A review of Topological Data Analysis and Machine Learning

Sura I. Mohammed Ali[1,*], Nidaa Hasan Haji[1], Hussien Sharaf[2]

[1]*Department of Mathematics and Computer Application, Collage of Science, Al-Muthanna University, Iraq,*

[1]*Department of Mathematics and Computer Application, Collage of Science, Al-Muthanna University, Iraq,*
[2]*Department of Computer Science, Faculty of Computers and Information, Suez University, Suez, Egypt*

[*]*Corresponding Author: suraibraheem@mu.edu.iq*

Received 22 Aug. 2024, Accepted 1 Oct. 2024, published 30 Dec. 2024.

DOI: 10.52113/2/11.02.2024/44-61

Abstract: During the past decade, it has been quite successful to observe how computational topology has incorporated some of the fundamental concepts and ideas from algebraic and differential topology into applications. This fusion of theories gave birth to a new field named Topological Data Analysis (TDA), which has a significant value in various fields, ranging from computational biology to personalized medicine and dynamic data analysis. Going beyond its foundational applications, TDA has enriched and complemented classical machine and deep learning frameworks in establishing what is now known as "topological machine learning." In this paper, we review the present landscape of this emerging field, emphasizing how it merges with machine learning algorithms such as deep neural networks. Each method confers special advantages, targeting areas like machine learning integration, network reconstruction, classification of network regimes, or reduction of noisy data. We described common methodologies, discussed current implementations, and anticipate future challenges in topological machine learning.

Keywords: TDA, persistent homology, machine learning, topology, topological machine learning.

1. Introduction

TDA is one of the most important tools in machine learning, particularly in Physics and real-world applications [1]. Topological data analysis excels in studying high-dimensional data and is robust against noise. It has been applied and proved success in a variety of fields, from physics phase transition detection to the classification of prediction. TDA methods have also been extended in their applicability to multivariate time series data. This is achieved through persistent homology and Wasserstein distances for mapping data into point clouds for analysis. Such a method has proven effective in tasks like room occupancy prediction and activity recognition. Moreover, TDA combined with machine learning already found an

application in the following algorithm of climate science: automatic recognition of atmospheric rivers in the climate data, hence informing extreme weather events and climate change scenarios [2]. Topological mathematics is that part of mathematics that deals with topological properties, those properties of space preserved under continuous transformations.

This, in combination with machine learning, forms very powerful tools and methodologies for the understanding of complex data structures. Topological mathematics gives an extremely different view on understanding data; it uncovers structures that, under traditional methods, might become invisible. By fusion of topological insight with machine learning, we will be able to construct models that are at the same time more robust, interpretable, and powerful in modelling complexity in real-world data. The interplay between Topology and Machine Learning thus establishes a new frontier of research and applications of importance to both data scientists and mathematicians. In[3] many examples in this domain include Persistent Homology in Image Analysis, whereby the topology of images is analyzed to help recognize shapes and patterns, which is useful during the classification and segmentation of images;

Time-series data's topological features—using TDA, cyclic patterns may be captured

in time series data that would aid in forecasting and anomaly detection; and Mapper Algorithm for Data Visualization [4], which allows for a simplified means of visualizing complex data sets and brings to the fore clusters and other structures not otherwise apparent within the raw data.

2. A Brief Overview of Topological Mathematical Methods

Topological and mathematical methods for the analysis of data, along with machine learning and a number of basic mathematical notions and formulae, are used to understand complex structures. Some of the main basic equations and notions that find broad application include:

1) **Persistent homology**: A technique to study the shape and structure of high-dimensional data; it is calculated by constructing a set of geometric shapes (such as simple complexes) and then studying their topological properties as the size changes. Equations involved in computing persistent homology as mentioned in [1]:

Chain Complexes and Homology: One common framework in which one can compute these homological invariants from topological spaces is chain complexes. This could be such invariants like homology groups, which represent dimensions of

cycles in data that cannot be continuously transformed into each other.

Equations related to the boundary maps : $\partial_k(\sigma) = \sum_i \sigma \circ F_i^k$ (1)

Where F_i^k in equation 1 represents the face maps of the simplex and ∂_k Are the boundary operators that connect k-simplices to (k−1) -simplices

Vietoris-Rips Complex

For a point cloud X, a Vietoris-Rips complex VRϵ(X) VR at scale ϵ\epsilonϵ includes:

- **Vertices**: All points in X.
- **Edges**: An edge between points xxx and y if d(x,y)≤ ϵ d(x, y) where d is a distance metric (e.g., Euclidean distance).
- **Higher-dimensional simplices**: A k-simplex is formed if an edge connects every pair of vertices in the simplex.

Betti Numbers:

The Betti number β_k at scale ϵ is the rank of H_K: $\beta_k(\epsilon)$= rank($H_k(VR_\epsilon(X))$). This counts the number of k-dimensional holes (like loops or voids) in VRϵ(X)

Homology Groups

The kth homology group H_k of complex K captures k-dimensional holes and is computed as:

$H_k(K) = \ker(\partial_k)/\mathrm{im}(\partial_k+1)$ (2)

Where ker (∂_k) in equation 2 is the kernel of ∂_k (cycles) and im(∂_k+1)\ is the image of ∂_k+1 (boundaries).

2) **Jaccard Similarity Coefficients**: Used to measure the similarity between data sets:

$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$ (3)

Where A and B in equation (3) two sets of elements.

3) **Fractal Dimension Techniques**: The fractal dimension D, D is given by

$D = \frac{\log N(\epsilon)}{\log(1/\epsilon)}$ (4)

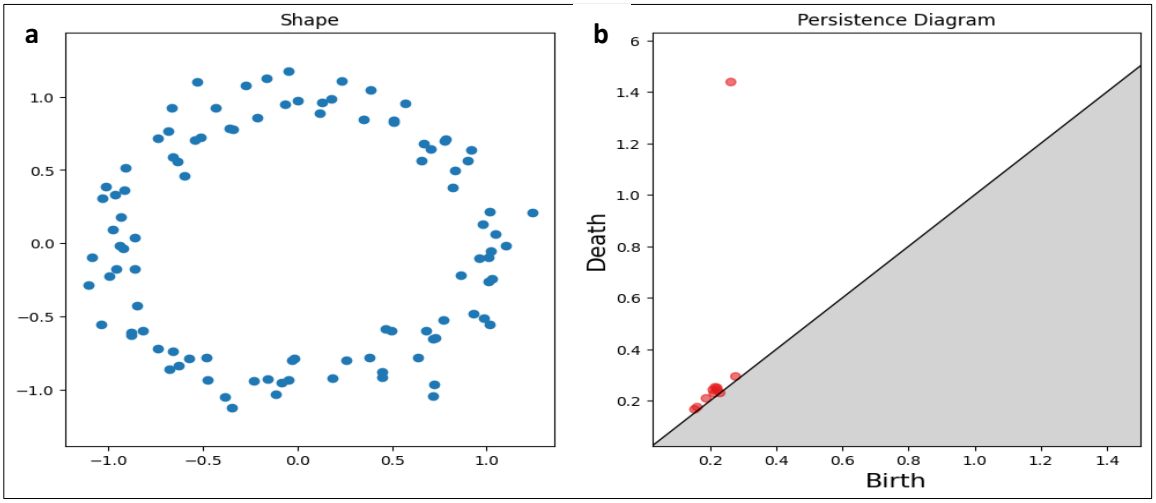

Fig (1): (a) A simple example of representing shape and (b) analyzing it using the persistent diagram

Where $1/\epsilon$ equation (4) is the scale, and N (ϵ) is the number of balls needed to cover the data. A persistent diagram is a topological tool used to analyze data and understand complicated structures and their evolution across several scales. This kind of analysis is mainly used within the persistent method, which is part of data topology. They all start from a set of data by making a simple polynomial complex, which will grow aggressively as the scaling parameter increases. In this structure, the creation and disappearance of holes are tracked as ϵ varies. The time (in terms of ϵ) at which each hole is born and dies—that is when it appears and disappearsin the structure that is recorded. These birth and death events are represented by points in a persistence diagram, which shows the horizontal axis of birth and the vertical axis of death. The points in the diagram are further analyzed to recognize the key topological features from the data. Points close to the line $y= x$ are considered unimportant or noisy features because they do not persist for long periods. Points that are far from the line $y=x$ indicate features that persist for a long time and are considered structurally significant in the data. Fig(1) shows a simple example of representing shape and analyzing it using the persistent diagram. It contains two parts: the first part shows the "shape" consisting of distributed points forming a rough circle, and the second part is the "diagram," which is used to analyze the topological properties of the shape.

The points form a circle-like shape. This can indicate a circular cluster in the data or represent a specific distribution of some variables around a certain center. This diagram shows the persistent persistence of topological features in the figure. Red Dots: Each dot represents a topological hole (like the holes in a doughnut or the segments of a circle in this case). The birth axis shows when the topological feature (hole) appeared in the figure. The death axis shows when the feature disappeared. The top red dot represents a more significant or persistent topological feature, appearing early and persisting for a long time before disappearing. Points within this area show topological features that have short persistence (appear and disappear rapidly) and are often considered to be the result of noise or small variations in the data.

The cyclic graph shows the presence of topological holes of different permanence in circular shapes and highlights the basic properties of the analyzed data structure. This technique is used to evaluate how topological features can contribute to a deeper and more useful understanding of the data.

3. Literature Review

3.1 Paper Selection Methodology

The papers represent state-of-the-art computational and mathematical tools applied in data analysis and network topology within several fields [2] considering issues of topological data analysis for physics, where graph theory is implemented to understand complex data structures.

Tis study [5]discusses the issues of reconstructing network topologies through matrix decomposition to pull out hidden structures in dynamical systems. This study [6] classifies the dynamics of neuronal networks and analyze these structures more in-depth by means of persistent homology. This literature [7]enhances the process development with graph neural networks to ensure data security in the analysis of plant topology data.

The proposal of [8] suggests a Bayesian approach to learning graphs from noisy data toward the simultaneous learning of the structure and noise variance. In[9]developed a topological machine-learning pipeline with a systematic selection of methods to transform Persistence Diagrams for applications in machine learning to improve accuracy in data classification Table 1. Shows the comparison of the methodologies of each study in detail:

Table 1: Comparison of the used methodologies and effectiveness of each study in detail.

Ref	Year	Method	Advantage	Disadvantage
Doddi and Salapaka[5]	2019	Matrix Decomposition Techniques	Effective in network topology reconstruction, precise with unique decomposition strategies for complex networks.	Relies heavily on the quality and quantity of data samples, which may not always be sufficient.
Bardin, Spreemann, and Hess[3]	2019	Persistent Homology	Classifies network regimes effectively, providing deep insights into network dynamics and structure.	The specificity of application to neuronal networks might not generalize across other types of networks.
S. Sardellitti, S. Barbarossa, and P. Di Lorenzo[14]	2019	Graph Topology Inference Based on a Laplacian matrix	Laplacian Pooling in Graph Neural Networks (GNNs) can facilitate learning sparse graph representations by focusing on important features while reducing less significant connections.	Lead to challenges in accurately modeling relationships in data, particularly in complex networks.
Mattioli et al.[10]	2019	Genetic Algorithms for DNN Topology Selection	Efficiently explores large parameter spaces to optimize DNN architectures with minimal human intervention.	The effectiveness of GAs can vary and might require substantial computational efforts to achieve optimal results.
N. L. Holanda and M. A. R. Griffith[13]	2020	Supervised learning algorithm to learn topological phases	Ability of machine learning to advance the research on exotic quantum materials with properties	Requirement for a sufficient amount of labeled training data, which can be challenging to obtain in experimental settings
Riihimäki et al.[11]	2020	TDA-based Classifier for	Handles multiple measurements	Underperforms in certain scenarios

		Repeated Measurements	effectively and is useful in biological and ecological data analysis.	compared to standard methods like SVM.
Sultana and Tamanna[8]	2021	Bayesian Framework and Minimum Mean Square Estimation	It provides a robust method for learning graph structures and reducing noise from noisy data.	It may require complex computational resources and a deep understanding of Bayesian methods.
C. Wu and C. A. Hargreaves[15]	2021	TopMix method represents an innovative approach to topological machine learning	It utilizes principles from topology to effectively handle both categorical and continuous data types, improving classification performance in complex datasets.	Limitations, particularly in handling mixed data types effectively.
B. Narayan and A. Narayan[16]	2021	non-Hermitian Su-Schrieffer-Heeger (SSH) model and a non-Hermitian nodal line semimetal	Improved accuracy, successfully learning to predict complex topological phases	Focuses on materials with nodal lines in their band structure, where the conduction and valence bands touch along a line in momentum space.
A. Kerr, G. Jose, C. Riggert, and K. Mullen[17]	2021	(TQPTs) are identified by computing the topological index as a function of system parameters	This approach helps to map out regions in parameter space where the properties of the quantum state change, indicating different phases	Process involves complex calculations and interpretations, making it a challenging task in the area of mathematical chemistry and quantum physics.
Conti, Moroni, and Pascali[9]	2022	Topological Machine Learning Pipeline	A systematic approach with a focus on selecting suitable filtrations and transformations for machine learning applications.	The success of the method hinges on the choice of appropriate filtrations and representations, which can be challenging to determine.
Singh et al.[12]	2022	Algebraic Topology-based	Effective in predicting hepatic	The method's performance and

		Machine Learning on MRI Data	decompensation in PSC patients using advanced topological data analysis of MRI features.	generalizability may depend heavily on the quality and consistency of the MRI data.
Leykam and Angelakis[2]	2023	Topological Data Analysis (TDA)	It provides a systematic approach to defining the shape of data and is useful for phase detection and machine learning integration.	The complexity of graph theory concepts may limit accessibility for those without a background in the field.
Jonas Oeing and Kevin Brandt[7]	2023	Graph Learning with Graph Neural Networks	Enhances process development by analyzing machine-readable plant topology data, implemented in a secure local environment.	Limited to the specific software and may not extend to other platforms or broader contexts without adaptation.

3.2 Recent Studies of Topological Data Analysis (TDA) and Machine Learning

TDA encompasses computational methods [2] dedicated to rigorously defining and analyzing the "shape" of complex, discrete data within high-dimensional settings. There is increasing interest in TDA from physicists, particularly those studying topological materials or incorporating machine learning into their work. This review seeks to explore the forefront applications of TDA in physics, offer an accessible explanation of its fundamental techniques, and highlight potential areas for further investigation. TDA employs concepts from graph theory to measure the intuitive geometries of discrete data sets, such as point clouds. The approach involves constructing a graph by linking sufficiently proximate points and then calculating the graph's topological invariants, like the number of k-dimensional holes, thereby simplifying the task of shape measurement to basic linear algebra operations on the graph's vertices and edges. This document provides a brief overview of TDA's core principles and its specific uses in physics, emphasizing phase detection and integration with machine learning, and proposes future research possibilities at the confluence of TDA and physics.

The research paper [5] focuses on

reconstructing the network topology of linear dynamical systems with latent nodes, allowing directed loops and bi-directed edges. Matrix decomposition techniques are used to extract the structure of the network from observed nodes involving sparse and low-rank matrices. It discusses conditions and methods for decomposing the Inverse of the Power Spectral Density Matrix (IPSDM) into sparse and low-rank components to identify the moral graph and Markov Blanket of hidden nodes. It explores the unique decomposition of skew-symmetric matrices into sparse and low-rank components, which is essential for topology learning. Addresses the reconstruction of exact network topology using IPSDM and provides concentration bounds on the error between estimated and true IPSDM from limited data samples. Techniques of Sparse Plus Low-rank Matrix Decomposition Matrix decompositions are used to decompose skew-symmetric matrices into sparse and low-rank components. Here, the IPSDM would be decomposed into sparse and low-rank matrices that help to uncover the network structure.

The process of decomposition helps in the identification of a moral graph associated with observed nodes and the Markov Blanket of hidden nodes. Conditions and methods are given for the unique decomposition of skew-symmetric matrices into sparse and low-rank skew-symmetric matrices crucial for topology learning. Significance of Unique Decomposition for Topology Learning in a scenario where, many a time, only some nodes are observable, unique decomposition is necessary to rebuild the exact network topology from observed nodes. It provides the moral graph corresponding to observed nodes and the Markov blanket of latent nodes and provides insight into how a network understands relationships. The decomposition process allows for the distinguishing of spurious links in the moral graph formed by observed nodes, hence improving the accuracy of the network structure by guaranteeing a unique decomposition, which in turn enables the exact recovery of a topology of the network.

The method adopted [6] in the paper is persistent homology, and it comes from algebraic topology. Persistent homology has been used in this work to study topological aspects of neuronal network activity. This contribution of applying persistent homology in this context lies within classification by network regimes using spike train distances. It provides further insight into the global properties of network

structure and dynamics. The full document addresses all specific parameters, configurations, measurements, and findings for topological exploration in an artificial neuronal network using algebraic topology. This contains network topology, connectivity patterns, neurone models, synapse models, plasticity, sources of input, measurements taken, and classification results and analysis based on persistent homology.

Furthermore, a section of acknowledgements and ancillary information regarding the neural network model of the study and its computational resources can be found in this document. The equations used in the document include equations related to the leaky integrate-and-fire neuron model, synaptic strength, synaptic delay, relative synaptic efficiency (g), and relative external rate. Specific equations for these parameters were not explicitly mentioned in the provided text. Still, they are likely part of the mathematical models used to simulate artificial neuronal networks and classify network regimes based on spike train distances. Additional details on the specific equations used for these parameters may be found in the original study referenced in the document.

The research paper [7] utilizes graph learning techniques, focusing on the use of graph neural networks (GNNs) to analyze machine-readable plant topology data. The model used in the Plant Engineer P&ID software is trained based on the provided data. Details of integrating the model into the software and its execution through a Docker container containing the pre-trained GNN model are explained. Access to P&ID data in Graph ML format is done through the Docker container, and the results of the AI-based consistency check are presented in a JSON file used as a communication file. The consistency check results are visualized in Plant Engineer, highlighting inconsistencies in components or mismatched links in red. The advantages of containerization, such as easy accessibility and model encapsulation, are emphasized. It is noted that the model can be easily replaced by swapping the container. The deliberate avoidance of a cloud solution is to ensure the data security of P&ID data through local processing. The paper underscores the importance of machine-readable plant topologies and demonstrates their potential for future process development. A comprehensive overview of designing machine-readable P&IDs is provided, outlining the key benefits of this

technology in enhancing engineering processes and development.

The study addresses [8] the problem of learning the structure of graphical models from noisy multivariate data. The researchers aim to represent the unique graph structure and estimate the noise variance simultaneously. Methodology: A Bayesian framework was used to formulate the problem, and a minimum mean square estimation approach was proposed for data demising. Contributions: 1. A Bayesian approach was introduced to learn graph structure and noise removal from the data. 2. A minimum mean square estimation approach was proposed for data demising. 3. Mathematical equations were developed to estimate the noise variance and improve data quality. This paper used equations to illustrate the relationship between noise estimates and variance and mathematical calculations for estimating the required values. These points represent some of the information present in the study, highlighting the methodology used, key contributions, and important mathematical equations employed in the research.

This paper [9] introduces a topological machine learning pipeline, examining the theoretical underpinnings that guide the selection of specific methodologies, aiming to systematize the use of topological data analysis in classification tasks. The proposed pipeline selects suitable filtrations to link Persistence Diagrams (PDs) with digital data, employs various methods to convert these PDs into vector formats amenable to machine learning applications, and assesses the pipeline's effectiveness by comparing these methods across standard datasets. The objective is to establish a pipeline that connects digital data with PDs through appropriate filtrations and converts these PDs into forms usable by machine learning algorithms. The primary challenges involved representing digital data as an algebraic structure with the correct filtration to generate its topological summary, the Persistence Diagram, and converting the PD into a format that can be integrated into machine learning algorithms. The paper aims to create a straightforward, ready-to-use pipeline for data classification leveraging persistent homology and machine learning while also exploring why certain combinations of filtration and topological representation might be more effective for specific datasets and tasks. It provides an in-depth discussion of the mathematical foundations of algebraic topology and

persistent homology, which are vital to understanding topological data analysis.

This paper[11] discusses a classification approach using (TDA) designed to manage multiple measurements. The research introduces a classifier developed on the principles of TDA for analyzing repeated measurement data by sampling from the data space and forming a network graph reflective of the data's topological structure. This classifier was evaluated through three case studies involving tree species classification, random point processes, and neuron spiking data, demonstrating superior performance in most instances compared to a conventional support vector machine (SVM) voting model. Additionally, the TDA classifier offers extra advantages, like identifying data subsets of high purity and important feature values. TDA employs geometric and algebraic topology techniques to analyze complex data relationships at a large scale. Standard TDA tools typically handle only single measurements per sample, but biological data often requires analysis of multiple measurements per sample. The study proposes a TDA-based algorithm capable of processing such repeated measurement data, inspired by the Mapper algorithm and integrated into a classifier using the Mapper graph. This algorithm incorporates internal cross-validation and multiple bootstraps to enhance partition robustness and reduce overfitting risks. It generates subgroupings of relevant classes for additional analysis. While the TDA classifier outperformed the SVM model in tree species and random point process analyses, it performed comparably to SVM in one neuron spiking dataset scenario and less effectively in another. The paper concludes that the algorithm and its software could significantly aid biological research involving repeated measurement data, serving dual roles as an effective classifier and a tool for feature selection.

In [10] explores the application of genetic algorithms (GA) to determine the configurations of deep neural networks (DNN). Designing a DNN topology appropriate for specific problems is essentially an optimization challenge, given the extensive array of parameters involved, such as the number of layers, the number of neurons per layer, activation functions, and learning algorithms. Due to the immense size of the parameter space, a comprehensive search is unfeasible, which necessitates robust optimization techniques. Crafting an effective DNN topology requires significant domain knowledge and

experience. Traditional methods like random search, grid search, and transfer learning fall short in identifying the best topologies. The paper advocates for the development of techniques that facilitate the creation of new DNN topologies with minimal human input and limited computational demands. The authors suggest employing genetic algorithms, a type of evolutionary computation algorithm that efficiently navigates through promising areas of the search space and avoids local optima, thus potentially lowering the computational effort compared to a full search. They assess the effectiveness of using GA for DNN topology selection, employing a fitness function that measures the performance of a DNN topology. This GA-based strategy aims to find an optimal or nearly optimal topology that addresses the problem efficiently with reasonable computational resources.

The paper [12] considers the application of algebraic topology-based machine learning on magnetic resonance imaging (MRI) data in estimating hepatic decompensation risk among subjects with primary sclerosing cholangitis. (PSC)is one of the challenging chronic cholestatic liver diseases that can lead to cirrhosis and hepatic decompensation. It has been very difficult to predict the outcomes of PSC patients. The challenge of the paper is to predict future outcomes, particularly the development of hepatic decompensation in patients with primary sclerosing cholangitis. In this work, inspired by the topological data analysis, nonlinear methods are used to extract features from MRI in order to predict the 1-year risk of hepatic decompensation. The training was based on the one-center derivation and then tested in another independent multi-center validation cohort. An area under the receiver-operating characteristic curve of 0.84 was recorded after applying the model to an unbiased validation cohort. The authors developed a machine learning approach, which they then independently validated using algebraic topology analysis of magnetic resonance imaging data in predicting the risk of early hepatic decompensations in patients with PSC. The challenge in this regard is that the traditional approaches, which include qualitative MRI/MRCP prognostic scoring systems and quantitative MRCP metrics, are limited by their performance, reproducibility, and generalizability. The authors hypothesized that an algebraic topology-based machine learning approach could extract more informative features

from MRI data to predict clinically relevant outcomes in PSC patients better.

This paper [13] presents a supervised machine-learning algorithm capable of learning topological phases from the real lattice data for finite condensed matter systems. The algorithm uses diagonalization in real space along with any supervised learning algorithm to learn topological phases through an eigenvector ensemble procedure. The authors combine their algorithm with decision trees and random forests to successfully recover the topological phase diagrams of Su-Schrieffer-Heeger (SSH) models from real-space lattice data. They show how the Shannon information entropy of ensembles of lattice eigenvectors can be used to retrieve a signal detailing how the topological information is distributed in bulk. They also explore the theoretical possibility of interpreting these topological information entropy signatures in terms of emergent information entropy wave functions, which leads them to Heisenberg and Hirschman uncertainty relations for topological phase transitions. This illustrates how the model explains the ability of machine learning to advance the research on exotic quantum materials with properties that may power future technological applications such as qubit engineering for quantum computing.

The paper [14] introduces a technique for deducing the topology of a graph from a provided dataset. The objective is to establish a block-sparse representation of the graph signal. A modular graph structure characterized by a Laplacian matrix, whose eigenvectors are the identified sparsifying dictionary, is the result. The approach involves two optimization phases: i) deriving an orthonormal sparsifying transform Word Wide Web (WWW directly from the dataset. ii) Determining the Laplacian matrix L using the previously learnt transform WWW. Initially, the transform WWW is developed by addressing an optimization problem that aims to reduce the reconstruction error while ensuring WWW remains orthonormal. Subsequently, the Laplacian matrix L is derived through a convex optimization problem designed to minimize the discrepancy between the transform WWW and the eigenvectors U of the Laplacian, with the condition that L conforms to the properties of a valid Laplacian matrix (symmetric, positive semi-definite with zero row sums). This method has been validated using both synthetic and actual brain data to infer brain functionality networks in epilepsy patients. The findings affirm the

method's capability to reconstruct the underlying graph topology accurately. The primary contribution of this paper is its dual-step framework that concurrently learns a sparse graph representation and the corresponding Laplacian matrix from the data.

The newly [15] introduced TopMix method represents an innovative approach to topological machine learning, specifically designed for processing mixed numeric and categorical data. It incorporates TDA concepts such as persistent homology, persistence diagrams, and Wasserstein distance. Applied to a heart disease prediction model, TopMix surpasses competing advanced algorithms in effectiveness. TDA, a modern discipline, excels in tackling high-dimensional and noisy datasets that include both numeric and categorical elements, presenting challenges for conventional machine learning techniques. Currently, there is a standardized framework for using TDA to classify such mixed data. A symmetry-breaking technique is implemented to adapt TDA to diverse feature types. Data points are transformed into point clouds using multiple projection maps, and from these clouds, persistence diagrams are created. The Wasserstein distance measures the disparity between these diagrams. Classification is then performed using the k-nearest neighbors (k-NN) algorithm. Within TDA, persistent homology serves as a key tool to analyze the "shape" of data and extract essential traits. A persistence diagram visually maps out a dataset’s topological attributes created through persistent homology. Wasserstein distance is a metric for comparing these diagrams, quantifying the differences in topological features across datasets. Symmetry breaking is employed to adapt TDA to handle varied features effectively, where each coordinate denotes a distinct attribute. One-hot encoding is utilized to transform categorical data into binary form, with each category represented by a separate column. The TopMix method, tested on a dataset featuring both numeric and categorical variables related to heart disease, aims to predict the presence of significant cardiac conditions (> 50% luminal narrowing in any major pericardial vessel). Experimental outcomes demonstrate that TopMix outperforms several leading algorithms in diagnosing heart disease.

The paper [16] explores the application of machine learning techniques to identify and predict non-Hermitian topological phases of matter, which exhibit unique properties not

found in Hermitian systems. The study focuses on models such as the non-Hermitian Su-Schrieffer-Heeger (SSH) model and a non-Hermitian nodal line semimetal, demonstrating the use of neural networks to classify these phases accurately based on their winding number. The model demonstrates robustness even with the introduction of disorder, maintaining high accuracy, which suggests potential for experimental data analysis where noise is inevitable. For a three-dimensional case, a convolutional neural network (CNN) was employed due to the fully connected network's inadequacy in handling higher-dimensional data. The CNN significantly improved accuracy, successfully learning to predict complex topological phases and transitions with about 99.95% accuracy.

The paper [17] discusses a method for identifying topological phase transitions using machine learning, specifically through diffusion maps. Topological phase transitions are different from conventional phase transitions as they are not characterized by local order parameters but by global topological indices. The authors introduce a heuristic that automates the process of selecting hyperparameters for diffusion maps, allowing for the unsupervised identification of topological phase boundaries without prior knowledge of the underlying topological invariant. Topological quantum phase transitions (TQPTs) differ from classical phase transitions as they are indicated by a change in a topological invariant, not by spontaneous symmetry breaking and local order parameters. TQPTs are identified by computing the topological index as a function of system parameters and mapping regions with different indices. The authors propose a heuristic that automatically determines the optimal hyperparameters for diffusion maps, eliminating the need for user intervention. This heuristic optimizes the resolution parameter by minimizing the mean squared error (MSE) between the similarity matrix derived from the data and an ideal similarity matrix.

4. Conclusions

This study reviews several state-of-the-art computational techniques and their integration into various scientific and technological fields during the period 2019 to 2023. These include Topological Data Analysis, Matrix Decomposition Techniques, Persistent Homology, Graph Learning with Graph Neural Networks, Bayesian Framework and Minimum Mean Square Estimation, and a Topological Machine Learning Pipeline. Each of these

methods confers special advantages, targeting areas like machine learning integration, network reconstruction, classification of network regimes, or reduction of noisy data. However, these techniques also, , come along with several challenges and limitations such as: difficulty in TDA concepts, quality, and quantity of data required for matrix decomposition, specificity of applications by persistent homology, limitations in software for graph learning, complexity of Bayesian methods, and dependence upon suitable representations for topological machine learning pipelines.

References

[1] Hensel, F., Moor, M., Rieck, B., 2021, A Survey of Topological Machine Learning Methods, Front. Artif. Intell., 4. doi: 10.3389/frai.2021.681108.

[2] Leykam, D., Angelakis, D. G., 2023, Topological data analysis and machine learning, Adv. Phys. X, 8(1), 1–15. doi: 10.1080/23746149.2023.2202331.

[3] Latha, B. A., Jagan, S., Ajitha, G., Radhakrishna, D., Hemavathi, S., 2022, Topological Machine Learning Data Analysis for the Extraction of Robust Geometric Information, Data Eng. Intell. Comput. Proc. 5th ICICC 2021, Vol. 1. Singapore Springer Nat. Singapore, 1, 167–177, 2022, doi: 10.1007/978-981-19-1559-8.

[4] Oulhaj, Z., Antipolis, S., Michel, B., 2024, D IFFERENTIABLE M APPER FOR T OPOLOGICAL O PTIMIZATION arXiv Preprint. arXiv2402.12854 (2024).

[5] Doddi, M. S. V., Salapaka, M. V., 2021, Topology Identification with Latent Nodes using Matrix Decomposition, IEEE Trans. Autom. Control 67.11, 5746–5761, [Online]. Available at :http://arxiv.org/abs/1912.07152.

[6] Bardin, J. B., Spreemann, G., Hess, K., 2019, Topological exploration of artificial neuronal network dynamics, Netw. Neurosci., 3 (3), 725–743. doi: 10.1162/netn_a_00080.

[7] Oeing, J., Brandt, K., Wiedau, M., Tolksdorf, G., Welscher, W., Kockmann, N., 2023, Graph Learning in Machine-Readable Plant Topology Data, Chemie Ing. Tech. , 95 (7), 1049-1060.

[8] Sultana, N., Tamanna, M., 2021, Discover internet of things., Springer Nat. OA Free Journals, 1(1), 1–2.

[9] Conti, F., Moroni, D., Pascali, M. A., 2022, A Topological Machine Learning Pipeline for Classification, Mathematics, 10(17), 1–33. doi: 10.3390/math10173086.

[10] Mattioli, F., Caetano, D., Cardoso, A., Naves, E., Lamounier, E., 2019, An experiment on the use of genetic algorithms for topology selection in deep learning, J. Electr. Comput. Eng., 2019. doi: 10.1155/2019/3217542.

[11] Riihimäki, H., Chachólski, W., Theorell, J., Hillert, J., Ramanujam, R., 2020, A topological data analysis based classification method for multiple measurements, BMC Bioinformatics, 21, (1), 1–18, doi: 10.1186/s12859-020-03659-3.

[12] Singh, Y., 2022, Algebraic topology-based machine learning using MRI predicts outcomes in primary sclerosing cholangitis, Eur. Radiol. Exp., 6 (1). doi: 10.1186/s41747-022-00312-x.

[13] Holanda, N. L., Griffith, M. A. R., 2020, Machine learning topological phases in real space, Phys. Rev. B, 102 (5), 1–13. doi: 10.1103/PhysRevB.102.054107.

[14] Sardellitti, S., Barbarossa, S., Di Lorenzo, P., 2019, Graph Topology Inference Based on Sparsifying Transform Learning, IEEE Trans. Signal Process., 67 (7), 1712–1727, doi: 10.1109/TSP.2019.2896229.

[15] Wu, C., Hargreaves, C. A., 2021, Topological machine learning for mixed numeric and categorical data, Int. J. Artif. Intell. Tools, 30, (5). doi: 10.1142/S0218213021500251.

[16] Narayan, B., Narayan, A., 2021, Machine learning non-Hermitian topological phases, Phys. Rev. B, vol. 103 (3), 1–6, doi: 10.1103/PhysRevB.103.035413.

[17] Kerr, A., Jose, G., Riggert, C., Mullen, K., Automatic learning of topological phase boundaries, *Phys. Rev. E*, 103 (2), 1–8, 2021, doi: 10.1103/PhysRevE.103.023310.