World Journal of Advanced Engineering Technology and Sciences
eISSN: 2582-8266
Cross Ref DOI: 10.30574/wjaets
Journal homepage: https://wjaets.com/

(RESEARCH ARTICLE)

Check for updates

# Anomaly detection in financial time series data via mapper algorithm and DBSCAN clustering

Md. Morshed Bin Shiraj *, Md. Mizanur Rahman, Md. Al-Imran, Mst Zinia Afroz Liza, Md. Masum Murshed and Nasima Akhter

*Department of Mathematics, University of Rajshahi, Rajshahi-6205, Bangladesh.*

World Journal of Advanced Engineering Technology and Sciences, 2024, 13(01), 070–084

Publication history: Received on 30 July 2024; revised on 07 September 2024; accepted on 09 September 2024

Article DOI: https://doi.org/10.30574/wjaets.2024.13.1.0396

## Abstract

Topological Data Analysis (TDA) has proven to be a powerful framework for uncovering hidden structures in high-dimensional data. This study investigates the integration of the Mapper algorithm with DBSCAN clustering to detect anomalies in financial time series data, specifically using daily price data from the Dhaka Stock Exchange. The methodology involves projecting the data into a lower-dimensional space using a filter function, covering this space with overlapping intervals, and applying DBSCAN to identify clusters within each subset. The resulting Mapper graph visualizes the relationships between clusters, with anomalies detected as unclustered points, isolated clusters, or small disconnected nodes. A total of 44 data points were identified as anomalies, which correspond to extreme price movements in the time series data. This combination of TDA and clustering provides a robust framework for anomaly detection, particularly in high-dimensional data where traditional clustering methods often fail to capture the full structure. Validation through SVM confirmed anomalies in the data, but the Mapper-DBSCAN approach demonstrated clearer separation of normal data and anomalies. The results demonstrate the potential of this approach for identifying anomalous behaviors in complex financial data.

**Keywords:** Anomaly Detection; DBSCAN Clustering; Mapper Algorithm; Persistent Homology; Support Vector Machine (SVM); Topological Data Analysis.

## 1. Introduction

Time Series Analysis is popular in different field of study including economics, physics, manufacturing, population, dynamics, Regression Analysis, Correlation Analysis, Spectral Analysis etc. [1]. Time Series Analysis has been used in environmental studies industries, Fisheries, medical [2] and many other diverse areas. Traditionally, statistical analysis techniques have been used in time series analysis [1]. There are many different problems of time series analysis, Anomaly Detection is one of the popular problem of them.

Anomaly detection has been used in diverse field of study with different methodologies [3]. Anomaly detection is a study of finding unusual behavior/pattern of a time series. Outliers and anomalies two different unusual behaviors of a time series but in the field of Anomaly detection they are considered as anomalies [3, 4]. Anomaly detection has been used in detecting fraudulent transactions using credit cards, insurance, medical, cyber security, difference systems of a country like, military, crime resilience unit etc. [3]. Anomalies may have significant indication of unusual activities which is important in many different sectors for taking some action against it. For example, unusual pattern in manufacturing indicates some error in the production or anomalies of stock exchanges daily closed price data may refer to some malicious transections/news spread out in the market.

* Corresponding author: Morshed Bin Shiraj (mbshraj@gmail.com)

Point Anomalies are some individual points which are abnormal with respect to the rest of the data. Most of the anomaly detection researches are focused on detecting point anomalies of some time series [3, 4]. Different Data Mining techniques, Statistics, Machine Learning techniques, Spectral theories, and information theories have been studied to detect anomalies. Authors of [3] found 6 types of techniques such as Statistical, Nearest neighbor based/density based, Spectral, Information theoretic, clustering based, classification based. There are different classification-based anomaly detection techniques have been studied like, neural network based, Bayesian network based, rule based and Support Vector Machine (SVM) based [3]. One-class SVM is the machine learning technique considering one-class learning which is faster than many other techniques. In Clustering-based Anomaly Detection technique [3, 4], similar data has been considered into one cluster and building clusters among a data set can induce anomalies which are those data points that are not belong to any clusters or belong to small clusters. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is one of the several clustering algorithms which can be used for anomaly detection. Clustering-based anomaly detection technique is vulnerable to clustering Algorithm.

Anomaly detection is one of the popular sector/field of data analysis which has been reviewed in several surveys [3] – [10]. Among of all them [4] has been studied exclusively in financial domain. In the study, the following three important problems of existing anomaly detection techniques have been mentioned**:**

- There is no global anomaly detection technique since there is no exact normal behavior of all data sets in general [3].
- Normal behavior can be changed time to time. Thus, it is hard to specify a specific rule to define normal data point and thus to define anomalies. For example, Small fluctuation in medical data can be considered as anomaly but similar fluctuation in stock exchange daily price data is normal [3].
- Fraudulent experts keep their activities look like normal incident and thus very hard to detect anomalies in this case.

For this reason, Topological data analysis (TDA) is also getting popular even in industrial manufacturing and production as one of the most potential tools that can play significant role in forming industries 4.0 in the $4^{th}$ industrial revolution period after 2010 [11]. Persistent Homology, UMAP and Mapper Algorithm are the most popular tools of TDA that can trigger this journey of the revolution [11]. In Stock market industry, those tools can also be employed to investigate a major data analysis problem like anomaly detection [11]. Different statistical based anomaly detection techniques are popular in scientific community, but statistical methodology needs lower dimensional embedding which leads to a certain amount of data lose [12]. In this content TDA based machine learning approach has been introduced for classification and anomaly detection of time series data. To do so persistent homology has been used on a transformed higher dimensional point cloud data of the inputted time series data to calculate Betti numbers and compared with the reference data which had been fixed earlier. Taken's Time Delay Embedding can transform a time series in a higher dimensional point cloud data and then periodicity of a time series can be tested [13]. TDA based feature extraction method has been introduced using Persistent Landscape and Silhouettes [14] Since TDA features are robust to noise, the study approach become effective. TDA based feature extraction method has been introduced based on persistent diagrams [15]. TDA based classification method has been proposed in [16] using persistent homology on the point cloud data processed by Time Delay Embedding. A temporal filtration has been constructed based on simplicial complex and then TDA tools like Persistent Homology, Mapper have been used to extract necessary information [17, 18]. Another TDA based classification method has been proposed based on filtered simplicial complex by recovering meaningful sub complex from the filtration [18]. Time series classification technique has been proposed based on persistent homology which has been vectorized in Betti series [19]. A new embedding has been introduced to do time series analysis tasks using TDA [20]. Topological attention has been defined for a particular filtration step calculating persistent homology to forecast time series [21]. Time series clustering algorithm has been introduced based on persistent homology by computing topological similarities among persistent diagram [22]. TDA based feature selection procedure has been explained mathematically using persistent homology in [23]. Another clustering methodology has been proposed combining spectral properties of simplicial complex after calculating eigenvectors from induced Hode-Laplacian of the simplicial complex [24]. TDA based clustering found effective through it in more expensive than other clustering methods and Mapper Algorithm is a unique visualization tool using Kepler Mapper Module of python [25].

Topological Data Analysis (TDA) has been applied to identify unusual patterns such as change points in studies like [26] – [29], chaotic behavior in [30], and failure data in software evolution using persistent homology techniques in [31]. An interval-based algorithm introduced in [32] detects anomalies in time series by measuring similarities between different segments of the data. In the financial domain, TDA has been used to analyze markets in [33], showing that efficient markets exhibit more persistence than volatile ones. Financial market crashes from January 2010 to June 2020 were examined in [34] using persistent landscapes to identify unusual daily return patterns across four US stock markets, with findings supporting the effectiveness of TDA in financial analysis. Another study [35] employed TDA tools

to investigate stock market behavior by tracking topological features such as connected components ($\beta_0$) and loops ($\beta_1$) across different filtration steps. Persistent diagrams and barcodes were computed using Python and Javaplex, although this study relied on Vietoris-Rips complexes rather than clustering-based methods for its analysis.

TDA has been used in financial data analysis to detect anomalies/malicious transections [13] and in bitcoin price data analysis concerning positive and negative financial bubbles [36]. TDA based clustering and classification of stock price data has been studied in [37], TDA has also been introduced in detecting topological shape [38] to follow stock market anomalies ([39] – [53]) for risk analysis, investment decision, analyzing past market crashes, finding doubtful transection and analyzing market dynamics.

Mapper Algorithm has been successfully studied in clustering ([54] – [56]). The big advantage of Mapper is that it can detect original data structure even if much noise appears. That's why Mapper is getting popular as anomaly detecting tool in different sectors. In [55], DBSCAN clustering has been used with Mapper Algorithm using Kepler Mapper python package to predict anomalies as fraud transection using credit cards. An exclusive explanation of Mapper graph and Mapper Algorithm with an example in RNA sample data has been disclosed in [56] where cluster nodes having larger size mean greater percentages in the overall data structure and links between the nodes refers connectivity among those clusters. Totally disconnected nodes can be treated as unique and unparalleled behavior that the other nodes [57]. Mapper Algorithm has been applied to study stock returns in [57] where anomalies have been visualized as Mapper cluster nodes which are isolated from the connected cluster nodes. But the study didn't consider those points which are not belonging to any of the clusters. In this study, Kepler Mapper has been used to visualize clusters using DBSCAN clustering.

Our domain of interest is Dhaka Stock Exchange (DSE). All of the studies reviewed earlier are in different financial domain of interests including US, China, Singapore, Taiwan, India etc. But DSE has not been studied well. The recent 2009-2010 Bangladesh's stock market crash has been happened due to mismanagement unethical interpret of authorities [58] and not having investment knowledge and analysis for the investors [59]. Still DSE is unstable while the market is unparallel and unpredictable comparing to the world's stock market [60].

There are several TDA tools that have been applied to financial time series data, but our focus is specifically on clustering-based techniques. For this purpose, we use the Mapper Algorithm combined with DBSCAN clustering to detect anomalies. To validate the results, we will also apply a popular classification-based anomaly detection method using Support Vector Machine (SVM) and compare its performance with the Mapper-DBSCAN approach.

### 1.1. Preliminaries

In this section, we discuss the key mathematical concepts behind Topological Data Analysis (TDA), the Mapper algorithm, clustering methods (specifically DBSCAN), and their role in anomaly detection.

#### *1.1.1. Topological Data Analysis (TDA)*

Topological Data Analysis (TDA) applies tools from algebraic topology to extract meaningful structures from high-dimensional data. The core objective of TDA is to compute topological invariants, such as connected components, loops, and voids, which persist across different scales in the data.

*Simplicial Complexes*: Given a data set $X \subseteq \mathbb{R}^n$, the first step is to construct a simplicial complex that encodes the structure of $X$. A simplicial complex $\mathcal{K}$ is defined as a collection of simplices (vertices, edges, triangles, and their higher-dimensional analogues) such that any face of a simplex in $\mathcal{K}$ is also in $\mathcal{K}$, and the intersection of any two simplices in $\mathcal{K}$ is either empty or a lower-dimensional simplex.

For instance, a 0-simplex represents a point, a 1-simplex represents an edge between two points, a 2-simplex represents a triangle, and so on.

*Persistent Homology*: The Vietoris-Rips complex $\mathrm{VR}_\epsilon(X)$ is commonly used in TDA. For a given scale parameter $\epsilon > 0$, $\mathrm{VR}_\epsilon(X)$ consists of simplices whose vertices are within $\epsilon$-distance of each other. The key idea of persistent homology is to track the changes in topological features (e.g., connected components, holes) as $\epsilon$ increases. Homology groups $H_k$ describe $k$ -dimensional features: $H_0$ describes connected components, $H_1$ describes loops, $H_2$ describes voids, and so on. The persistence of these features as $\epsilon$ increases provide insight into the underlying shape of the data.

*1.1.2. Mapper Algorithm*

The Mapper algorithm is a topological tool that visualizes the structure of high-dimensional data by creating a simplified representation in the form of a graph. Formally, the Mapper algorithm can be broken down into the following mathematical steps:

*Filter Function*: Let $f: X \to \mathbb{R}$ be a continuous filter function defined on the data $X \subseteq \mathbb{R}^n$. The filter function projects the data onto a lower-dimensional space (often $\mathbb{R}^1$ or $\mathbb{R}^2$) while preserving topological properties. Common choices for $f$ include projection onto principal components, distance functions, or density estimates.

*Covering*: The range of the filter function $f(X)$ is covered by a collection of overlapping intervals $\{I_i\}_{i=1}^m$. Mathematically, let $I_i = [a_i, b_i]$ be an interval, where $a_i \leq b_i$ and $a_{i+1} < b_i$. The overlap between intervals is defined by a parameter $\alpha$, which controls the degree of overlap.

*Clustering*: For each interval $I_i$, consider the subset $X_i = f^{-1}(I_i) \subseteq X$. A clustering algorithm, such as DBSCAN, is applied to each subset $X_i$, forming clusters $C_{i1}, C_{i2}, \ldots, C_{ik}$ within each interval.

*Graph Construction*: The Mapper graph $G = (V, E)$ is constructed where:

$$V \text{ represents the clusters } C_{ij}.$$

An edge $(v_1, v_2) \in E$ is drawn if the corresponding clusters $C_{i1}$ and $C_{j2}$ share data points.

This graph captures the global shape of the data in terms of its connectivity, and the nodes in the graph represent clusters of points from the data set.

*1.1.3. Clustering with DBSCAN*

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a clustering algorithm that classifies points based on density. DBSCAN groups points that are densely packed together and identifies points that lie in low-density regions as noise or outliers. The core parameters in DBSCAN are:

$\epsilon$ *(eps)*: The maximum distance between two points for them to be considered neighbors.

*min_samples*: The minimum number of points required to form a dense region.

Mathematically, for a point $x \in X$, the $\epsilon$-neighborhood of $x$ is defined as:

$$N_\epsilon(x) = \{ y \in X \mid \text{dist}(x, y) \leq \epsilon \}.$$

A point $x$ is considered a core point if $|N_\epsilon(x)| \geq \textit{min\_samples}$. Clusters are formed by connecting core points and their neighbors. DBSCAN is effective for identifying outliers, as points that do not meet the minimum density requirement are marked as noise.

*1.1.4. Anomaly Detection using Mapper and DBSCAN*

Anomalies, or outliers, are data points that deviate significantly from the general pattern. In the context of the Mapper algorithm combined with DBSCAN, anomalies can be detected as follows:

*Points not belonging to any cluster*: DBSCAN marks points as noise if they do not belong to a dense region. These points are often anomalies because they do not form part of any coherent structure in the data.

*Small or isolated clusters*: In the Mapper graph, small or disconnected nodes often represent anomalies. If a cluster has very few data points or is not well connected to other nodes, it may indicate an anomalous pattern.

The formal criteria for anomaly detection can be expressed as:

*Unclustered points*: $x \in X$ where $x$ is not part of any node in the Mapper graph.

*Isolated nodes*: Nodes $v \in V$ where $\text{degree}(v)$ is low or zero.

*1.1.5. Graph Theory in Mapper*

In the Mapper algorithm, the graph $G = (V, E)$ represents the topological structure of the data:

$V$ (vertices) correspond to clusters of points,

$E$ (edges) indicate overlapping clusters.

Mathematically, the edge set $E$ is constructed such that $(v_1, v_2) \in E$ if $C_1 \cap C_2 \neq \emptyset$, meaning the clusters share common points. The degree of a node $v$, denoted $\deg(v)$, gives insight into the connectivity of the data. Isolated nodes (with low or zero degree) often signify outliers or unusual patterns in the data.

## 2. Methodology Overview

### 2.1. Definition (Time Delay Embedding)

Given a time series $\{x_t\}_{t=1}^{n}$, we define the time delay embedding of the series as a set of vectors $\mathbf{y}_t \in \mathbb{R}^m$, where each vector is constructed as:

$$\mathbf{y}_t = \left[x_t, x_{t+\tau}, x_{t+2\tau}, \ldots, x_{t+(m-1)\tau}\right]$$

where $\tau$ is the time delay and $m$ is the embedding dimension. This process transforms the 1-dimensional time series into a set of vectors in $\mathbb{R}^m$, preserving the underlying dynamical structure of the series.

### 2.2. Definition (Mapper Construction)

Let $X = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\}$ be the set of time delay embedded vectors in $\mathbb{R}^m$. The Mapper algorithm creates a topological representation of $X$ as a graph $G = (V, E)$, where:

$V$ is the set of nodes (clusters) formed by applying a clustering algorithm (e.g., DBSCAN) on overlapping subsets of $X$ defined by a filter function (e.g., projection to a lower dimension).

$E$ is the set of edges between nodes, representing overlap between clusters.

The Mapper output is a graph $G$ that captures the topological structure of $X$, with each node representing a cluster of points from $X$.

### 2.3. Definition (Anomaly)

A point not included in any Mapper node is defined as a data point $\mathbf{y}_t \in X$ that does not belong to any of the clusters $V$ in the Mapper graph $G$. These points are considered potential anomalies, as they represent behavior not captured by the topological clusters.

### 2.4. Theorem (Anomaly Detection via Mapper Graph)

*Let $X = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N\}$ be a set of points obtained from time delay embedding of a time series, and let $G = (V, E)$ be the Mapper graph constructed from $X$. A data point $\boldsymbol{y}_i \in X$ is an anomaly if and only if it is not included in any cluster (node) in $G$.*

**Proof**

*Step 1: Time Delay Embedding*

Let $X = \{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_N\} \subset \mathbb{R}^m$ be the set of time delay embedded vectors generated from the time series $\{x_t\}_{t=1}^{n}$. By Takens' Embedding Theorem, this embedding preserves the dynamics of the original system, allowing us to analyze the time series in $\mathbb{R}^m$.

*Step 2: Mapper Graph Construction*

The Mapper algorithm defines a filter function $f: X \to \mathbb{R}^k$ (e.g., projection to principal components or other lens functions) and covers the range of $f(X)$ with overlapping intervals.

For each interval $U_i$, we apply a clustering algorithm (e.g., DBSCAN) to the pre-image $f^{-1}(U_i) \subseteq X$, forming clusters. Each cluster becomes a node in the Mapper graph $G$, and edges are created between nodes if clusters share common points (i.e., overlapping subsets of $X$).

*Step 3: Node Membership*

Each node $v \in V$ in the Mapper graph represents a cluster of data points from $X$. Specifically, each cluster is a subset $C_v \subseteq X$.

The set of all data points that belong to some node in the Mapper graph is given by:

$$\mathcal{C} = \bigcup_{v \in V} C_v$$

Thus, a data point $\mathbf{y}_i$ belongs to the Mapper graph if and only if $\mathbf{y}_i \in \mathcal{C}$.

*Step 4: Anomalous Points (Outliers)*

A data point $\mathbf{y}_i \in X$ is considered an **anomaly** if it does not belong to any node in the Mapper graph. Mathematically, this is expressed as:

$$\mathbf{y}_i \notin \mathcal{C}.$$

Therefore, the set of anomalies $A \subseteq X$ is:

$$A = X \setminus \mathcal{C}.$$

This set $A$ contains all the points that are not captured by any cluster in the Mapper graph, indicating that they are structurally different from the main population of points.

*Step 5: Anomaly Detection*

Since $\mathcal{C}$ contains all the points that belong to the nodes in the Mapper graph, the complement $A = X \setminus \mathcal{C}$ represents the points that are not part of any cluster. By definition, these points are anomalies, as they do not conform to the topological structure of the data represented by $G$.

Thus, any point $\mathbf{y}_i$ that is not part of any node in $G$ is identified as an anomaly.

### 2.5. Working Steps

Therefore, the Mapper algorithm serves as a tool for anomaly detection by identifying points that deviate from the clustered topological structure of the dataset.

We follow the above procedure in few steps which can be simplified as:

- Firstly, transform time series data into higher dimensions using time delay embedding to reveal hidden patterns.
- Then, standardize the data to ensure equal contribution from all variables by normalizing them.
- After that, use the Mapper algorithm to create a simplified topological structure by clustering similar points.
- Identify anomalies as points that do not belong to any cluster in the Mapper graph.
- Finally, visualize the results and fine-tune the model parameters to improve the detection of anomalies.

### 2.6. Data Source

The process of scraping financial data from online archives, such as the Dhaka Stock Exchange, involves several key steps (See Fig. 1). First, you need to identify and access the specific URL where the desired stock data, such as daily prices or volumes, is available. Once the webpage is accessed, the structure of the webpage must be analyzed to locate the elements that contain the data, such as HTML tables or dynamically loaded content. After extracting relevant data, including dates, closing prices, and volumes, the data is organized and stored in a structured format like CSV or Excel for further analysis. If the data is dynamically loaded, additional tools or API access may be required. Automation of the

process can be implemented for multiple queries, such as different stock symbols or date ranges. Throughout the process, it's essential to ensure compliance with the website's terms of service and legal requirements.

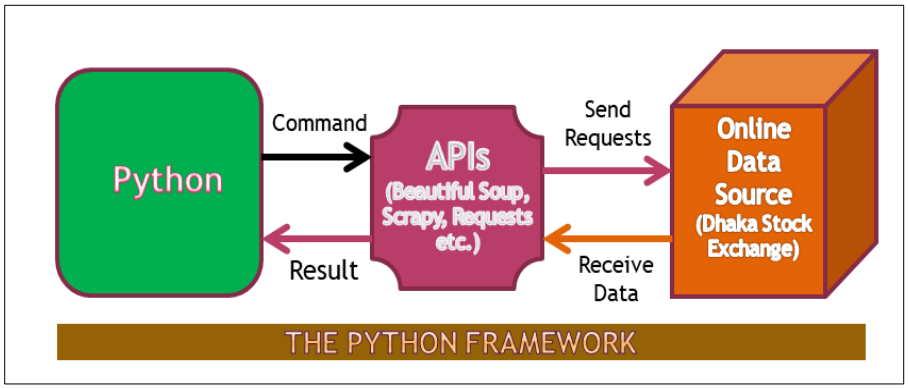


**Figure 1** An overview of data scraping from Dhaka stock exchange.

## 3. Results and discussion

In this study, a stock named INTRACO (Intraco Refueling Station PLC) of Dhaka stock exchange has been collected using the python framework (Fig. 1) and it has been displayed in Fig. 2 where closed prices of INTRACO lies between the time period September 1, 2022 and August 30, 2024.

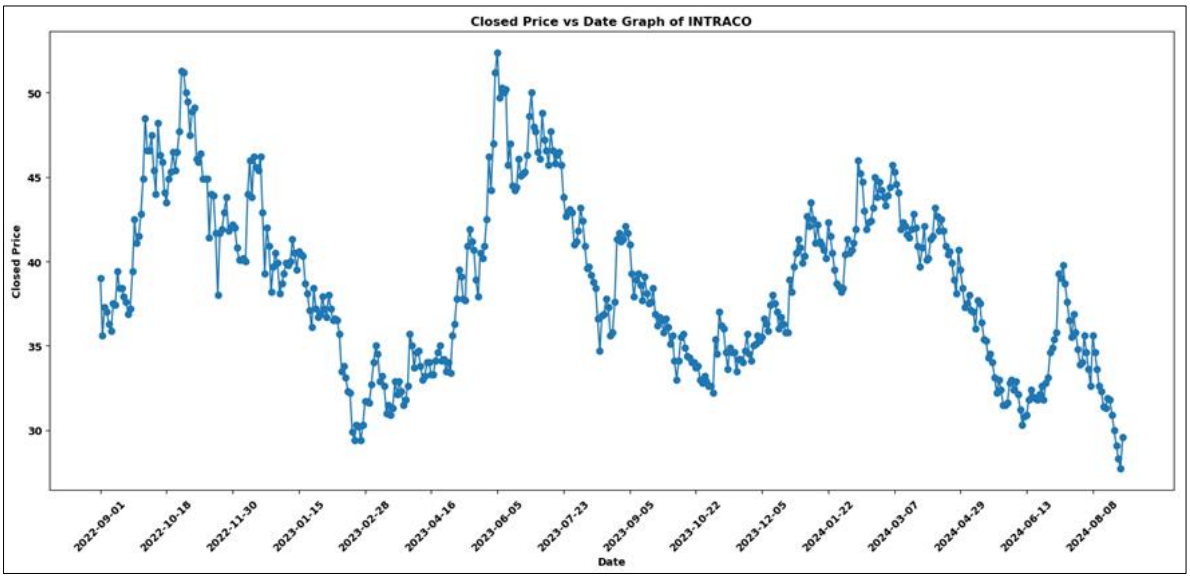


**Figure 2** Closed prizes of INTRACO from 01.09.2022 to 30.08.2024.

Persistent Diagram of the price data points (Fig. 3) has been checked to follow the topological properties of the data set. From the figure it is clear that there is no such significant loop indicated in orange dots. So, we focused on the connected components in blue dots where after a certain time (death) around 2.66 the death of connected components became irregular than bellow 2.66 and few gaps have been generated. The same gap can be found in Fig. 4 after $\varepsilon = 2.66$. Fig. 4 shows the relation of number of connected components against $\varepsilon$ filtration steps.

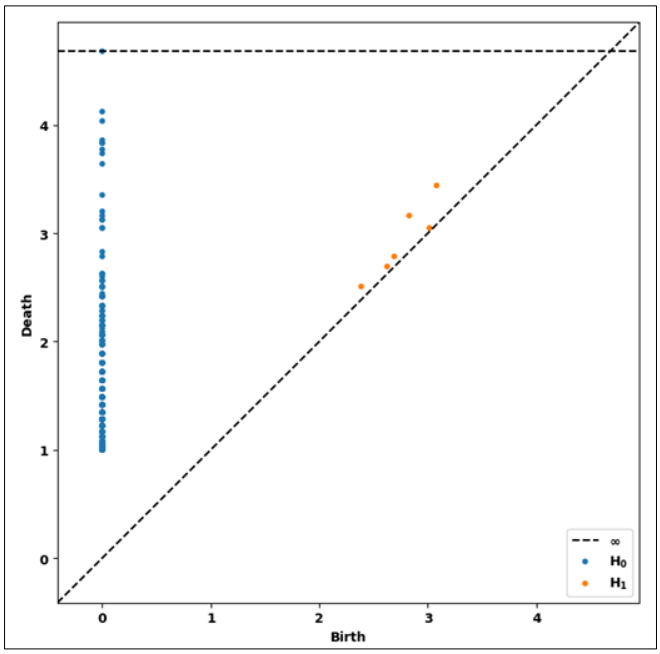


**Figure 3** Persistent Diagram of the INTRACO daily closed price data points.

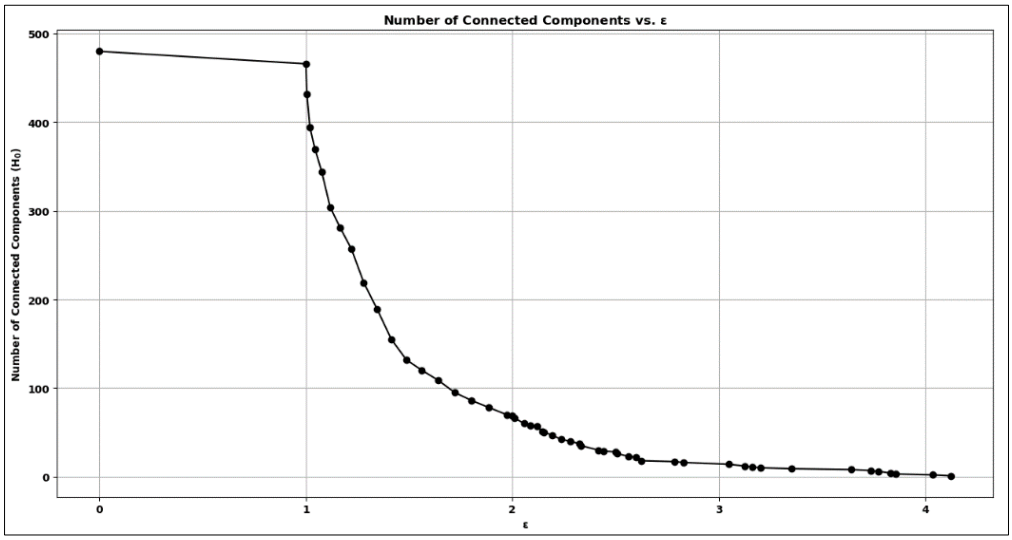


**Figure 4** No. of connected components against $\boldsymbol{\varepsilon}$ filtration steps of the data points.

The Fig. 5 shows the results of detecting anomalies in the INTRACO stock's closing prices using One-Class SVM with a parameter $nu = 0.04$. The blue dots represent the original data points of the closing prices over time, while the red crosses indicate the detected anomalies. These anomalies represent significant deviations from the general trend in the dataset, potentially identifying irregular or unusual price movements. The anomalies appear scattered throughout different time periods, particularly during peaks or steep fluctuations, which could indicate market volatility or significant external events influencing the stock price.

The Mapper graph shown in Fig. 6 represents the topological structure of the data using Kepler Mapper visualization [61]. Each node in the graph corresponds to a cluster of points, and the edges represent connections between clusters.

- Nodes: Each circular node represents a group of data points that share similar characteristics. There are a total of 9 nodes as indicated in the Mapper summary.
- Node 1 is isolated, suggesting a group of points that are significantly different from others. This might represent a small cluster of unusual or anomalous points.
- Nodes 8 and 9 are closely connected and may indicate another substructure or pattern within the data.

- Edges: The edges connecting the nodes represent transitions or relationships between different clusters. In this case, the graph suggests that some groups of data are more connected or related to each other than others.
- This graph helps identify patterns and anomalies based on the topological structure of the data.

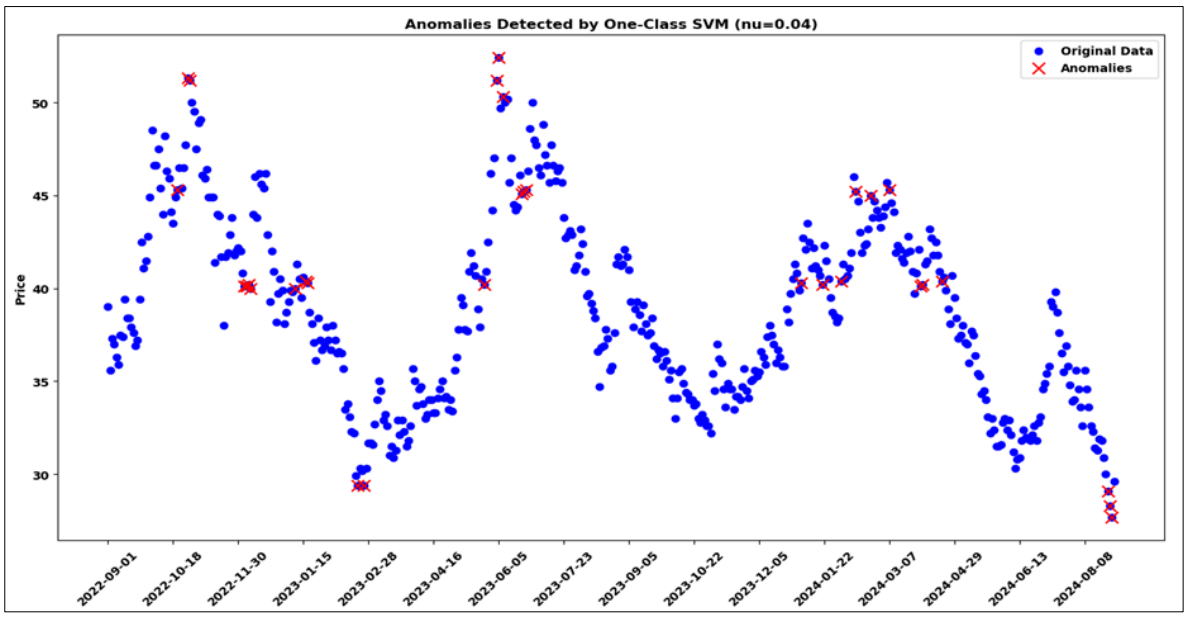


**Figure 5** Anomalies (red cross) detected using One-class SVM choosing $\boldsymbol{nu = 0.04}$.

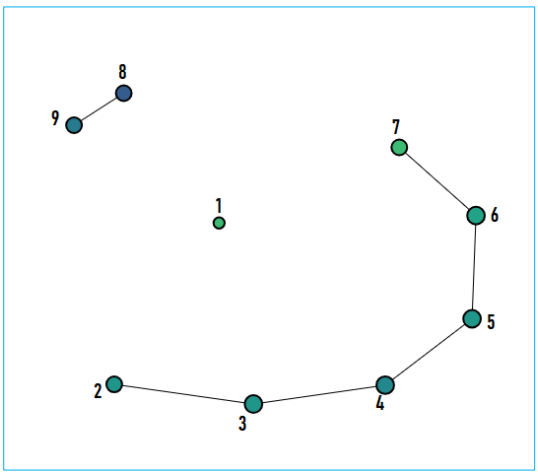


**Figure 6** Mapper Graph using DBSCAN Clustering while $\boldsymbol{eps = 0.075}$ and *min_samples* $= \mathbf{5}$.

The Mapper summary as displayed in Fig. 7 obtained from the Kepler Mapper visualization provides key details about the algorithm's configuration and the results:

- *Projection*: Custom projection, meaning that a specific method (e.g., time series-specific projection) was applied to transform the data before clustering.
- *Number of Cubes (N_Cubes)*: The data space was divided into 10 intervals (cubes) to apply the Mapper algorithm. This division allows for a finer resolution of data structure.
- *Percentage Overlap (PERC_OVERLAP)*: 10% overlap between the cubes, allowing for smoother transitions between clusters.
- *Clusterer*: DBSCAN was used with parameters eps=0.04 and min_samples=6 to form clusters. DBSCAN is effective in identifying dense regions of data and isolating noise or anomalies.
- *Nodes*: A total of 9 nodes were identified in the data.
- *Edges*: 6 edges were found, indicating relationships between clusters.

- *Total Samples*: 486 data points were analyzed, with 442 unique samples mapped to nodes. This indicates that 44 samples (486 - 442) were not included in any nodes, and these are the potential anomalies.
- *Node Distribution*: The graph also provides a color distribution of the nodes based on the data. The green and blue colors indicate the size of the nodes, with larger nodes representing more data points.

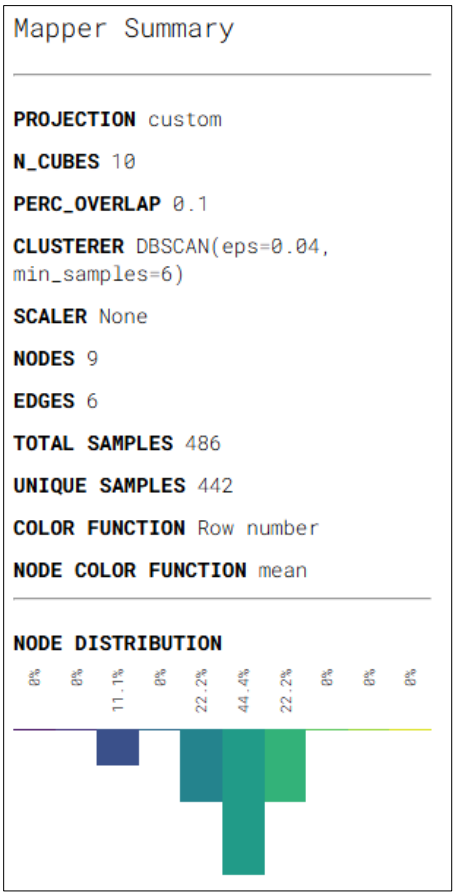


**Figure 7** The Mapper Summary obtained from the visualization of the mapper graph of the data using Kepler Mapper [61].

The size of the 9 clusters detected for the data set has been tabulated in Table 1. The isolated cluster 1 contains 5 data points and Clusters 3 and 5 are the largest, each containing 90 data points.

**Table 1** Size of the detected clusters

| Cluster No. | Size |
|---|---|
| 1 | 5 |
| 2 | 23 |
| 3 | 90 |
| 4 | 80 |
| 5 | 90 |
| 6 | 87 |
| 7 | 40 |
| 8 | 43 |
| 9 | 28 |

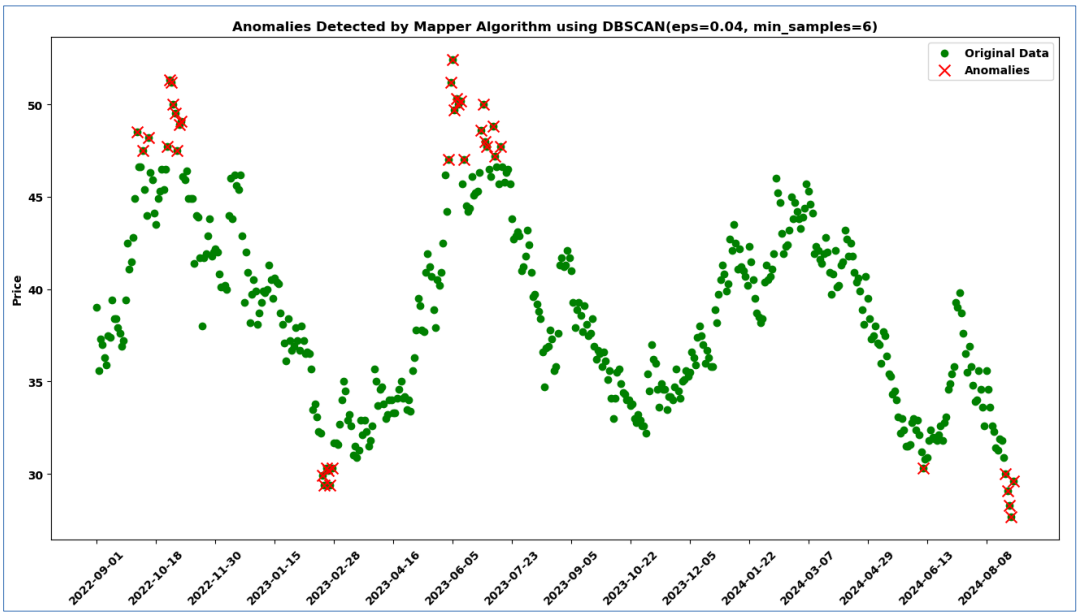


**Figure 8** Anomalies calculated by Mapper Algorithm using DBSCAN Clustering (while *eps=0.04* and *min_samples=6*) have been marked with red crosses.

This plot shows the time series data points (in green) along with the detected anomalies (marked in red). The anomalies were identified by applying the Mapper algorithm in combination with the DBSCAN clustering algorithm.

- *Green Points*: These represent the original data points, which form the bulk of the data set and indicate normal behavior.
- *Red Crosses*: These are the data points that were not included in any Mapper nodes, and are thus identified as potential anomalies. These points deviate from the overall data structure as revealed by the Mapper clustering.

Key observations:

- *Anomalies at the extremes*: Notice that the red crosses tend to occur at the peaks and troughs of the time series, such as around the price values of 50 and 30. These could represent sharp deviations or unusual events in the data.
- *Clustering of anomalies*: The anomalies appear in groups, indicating that certain periods in the time series exhibit abnormal patterns (e.g., sharp rises or falls).

In the context of detecting anomalies in the INTRACO stock's closing prices, the Mapper algorithm offers several advantages over the One-Class SVM approach. While both methods are capable of identifying anomalies, Mapper provides a topological view of the data, capturing its underlying structure through clustering. This allows Mapper to detect anomalies based on deviations from the overall data shape, particularly in regions with complex or non-linear behavior. Unlike SVM, which focuses on separating normal and anomalous points based on a predefined boundary (with $nu = 0.04$ in this case), Mapper can reveal hidden patterns in the data, particularly during periods of market volatility or extreme price fluctuations. The Mapper-based anomalies tend to cluster in groups, highlighting sustained abnormal periods, while SVM anomalies may appear scattered, missing this temporal correlation. Mapper's reliance on DBSCAN clustering also ensures that anomalies are identified based on data density, making it robust in detecting sharp rises or falls in price that might be overlooked by SVM. Thus, Mapper is more effective in capturing the time series' structural complexity and detecting contextually relevant anomalies.

## 4. Conclusion

In this study, we successfully demonstrated the application of the Mapper algorithm combined with DBSCAN clustering for anomaly detection in financial time series data, using daily stock prices from the Dhaka Stock Exchange as a case study. The integration of Topological Data Analysis (TDA) with clustering enabled the identification of anomalies by visualizing the global structure of the data, uncovering unclustered points, isolated clusters, or disconnected nodes as potential outliers.

The Mapper-DBSCAN framework identified 44 anomalous data points, corresponding to extreme price movements that did not belong to any node in the Mapper graph. The graph revealed key structures and patterns in the data, with smaller isolated nodes suggesting unique price behaviors or outliers. The validation using SVM further corroborated these findings, although the Mapper-DBSCAN approach provided clearer visual separations between normal data and anomalies.

The Mapper-DBSCAN method is particularly well-suited for high-dimensional data, where traditional clustering techniques struggle to capture the full complexity of the data's structure. By leveraging both topological summaries and density-based clustering, the method can identify not only local anomalies but also patterns that may be globally significant. The Mapper graph offers a clear and interpretable representation of the data's structure, allowing for more intuitive anomaly detection.

The results of the Mapper algorithm depend on the choice of filter function, number of intervals, and overlap percentage, making parameter selection critical for obtaining meaningful results. While Mapper provides a visual overview, interpreting the significance of smaller clusters or isolated nodes may require expert knowledge of the data domain. Although suitable for high-dimensional data, the computational complexity of both the Mapper algorithm and DBSCAN increases with large datasets.

This study can be conduct further for developing methods to optimize the selection of Mapper and DBSCAN parameters for more accurate anomaly detection across different data sets, for experimenting with more advanced filter functions, such as those based on machine learning or domain-specific features, to improve anomaly detection accuracy, and for conducting a comparative study of Mapper with other TDA-based tools like persistence homology to assess their relative strengths and weaknesses for anomaly detection.

This approach has several promising applications:

- *Financial Market Monitoring*: Mapper-DBSCAN could be used to detect market anomalies, such as sudden price fluctuations, outlier trading patterns, or early signs of market crashes.
- *Fraud Detection*: The framework can be adapted to identify fraudulent activities in transaction data, such as outlier transactions or abnormal sequences of trades.
- *Healthcare Data*: In the medical field, it could be applied to detect abnormal patient records in high-dimensional health data, potentially flagging rare diseases or unusual responses to treatment.
- *Cybersecurity*: This method could be useful in identifying anomalous behavior in network traffic or security logs, detecting intrusions or irregular access patterns

## Compliance with ethical standards

### *Acknowledgments*

The Authors are deeply grateful to the Dean of Faculty of Science, University of Rajshahi for funding this study via Prof. Dr. Nasima Akhter (Grant No. A-1753/5/52/RU./Science-01/2023-2024) and to the Ministry of Science and Technology, Bangladesh for providing fellowship to Md. Morshed Bin Shiraj.

### *Disclosure of conflict of interest*

There is no conflict of interest to publish this study among the authors.

## References

[1] Shumway, R. H., Stoffer, D. S., & Stoffer, D. S. (2000). Time series analysis and its applications (Vol. 3). New York: springer.

[2] Chatfield, C. (2013). The analysis of time series: theory and practice. Springer.

[3] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM computing surveys (CSUR), 41(3), 1-58.

[4] Agrawal, S., & Agrawal, J. (2015). Survey on anomaly detection using data mining techniques. Procedia Computer Science, 60, 708-713.

[5] Xu, X., Liu, H., & Yao, M. (2019). Recent progress of anomaly detection. Complexity, 2019.

[6] Patcha, A., & Park, J. M. (2007). An overview of anomaly detection techniques: Existing solutions and latest technological trends. Computer networks, 51(12), 3448-3470.

[7] Pang, G., Shen, C., Cao, L., & Hengel, A. V. D. (2021). Deep learning for anomaly detection: A review. ACM computing surveys (CSUR), 54(2), 1-38.

[8] Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. Journal of Network and Computer Applications, 60, 19-31.

[9] Prasad, N. R., Almanza-Garcia, S., & Lu, T. T. (2010). Anomaly detection. Computers, Materials, & Continua, 14(1), 1-22.

[10] Ahmed, M., Mahmood, A. N., & Islam, M. R. (2016). A survey of anomaly detection techniques in financial domain. Future Generation Computer Systems, 55, 278-288.

[11] Uray, M., Giunti, B., Kerber, M., & Huber, S. (2023). Topological Data Analysis in smart manufacturing processes--A survey on the state of the art. arXiv preprint arXiv:2310.09319.

[12] Umeda, Y., Kaneko, J., & Kikuchi, H. (2019). Topological data analysis and its application to time-series data analysis. Fujitsu Scientific & Technical Journal, 55(2), 65-71.

[13] Ravishanker, N., & Chen, R. (2019). Topological data analysis (TDA) for time series. arXiv preprint arXiv:1909.10604.

[14] Kim, K., Kim, J., & Rinaldo, A. (2018). Time series featurization via topological data analysis. arXiv preprint arXiv:1812.02987.

[15] Singh, M. K., Chaube, S., Pant, S., Singh, S. K., & Kumar, A. (2023). An integrated image visibility graph and topological data analysis for extracting time series features. Decision Analytics Journal, 8, 100253.

[16] Karan, A., & Kaygun, A. (2021). Time series classification via topological data analysis. Expert Systems with Applications, 183, 115326.

[17] Lin, B. (2022). Topological data analysis in time series: Temporal filtration and application to single-cell genomics. Algorithms, 15(10), 371.

[18] Kindelan, R., Frías, J., Cerda, M., & Hitschfeld, N. (2023). A topological data analysis based classifier. Advances in Data Analysis and Classification, 1-46.

[19] Pilyugina, P., Rivera-Castro, R., & Burnaev, E. (2020). TOTOPO: Classifying univariate and multivariate time series with Topological Data Analysis. arXiv preprint arXiv:2010.05056.

[20] Tran, Q. H., & Hasegawa, Y. (2019). Topological time-series analysis with delay-variant embedding. Physical Review E, 99(3), 032209.

[21] Zeng, S., Graf, F., Hofer, C., & Kwitt, R. (2021). Topological attention for time series forecasting. Advances in neural information processing systems, 34, 24871-24882.

[22] Zhang, Y., Shi, Q., Zhu, J., Peng, J., & Li, H. (2021). Time series clustering with topological and geometric mixed distance. Mathematics, 9(9), 1046.

[23] Bubenik, P., & Bush, J. (2023). Topological feature selection for time series data. arXiv preprint arXiv:2310.17494.

[24] Grande, V. P., & Schaub, M. T. (2023). Topological point cloud clustering. arXiv preprint arXiv:2303.16716.

[25] Combs, K., & Bihl, T. (2023). Clustering and Topological Data Analysis: Comparison and Application.

[26] Islambekov, U., Yuvaraj, M., & Gel, Y. R. (2020). Harnessing the power of topological data analysis to detect change points. Environmetrics, 31(1), e2612.

[27] Zheng, X., Mak, S., & Xie, Y. (2021). Online high-dimensional change-point detection using topological data analysis. arXiv preprint arXiv:2103.00117.

[28] Zheng, X., Mak, S., Xie, L., & Xie, Y. (2023). Percept: a new online change-point detection method using topological data analysis. Technometrics, 65(2), 162-178.

[29] Islambekov, U., Pathirana, H., Khormali, O., Akçora, C., & Smirnova, E. (2024). A topological approach for capturing high-order interactions in graph data with applications to anomaly detection in time-varying cryptocurrency transaction graphs. Foundations of Data Science, 0-0.

[30] Tempelman, J. R., & Khasawneh, F. A. (2020). A look into chaos detection through topological data analysis. Physica D: Nonlinear Phenomena, 406, 132446.

[31] Costa, J. P., & Grbac, T. G. The topological data analysis of time series failure data during the software evolution.

[32] Zhou, Y., Ren, H., Li, Z., & Pedrycz, W. (2021). An anomaly detection framework for time series data: An interval-based approach. Knowledge-Based Systems, 228, 107153.

[33] Turiel, J., Barucca, P., & Aste, T. (2022). Simplicial Persistence of Financial Markets: Filtering, Generative Processes and Structural Risk. Entropy, 24(10), 1482.

[34] Aguilar, A., & Ensor, K. (2020). Topology data analysis using mean persistence landscapes in financial crashes. Journal of Mathematical Finance, 10(4), 648-678.

[35] Yen, P. T. W., & Cheong, S. A. (2021). Using topological data analysis (TDA) and persistent homology to analyze the stock markets in Singapore and Taiwan. Frontiers in Physics, 9, 572216.

[36] Akingbade, S. W., Gidea, M., Manzi, M., & Nateghi, V. (2024). Why topological data analysis detects financial bubbles?. Communications in Nonlinear Science and Numerical Simulation, 128, 107665.

[37] Majumdar, S., & Laha, A. K. (2020). Clustering and classification of time series using topological data analysis with applications to finance. Expert Systems with Applications, 162, 113868.

[38] Goel, A., Pasricha, P., & Mehra, A. (2020). Topological data analysis in investment decisions. Expert Systems with Applications, 147, 113222.

[39] Kulkarni, S., Pharasi, H. K., Vijayaraghavan, S., Kumar, S., Chakraborti, A., & Samal, A. (2024). Investigation of Indian stock markets using topological data analysis and geometry-inspired network measures. Physica A: Statistical Mechanics and its Applications, 643, 129785.

[40] Guo, H., Ming, Z., & Xing, B. (2023). Topological data analysis of Chinese stocks' dynamic correlations under major public events. Frontiers in Physics, 11, 1253953.

[41] Guo, H., Yu, H., An, Q., & Zhang, X. (2022). Risk analysis of China's stock markets based on topological data structures. Procedia Computer Science, 202, 203-216.

[42] Prabowo, N. A., Widyanto, R. A., Hanafi, M., Pujiarto, B., & Avizenna, M. H. (2021). With topological data analysis, predicting stock market crashes. International Journal of Informatics and Information Systems, 4(1), 63-70.

[43] Oseko, N. N., Omondi, A. G., Onyango, H. D., Olwa, D. A., Maina, G., Morara, M. O., & Thiong'o, K. M. (2024). Forecasting Financial Crisis using Topological Data Analysis Approach. African Scientific Annual Review, 1(Mathematics 1), 1-17.

[44] Gidea, M., & Katz, Y. (2018). Topological data analysis of financial time series: Landscapes of crashes. Physica A: Statistical Mechanics and its Applications, 491, 820-834.

[45] Guo, H., Xia, S., An, Q., Zhang, X., Sun, W., & Zhao, X. (2020). Empirical study of financial crises based on topological data analysis. Physica A: Statistical Mechanics and its Applications, 558, 124956.

[46] Gong, X., Tian, W., & Li, B. (2021). Warning Ahead of Market Crashes: The Application of Topological Data Analysis. Available at SSRN 3878119.

[47] Miranda, V., & Zhao, L. (2020). Topological data analysis for time series changing point detection. In Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery: Volume 2 (pp. 194-203). Springer International Publishing.

[48] Islambekov, U., Yuvaraj, M., & Gel, Y. R. (2019). Harnessing the power of Topological Data Analysis to detect change points in time series. arXiv preprint arXiv:1910.12939.

[49] Rai, A., Sharma, B. N., Luwang, S. R., Nurujjaman, M., & Majhi, S. (2024). Identifying Extreme Events in the Stock Market: A Topological Data Analysis. arXiv preprint arXiv:2405.16052.

[50] Gidea, M. (2017). Topological data analysis of critical transitions in financial networks. In 3rd International Winter School and Conference on Network Science: NetSci-X 2017 3 (pp. 47-59). Springer International Publishing.

[51] Gidea, M., Goldsmith, D., Katz, Y., Roldan, P., & Shmalo, Y. (2020). Topological recognition of critical transitions in time series of cryptocurrencies. Physica A: Statistical mechanics and its applications, 548, 123843.

[52] Davies, T. (2022). Topological Data Analysis for Anomaly Detection in Host-Based Logs. arXiv preprint arXiv:2204.12919.

[53] Nguyen, N. K. K., & Bui, M. (2021). Detecting anomalies in the dynamics of a market index with topological data analysis. International Journal of Systematic Innovation, 6(6), 37-50.

[54] Sun, C. (2020, August). Exploration of mapper-a method for topological data analysis. In 2020 International Conference on Information Science, Parallel and Distributed Systems (ISPDS) (pp. 142-145). IEEE.

[55] Lee, D., & Jung, J. H. (2023). A Node Prediction Algorithm with the Mapper Method Based on DBSCAN and GIOTTO-TDA. Journal of the Korean Society for Industrial and Applied Mathematics, 27(4), 324-341.

[56] Carlsson, G. (2020). Topological methods for data modelling. Nature Reviews Physics, 2(12), 697-708.

[57] Dłotko, P., Qiu, W., & Rudkin, S. T. (2024). Financial ratios and stock returns reappraised through a topological data analysis lens. The European Journal of Finance, 30(1), 53-77.

[58] Saha, S. (2012). Stock market crash of Bangladesh in 2010-11: Reasons & roles of regulators.

[59] Molla, Md. (2019). Cause and Effect Analysis of Stock Market Crash during 2010 11 and Investors Impression The Case of Dhaka Stock Exchange. 29. 1-15. 10.5281/ZENODO.3734784.

[60] Habib, A. (2023, January 10). How Bangladesh's stock market remains an outlier. The Daily Star. Retrieved from https://www.thedailystar.net/business/economy/news/how-bangladeshs-stock-market-remains-outlier-3216941

[61] Van Veen, H. J., Saul, N., Eargle, D., & Mangham, S. W. (2019, October 14). Kepler Mapper: A flexible Python implementation of the Mapper algorithm (Version 1.4.1). Zenodo. http://doi.org/10.5281/zenodo.4077395