Tlife-GDN: Detecting and Forecasting Spatio-Temporal Anomalies via Persistent Homology and Geometric Deep Learning

Zhiwei Zhen[1], Yuzhou Chen[2,3], Ignacio Segovia-Dominguez[1,4], and Yulia R. Gel[1,5]

[1] The University of Texas at Dallas, Richardson, TX 75080, USA
{Zhiwei.Zhen, Ignacio.SegoviaDominguez, ygl}@utdallas.edu
[2] Princeton University, Princeton, NJ 08544, USA
yc0774@princeton.edu
[3] Lawrence Berkeley National Laboratory, Berkeley, CA 94720
[4] Jet Propulsion Laboratory, Caltech, Pasadena, CA 91109, USA
[5] National Science Foundation, Alexandria, VA 22314

Abstract. Most recently, the tools of geometric deep learning (GDL) and, in particular, graph neural networks emerge as a promising new alternative in unsupervised anomaly detection problems where the data exhibit a sophisticated nonlinear dependence structure such as various geospatial surveillance systems. However, prevailing GDL-based methods for anomaly detection tend to exhibit limited capabilities to capture multiscale spatio-temporal variability which is ubiquitous in many applications, particularly, related to biosurveillance and biothreats. Motivated by the problem of assessing COVID-19 severity, we develop a novel approach to unsupervised anomaly detection in spatio-temporal data by fusing the notion of GDL with the emerging direction of persistent homologies and topological data analysis. In particular, our key idea is to bolster the GDL performance by leveraging the complementary insight on the intrinsic multiscale data organization which topological descriptors can provide. We also go one step further and show how our ideas at the interface of topological and geometric deep learning can be used not only for detection but for prediction of future anomalies. We show the utility of the new approach to detecting, forecasting and interpreting risks in COVID-19 clinical severity, measured in terms of hospitalization rates, in three U.S. states: California, Texas, and Pennsylvania.

Keywords: Anomaly Detection · Geometric Deep Learning · Persistent Homology · COVID-19

1 Introduction

Efficient identification of data instances which differ noticeably from the expected baselines is the core behind such diverse tasks as combating money laundering on blockchain, river water-quality monitoring, and defending information systems against breaches of cybersecurity. With a long history in robust statistics and continually emerging new types of threats, anomaly detection remains

one of the most actively developing fields at the nexus of machine learning and statistical sciences. Efficient detection of anomalies in dynamic settings such as biological and cyber threats is further exacerbated, first, by the limited or even non-existing records of labeled attack examples and, second, by a sophisticated dependence structure among entities of the underlying time-evolving object. For instance, transmission of many pathogens exhibit complex spatio-temporal interactions with atmospheric conditions, and moreover, pathogenicity of biothreats mat vary across spatial and temporal scales [28, 29, 32, 41].

To address the first challenge, anomaly detection is often viewed as an unsupervised problem. Among some most widely used unsupervised tools for anomaly detection are Connectivity-based Outlier Factor (COF) [42] and Influenced Outlierness (INFLO) [23]. However, such approaches tend to focus on linear relationships among system entities, and as a result, show limited ability to account for early warning signals induced by nonlinear interactions exhibited by most complex real-world systems. Various deep learning (DL) tools such as variational autoencoders (VAE) [25], Long short-term memory (LSTM) [31], and Generative adversarial Networks (GANs) [26] partially mitigate this problem and are found to be promising approaches for anomaly detection in high-dimensional settings.

However, such DL methods are restricted in their ability to learn multiple types of interactions among system entities in dynamic settings, e.g., georeferencing. As such, in the last couple of years, there has been a spike of interest in bringing the tools of Graph Neural Networks (GNNs) [10] and other methods of geometric deep learning (GDL) to anomaly detection tasks [18]. Indeed, GDL offers a systematic framework for learning non-Euclidean objects such as graphs and manifolds, and hence, GDL allows us for more flexible modeling of complex interactions among entities in a broad range of complex data structures, including multivariate time series and dynamic networks.

Our goal here is to further enhance this emerging GDL direction in anomaly detection and to bolster its performance by leveraging the power of data topological (or shape) descriptors. By topological descriptors, we broadly understand data characteristics that are preserved under continuous transformations such as bending, twisting, and stretching. In turn, a few most recent studies show that integration of topological summaries of time-evolving structures such as spatio-temporal processes into DL, either in a form of a topological layer or as additional data attributes, can noticeably improve forecasting performance [12, 13, 47]. This phenomenon can be explained by the complementary information on the underlying intrinsic system organization at multiple scales which topological descriptors (or more precisely, tools of persistent homology) can deliver. Motivated by biothreat applications where variation of pathogenicity is ubiquitous across spatio-temporal scales, we believe that integration of topological summaries into GNNs may enhance not only anomaly detection performance but bring an invaluable insight about various hidden mechanisms behind anomaly formation. To investigate this hypothesis, we consider anomaly detection in COVID-19 clinical severity, measured in terms of hospitalization rates, in three U.S. states: California, Texas, and Pennsylvania, Moreover, we make a step forward in not only *detecting the existing anomalies* but *forecasting the future anomalies*. While as-

sessing future anomalous patterns is the core behind proactive risk mitigation, especially, in healthcare analytics such as during COVID-19 pandemic, to the best of our knowledge, neither GDL nor any other DL tools have ever been used for spatio-temporal forecasting of anomalies.

The key novelty and contributions of this paper are summarized as follows:

- We are the first to integrate topological descriptors within GDL for anomaly detection tasks. Our Tlife-GDN model with a fully trainable topological layer within GNN shows competitive performance against existing state-of-the-art approaches and allows improving tractability of the latent mechanisms behind emergence of anomalies.
- This is the first paper to address the problem of *future* anomaly forecasting with GDL, which is the key behind developing proactive risk mitigation strategies.
- This is the first approach to assess evolution of existing and future spatio-temporal anomalies in COVID-19 clinical severity, measured in terms of hospitalization rates.

2 Related work

Anomaly Detection in Time-Evolving Processes Traditional tools for this task include Principal Component Analysis [39] and K Nearest Neighbors (KNN) [5]. Most recently, there has been suggested a number of approaches that leverage topological descriptors for anomaly detection within statistical algorithms. For instance, [22] proposes to detect change points in topological summaries of the observed data instead of analyzing the observed data directly, as in prevailing tools. In turn, [43] considers topological summaries as a supplement to observed data as the input for arrhythmia detection. Finally, [27] and [33] propose anomaly detection in Ethereum blockchain graphs based on assessing similarity among the topological summaries of the data at adjacent time snapshots.

Most recently, DL tools emerge as powerful alternatives to address anomaly detection in spatio-temporal processes. Among such notable DL approaches are Autoencoders (AE) of [2] based on the idea of reconstruction errors; Deep Autoencoding Gaussian Model (DAGMM) of [48] which expands AE with the Gaussian Mixture Model, and Variational Autoencoders (VAE) of [25] with regularized encoding's distribution. Furthermore, inspired by the Support Vector Data Description (SVDD) [35], [34] proposes a Deep Support Vector Data Description (DEEP-SVDD) for anomaly detection tasks which is capable of learning the nodes' representation and hypersphere center of the data simultaneously.

Finally, in the last couple of years, there has been a spike of interest in bringing the power of GNNs to anomaly detection tasks on spatio-temporal data [30]. For instance, most recently [14] proposes a Graph Neural Network-Based Anomaly Detection tool based on the approach of graph attention mechanism with location embedding and structure learning. Although all those methods intend to discover the hidden relationships between system entities, to our best

knowledge, there exists *no GNN which have explored the power of topological data descriptors for enhancing anomaly detection* in time-evolving processes.

Different from the anomaly detection task, anomaly prediction is the task of recognizing *future* abnormal instances relative to the currently recorded data patterns. The problem of anomaly prediction is noticeably more challenging due to elevated uncertainty of forecasting and, while playing a key role in efficient and proactive management of emergency preparedness, remains largely understudied. Previous works in this filed include applications of machine learning tools like Support Vector Machines (SVMs) [45] and epsilon-Support Vector Regression (ϵ-SVR) [6] in software programs [4], water pipeline [46]. However, to the best of our knowledge, neither the utility of GNNs nor DL tools, in general, has been explored before for anomaly prediction in conjunction with analysis of time-evolving processes.

COVID-19 Severity Prediction Many recent studies have analyzed the risk factors for the severe acute respiratory syndrome coronavirus 2 (i.e., SARS-CoV-2, the virus which causes COVID-19). For example, [8] examines the possible correlation between obesity and COVID-19 clinical severity by surveying patients in a hospital, while [7] considers the linkage between anticancer therapy and COVID-19. More generally, [16] reviews the factors in demographics, comorbidities, hypoxia and radiographic features that might worse COVID-19 outcomes. However, the majority of the COVID-19 severity research focuses on the patients' clinical features rather than on the severity in a certain geographical area.

Two notable studies on spatio-temporal anomaly detection in conjunction with COVID-19 are [19] and [24] who consider topological data analysis (TDA) and the deep hybrid autoencoder networks for assessing daily new cases, respectively. Furthermore, [36–38] consider various GDL and LSTM models, coupled with topological descriptors for tracking COVID-19 hospitalizations and number of cases, but do not address the problem of spatio-temporal anomaly detection in COVID-19 clinical severity. As such, spatio-temporal anomalies in COVID-19 clinical severity and, particularly, anomalies in hospitalization rates remain largely under-explored. To the best of our knowledge, there exists *no current method assessing risk scoring in COVID-19 clinical severity using GNNs or TDA based on hospitalization data.* Our paper aims to take advantage of GNNs with topological descriptors to improve the performance and tractability of the unsupervised spatio-temporal anomaly detection and anomaly prediction for COVID-19 hospitalization rates.

3 Preliminaries on Persistent Homology

Persistent homology (PH) is a methodology under the framework of topological data analysis, which aims to study the most inherent shape characteristics of the observed data. The PH machinery is applicable to a broad range of data types, e.g., point clouds in Euclidean spaces, images, graphs, and more generally, objects in metric spaces. Here, we primarily focus on shape characteristics of the graph $\mathcal{G}$ generated from spatio-temporal time series[6] [9, 11]. The approach

[6] Generation details are available in Algorithm 1

consists of the three main steps. First, we convert $\mathcal{G}$ into a filtration of graphs $\mathcal{G}_1 \subseteq \mathcal{G}_2 \subseteq \ldots \subseteq \mathcal{G}_k = \mathcal{G}$. We can now track evolution of various patterns in this graph filtration, which ought to reveal the underlying structure of $\mathcal{G}$ at different scales. Second, to make the tracking process systematic and efficient, we build a simplicial complex $\mathscr{C}$ on top of $\mathcal{G}$ and, as such, our graph filtration is now associated with a nested sequence of complexes $\mathscr{C}(\mathcal{G}_1) \subseteq \mathscr{C}(\mathcal{G}_2) \subseteq \ldots \subseteq \mathscr{C}(\mathcal{G}_n)$. That is, we can now compute simplicial homologies and record which shape characteristics, for example, connected components, loops, and cavities, appear in the filtration of complexes. In particular, we say that a topological feature is born at i_b if $\mathscr{C}(\mathcal{G}_{i_b})$ is the complex where we first observe it. In turn, we record death of a topological feature at j_d if this feature is last seen in $\mathscr{C}(\mathcal{G}_{j_d})$. The longer the lifespan $j_d - i_b$ of the topological feature is, the likelier this feature contains important structural information on $\mathcal{G}$. Features with longer lifespans are also said to persist, while features with shorter lifespans are sometimes referred to as topological noise. Finally, in our third step, we summarize all the extracted topological features in a form of a multi-set $\mathcal{D} = \{(i_b, j_d) \in \mathbb{R}^2 | i_b < j_d\}$, called *persistence diagram* (PD). Since lifespan $j_d - i_b \geq 0$, all points in $\mathcal{D}$ are in the half-space on or above $y = x$. Finally, there exists multiple options to construct graph filtrations [20]). For instance, consider a continuous function $f : \mathcal{V} \to \mathbb{R}$ acting on nodes of $\mathcal{G}$ and a sequence of non-negative scales $\xi_1 < \xi_2 < \ldots < \xi_n$. Then, we can define the corresponding simplicial complex as $\mathscr{C}_i = \{\sigma \in \mathscr{C} : \max_{v \in \sigma} f(v) \leq \xi_i\}$. Similarly, filtration can be defined as $\mathcal{E}$ of $\mathcal{G}$. In this paper, we consider the weight rank clique filtration [40] and Vietoris-Rips abstract simplicial complexes [15], due to their computational benefits.

Since $\mathcal{D}$ is a multi-set, we cannot directly feed it into DL framework. As such, we use its vectorized representation, i.e., *persistence image (PI)* [1]. To construct PI, we first map $\mathcal{D}$ to an integrable function $\rho_{\mathcal{D}} : \mathbb{R} \to \mathbb{R}^2$, which is referred to as *the persistence surface* and which is given by sums of weighted Gaussian functions centered at each point in $\mathcal{D}$. We then integrate $\rho_{\mathcal{D}}$ over each grid box to obtain PI such that the value of each pixel z is given by

$$\mathrm{PI}(z) = \iint_z \sum_{\mu \in T(\mathcal{D})} \frac{g(\mu)}{2\pi\delta_x\delta_y} e^{-\left(\frac{(x-\mu_x)^2}{2\delta_x^2} + \frac{(y-\mu_y)^2}{2\delta_y^2}\right)} dydx. \tag{1}$$

Here $T(\mathcal{D})$ is the transformed PD $\mathcal{D}$ (i.e., $T(x, y) = (x, y-x)$), $g(\mu)$ is a weighting function, where $\mu = (\mu_x, \mu_y) \in \mathbb{R}^2$), while μ_x and δ_x and μ_y and δ_y are the mean and the standard deviation of the Gaussians in x and y direction, respectively.

Graph Neural Network-Based Anomaly Detection The GDN architecture addresses the structure learning process with graph neural networks and combines it with attention weights to detect anomaly. The GDN model learns the vector embedding for each location during the training process and uses the similarity between vectors to build the connection relationships. The observed data at time t is $\mathbf{s^{(t)}}$. When the size of the sliding window is w, the input $\mathbf{x}^{(t)}$ is $\mathbf{x}^{(t)} = \left[\mathbf{s^{(t-w)}}, \mathbf{s^{(t-w+1)}}, \cdots, \mathbf{s^{(t-1)}}\right]$. Based on the learned graph structure, the

aggregated representation of node is computed as

$$\mathbf{z}_i^{(t)} = \mathrm{ReLU}(\alpha_{i,i}\mathbf{W}\mathbf{x}_i^{(t)} + \sum_{j \in \mathcal{N}(i)} \alpha_{i,j}\mathbf{W}\mathbf{x}_j^{(t)}), \tag{2}$$

where $\mathbf{W}$ is a weighted matrix, $\mathbf{x}_i^{(t)}$ is the input feature of node i, $\mathcal{N}(i)$ denotes the neighbors of node i from structure learning, and $\alpha_{i,j}$ is attention coefficient.

Then the GDN model [14] utilizes the representation of the node i, i.e., $\mathbf{z}_i^{(t)}$ and embeds the corresponding vector $\mathbf{v}_i$ to predict the current value. Lastly, GDN generates the anomaly score and identify anomaly.

4 Topological Lifespan Graph Neural Network-Based Anomaly Detection Approach(Tlife-GDN)

Problem Statement Mathematically, the anomaly detection problem can be formulated as follows. Let $\mathbf{s}^{(t)}$ be records (e.g., COVID-19 hospitalizations) from N locations, where $t = \{1, 2, \ldots, T\}$. Let $l^{(t)}$ be the binary anomaly status at time t, e.g., $l^{(t)} = 0$ represents a normal behaviour, whilst $l^{(t)} = 1$ when some abnormality occurs. Let $\mathcal{G}^{(t)} = (V, E, \omega^{(t)})$ be a weighted connectivity network among locations $\mathbf{s}^{(t)}$, with node set $V = \{v_1, v_2, \ldots, v_N\}$, i.e., each node represents a location, edge set $E \in V \times V$ and the non-negative symmetric edge-weight matrix $\omega^{(t)}$ with entries $\{\omega_{ij}^{(t)}\}_{1 \le i,j \le N}$. In this paper, we focus on two problems: 1) current anomaly prediction and 2) forecasting of future anomalies.

`Problem 1`:To learn a mapping function $\mathcal{H}(\{\mathbf{s}^{(t)}\}_{t=1}^{T-1}, \{\mathcal{G}^{(t)}\}_{t=1}^{T-1})$ which maps the records to a binary anomaly output $l^{(t)}$.

`Problem 2`: Given an ahead horizon h, our goal is to learn a mapping function $\mathcal{H}(\{\mathbf{s}^{(t)}\}_{t=1}^{T-1}, \{\mathcal{G}^{(t)}\}_{t=1}^{T-1})$ which maps the records to a binary anomaly output $l^{(t+h)}$.

In order to capture the complex topological features of the spatio-temporal data, we construct dynamic networks, and extract the n-dimensional features in the form of persistence diagram and vectorize the persistence diagram to obtain persistence image. Then, we integrate the persistence image into the GNN framework for detection of existing anomalies and prediction of future anomalies.

4.1 Topological Features of Dynamic Networks

Topological features provide a way to systematically describe the graph structure and track the evolution of hidden patterns of data. In this paper, we make use of lifespans of those topological features from different nodes in the dynamic network. Specially, with records $\{\mathbf{s}_i^{(t)}\}_{i=1}^{N}$, we calculate the L_1 distance matrix $H^{(t)}$ of record values $\{\mathbf{s}_i^{(t)}\}_{i=1}^{N}$ to build connections between locations (e.g., counties) as shown in Algorithm 1. The locations with close values are considered to have similar patterns. In our study, the counties with similar COVID-19 cases rate may have similar geometric structure information regarding the COVID-19

Algorithm 1 Topological Features from Dynamic Networks

INPUT: Location Records $\{\mathbf{s}_i^{(t)}\}_{i=1}^{N}, t = \{1, 2, \ldots, T\}$
OUTPUT: Topological summaries
for $t \leftarrow 1 : T$ **do**
    for $i \leftarrow 1 : N - 1$ **do**
        for $j \leftarrow i + 1 : N$ **do**
            Compute $H_{ij}^{(t)} = |\mathbf{s}_i^{(t)} - \mathbf{s}_j^{(t)}|$
    Keep only bottom-m values in $H^{(t)}$
    Compute $\omega^{(t)}(e) = 1 - H^t / \max(H^{(t)})$
    Generate $\mathcal{G}^{(t)}$ based on $\omega^{(t)}(e)$
Apply persistent homology on dynamic networks $\mathbb{G} = \{\mathcal{G}^{(t)}\}_{t=1}^{T}$ for different dimensions and generate persistence diagram(PD) for each timestamp t
Apply equation 1 in section 3 to generate persistence image (PI) from PD

transmission, empty ICU beds and hospitalization severity. We take the lowest-m values in the connection matrix, where m is a predefined number based on the dataset. Then, we generate an edge weight matrix $\omega(e)$ by taking 1 minus the standardized $H^{(t)}$ and get its corresponding weighted graph $\mathcal{G}^{(t)}$. The next step is to use persistent homology to track the invariant structure features, and compute a persistence diagram (PD) for each network $\mathcal{G}^{(t)}$ and its corresponding lifespan information. Finally, we generate the vectorized represented persistent image (PI) defined in Section 3 as the topological features from the location's dynamic networks.

4.2 Tlife-GDN Architecture

With the spatio-temporal dataset $\mathbf{s}^{(t)}$ (where $t = \{1, 2, \ldots, T\}$), we capture the topology features PI defined in Section 4.1. Then we train our topology-based GDN to capture the hidden structure between different locations. Equation 3 shows the implementation of persistence image $\mathrm{PI}^{(t-1)}$ in the graph neural networks framework

$$\mathbf{z}_i^{(t)} = \mathrm{ReLU}\left((\alpha_{i,i}\mathbf{W}\mathbf{x}_i^{(t)} + \sum_{j \in \mathcal{N}(i)} \alpha_{i,j}\mathbf{W}\mathbf{x}_j^{(t)})\mathbf{Q}^{(t)}\right), \tag{3}$$

where $\mathbf{z}_i^{(t)}$ denotes the latent representation of the node i at timestamp t. $\mathbf{Q}^{(t)} \in \mathbb{R}^d$ is the topological representation from the CNN based model (where d is the length of embedding vector for each location), which is formulated as $\mathbf{Q}^{(t)} = f_{cnn}(\mathrm{PI}^{(t-1)})$, where f_{cnn} is a CNN-based model and $\mathrm{PI}^{(t-1)}$ denotes the PI for the network at $(t-1)$ timestamp. Then, we add the latent nodes' representation into the graph detection network architecture to predict the location's value. For anomaly detection/prediction, we use the loss function and error score as

$$L_{\mathrm{MSE}} = \frac{1}{T-w}\sum_{t=w+1}^{T-h}\left\|\hat{\mathbf{s}}^{(\mathbf{t+h})} - \mathbf{s}^{(\mathbf{t})}\right\|_2^2, \quad \mathrm{Err}_i(t+h) = \left|\mathbf{s}_\mathbf{i}^{(\mathbf{t})} - \hat{\mathbf{s}}_\mathbf{i}^{(\mathbf{t+h})}\right|, \tag{4}$$

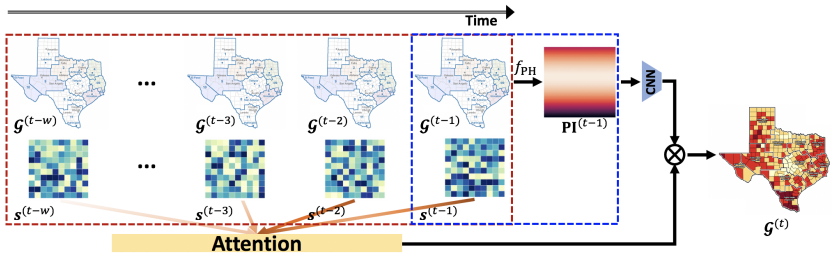


Fig. 1: Architecture overview of the Tlife-GDN model, where $(\mathcal{G}^{(t-w)}, \ldots, \mathcal{G}^{(t-2)}, \mathcal{G}^{(t-1)})$ and $(\mathbf{s}^{(t-w)}, \ldots, \mathbf{s}^{(t-2)}, \mathbf{s}^{(t-1)})^\top \in \mathbb{R}^{N \times F \times w}$ denote all graph structures and values of all features for each node over w time slices, respectively.

where h is the prediction window and $h = 0$ correspond with the detection task. For both detection and prediction tasks, the anomalousness score at time t is the maximal score across locations $A(t) = \max_i a_i(t)$ where $a_i(t)$ is the standardized error score.

The overall architecture is shown in Figure 1. The intuition here is to combine the topological features along with the records (e.g., COVID-19 hospitalization rates) as the input for the topology-based GDN model. At timestamp t, we generate PIs for the latest timestamp $t-1$, as the topological summaries and use a CNN-based model to learn its representation. With the enriched input data, we use equation 3 to get the latent node's representation. Although different DL methods have been proposed to improve the anomaly detection accuracy, PIs have not been incorporated into this task. Furthermore, regrading COVID-19 spreading, topological summaries can help the learning grasp on the persistent hidden features behind the progression process caused by environmental or social-demographic variables. As a result, Tlife-GDN model extracts the complex spatio-temporal dependence properties which are inaccessible with other GDL tools.

5 Experiment

5.1 Datasets, experiment setup and evaluation metrics

We conduct experiments on 5 datasets: COVID-19 records in Texas (TX), California (CA) and Pennsylvania (PA), Curiosity Rover on Mars (MSL) and Water Distribution (WADI). Table 1 summarize the properties of each dataset. The daily records for COVID-19 cases and hospitalizations come from CovidActNow project[7] and Johns Hopkins University[8]. These data sources contain COVID-19 time series from official state and county websites. We take 2 per thousand people as the anomaly threshold for hospitalization rate at state level. New

[7] Available at https://covidactnow.org/?s=24821397
[8] Available at https://github.com/CSSEGISandData/COVID-19

Table 1: Summary of the datasets. The anomaly rate is the ratio of true anomaly in the testing set.

Statistics	**MSL**	**WADI**	**TX**	**CA**	**PA**
Number of Variables	28	128	252	56	61
Training Size	1565	1784	200	200	200
Testing Size	500	577	175	175	175
Anomaly rate	20.24%	5.55%	56.57%	30.86%	31.42%

cases rate at county level, which indicates the spread of COVID-19, is used for training and prediction. The Curiosity Rover on Mars (MSL) is an expert-labeled telemetry anomaly data which originally comes from Incident Surprise, Anomaly (ISA) [21]. The reports assists in reducing the risk of the unexpected events which influence the post lunch operations. In our study, we use a public available sub-set[9]. The anomaly ratio in the MSL test dataset is 78.13%, to make the data more balanced, we use the first 500 observations, which has anomaly ratio 20.24%. Water Distribution (WADI) is a sensor-based dataset derived from a distribution system comprising numerous pipelines[10] [17]. Here, a test with size 16 days is conducted, with 14 days under normal operation which are used as training data and 2 days under controlled attack scenarios which is our test set.

We conduct our experiments using a Google colab sever with Intel(R) Xeon(R) CPU @ 2.20GHz, 52 GB RAM, K80,T4 and P100 graphic cards. All models are trained under ADAM optimizer with learning rate 1×10^{-6} and no decay rate. We perform 10 runs, train the models using 100 epochs, and use early stopping of 10. For GDN and Tlife-GDN, we use 128 as the length of embedding vectors and the number of neurons for all datasets. For COVID-19 anomaly prediction, we set similar setting of parameters as in the detection task and set the prediction window h to 7, and the validation ratio to 0.2.

To evaluate the performance of anomaly detection, we use the metrics: F1-Score (F1) and the area under the receiver operating characteristic curve (AUC). As the anomaly score range and the way to choose a suitable threshold is different from method to method, in order to keep the comparison fair for different detection baselines, we set the threshold to be the one which maximizes F1 score for all baselines. The scores above the threshold are considered as anomaly. Our source codes are publicly available in Github[11]

5.2 Experimental Results

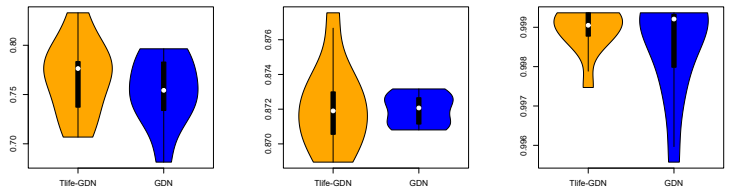
Are persistent images really helpful for COVID-19 anomaly detection and prediction? The anomaly detection results for COVID-19 datasets in TX, CA, and PA are shown in Table 2. For all baselines, we take the average value

[9] Available at https://github.com/d-ailin/GDN/tree/main/data/msl
[10] Further details at https://itrust.sutd.edu.sg/testbeds/water-distribution-wadi/ [3].
[11] https://github.com/ZhiweiZhen/Tlife-GDN

Table 2: Average F1 and AUC scores on COVID-19 datasets in 10 runs. For each metric, the best result is highlighted in yellow.

Model	TX		CA		PA	
	F1	AUC	F1	AUC	F1	AUC
PCA [39]	0.570 (< 0.0001)	0.739 (< 0.0001)	0.550 (< 0.0001)	0.498 (< 0.0001)	0.536 (< 0.0001)	0.498 (< 0.0001)
KNN [5]	0.640 (< 0.0001)	0.757 (< 0.0001)	0.767 (< 0.0001)	0.663 (< 0.0001)	0.631 (< 0.0001)	0.570 (< 0.0001)
AE [2]	0.729 (0.0001)	0.739 (0.0002)	0.550 (0.0022)	0.498 (0.0010)	0.534 (0.0023)	0.495 (< 0.0001)
DAGMM [48]	0.525 (0.0171)	0.710 (0.0422)	0.680 (0.0697)	0.6390 (0.0422)	0.875 (0.0443)	0.533 (0.0443)
VAE [25]	0.565 (0.0050)	0.519 (0.0016)	0.535 (0.0059)	0.484 (0.0026)	0.531 (0.0032)	0.516 (< 0.0001)
DEEP-SVDD [34]	0.675 (0.0156)	0.739 (0.0122)	0.776 (0.0129)	0.436 (0.0189)	0.960 (0.0242)	0.492 (0.0112)
GDN [14]	0.754 (0.0352)	0.742 (0.0122)	0.928 (0.0015)	0.743 (0.0008)	0.994 (0.0013)	0.975 (0.0002)
Tlife-GDN	0.767 (0.0374)	0.759 (0.0092)	0.962 (0.0020)	0.754 (0.0006)	0.995 (0.0220)	0.976 (0.0001)

for F1 score and AUC score in 10 runs, and the standard deviation is shown in parenthesis. From the result, we can see that Tlife-GDN outperforms all baselines across both F1 score and AUC on all 3 states. The topological features extracted from the counties tend to improve the detection performance through comparing Tlife-GDN with GDN model (which is the best baseline). In addition, Table 2 also indicates that integrating topological summaries into GDN model will not increase standard deviation of F1 score and AUC score. Furthermore, Figure 2 shows the box-plot of AUC score for Tlife-GDN and GDN, from which we conclude that Tlife-GDN exhibits high stability.



(a) (b) (c)

Fig. 2: Box plot of AUC scores in 10 runs from Tlife-GDN and GDN in (a) Texas (b) California (c) Pennsylvania.

In addition, for the traditional anomaly detection problem, we utilize Tlife-GDN and GDN (i.e., the best baseline) to predict future anomalies and verify the significance of topological features. Figure 3 shows that Tlife-GDN achieves a better performance on TX and CA. On PA, both GDN and Tlife-GDN perform well. We can see that the complex hidden topological relationships between counties have a profound impact on future hospitalization anomalies as it may contain the information about the COVID-19 transmission at that moment.

What is the performance of Tlife-GDN on MSL and WADI datasets? To verify the value added by topological summaries for different types of anomaly detection problems, we also evaluate the performance of our Tlife-GDN model on MSL and WADI datasets. The results are shown in Table 4. We find that Tlife-GDN outperforms all baselines in terms of both F1 score and AUC score

Table 3: Average precision, recall, and F1 score on COVID-19 datasets for one-week ahead anomaly prediction in 10 runs based on GDN and Tlife-GDN.

Model	TX		CA		PA	
	F1	AUC	F1	AUC	F1	AUC
GDN	0.728 (0.0647)	0.762 (0.0521)	0.869 (0.0353)	0.940 (0.1028)	0.927 (0.0398)	0.994 (0.0315)
Tlife-GDN	0.741 (0.0321)	0.765 (0.0198)	0.894 (0.0265)	0.962 (0.0374)	0.927 (0.0379)	0.993 (0.0452)

for WADI. For MSL, Tlife-GDN achieves the best result in F1 score and also competitive result in AUC.

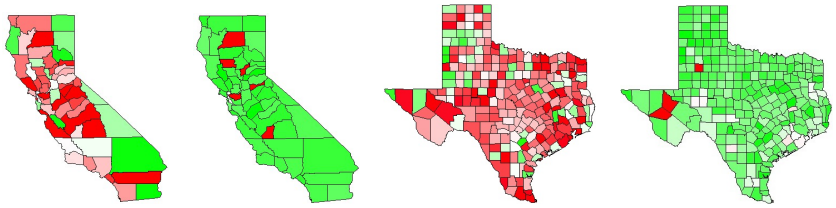
Table 4: Average F1 and AUC scores on MSL and WADI datasets in 10 runs. For each metric, the best result is highlighted in yellow.The results from Tlife-GDN is highlighted in blue if there is improvement compared to GDN.

Model	MSL		WADI	
	F1	AUC	F1	AUC
PCA	0.151 (< 0.0001)	0.533 (< 0.0001)	0.120 (< 0.0001)	0.504 (< 0.0001)
KNN	0.109 (< 0.0001)	0.664 (< 0.0001)	0.119 (< 0.0001)	0.475 (< 0.0001)
AE	0.152 (< 0.0001)	0.553 (0.06187)	0.120 (< 0.0001)	0.503 (0.02546)
DAGMM	0.361 (0.0549)	0.631 (0.0708)	0.289 (0.0250)	0.603 (0.0603)
VAE	0.120 (0.2210)	0.553 (0.2317)	0.148 (0.1376)	0.503 (0.0557)
DEEP-SVDD	0.337 (0.0555)	0.665 (0.1003)	0.100 (0.0483)	0.477 (0.0019)
GDN	0.407 (0.0125)	0.496 (0.0267)	0.356 (0.0745)	0.785 (0.0632)
Tlife-GDN	0.419 (0.0198)	0.563 (0.1054)	0.371 (00319)	0.797 (0.0480)

Possible linkage between detection results and environment In this study, we also explore the impact of topological features on the detection results. We investigate the timestamps where Tlife-GDN achieves the accurate anomaly detection performance compared with GDN. We believe that those timestamps may share some similarity in terms of environmental variables. Figure 3 shows the Aerosol Optical Depth (AOD) values, a measure of light extinction by aerosol in the atmospheric column above the earth's surface [44], in TX and CA whenever Tlife-GDN outperforms GDN at county level. In addition, Fig. 3 suggests that topological features can improve the ability of non-anomaly detection when AOD is low and help detect anomalies when AOD is high. Furthermore, we can find that the hospitalization rate can be well reflected by the AOD values, which can be used to define anomalies in the anomaly detection task.

6 Conclusion

In this paper, we introduce a new topology-based graph neural network, i.e., Tlife-GDN to detect and predict anomaly. The experimental results show that Tlife-GDN provides more accurate detection and prediction for COVID-19 hospitalization anomalies in Texas, California, and Pennsylvania, which is critical to forecast pandemic trend, announce travel warnings and help local government prepare potential waves in advance. In the future, we can take pre-existing



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

Fig. 3: Aerosol Optical Depth (AOD) values in CA and TX. The color goes from red to green as AOD increase. (a) Non-anomaly CA. (b) Anomaly CA. (c) Non-anomaly TX. (d) Anomaly TX.

health conditions, distribution of medical resources and demographic variables into consideration and extend the application of Tlife-GDN to anomaly regarding network defense and national cyber security.

Acknowledgments

This work has been supported in part by grants NSF DMS 1925346, NSF ECCS 2039701, NASA 20-RRNES20-0021, and the Department of the Navy, Office of Naval Research under ONR award number N00014-21-1-2530. Part of this material is also based upon work supported by (while serving at) the National Science Foundation. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation and/or the Office of Naval Research. The authors are grateful to Huikyo Lee, NASA's Jet Propulsion Lab for the motivating discussion.

References

1. Adams, H., Emerson, T., Kirby, M., Neville, R., Peterson, C., Shipman, P., Chepushtanova, S., Hanson, E., Motta, F., Ziegelmeier, L.: Persistence images: A stable vector representation of persistent homology. JMLR **18** (2017)
2. Aggarwal, C.C.: Data Mining: The Textbook. Springer, Cham (2015)
3. Ahmed, C.M., Palleti, V.R., Mathur, A.P.: WADI: a water distribution testbed for research in the design of secure cyber physical systems. In: CySWATER (2017)
4. Alonso, J., Belanche, L., Avresky, D.R.: Predicting software anomalies using machine learning techniques. In: IEEE NCA. pp. 163–170 (2011)
5. Angiulli, F., Pizzuti, C.: Fast outlier detection in high dimensional spaces. In: ECML PKDD (2002)
6. Basak, D., Pal, S., Patranabis, D.C.: Support vector regression. Neural Information Processing – Letters and Reviews **11**(10) (2007)
7. Brar, G., Pinheiro, L.C., Shusterman, M., Swed, B., Reshetnyak, E., Soroka, O., Chen, F., Yamshon, S., Vaughn, J., Martin, P., et al.: COVID-19 severity and outcomes in patients with cancer: A matched cohort study. J Cl. Oncol pp. 3914–3924 (2020)

8. Cai, Q., Chen, F., Wang, T., Luo, F., Liu, X., Wu, Q., He, Q., Wang, Z., Liu, Y., Liu, L., et al.: Obesity and COVID-19 severity in a designated hospital in Shenzhen, China. Diabetes care **43**(7), 1392–1398 (2020)
9. Carlsson, G.: Topology and data. BAMS **46**(2), 255–308 (2009)
10. Chaudhary, A., Mittal, H., Arora, A.: Anomaly detection using graph neural networks. In: COMITCon. pp. 346–350. IEEE (2019)
11. Chazal, F., Michel, B.: An introduction to topological data analysis: fundamental and practical aspects for data scientists. Frontiers in Artificial Intelligence **4** (2021)
12. Chen, Y., Segovia-Dominguez, I., Coskunuzer, B., Gel, Y.R.: TAMP-S2GCNets: Coupling time-aware multipersistence knowledge representation with spatio-supra graph convolutional networks for time-series forecasting. In: ICLR (2022)
13. Chen, Y., Segovia-Dominguez, I., Gel, Y.R.: Z-GCNETs: Time zigzags at graph convolutional networks for time series forecasting. In: ICML (2021)
14. Deng, A., Hooi, B.: Graph neural network-based anomaly detection in multivariate time series. In: AAAI (2021)
15. Dey, T.K., Wang, Y.: Computational Topology for Data Analysis. Cambridge University Press (2022)
16. Gallo Marin, B., Aghagoli, G., Lavine, K., Yang, L., Siff, E.J., Chiang, S.S., Salazar-Mather, T.P., Dumenco, L., Savaria, M.C., Aung, S.N., et al.: Predictors of COVID-19 severity: A literature review. Rev. in medical virology **31**(1), 1–10 (2021)
17. Goh, J., Adepu, S., Junejo, K.N., Mathur, A.P.: A dataset to support research in the design of secure water treatment systems. In: CRITIS (2016)
18. Golan, I., El-Yaniv, R.: Deep anomaly detection using geometric transformations. arXiv:1805.10917 (2018)
19. Hickok, A., Needell, D., Porter, M.A.: Analysis of spatiotemporal anomalies using persistent homology: case studies with COVID-19 data. arXiv:2107.09188 (2021)
20. Hofer, C.D., Graf, F., Rieck, B., Niethammer, M., Kwitt, R.: Graph filtration learning. In: ICML. vol. 119, pp. 4314–4323. PMLR (2020)
21. Hundman, K., Constantinou, V., Laporte, C., Colwell, I., Soderstrom, T.: Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. arXiv:1802.04431 (2018)
22. Islambekov, U., Yuvaraj, M., Gel, Y.R.: Harnessing the power of topological data analysis to detect change points in time series. Environmetrics **31**(1) (2020)
23. Jin, W., Tung, A.K., Han, J., Wang, W.: Ranking outliers using symmetric neighborhood relationship. In: PAKDD. pp. 577–593. Springer (2006)
24. Karadayi, Y., Aydin, M.N., Öğrenci, A.S.: Unsupervised anomaly detection in multivariate spatio-temporal data using deep learning: Early detection of COVID-19 outbreak in italy. IEEE Access **8**, 164155–164177 (2020)
25. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv:1312.6114 (2013)
26. Li, D., Chen, D., Goh, J., Ng, S.k.: Anomaly detection with generative adversarial networks for multivariate time series. arXiv:1809.04758 (2018)
27. Li, Y., Islambekov, U., Akcora, C., Smirnova, E., Gel, Y.R., Kantarcioglu, M.: Dissecting ethereum blockchain analytics: What we learn from topology and geometry of the ethereum graph? In: SDM. pp. 523–531. SIAM (2020)
28. Liang, L., Gong, P.: Climate change and human infectious diseases: A synthesis of research findings from global and spatio-temporal perspectives. Environment international **103**, 99–108 (2017)
29. Liu, D., Veeramachaneni, K., Geiger, A., Li, V.O.K., Qu, H.: AQEyes: visual analytics for anomaly detection and examination of air quality data. arXiv:2103.12910 (2021)

30. Ma, X., Wu, J., Xue, S., Yang, J., Zhou, C., Sheng, Q.Z., Xiong, H., Akoglu, L.: A comprehensive survey on graph anomaly detection with deep learning. IEEE Trans. Knowl. Data Eng. (2021)
31. Malhotra, P., Vig, L., Shroff, G., Agarwal, P.: Long short term memory networks for anomaly detection in time series. In: ESANN. vol. 89, pp. 89–94 (2015)
32. Moore, M., Landree, E., Hottes, A.K., Shelton, S.R.: Environmental biodetection and human biosurveillance research and development for national security. Tech. rep., Homeland Security Operational Analysis Center, RAND Corp. (2018)
33. Ofori-Boateng, D., Dominguez, I.S., Kantarcioglu, M., Akcora, C.G., Gel, Y.R.: Topological anomaly detection in dynamic multilayer blockchain networks. In: ECML (2021)
34. Ruff, L., Vandermeulen, R.A., Görnitz, N., Deecke, L., Siddiqui, S.A., Binder, A., Müller, E., Kloft, M.: Deep one-class classification. In: ICML. vol. 80, pp. 4393–4402 (2018)
35. Sanchez-Hernandez, C., Boyd, D.S., Foody, G.M.: One-class classification for mapping a specific land-cover class: SVDD classification of fenland. GRSS-IEEE **45**(4), 1061–1073 (2007)
36. Segovia Dominguez, I., Lee, H., Chen, Y., Garay, M., Gorski, K.M., Gel, Y.R.: Does air quality really impact COVID-19 clinical severity: coupling NASA satellite datasets with geometric deep learning. In: ACM SIGKDD. pp. 3540–3548 (2021)
37. Segovia-Dominguez, I., Lee, H., Zhen, Z., Chen, Y., Garay, M., Crichton, D., Wagh, R., Gel, Y.R.: Using NASA satellite data sources and geometric deep learning to uncover hidden patterns in COVID-19 clinical severity. arXiv:2110.10849 (2021)
38. Segovia-Dominguez, I., Zhen, Z., Wagh, R., Lee, H., Gel, Y.R.: TLife-LSTM: Forecasting future COVID-19 progression with topological signatures of atmospheric conditions. In: PAKDD. pp. 201–212 (2021)
39. Shyu, M.L., Chen, S.C., Sarinnapakorn, K., Chang, L.: A novel anomaly detection scheme based on principal component classifier. Tech. rep., Miami Univ Coral Gables Fl Dept of Electrical and Computer Engineering (2003)
40. Stolz, B.J., Harrington, H.A., Porter, M.A.: Persistent homology of time-dependent functional networks constructed from coupled time series. Chaos **27**(4), 047410 (2017)
41. Tack, A.J., Thrall, P.H., Barrett, L.G., Burdon, J.J., Laine, A.L.: Variation in infectivity and aggressiveness in space and time in wild host–pathogen systems: causes and consequences. Journal of evolutionary biology **25**(10), 1918–1936 (2012)
42. Tang, J., Chen, Z., Fu, A.W.C., Cheung, D.W.: Enhancing effectiveness of outlier detections for low density patterns. In: PAKDD. pp. 535–548. Springer (2002)
43. Umeda, Y., Kaneko, J., Kikuchi, H.: Topological data analysis and its application to time-series data analysis. Fujitsu Sci. & Technical J. (2019)
44. Van Donkelaar, A., Martin, R.V., Brauer, M., Kahn, R., Levy, R., Verduzco, C., Villeneuve, P.J.: Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: development and application. Environmental health perspectives **118**(6), 847–855 (2010)
45. Vapnik, V.N.: An overview of statistical learning theory. IEEE transactions on neural networks **10**(5), 988–999 (1999)
46. Vries, D., Van Den Akker, B., Vonk, E., De Jong, W., Van Summeren, J.: Application of machine learning techniques to predict anomalies in water supply networks. Water Sci. Technol. **16**(6), 1528–1535 (2016)
47. Zeng, S., Graf, F., Hofer, C., Kwitt, R.: Topological attention for time series forecasting. In: NeurIPS (2021)
48. Zong, B., Song, Q., Min, M.R., Cheng, W., Lumezanu, C., Cho, D., Chen, H.: Deep autoencoding gaussian mixture model for unsupervised anomaly detection. In: ICLR (2018)