley, A. J. & Mauri, L. Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models. Circulation 115, 654–657 (2007).

24. Bendich, P., Marron, J. S., Miller, E., Pieloch, A. & Skwerer, S. Persistent Homology Analysis of Brain Artery Trees. Ann Appl Stat 10, 198–218 (2016).

25. Bubenik, P. The Persistence Landscape and Some of Its Properties. 15, 97–117 (2020).

26. Chazal, F., Fasy, B. T., Lecci, F., Rinaldo, A. & Wasserman, L. Stochastic convergence of persistence landscapes and silhouettes. Journal of Computational Geometry 6, 140–161 (2015).

27. Adams, H. et al. Persistence Images: A Stable Vector Representation of Persistent Homology. Journal of Machine Learning Research 18, 1–35 (2017).

28. Rucco, M., Castiglione, F., Merelli, E. & Pettini, M. Characterisation of the idiotypic immune network through persistent entropy. in Springer Proceedings in Complexity (Springer, Cham, 2015). doi:https://doi.org/10.1007/978-3-319-29228-1_11.

29. Mitchell, T. M. Machine Learning. (McGraw-Hill, 1997).

30. Bartolomei, F. et al. Defining epileptogenic networks: Contribution of SEEG and signal analysis. Epilepsia 58, 1131–1147 (2017).

31. Wu, S. et al. Role of ictal baseline shifts and ictal high-frequency oscillations in stereo-electroencephalography analysis of mesial temporal lobe seizures. Epilepsia 55, 690–698 (2014).

32. Diehl, B. & Lüders, H. O. Temporal Lobe Epilepsy: When Are Invasive Recordings Needed? Epilepsia 41, S61–S74 (2000).