Topological Data Analysis for Scalp EEG Signal Processing

Jingyi Zheng
Department of Mathematics and Statistics
Auburn University
Auburn, AL, USA
jingyi.zheng@auburn.edu

Ziqin Feng
Department of Mathematics and Statistics
Auburn University
Auburn, AL, USA
zzf0006@auburn.edu

Yuexin Li
Department of Mathematics and Statistics
Auburn University
Auburn, AL, USA
yzl0233@auburn.edu

Fan Liang
Department of Computer Science
Sam Houston State University
Huntsville, TX, USA
fxl027@shsu.edu

Xuan Cao
Department of Mathematical Sciences
University of Cincinnati
Cincinnati, OH, USA
xuan.cao@uc.edu

Linqiang Ge
TSYS School of Computer Science
Columbus State University
Columbus, GA, USA
ge_linqiang@columbusstate.edu

2023 8th International Conference on Signal and Image Processing (ICSIP) | 979-8-3503-9793-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICSIP57908.2023.10270899

***Abstract*—Topological Data Analysis is a fast-growing and promising approach that recently gains popularity in the data science field. It utilizes topological and geometric measurements to describe the structure, for example the shape, of complex data, which is fundamental and important for modeling the data. Scalp Electroencephalography (EEG) is widely used in clinical trials and scientific research to measure the brain activities. However, analyzing and modeling scalp EEG signals is still an open field due to the complex and non-stationary nature of the EEG signal itself as well as the transformed signals. Therefore, in this paper, we propose a topological-based processing pipeline that utilizes persistent homology to capture the underlying system dynamic of the transformed EEG signals and further construct machine learning classifiers. A public available scalp EEG data is used to validate our algorithms, and the results show that the topological features successfully capture the subtle changes in the time-frequency representations revealed by Hilbert-Huang Transformation, with area under ROC curve reaching** 0.96**.**

***Index Terms*—Topological Data Analysis, Scalp EEG, Hilbert-Huang Transformation, Machine Learning**

I. Introduction

Scalp Electroencephalography (EEG) has been widely used to record the electrical activity of the brain from the electrodes placed on the scalp. Clinicians and researchers have been using scalp EEG to diagnose or treat epilepsy, sleep disorders, brain damage, etc. The analysis of scalp EEG are generally categorised in event-related and spectral-related study. The former usually investigates scalp EEG from temporal domain, for example detection of the onset of stimulus or epileptic seizure [1]. The latter studies the spectral content of EEG in the frequency domain and focuses on the neural oscillations (i.e., brain waves) recorded in EEG signals. The spectral-related scalp EEG study has been extensively used in neuroscience, cognitive science, and cognitive psychology research [2], [3], [4]. The brain waves are conventionally categorized as delta, theta, alpha, beta, and gamma waves based on their frequency ranges: delta (1-4 Hz), theta (4-8 Hz), alpha (8-12 Hz), beta (12-30 Hz), low gamma (30-60 Hz) [5], [6]. To study the spectral content of EEG signals, numerous research [3], [7], [8], [9], [10] have been conducted to extract representative information from brain waves. However, the data can be complex once transformed into frequency or time-frequency domain, which requires a powerful tool for analyzing such complex data.

Topological data analysis (TDA) is a powerful tool for analyzing complex data which has shown great promise in extracting the topological features of data. However, few work has been conducted using TDA to analyze scalp EEG signals. [11] calculates the area of a 1-dimensional Betti curve of EEG signal as TDA score. [12] utilizes Betti number from raw EEG signals as TDA features for kNN classifier. [13] extracts persistent entropy from single trial EEG to classify Autism Spectrum Disorder via support vector machine. [14] leverages persistence landscape of Fourier transformed EEG signals as TDA features. More information can be revealed by transforming EEG signals into time-frequency-representation. However, no study has been conducted to extract TDA features from the time-frequency-representation of EEG signals.

Therefore, we propose a topological-based processing pipeline to analyze scalp EEG signals. In the pipeline, we first leverage Hilbert-Huang Transform (HHT) to decompose the scalp EEG signal into a collection of sub-signals that correspond to different brain waves, and acquire their time-frequency information. Then, TDA features are extracted via utilizing persistent homology to the time-delay embedding that captures the underlying system dynamics of the time-frequency representation of brain waves. TDA score is then calculated as the area of a 1-dimensional Betti curve, one of the outputs from persistent homology, that represents irregularity of the time series. The TDA features are then served as input for the statistical and machine learning models to classify EEG signals. The processing pipeline is validated using a scalp EEG data, and results in promising performance.

The remainder of this paper is organized as follows: In Sec-

tion II, we introduce the scalp EEG data and the notation. In Section III, we present our approach in details. In Section IV, we summarize the evaluation results. Finally, we conclude the paper in Section V.

II. Scalp EEG data description

A. Scalp EEG Study

The experiment recruited 19 adults (7 females, 12 males) from the University of Arizona community[1]. Participants were asked to complete a spatial distance monitoring task to identify two possible outcomes, short vs long distance, while standing in an immersive virtual environment. Each trial lasts around 5.656 seconds, and each participants completed 48 trials in total with 24 short and 24 long distance. The scalp EEG data were recorded over 64 electrodes placed on the scalp, with a sampling rate of 500 Hz. All data were processed by 1-50 Hz bandpass filtering, artifact amelioration, and eye/muscle artifacts removal. The data is available at https://osf.io/3vxkn/, and the details of experimental design and data pre-processing is discussed in [15].

B. EEG Notation

For the illustration purpose, we take the EEG recording of one participant as an example. Denote the scalp EEG signal recorded from the j^{th} electrode channel during the k^{th} trial as $x_j^k(t)$. For the spatial task, the length of the signal is 2828 (around 5.656 seconds) and the label of $x_j^k(t)$ is binary (i.e., short or long distance). Our process unit is $x_j^k(t), j = 1, \ldots, 64, k = 1, \ldots, K$ in the following sections. The number of trials K is 48 for all subjects. In the following, each $x_j^k(t)$ for all j and k will go through the processing pipeline. Therefore, to avoid any further confusion caused by more subscripts, we will simply use $x(t)$ to denote $x_j^k(t)$ for illustration purpose.

III. Our Approach

In this section, we introduce the proposed topological-based processing pipeline in details. The pipeline is composed of three major steps: signal transformation via HHT to reveal time-frequency representation, topological feature extraction, and classification. A graphical summary of the pipelie is shown in Figure 1.

A. Hilbert-Huang Transformation (HHT)

Scalp EEG signal is nonlinear and non-stationary by nature, which introduces difficulties for EEG analysis. Fourier transform works for stationary signal and estimates a constant power for each frequency over time. However, the frequency and the corresponding power changes along with time. To better capture the time-varying frequency and power, Hilbert-Huang Transformation (HHT) is utilized. HHT, proposed by [16], is a data-driven approach for analyzing non-stationary signals. HHT is composed of two steps: Empirical Mode Decomposition (EMD) and Hilbert Transform (HT).

[1]Written informed consent was obtained in accordance with the Institutional Review Board at the University of Arizona.

EMD decomposes the signal $x(t)$ into a collection of sub-signals $\{c_i(t), i = 1, \ldots, n\}$ named Intrinsic Mode Functions (IMFs) via a sifting processing.

$$x(t) = \sum_{i=1}^{n} c_i(t) + r(t). \tag{1}$$

The Hilbert transform is then applied on each of the IMFs to reveal its instantaneous frequency $\omega_i(t)$ and instantaneous amplitude $a_i(t)$. By replacing $c_i(t)$ using its analytical expression, we can rewrite Equation (1) as:

$$x(t) = \sum_{i=1}^{n} Re\{a_i(t)exp(i \int \omega_i(t)dt)\} + r(t), \tag{2}$$

$$x(t) = Re\{\sum_{i=1}^{\infty} a_i e^{i\omega_i t}\}. \tag{3}$$

Different from Fourier transform which estimates constant power a_i for each frequency ω_i in Equation (3), HHT reveals the dynamic frequency and power, which is important especially for non-stationary signals. For scalp EEG signals, the dynamic frequency and power are used to reveal the subtle changes in the underlying dynamic structure. Besides, the IMFs are in descending frequency ranges with the first IMF carrying the highest frequency components. IMFs can also be used to represent different brain waves including delta, theta, alpha, and beta waves. In this study, we use the first four IMFs to represent the aforementioned four brain oscillations.

Though the instantaneous frequency $\omega_i(t)$ and amplitude $a_i(t)$ contain the important information about the dynamic structure of EEG signals, the dimension and amount of data is much larger. For one subject during one trial, the scalp EEG is 64×2828 with signal length being 2828 from each of the 64 electrode channels. After HHT, the dimension of the data becomes $64 \times 4 \times 2 \times 2828$ if we keep the first four IMFs and both instantaneous frequency and amplitude. Therefore, TDA is chosen to efficiently and effectively extract features from the time-frequency representation of EEG signals in the following.

B. Topological Data Analysis (TDA)

Both the instantaneous frequency $\{\omega_i(t), i = 1, \ldots, n\}$ and the instantaneous amplitude $\{a_i(t), i = 1, \ldots, n\}$ on each electronic channel can be treated as scalar time series. Based on Taken's embedding theorem, time delay embedding is a very useful way to reconstruct state space from single signal source. Here we use $\omega_i(t)$ as an example. The Taken's Embedding theorem ([17], [18]) asserts that when the states of a dynamical system lies on a low-dimensional manifold, the complete information about the states can be preserved in the time-series output and the states can be reconstructed through an embedding map, namely, a delay coordinate mapping. Hence, we embed $\omega_i(t)$ into an N-dimensional state space with Δt time delay. The coordinates of the corresponding points in the N-dimensional state space can be represented by

$$y_i(t) = (\omega_i(t), \omega_i(t + \Delta t), \ldots, \omega_i(t + (N-1)\Delta t)). \tag{4}$$

The parameters N and Δt determine the embedding. We set $N = 3$ and $\Delta t = 1$ in processing the instantaneous frequency and the instantaneous amplitude.

Homologies are used to capture the topological structures of simplicial complexes such as components, circles, and voids. Vietoris-Rips filtration is one of the methods to extract a filtration of simplicial complexes from a state space.This filtration starts with expanding each point in the state space to a disk with radius zero. The radii of the disks grow uniformly, and then the procedure ends when they reach a predetermined value. The predetermined value in our calculation is the one such that the resulting simplicial complex loses all the topological structures, i.e., homotopy equivalent to a singleton. For each radius, a graph is formed using the points and edges between any two points when the associated disks intersect with each other. The clique complexes (flag complexes) of such graphs yield a filtration of nested simplicial complexes, $K_0, K_1, \ldots, K_n$ with $K_i \subset K_j$ when $i < j$.

Let K be a simplicial complex. For each integer $p \geq 0$, the pth simplicial homology group $H_p(K)$ with integer coefficients is abelian and the rank of $H_p(K)$ gives the count of p-dimensional 'hole' in K, which is also called the pth Betti number $\beta_p(K)$. Topologically, a 1-dimensional 'hole' is a circle and a 2-dimensional 'hole' is a void. The rank of the 0th homology group $H_0(K)$ is the number of components in the simplicial complex K. The Betti numbers of the state spaces in our discussion are β_0, β_1, and β_2. In this paper, we choose the 1st Betti number β_1 of the corresponding state spaces to represent the information about whether a participant monitors short or long distances. Each inclusion map from a simlicial complex K_i to another K_j with $i < j$ induces a homomorphism ϕ from the homology group $H_p(K_i)$ to $H_p(L_j)$ for each $p \geq 0$.

The persistent homology of a filtration of simplicial complexes $\{K_i : i = 1, \ldots, n\}$ is the homology groups $\{H_p(K_i) : p \geq 0 \text{ and } i = 1, \ldots, n\}$ and the homomorphisms $\{\phi_p^{i,j} : p \geq 0 \text{ and } 1 \leq i < j \leq n\}$, where $\phi_p^{i,j}$ is the homomorphism from $H_p(K_i)$ to $H_p(K_j)$ induced by the inclusion map. As the radii of disks increase, some topological invariants (components, circles, or voids) persist longer in these simplicial complexes, while others disappear quickly. Hence, the persistent homology yields p-dimensional Betti interval, $[t_{\text{birth}}, t_{\text{death}})$, which defines the time at which a p-dimensional hole appears in the simplicial complex $K_{t_{\text{birth}}}$, while dies in the simplicial complex $K_{t_{\text{death}}}$.

A graphical representation of those intervals of a state space is called a persistence barcode and it is associated to the Vietoris-Rips filtration. A persistence barcode can also be represented by a persistence diagram. The simplicies involved in the p-dimensional holes, which are the generators of the p-dimensional homology group, can also be obtained in the computation of persistent homology.

We use B_p to represent the collection of p-dimensional Betti intervals of a state space with a Vietoris-Rips filtration. The element in B_p is in the interval form $[x, y)$. Let $B_p = \{[x_k, y_k) : k \in I\}$ where I is a finite index set. Then we calculate a p-dimensional TDA score of a state space

$$C_p = \sum_{k \in I} |y_k - x_k|. \tag{5}$$

The p-dimensional TDA score roughly measure the magnitude of the persistence of all p-dimensional 'hole'. To summarize, for each p, we obtain a p-dimensional TDA score of the state space from each instantaneous frequency or instantaneous amplitude.

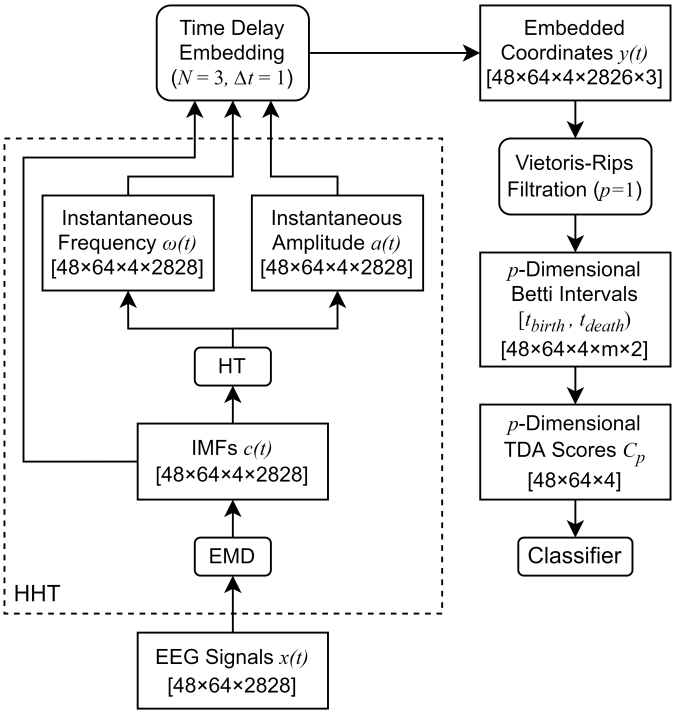

Fig. 1. A graphical summary of the proposed topological-based processing pipeline. The time delay embedding module takes one of the three time series (i.e., $c(t)$, $\omega(t)$, or $a(t)$) as input. m is the number of Betti intervals obtained by Vietoris-Rips filtration and is decided by the time series input.

C. Classification

Given a EEG signal $x(t)$, the pipeline starts with HHT in Section III-A. $x(t)$ is decomposed into a set of IMFs $\{c_i(t), i = 1, \ldots, 4\}$ through EMD. In our study, we choose the first 4 IMFs based on their Hilbert spectrum. Then the instantaneous frequency $\{\omega_i(t), i = 1, \ldots, 4\}$ and instantaneous amplitude $\{a_i(t), i = 1, \ldots, 4\}$ are obtained via HT. For each time series (i.e., $c_i(t)$, $\omega_i(t)$, or $a_i(t)$), TDA features are then extracted according to Section III-B. In this study, we set $p = 1$ and extract C_1 for each series. Taking $c_i(t)$ as an example, the TDA feature is created from embedding $c_i(t)$ into a state space and calculate the 1-dimensional TDA score C_1 using the 1-dimensional Betti intervals from the persistent homology of the state space. Similarly, the instantaneous frequency and instantaneous amplitude are also used as signal sources from which the state space is reconstructed, and the corresponding TDA features form the second and third feature set, respectively. Lastly, the second and third feature set are

combined to construct a fourth set of features. For each subject, the dimension of the scalp EEG data is $48 \times 64 \times 2828$. The entire TDA feature extraction pipeline, as well as the dimension of the input and output at each step, are presented in Figure 1, and the 4 extracted feature sets are summarized in Table I.

TABLE I
SUMMARY OF 4 FEATURE SETS FOR EACH SUBJECT

Feature Set	Signal Source for TDA	Dimension
1	$\{c_i(t)\|i = 1, \ldots, 4\}$	$48 \times 64 \times 4$
2	$\{\omega_i(t)\|i = 1, \ldots, 4\}$	$48 \times 64 \times 4$
3	$\{a_i(t)\|i = 1, \ldots, 4\}$	$48 \times 64 \times 4$
4	$\{(\omega_i(t), a_i(t))\|i = 1, \ldots, 4\}$	$48 \times 64 \times 8$

The extracted TDA features are then fed into a variety of machine learning classifiers including random forest (RF), support vector machine (SVM), extreme gradient boosting (XGBoost), and least absolute shrinkage and selection operator (LASSO) to evaluate the effectiveness of the proposed TDA features. For each of the 19 subjects, the number of features are much larger than the number of trials. To avoid overfitting, we conduct feature selection to select the most important features for each classifier. For RF and XGBoost, the importance of each feature is measured by the mean decrease in Gini coefficient, which is an indication of how much each feature contributes to the homogeneity of the nodes and leaves in a tree-based model. For SVM, the importance of each feature is measured by the area under receiver operating characteristic curve (AUC) value when only that feature is present in the model. For RF, SVM, and XGBoost, a full model is fitted with the entire feature set first, and the normalized importance score of each feature is obtained. Recursive feature elimination (RFE) is then performed, starting with the top 40 strongest features from each set and eliminating 5 weakest features at each step. The reduced model with the highest AUC value is chosen as the final model, and the features used to produce the final model is considered the optimal selection. For LASSO, since the L1 penalty already equips it with feature selection and regularization, no additional feature selection is performed. For each subject and each model, repeated leave-group-out cross-validation (LGOCV), which uses 75% trials as training set and the rest 25% as testing set, is performed as well as hyperparameter tuning to maximize model performance.

In addition, to demonstrate the competitiveness of the proposed approach against existing deep-learning-based feature extraction methods, we built a simple neural network with 2 dense layers with ReLU activation, 2 dropouts with rate 0.5 and 0.3, and a dense layer with sigmoid activation for binary classification at the end. For comparison, $c_i(t)$, $\omega_i(t)$, and $a_i(t)$ are used to train this network without passing through the TDA feature extraction process.

IV. RESULTS

To evaluate the model performance, we use area under receiver operating characteristic curve (AUC) as the metric.

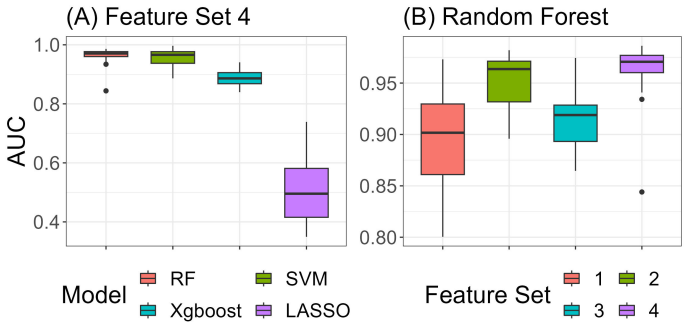

Fig. 2. AUC of Final Models among 19 Patients. (A) with feature set 4, the performance of the four Classifiers. (B) with RF, the performance of the four feature sets.

Figure 2 shows the model performance across all subjects. Panel (A) shows the comparison of classifiers using the same feature set while Panel (B) compares the four features using the same classifier. Combining four feature sets and four classifiers, 16 models are trained in total, and the average performance of each model over 19 subjects are summarized in Table II. The performance of the 4 neural networks trained with the corresponding time series without extracting TDA features are summarized in Table III. In general, the machine learning classifiers coupled with TDA features outperform the neural network, except for LASSO. Overall, the fourth feature set which contains the TDA features extracted from both instantaneous frequency and amplitude shows the best prediction performance, and RF is the best classifiers regardless of the feature sets.

TABLE II
AVERAGE AUC OF FINAL MODELS AMONG 19 PATIENTS

	Feature Set			
Model	1	2	3	4
RF	**0.8963**	**0.9499**	**0.9132**	**0.9612**
SVM	0.8663	0.9445	0.8778	**0.9540**
XGBoost	0.8440	0.8536	0.8615	**0.8869**
LASSO	0.5216	0.5126	**0.5221**	0.5011

TABLE III
AVERAGE AUC OF NEURAL NETWORK AMONG 19 PATIENTS

Feature	$c_i(t)$	$w_i(t)$	$a_i(t)$	$w_i(t), a_i(t)$
Average AUC	0.5385	0.5036	0.5122	0.5317

V. CONCLUSION

In this paper, we propose a topological-based processing pipeline for scalp EEG signal analysis. The pipeline is capable of analyzing both stationary and non-stationary signals, and can be applied to signals in various fields including agriculture, computer vision, and biomedical imaging. There are mainly three steps in the processing pipeline: HHT to reveal the time-frequency-representation of signals, TDA to extract

topological features from the Hilbert spectrum, and classifiers to classify signals. The effectiveness and competitiveness of the proposed pipeline is validated using real scalp EEG data, and compared with deep learning-based methods. Besides extracting the proposed TDA features from Hilbert spectrum, we are also exploring other TDA features coupled with wavelet transform.

REFERENCES

[1] S. M. Usman, S. Khalid, R. Akhtar, Z. Bortolotto, Z. Bashir, and H. Qiu, "Using scalp eeg and intracranial eeg signals for predicting epileptic seizures: Review of available methodologies," *Seizure*, vol. 71, pp. 258–269, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1059131119302213

[2] V. K. Harpale and V. K. Bairagi, "Time and frequency domain analysis of eeg signals for seizure detection: A review," in *2016 International Conference on Microelectronics, Computing and Communications (MicroCom)*, 2016, pp. 1–6.

[3] J. Zheng, M. Liang, S. Sinha, L. Ge, W. Yu, A. Ekstrom, and F. Hsieh, "Time-frequency analysis of scalp eeg with hilbert-huang transform and deep learning," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 4, pp. 1549–1559, 2022.

[4] G. Pfurtscheller and A. Aranibar, "Event-related cortical desynchronization detected by power measurements of scalp eeg," *Electroencephalography and Clinical Neurophysiology*, vol. 42, no. 6, pp. 817–826, 1977. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0013469477902358

[5] E. Niedermeyer and F. L. da Silva, *Electroencephalography: basic principles, clinical applications, and related fields*. Lippincott Williams & Wilkins, 2005.

[6] F. L. da Silva, "Eeg and meg: relevance to neuroscience," *Neuron*, vol. 80, no. 5, pp. 1112–1128, 2013.

[7] P. Garc and C. Pe, "Analysis of eeg signals using nonlinear dynamics and chaos: A review," 2015.

[8] Y. Yi, N. Billor, M. Liang, X. Cao, A. Ekstrom, and J. Zheng, "Classification of eeg signals: An interpretable approach using functional data analysis," *Journal of Neuroscience Methods*, vol. 376, p. 109609, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165027022001364

[9] J. Zheng, M. Liang, A. Ekstrom, L. Ge, W. Yu, and F. Hsieh, "On association study of scalp eeg data channels under different circumstances," in *Wireless Algorithms, Systems, and Applications*, S. Chellappan, W. Cheng, and W. Li, Eds. Cham: Springer International Publishing, 2018, pp. 683–695.

[10] M. Soleymani, S. Asghari-Esfeden, Y. Fu, and M. Pantic, "Analysis of eeg signals and facial expressions for continuous emotion detection," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 17–28, 2016.

[11] T. Yamanashi, M. Kajitani, and e. a. Masaaki Iwata, "Topological data analysis (tda) enhances bispectral eeg (bseeg) algorithm for detection of delirium," *Scientific Reports*, vol. 11, no. 304, 2021.

[12] F. Altindis, B. Yilmaz, S. Borisenok, and K. Icoz, "Use of topological data analysis in motor intention based brain-computer interfaces," in *2018 26th European Signal Processing Conference (EUSIPCO)*, 2018, pp. 1695–1699.

[13] S. Majumder, F. Apicella, F. Muratori, and K. Das, "Detecting autism spectrum disorder using topological data analysis," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 1210–1214.

[14] Y. Wang, H. Ombao, and M. K. Chung, "Statistical persistent homology of brain signals," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 1125–1129.

[15] M. Liang, J. Zheng, E. Isham, and A. Ekstrom, "Common and distinct roles of frontal midline theta and occipital alpha oscillations in coding temporal intervals and spatial distances," *bioRxiv*, p. 2020.08.05.237677, Jun 2021.

[16] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences*, vol. 454, no. 1971, pp. 903–995, 1998.

[17] R. D. Takens, F. and L.-S. Young, ""detecting strange attractors in turbulence" in dynamical systems and turbulence warwick 1980 sere," *Lecture Notes in Mathe-matics*, vol. 898.

[18] Y. J. A. Sauer, Tim and M. Casdagli, "Embedology," *Journal of Statistical Physics*, vol. 65.