# Financial Scam/Fraud Detection Using Topological Data Analysis

Taiwo Onitiju
*School of Computing*
*University of North Florida*
Jacksonville, Florida, USA
N01578746@unf.edu

This research plan aims to investigate the use of Topological Data Analysis (TDA) in detecting financial scams. Conventional fraud detection techniques find difficulty in identifying convoluted patterns in transaction data which results in individual and organizational loss in finances. This study offers an evaluation of the effectiveness of TDA as a new approach to locating hidden abnormalities that would be a sign of fraudulent activities, comparing its performance to traditional machine learning models. By using publicly available financial transaction data, this research will explore how TDA can enhance detection precision as well as reduce the number of false positives thus contributing needed insight and improving financial security in an ever-growing financial landscape [1].

***Index Terms*—Topological Data Analysis, Fraud Detection, Security, Finance, AI Machine Learning**

## I. Introduction:

An evergrowing concern in the present-day world is scams, particularly financial scams which appear to trend alongside the rise of online financial transactions and operations. Due to this, the need to have up-to-date systems and techniques for detecting these scams is crucial for protecting companies and individuals from financial loss, keeping trust and faith in the global financial system, and dissuading more individuals and groups from viewing scamming as an easy avenue for profit gain. Conventional techniques of scam detection often rely on statistical techniques in combination with machine learning models which have proven effective, but many of these approaches often overlook complex patterns hidden within the data.

Topological Data Analysis (TDA) is an emerging sector in data science that brings a unique way to see such complex patterns by analyzing the shape of the data. TDA offers a collection of techniques that revolves around the idea of data having a shape that can visualize relationships other methods may miss [2]. This research dives into how TDA can be used to detect financial scams by flagging unusual data structures that could indicate fraudulent activity.

The main goal of this study is not to revolutionize financial security but rather to investigate the effectiveness of TDA in improving financial scam detection by taking advantage of its capability to analyze intricate forms within financial datasets. As more complex techniques are developed for committing financial scams as well as a practical need for new mechanisms that can aid existing techniques, creating more diverse detection methods is important. This research is especially important due to the increase in financial fraud cases as well as the increased complexity of scams that are getting better at bypassing traditional security measures, so thereby, utilizing TDA, this study will contribute to a better understanding of fraud patterns while also improving detection processes in an effective way [3].

## II. Background:

Financial scam detection is a wide-ranging field that crosses cybersecurity, fiance, and data science. It involves identifying fraudulent activities such as scams, phishing, identity theft, etc. within the financial zone that would result in financial loss. As e-transactions have snowballed through the popularity of e-commerce and online banking, the requirement for reliable fraud detection techniques has become more important [5].

Factually, financial scam detection has depended primarily on a rule-based system, which marked transactions based on preset thresholds such as transaction frequency, irregular geographic locations, or abnormally high amounts deposited or withdrawals. A drawback of such systems would be their restricted flexibility and often high reports of false positives. To address this, the incorporation of machine learning models has proven an optimal solution, taking advantage of large amounts of data to find patterns related to fraudulent activities [6].

Statistical models such as logistic regression and neural networks that are being used, while effective often over-rely on feature engineering where specialists need to specify important characteristics for detecting fraud. This can be restricting especially when complex fraud patterns come into play. In addition, such models are susceptible to hostile attacks, in which scammers adapt their behavior to prevent detection.

Topological Data Analysis is a rising technique stemming from the mathematical sector that aims to address some of the limitations discussed earlier. TDA studies the shape of data to answer questions through hidden patterns analysis. This approach is already being applied in different areas such as medicine, genomics, and natural language processing among many others, but its use in financial fraud detection is a fairly new approach with the potential to offer useful insight that would aid in effectively combating fraud attacks [2].

Compared to past research, there has rarely been a use for implementing TDA techniques to improve the detection of financial scams. Since TDA focuses on the shape and link between transactions in a manner that would aid machine learning models, it becomes a new viewpoint through which transaction data could be analyzed. This study's purpose is to make use of TDA to locate complex fraudulent patterns, stacking another layer/check that conventional detections methods may not catch.

## III. Objective:

The main goal of this research is to investigate the use of Topological Data Analysis in improving financial scam detection. Notably, this study looks to know if TDA can identify anomalies in financial transactions that are a sign of scams thus giving a favorable approach to conventional fraud detection techniques [4]. The research is structured by the following questions:

- How can Topological Data Analysis be used in financial transaction data to locate scam/fraud patterns: This will look at the techniques of using TDA in financial datasets and the steps needed to be taken in creating a topological representation of the data.
- What kinds of financial scams/fraud can be detected using TDA and how effective will they be as compared to traditional machine learning methods: This will look at knowing the limitations and strengths of TDA in detecting different kinds of financial fraud, as well as if it can collect patterns that are not seen by traditional methods.
- Does using TDA improve the accuracy of scam detection vs existing models: This will look to compare how well TDA detection methods are to current techniques and to see if it reduces false positives or catches scams that would otherwise go undetected.

The steps to answering the above questions would involve using real-world financial data to test and see how effective TDA is. By incorporating this method into present detection systems, this research will provide an in-depth understanding of how the shape of financial transaction data can bring to light hidden fraudulent activity, and if TDA can serve as an important tool for future financial fraud detection methods.

## IV. Methods:

Data Collection is the first step that will be taken towards answering the questions proposed in the previous section. Financial datasets that would include both legitimate and fraudulent transactions will be used, retrieved from publicly available databases such as Kaggle. The data will be preprocessed to remove irrelevant entries as well as to maintain a uniform structure. Key information such as transaction amount, geographical location, sender and receiver information, etc. will be extracted for use in feature engineering, using topological features and standard features for comparison [7].

The next step would be the application of Topological Data Analysis to the processed data which would be in the form suitable to TDA, likely involving mapping the transaction groups to high-dimentional spaces. Using TDA techniques such as Persistent Homology or Mapper Graph, the structure and shape of the data will be identified to observe for any abnormal behavior i.e. deviations from normal transaction patterns which would mean potential scams [8].

The next step would involve making use of the topological structure in combination with traditional features, passing it off to various supervised and unsupervised machine learning models such as Logistic Regression and Neural Networks alongside anomaly detection methods like Isolation Forest and Z-score to identify abnormalities.

Once completed, the models will be rated based on performance such as accuracy, and F1-score among others. The idea is to compare the performance of traditional techniques against those enhanced with TDA. Investigating the number of false positives will also be looked into and compared as it plays an important role in regards to its use in fraud detection. The study will also measure if TDA contributes to identifying complex patterns in fraud that are missed by conventional methods.

Utilizing Pythons libraries for Topological Data Analysis such as KeplerMapper, as well as basic machine learning libraries like TensorFlow for model evaluation, a solid foundation for implementation has been provided.

Through these techniques, the study can answer the research questions and determine the potential and impact scale of TDA in detecting financial scams by performance comparison, accuracy, and contribution to the field.

## V. Data:

As mentioned previously, the data needed for this research will mainly be made up of financial transaction datasets which is key for evaluating fraudulent patterns. The dataset will be sourced publicly from available databases and based on the test scenario, it will be modified to emulate a real-world financial transaction with fraudulent activities. The datasets that contain both fraudulent and non-fraudulent transactions will be acquired from reputable sources such as Kaggle and IEEE Dataport following any data-sharing regulation that may apply. The data would then be preprocessed i.e. undergo cleaning and parsing which can be accomplished using a MySQL database to capture relevant attributes needed for topological representation before then integrating into machine learning and finally a performance analysis of traditional machine learning approach that relies on just rule-based techniques vs TDA enhanced machine learning [10].

## VI. Execution Plan:

Given the limited time available, a comprehensive study that would involve extensive data collection, multiple model testing, etc. cannot be achieved, rather a condensed 5-week execution plan is laid out as follows:

1) Phase 1: Research Design and Data Collection (Week 1)

- Days 1-3: Summarize research objectives, formulate research plan, and retrieve publicly available financial transaction datasets.
- Days 4-7: Commence data preprocessing i.e. cleaning, retrieving key attributes, etc.

2) Phase 2: Topological Data Analysis Application (Week 2)
    - Days 8-10: Transform processed data for TDA as well as perform feature engineering for machine learning.
    - Days 11-14: Apply TDA persistent Homology or Mapper Graph Techniques.
3) Phase 3: Initial Testing and Model Development (Week 3)
    - Days 15-17: Train machine learning models
    - Days 18-21: Tests and evaluates the performance of both conventional and TDA-enhanced models.
4) Phase 4: Model Refinement(Week 4)
    - Days 22-24: Adjust the model based on performance i.e. accuracy, F1-score, etc.
    - Days 25-28: Finalize evaluation, access advantages and disadvantages of models. Visualize findings using visual aid tools.
5) Phase 5: Final Write-Up (Week 5)
    - Days 29-31: Organize results into a research report explaining the techniques used, data analysis, and major findings.
    - Days 32-35: Review and Revision before potential submission to an academic journal such as the Journal of Financial Crime, IEEE Transactions on Information Forensics and Security, or Journal of Cybersecurity.

This schedule serves to ensure focused and efficient progress while also delivering a robust analysis within the limited time frame. Addition time would be detected toward further improvement on the model through the implementation of additional TDA techniques, learning models, and anomaly detection methods. Further updates would also be dedicated to the research paper if needed.

## VII. Ethical Considerations:

Due to the nature of the data being used. Information present such as personal names, account numbers, etc. must be carefully handled. Luckily, many reputable databases that contain these real-world example datasets maintain anonymity by omitting or modifying the data to ensure individuals and organizations are not put at risk while also ensuring the data remains accurate for use in testing and analysis. Any restriction or licensing agreements imposed by the data provider will be adhered to and the dataset will be used in accordance with it [11]. Clear information on how the data is used will be laid out in the research from data retrieval to analysis, to ensure transparency and accountability for future researchers to build upon the work ethically.

## VIII. Expected Results:

This section aims to cover the current findings from the use of Topological Data Analysis (TDA) in conjunction with conventional machine learning models on various bank transaction data sets retrieved from Kaggle. The dataset contains two thousand extracted bank account statements from various banks with personal identifying information omitted to ensure data privacy. Each transaction is provided in the form (See Table 1) such that key factors can be retrieved and used for TDA and machine learning analysis.

TABLE I
Sample Bank Transaction

| Transaction | Sample Data |
|---|---|
| Account No | 1196711 |
| Date | 10/5/2015 |
| Transaction Details | CASHPMT/GURGAON/SELF |
| Check Number | 873917 |
| Value Date | 10/5/2015 |
| Withdrawal Amount | $100,000 |
| Deposit Amount | $0 |
| Balance Amount | $7,349,950 |

The sample data used is split into multiple forms with the outcomes displayed in a descriptive form backed up by both visual and statistical representation.

### A. *Sample Data Overview:*

#### 1) *Samplebank.csv (Baseline):*

- **Key Points:**
    - Number of Transactions: 119
    - Transaction Type: Mixed account numbers, amounts, details
    - Manually Inserted Anomalies: None

#### 2) *Samplebank-1.csv (Equivalent Transactions):*

- **Key Points:**
    - Number of Transactions: 119
    - Transaction Type: Uniform account numbers, amounts, deposit, and withdrawals
    - Manually Inserted Anomalies: None

#### 3) *Samplebank-2.csv (Anomaly Inserted Transactions):*

- **Key Points:**
    - Number of Transactions: 119
    - Transaction Type: Uniform account numbers, mostly similar deposit and withdrawal amounts, mostly identical transaction details
    - Manually Inserted Anomalies: 2 fraud transactions added

#### 4) *Samplebank-500.csv (Larger Mixed Data Set):*

- Key Points:
    - Number of Transaction: 500
    - Transaction Type: Mixed account numbers, amounts, details
    - Manually Inserted Anomalies: None

### B. *Expected Outcome:*

Implementing TDA analysis, I expect to locate transaction clusters that display similar attributes, while detected anomalies being the ones that deviate from the cluster. I also expect the machine learning models will achieve a higher accuracy in detecting anomalies when TDA is combined with it. Overall, a complete analysis should show the robustness of TDA as an added pre-processing tool for datasets based on the results obtained.

### C. *Results:*

*1) Applying TDA Analysis:* The TDA analysis of the Samplebank.csv file which included 119 transactions with mixed types and no manually added anomalies revealed various cluster groups (see fig 1) with the presence of minor noise or abnormal activity.

The analysis of the Samplebank-1.csv file which contained 119 identical account numbers and repeated transactions revealed a uniform topology, highlighting redundancy which would require a need for further investigation.

The Samplebank-2.csv file contained 117 transactions with two manually inserted fraud transactions which was presented as deviated cluster points outside the primary structure.

The Samplebank-500.csv file contained 500 mixed transactions to emulate Samplebank.csv but on a larger scale. Similar behavior was observed in which there was presence of minor noise with various cluster groups.

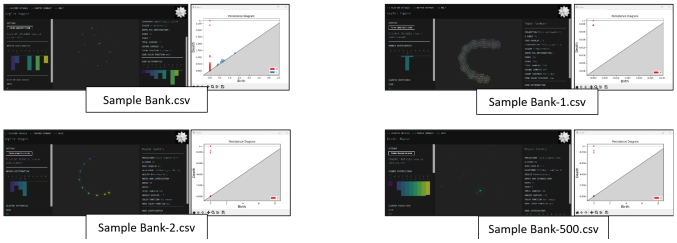

Fig. 1. TDA Analysis and Persistence Diagram of each sample dataset.

*2) Applying Machine Learning:* The goal was to further visualize and detect anomalies present within the transaction dataset using both a z-score distribution graph as well as a 3D Principal Component Analysis (PCA) model to reduce high dimensional data into a 3D space while keeping important features. From the displayed figures (see Fig. 2), Samplebank-1.csv and Samplebank-2.csv display an elongated spread that reflects the uniformity in the data which aligns with how the transaction data is structured. Samplebank.csv and Samplebank-500.csv share similarities in display as the data is spread out due to the mixed transactions present. Anomalies found are also highlighted for each with the z-score distribution graph displaying it in the form of transactions over the threshold (dashed red line).

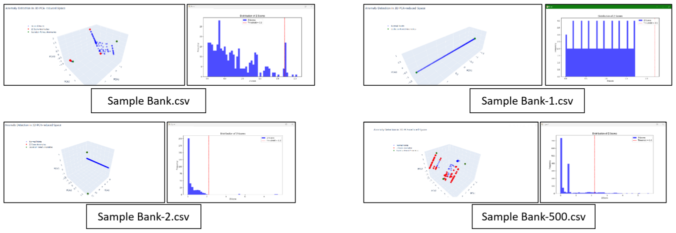

Fig. 2. Anomaly Detection in 3D PCA with Z-Score Distribution Graph.

*3) Final Report:* From the table (see Table II), the analysis of the various transaction datasets can be summed up as:

- **Samplebank.csv:**
  - Z-score detected 20 anomalies present, displaying a higher than average sensitivity to abnormalities which could lead to higher detection of false positives.
  - Isolation forest detected 2 anomalies indicating a less lenient detection system which could lead to missing true positives.
  - Logistic Regression detected 35 anomalies offering the highest sensitivity.
  - 1 anomaly was highlighted by TDA showing its main goal of displaying topological features that may go unseen by other models.
  - Using a combined approach, a total of 37 anomalies out of the 119 transactions were detected.
- **Samplebank-1.csv:**
  - Z-score detected no anomalies present indicating no deviations were found.
  - Isolation forest detected 2 anomalies indicating potential abnormalities.
  - Logistic Regression detected 0 anomalies matching the z-score.
  - 0 anomaly were detected by TDA showing no topological abnormalities detected.
- **Samplebank-2.csv:**
  - With the inclusion of 2 fraudulent transactions, all model detected the same anomalies indicating a uniform agreement.
- **Samplebank-500.csv:**
  - Z-score detected 77 anomalies present, indicating higher detection of false positives.
  - Isolation forest detected 5 anomalies also indicating less lenient detection system which could lead to missing true positives.
  - Logistic Regression detected 99 anomalies offering the highest sensitivity and likely capturing noise.
  - 2 anomalies were highlighted by TDA's ability to filter out noise and focus on the important features.
  - Using a combined approach, a total of 99 anomalies out of the 500 transactions were detected.

TABLE II
DATA

| | |
|---|---|
| **Sample Bank.csv** | |
| Shape of data after PCA | **(119, 3)** |
| Anomalies detected by Z-score | **20** |
| Anomalies detected by Isolation Forest | **2** |
| Anomalies detected by Logistic regression | **35** |
| Anomalies detected by TDA | **1** |
| Anomalies detected by Combined approach | **37** |
| **Sample Bank-1.csv** | |
| Shape of data after PCA | **(119, 3)** |
| Anomalies detected by Z-score | **0** |
| Anomalies detected by Isolation Forest | **2** |
| Anomalies detected by Logistic regression | **0** |
| Anomalies detected by TDA | **0** |
| Anomalies detected by Combined approach | **2** |
| **Sample Bank-2.csv** | |
| Shape of data after PCA | **(119, 3)** |
| Anomalies detected by Z-score | **2** |
| Anomalies detected by Isolation Forest | **2** |
| Anomalies detected by Logistic regression | **2** |
| Anomalies detected by TDA | **2** |
| Anomalies detected by Combined approach | **2** |
| **Sample Bank-500.csv** | |
| Shape of data after PCA | **(119, 3)** |
| Anomalies detected by Z-score | **72** |
| Anomalies detected by Isolation Forest | **5** |
| Anomalies detected by Logistic regression | **99** |
| Anomalies detected by TDA | **2** |
| Anomalies detected by Combined approach | **99** |

## IX. CONCLUSION:

The research on ”Financial Scam/Fraud Detection using Topological Data Analysis” offers an opportunity to improve fraudulent activity detection within financial systems. By using Topological Data Analysis, this study aims to offer an inventive solution to a serious issue that affects individuals, businesses, and financial institutions.

As fraudsters become increasingly complex in their tactics, conventional detection methods are slowly lagging behind in recognizing the patterns and anomalies related to fraudulent transactions. TDA offers a potential layer of security to locate such complex patterns not easily seen [9].

The results from using TDA are meant to show its unique ability to detect hidden anomalies through it's analysis of the shape. By complementing conventional methods such as Isolation Forest, Logistic Regression, and Z-score, hidden relationships that signal fraudulent behavior can be seen through this robust and modified detection system .

In conclusion, implementing TDA provides a favorable avenue in tackling increasingly complex scams. By highlighting the interaction between machine learning models and TDA, the importance of multi-layered detection methods with the potential to attune with evolving threat tactics is offered.

## REFERENCES

[1] G. Carlsson, “Topology and data,” *Bulletin of the American Mathematical Society*, vol. 46, pp. 255–308, 02 2009.

[2] A. Zomorodian and G. Carlsson, “Computing persistent homology,” *Discrete Computational Geometry*, vol. 33, pp. 249–274, 2005.

[3] C. Phua, V. Y. Lee, K. Smith, and R. Gayler, “A survey of data mining techniques for social network analysis,” *ACM Computing Surveys*, vol. 41, pp. 1–30, 2010.

[4] M. Ahmed, A. N. Mahmood, and J. Hu, “A survey of network anomaly detection techniques,” *Journal of Network and Computer Applications*, vol. 60, pp. 1–21, 2016.

[5] M. Fitting, “Financial fraud detection using machine learning: A review,” *International Journal of Computer Applications*, vol. 159, pp. 5–10, 2017.

[6] C. C. Albrecht and K. Garrison, “An overview of fraud detection: Techniques and applications,” *Journal of Economic Perspectives*, vol. 32, pp. 213–224, 2018.

[7] V. J. Hodge and J. Austin, “A survey of outlier detection methodologies,” *Artificial Intelligence Review*, vol. 22, pp. 85–126, 2004.

[8] B. Iglewicz and D. C. Hoaglin, *How to Detect and Handle Outliers*. SAGE Publications, 1993.

[9] H.-P. Kriegel, P. Kroger, and A. Zimek, “Geometric approaches to anomaly detection,” *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 4, pp. 31–39, 2011.

[10] S. Dey and M. Bhatia, “Anomaly detection in financial transactions using topological data analysis,” *International Journal of Information Management*, vol. 57, p. 102131, 2021.

[11] B. J. Zikmund-Fisher, A. Fagerlin, and P. A. Ubel, “Ethical considerations in the use of data for research,” *Health Affairs*, vol. 29, no. 3, pp. 540–546, 2010.

[1] [2] [3] [4] [5] [6] [7] [8] [9] [10] [11]