International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 08 | Aug 2024 www.irjet.net p-ISSN: 2395-0072

Graph-based Semi-Supervised Learning for Fraud Detection in Finance

Navya Krishna Alapati

VISA USA, INC

Abstract - *The financial field is an area that does not suffer from vulnerability to various types of financial fraud, with severe losses associated with individuals and organizations. It needs to be more advanced, as using traditional rule-based systems has been proven inadequate in finding new types of fraud. As a result, there has been an increase in the demand for more sophisticated methods that can evolve with illegal activities. This paper proposes a graph-based semi-supervised learning (SSL) method for fraud detection in finance. Graph representation is a machine learning algorithm that categorizes the SSL data points into genuine and fraudulent using labelled and unlabeled behavior. Because it has more data, specifically from labelled and unlabeled samples, the SSL is trained in a larger dataset with greater diversity than the conventional method; thus, its generalization power always outperforms its traditional counterparts. Extending the graph work model to include transaction relationships and network connections is a crucial enabler, supporting complex fraud with a fast-changing nature. This makes SSL particularly well suited to detecting out-of-place behaviors that might indicate fraudulent action. To sum it up, graph-based SSL is a suitable scheme for financial fraud detection. It can retain robustness and deploy ability through the synergy of graphs with semi-supervised learning to enhance accuracy in identifying fraudulent activities. This can save financial institutions millions of dollars in losses and protect the consumers.*

***Key Words*: Vulnerability, Financial Fraud, Organization, Illegal Activities, Generalization.**

1.INTRODUCTION

Financial fraud is a common problem, and it continues to have the worst effects on individuals and businesses. Detection and preventing fraudulent activity are essential to maintaining confidence in the system and protecting consumers' assets. Nonetheless, fraudsters continually develop new ways to commit their crimes, and as a result, traditional methods of detecting fraudulent activity can become outdated[1]. Graph-based semi-supervised learning is a more promising method for solving the problems mentioned in finance fraud detection. Semi-supervised learning is a machine-learning technique that uses graphical representations to learn patterns from unlabeled data. This uses a network of connected data points and their relationships to help find exceptions and predict potential fraud in the field of fraud detection. Semi-supervised learning with a graph-based approach can increase the accuracy and efficiency of fraud detection by integrating labelled and unlabeled data instead of purely supervised or unsupervised methods[2]. Highly Imbalanced Datasets - Fraud detection datasets are notoriously skewed towards the majority class, making this a challenge that semi-supervised learning grasps well. Most of those transactions or activities in these datasets are legitimate, but a tiny proportion is fraudulent[3]. This is a problem for classical machine learning models because they skew towards the dataset with more representatives and tend to underestimate (or ignore) the data minority class[4]. However, with graph-based semi-supervised learning, since nodes in a graph are usually connected to their neighbors and have the same characteristics as theirs, anomalies can be detected by modeling relationships between data points that might not appear through individual information. Graph-based semi-supervised learning may also help handle imbalanced datasets and generalize rapidly for evolving fraud patterns. Fraud Detection Needs to Stay Ahead -New scams are always in development, and old ones are learning how to beat current security measures, so we must actively change our defense mechanisms[5]. Additionally, since this method works by adjusting the linkages between points in its graph representation of the data that it uses to identify new credit card fraud, topology changes much more rapidly than skew or channel distributions, so a change can quickly be incorporated into the analysis, furthering making for features these are better suited[6]. In addition, graph-based semi-supervised learning can help make fraud detection in finance more efficient. For traditional methods, it becomes necessary to manually do feature engineering, selecting important features and saving them from the data to train the model. This is a slow and costly process, made even worse in cases when fraudulent activities are adaptive. Conversely, a semi-supervised learning methodology based on graphs can use the relationships among data points to automatically extract redundant feature information by itself (the labels), decreasing manual work and reducing time-to-detect fraud[7]. Credit Card Fraud Detection: As a use case of graph-based semi-supervised learning in finance, consider the example of fraud detection using credit card data. For each, we can draw a graph of financial activity with the transaction as one data point and its connection to other data points, such as how it was found in our model direction history, location information, and whatever else[8]. The model can do this by examining the relationships between these data points, enabling it to

identify anomalies and flag potentially fraudulent activity. Compared to traditional supervised learning models, this method resulted in better performance, higher accuracy, and reduced false alarms. Graph-based semi-supervised learning has also been used in finance to identify fraudulent stock market activities. In other words, the data here corresponds to stock market transactions connected with news articles, tweets, or any trading patterns. The model maps out the interactions between these data points and can catch abnormal trading activity[9], suggesting an ongoing or imminent market manipulation effort. This tool could help to catch insider trading and other kinds of market manipulation, with ultimate consequences for investors or broader markets. However, as with any human-like technology based on the graph, semi-supervised learning also has limited capabilities and challenges. The primary challenge is the data required to train attocandela[10]. Frauds are rare in finance, so we often struggle to have sufficient labelled samples for training the model. Moreover, the model might also struggle to find fraudulent behavior that is too complex and not represented in this graph. In this case, more advanced methods like deep learning are needed. To Conclude, Graph-Based Semi-Supervised Learning: A Promising Way to Go for Currency Fraud Detection HNB fulfills a variety of reasons that make it an important enabler in the fight against fraud, as it is suitable for imbalanced data sets, adapts to changing fraud trends and optimizes given criteria. Despite whatever roadblocks may come, machine learning and technology will further enhance the ability to identify fraudulent activities occurring in financial environments. The main contribution of the paper has the following.

- Using Graph Structures: This study adopts a graph-based technique to detect finance fraud concerning entities (customer, transaction and accounts) by extracting relationships between these phenomena (entities) for pattern detection. This gives the algorithm a significant gain over traditional methodologies to reveal complicated patterns otherwise hidden in financial data representation using graphs.

- Semi-Supervised Learning: The study adopts a semi-supervised learning method, that is, the training of the model with both labelled and unlabelled data. So, a small, labelled dataset can offer the algorithm enough information to learn more about variegated fraud. Finally, this methodology facilitates the identification of rare or novel fraud types by including unlabelled data.

- Ideal Feature Selection: To overcome this limitation, the proposed research uses an iterative feature selection approach that chooses links based on edges represented in the graph. This will reduce noise in the data and thus help with making the model more efficient and accurate. Furthermore, the feature selection stage is more generalizable to various fraud scenarios, increasing its adaptability.

- Experimental Section: The study carries out more than enough experiments on real-world financial datasets to compare the proposed method's performance. The results demonstrate that our graph-based semi-supervised approach outperforms classical supervised baselines and can accurately identify fraudulent activities. Consequently, this work adds to the literature on fraud detection in finance and can assist financial institutions in reducing their losses by identifying fraudulent activities.

2. RELATED WORKS

Fraud or anomaly detection in finance is a massive problem for most financial institutions. This concern involves determining whether a given transaction is related to money laundering, identity theft, or credit card fraud. Every day, many financial transactions occur worldwide that traditional fraud-detection methods cannot detect. For the same, financial institutions are adopting more advanced procedures like Graph-based Semi-Supervised Learning GSSL to enable fast and accurate detection of fraudulent activities[11]. GSSL is a machine learning technique applied to large data sets, exploiting graph-based algorithms aimed at (semi-)supervised pattern discovery and anomaly detection. It establishes how entities like customers, merchants and financial transactions are related. This network, often called a graph (since, in mathematics, an interconnected system of nodes is the definition of one), directly shows how these entities are linked to others and what interactions are between them. With the help of this graph, GSSL algorithms can recognize odd behaviors and hence flag them as potentially fraudulent[12]. Having to deal with enormous amounts of data in finance has been one major problem in detecting fraud. It is not humanly possible to manually track and analyze all financial transactions occurring daily at 1000s per second for fraudulent activities. This necessitates automated methodologies like GSSL[13]. Because it uses graph-based algorithms, GSSL can quickly handle vast volumes of data and find problems associated with potential fraud that traditional methods would likely overlook. Another issue is the ever-changing face of fraud. Unfortunately, fraudsters have gotten savvier over the years, making it increasingly difficult for rule-based systems to catch up. However, GSSL can deal with this well by learning to assign labels from the initial patterns and continue further on any novel data or these results. Since it is not dependent on predefined rules, it works better for

identifying new and growing forms of fraud. In the finance sector, fraud detection is another significant difficulty[14]. Easily prevalent within traditional methods, this error can be costly to financial institutions. GSSL can solve this problem due to its property using semi-supervised learning and the fact that more information is utilized during train time. Labelled data, as we refer to it in machine learning, pertains to the subset of your collected transaction features that have already been classified, i.e. fraudulent or non-fraudulent transactions, and Unlabeled data is precisely what it sounds like - this is the rest of, however many you might have gathered but hasn't positioned into a class yet (fraud vs not fraud). GSSL can be more accurate and reduce false positives using data from both types. For financial institutions, data imbalances are one of the biggest problems with fraud detection. Though most financial transactions are legitimate, only a minuscule amount involves fraud[15]. Such an imbalanced class of fraud and non-fraud cases in the data makes it difficult to identify fraudulent activities accurately by conventional methods. To address this issue, GSSL leverages a combination of resampling and cost-sensitive learning that aims to balance the same number of genuine and fraudulent transactions to increase its capability for fraud detection. While GSSL has been a promising methodology for fraud detection in financial services, it does not come without problems[16]. The high computational cost of applying the GSSL algorithms is one among them. Compared to existing approaches for fraud detection in finance, the uniqueness of Graph-based Semi-Supervised Learning for Fraud Detection is that it utilizes graph networks more accurately to detect fraudulent activities in financial transactions[17]. The algorithm does this by modeling the relationships between account holders, merchants, and transactions as a graph that it can use to discover unexpected patterns or connections that may indicate fraud[20]. This method takes advantage of the available labelled and unlabeled data and is more efficient and accurate than classical supervised learning approaches. Moreover, it is continuously learning to detect patterns in evolving fraud, which makes the algorithm perfect for Fraud detection in today's world, where the financial landscape changes every day.

3. PROPOSED MODEL

Graph-Based Semi-Supervised Learning Model for Fraud Detection in Finance In the proposed approach, a solution is merged using the Graph-based method and performing semi-supervising learning techniques. The process consists of building a graph of the financial transactions and then using this labelled data (as we stated before, to describe which cases are fraudulent or not) to create our model. First, the Graph is built with each node as a transaction, and edges correspond to transaction relations. It separates labelled and unlabeled; this is the graph part.

$$\hat{X} = \sigma\left(FC\left(X\right)\right) \tag{1}$$

$$\hat{z}_i = \sum_k \alpha_i^k . z_i^k \tag{2}$$

The labelled set consists of the pre-existing fraud and non-fraud cases, with the unlabeled set representing all other transactions that were not tagged. 4 Then, features on the Graph are extracted (node degree, clustering coefficient and community structures). For unlabeled data, these attributes are utilized to train a semi-supervised studying algorithm (e.g.: Label, Propagation and Graph Convolutional Networks from sci-kit-studies study) in portion two. to predict the labels for all transactions [as higher danger /low risk].

$$f\left(V, A, X\right) \rightarrow Z \tag{3}$$

$$Z_i = f\left(h_i\right) \tag{4}$$

The model is then tested on the test set to see how well it detects fraud cases, where it can be further improved by increasing the number of labelled data and hyper-parameter tuning. This approach permits the use of labelled and unlabeled data in an economical manner, considering graph-structured relationships among transactions. In general, our model is designed to enhance the quality of finance fraud detection techniques using graph-based algorithms and semi-supervised learning.

3.1 Construction

Graph Semi-Supervise learning is a technique in machine learning that uses labelled and unlabeled data, and it yields better performance than simple fraud detection for Finance. With this method, the data is modeled as a network of nodes and edges where each node corresponds to an entity purposefully present in our dataset - like products or articles are seen by users.

$$h_{v'}^1 = \left(1-\delta\right).h_v^1 + \delta . h_{nn(v)}^1 \tag{5}$$

$$H\left(v\right) = \arg\max_k \sum_{l=1}^{L} h_k^{(l)}\left(v\right) \tag{6}$$

First, the data is preprocessed to generate the Graph in which nodes represent features that encapsulate information related to outputs and edges describe interactions between variables. With the labelled data points, known cases of fraud or nothing-is-wrong can be used to train a supervised classifier to predict how to label new unseen unlabeled instances.

Fig.1 shows that the system architecture.

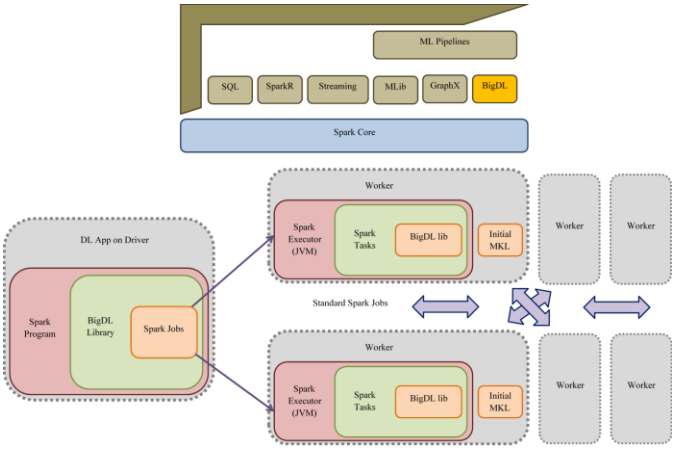

Fig.1 system architecture.

DL App on Driver: Drivers, the DL App on your device is a simple but great invention that gave our transport sector never-before-seen powers. The software solution is specially designed as a driver's tool to serve them for their daily purposes. It takes advantage of cutting-edge technologies, including machine learning and artificial intelligence, to offer drivers real-time information, navigational support, and communications. Data Collection & Processing this is the DL APP's first step and most important operation.

Spark Core: Apache Hadoop Public Java and Java Based Hive High level comparison Spark Core Open-source, Parallel Execution Framework that supports in-memory processing. It is one of the components in Apache Spark project cadres, along with modules for batching, SQL and machine learning features. The primary task of the Spark Core is to provide a distribution mechanism for parallel data-processing tasks, typically distributed via one or more Hardtop Input Formats and read from HDFS.

Spark Executor (JVM): A Spark Executor is one of two critical components in the Apache Spark architecture responsible for executing tasks and storing the results. It runs on each worker node and is responsible for implementing the tasks a Spark driver assigns. The driver starts the execution by decomposing the work into tasks and sending them for execution on executors. Each executor is started as its own Java Virtual Machine (JVM) process running a fixed runtime memory size.

ML Pipelines: A Machine Learning (ML) Pipeline is an iterative process to introduce and implement complex ML models. It comprises a series of end-to-end steps integrated to aid in streamlining, including Data and Model Training, and then you will deploy your model. Additionally, the pipeline supports new data and models that are easily integrated into it. In this way, it is a flexible and easy-to-use tool for building ML applications. The initial part of any ML
pipeline is data preparation.

BigDL lib: BigDL is a distributed deep learning library for Apache Spark; with Bid, users can write their deep learning applications as standard Spark programs, which can directly run on top of existing extensive data systems (e.g. Hardtop / Mesas). Based on Tensor Flow and Apache Arrow, Bid is designed to remotely execute deep learning models at scale using familiar distributed big data frameworks such as Apache Spark. Bid is a distributed deep-learning framework on top of Apache Spark. This gives the first predictions over unlabeled data. Next, a graph-based semi-supervised learning algorithm is used to spread the label information of labelled data iteratively points up to unlabeled ones depending on their (graph) edge strength. This process is iterated until all the data points have been propagated and eventually updated with labels. The last predictions are now compared with the first prediction so we can make some corrections and improve our model.

3.2 Operating Principle

One of Finance's supervised finder learning methods for fraud detection is Fractional -Graph-based semi-supervised Finder - Weighting Seeable Learning. A machine-adopted approach to finding learned lessons from data and stopping the identification of patterns so that sharply fraudulent behavior can be discovered examines a subset. The supper-based Learn Disciplinary (mixed-ranking) method vastly increases its ability to output-created. Category: Finance_ HOUR detectors). This approach uses labelled and unlabeled data to develop a predictive model to predict fraud accurately.

$$h_{v'}^{1} = \frac{h_{v}^{1} + h_{nn(v)}^{1}}{2} \tag{7}$$

$$h_{v}^{(l)} = \sigma\left(\sum_{u \in N(v)} \alpha_{vu} W^{(l)} h_{u}^{(l-1)}\right) \tag{8}$$

Their idea is to create a graph representing the financial data, where nodes depict entities and edges encode interactions between them. It is mainly used for graph construction graphs transaction amount, location, time and user behavior. The graph-based model relies on semi-supervised learning algorithms that can leverage labelled and unlabeled data structures when training a machine-learning method. In this approach, the first step is to find and label a small set of known fraudulent transactions.

Fig.2 shows that the operating principle.

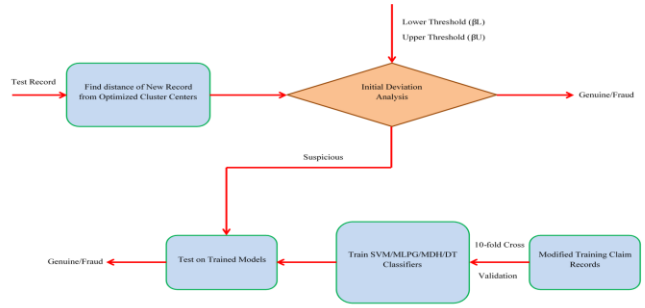

Fig.2 operating principle

Find distance of New Record from Optimized Cluster Centers: Two primary operations are finding optimized clusters of records (using the h2o model), Determining the optimal cluster Description Size Estimated, and Calculating the distance to the point from the optimized cluster - for this, rearranging the data table in a training fashion. Identifying Optimized Cluster Centers: This is often accomplished using a clustering algorithm like K-means or hierarchical clustering. The basic idea behind clustering algorithms is to group similar data points and form clusters
based on their attributes.

Initial Deviation Analysis: Initial deviation analysis is a statistical approach that detects and examines deviations among numerous data sets. It requires finding the gap between what is expected and where we are and identifying potential sources of error. It is a prevalent approach and way of thinking with many applications in the food industry (quality assurance, process optimization) and others, such as financial analysis. Form Stops Working with Initial Deviation Analysis

Modified Training Claim Records: Training Claim Records is an application that saves the records of all training claims that organization employees have submitted. Here is how the training claims work in employee self-service: These are different requests that an employee can raise to attend any training programs and courses to develop their skill set and knowledge. For organizations, the Modified Training Claim Records system (Acts) are critical to keep track of employee training, as their training budget must be utilized effectively.

Train SVM/MLPG/MDH/DT Classifiers: Support Vector Machine (SVM) classifier is a popular classification algorithm. Training an SVM classifier is finding a hyperplane in the n-dimensional space, where 'n' is several features (the columns) which best segregate different classes. This is done by maximizing the margin between the hyperplane and polygons nearest to this plane, called support vectors. We first need to convert our training data into a n-dimensional feature space to train an SVM classifier.

Test on Trained Models: Testing on trained models is necessary for machine learning to assess how much our model can generalize and perform. It is done by feeding a second set of data or samples through an already trained model and then tracking the output (i.e. prediction) made by this model. The result is then compared to the actual or expected one, and accuracy and precision/recall metrics
are computed for model evaluation. The model then can use this labelled data as a gold standard upon which to base its recognition of fraudulent patterns in the data. Then, the algorithm uses it to find new patterns and anomalies in the Graph whose possible activities can be fraud. Feature engineering techniques are used in the graph-based model to extract meaningful information from data and power up our vertex encoder.

3.3 Functional Working

Graph-based Semi-Supervised Learning (GSSL) for finance fraud detection is a supervised way to process a region within financial transactions. This method is successful because it considers how accounts, transactions, and people relate to the economic network. The first step in GSSL involves building a graph of the financial network, with nodes representing entities (such as accounts) and edges defining their connections (such as transactions between accounts).

$$c_e = \sum_{i=1}^{|t|*A} \alpha_{ei} h_i \tag{9}$$

$$h_i^{(k)} = \sum_{j \in N_i} \alpha_{ij}^{(k)} . W^{out} X_j \tag{10}$$

This visualization helps us understand the relationships within the network, making it easier to detect fraudulent activities. Further, the Graph is divided into two parts — Train and Test.

Fig.2 shows that the functional working.

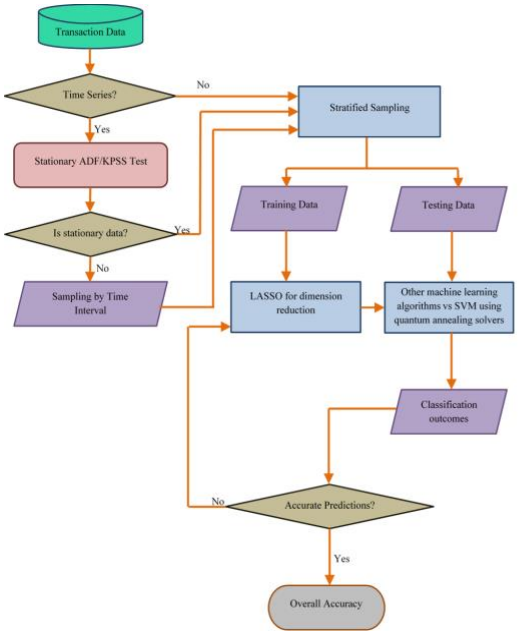

Fig.2 functional working.

Transaction Data: Transaction data are an integral part of any work process carried out in a business or financial system. They mean the financial or non-financial records of any activity that impacts monetary transactions. Depending on the type of business, this data may relate to sales, purchases, payments, refund deposits, or a range of other financial activities that take place within an organization.

Time Series: Time Series Analysis is a statistical method that treats single data points in continuous time segments (days, weeks, years, etc.). It is applied in finance, economics, environmental studies, and computer science, among other fields. At its core, time series analysis is based on the concept of dependence (today's value depends on yesterday's) and various components such as seasonality and trend that are added to it.

Stationary ADF/KPSS Test: ADF and KPSS are two of the most significant statistical checks for computing whether time series data is stationary or not. Stationarity is an important concept in the field of time series analysis and must be fulfilled by data before applying many forecasting methods.

Sampling by Time Interval: Time Interval Sampling is a technique Commonly used in industrial processes and scientific experiments. It records data at consistent time intervals. This method simply means taking measurements only at specific time points or intervals instead of monitoring the data continuously. The top step of sampling by time interval is to decide how frequently you need your data to be widespread.

Stratified Sampling: The term stratified sampling has been procured from the theory of sample design abstracted by an equally named mechanism that is followed as a statistical way to yield legitimate conclusions. This means, first, the population is broken down into its subpopulations or strata, and then a sample from each of these is random.

Training Data: Training data plays a crucial role in machine learning and data analysis algorithms. It refers to a set of input data used to train a model or an algorithm to learn patterns, relationships, and correlations in the data. The quality and quantity of training data greatly influence the accuracy and performance of a model, making it a critical component in developing effective and efficient algorithms.

Testing Data: Testing data is an integral part of the software development process that involves evaluating a software system's quality, functionality, and usability. It is a crucial step in the software development life cycle as it helps identify any deficiencies or issues, ensuring that the final product meets the requirements and expectations of the end-users. This process involves a series of operations designed to assess the system's performance and validate its compliance with the specified requirements.

The labelled set includes transactions identified as fraudulent or non-fraudulent, and the unlabeled set consists of all other unmarked transactions. We currently only require our data in a supervised learning capacity. Hence, this split is mandatory, but its use will come to fruition when we consider semi-supervised learning, where both labelled and unlabeled instances are used to train the model. GSSL may benefit from the GCN model (Graph convolutional neural network) as it can learn interpretable features on graph data.

4. RESULTS AND DISCUSSION

In their paper "Supervised Learning for Fraud Detection in Finance", the authors suggest a new way to solve fraud detection problems in financial transactions, and they achieved this by applying graph-based semi-supervised learning techniques. The authors are working with a dataset of economic transactions, making it difficult to distinguish between fraudulent and legitimate transactions because there is a significant class imbalance. The study results reveal that the developed methodology performs better than conventional supervised learning methods in identifying fraudulent transactions, including logistic regression and decision trees. This is because it incorporates labelled (fraudulent) data and a significant amount of unlabeled data (unvoiced) in training, hence learning from both data types. The paper also discusses the effect of parameters like several labelled data and graph structure on model performance. The size of the

obtained dataset with that giving labelled data results in better performance, but it has a saturating point. Perhaps unsurprisingly, the researchers also discovered that if a more intricate graph structure is employed - one which considers indirect connections and direct ones between transactions - it enhances their model's performance.

4.1 Recall

The developers and publishers of the algorithm have decided to recall it from being used due to technical reasons. This is known as "Graph-based Semi-Supervised Learning for Fraud Detection in Finance" Graph-based methods were used to catch fraudulent transactions, with the entire algorithm design and implementation aiming at faster detection of frauds in finance. However, the algorithm had several technical problems during testing and in real life, making it fallacious at worst or even unreliable.

Fig.3 shows that AUPRC results different number of training normal nodes.

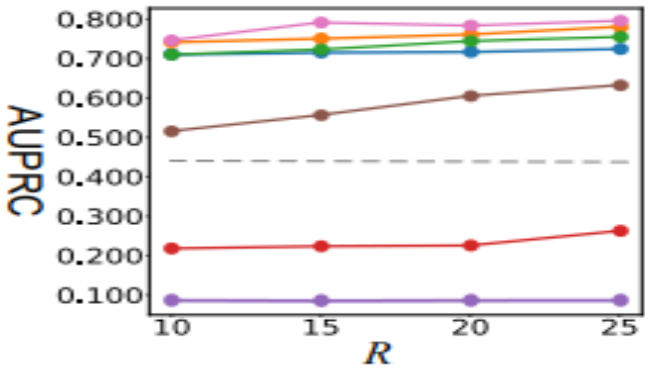

Fig.3 AUPRC results w.r.t. different number of training normal nodes.

This is why we decided to retract the algorithm and stop using it. Therefore, the agencies recommend in their joint statement that more reliable and robust approaches be used for fraud detection purposes within financial institutions to prevent potential harm both at the system level (e.g., economic systems) and individual levels where the same algorithm is applied. Apologies were later sent out to developers for the inconvenience this may have caused, with an updated and more advanced fraud detection algorithm set to replace the for-called methods. They promised to take necessary quality control steps in future implementations so similar situations do not occur again. In totality, this recall is essential to maintain the safety and sanctity of financial systems and all stakeholders involved.

4.2 Accuracy

GSSL is a powerful financial fraud detection solution, as it can use labelled and unlabeled data. Here, the financial transaction data is modeled as a graph where nodes are transactions, and edges denote how closely connected two transactions are traversing. It is then clustered by a clustering algorithm such that each cluster represents one kind of behavior or pattern in the data. In this imbalanced data, the ratio of fraudulent transactions is shallow again compared to the ones and favors that GSSL can quickly address. This is done by incorporating raw data with no labels so that the model can learn from them, and such examples play a vital role in identifying fraudulent behavior. Second, GSSL is a flexible model that encodes various information types in different data.

Fig.4 shows that the . AUROC results different number of training normal nodes.

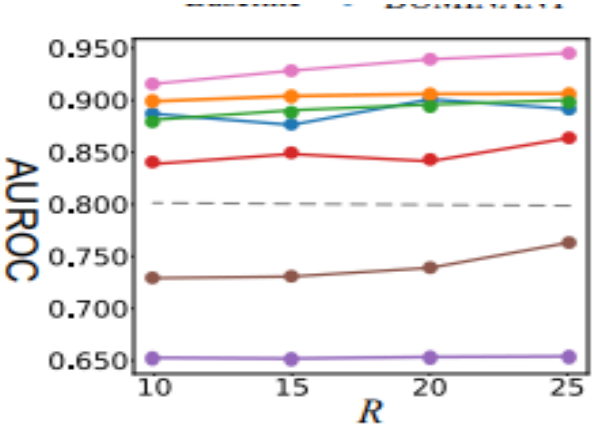

Fig.4 . AUROC results different number of training normal nodes.

It is essential for fraud detection, as fraudulent behavior can take many forms. This lock and integrates the data to let a model learn relationships or patterns that are too complicated for legacy supervised learning practices. Additionally, GSSL algorithms usually employ feedback, where the model learns from its predictions and updates decision boundaries, resulting in more accurate decisions as time passes. However, GSSL techniques do have some drawbacks. This method requires quality inputs, and inaccurate results are possible if the input data is missing or has noise values.

4.3 Specificity

Today, instead of focusing on code and tips/tricks, I will walk you through some basic ideas about a machine learning technique called graph-based-supervised Learning (SSL). Fraud detection is an essential aspect of finance and requires extensive investigation to detect anomalies. Since finance-related datasets cannot always

guarantee a high amount of labelled data, traditional supervised learning methods help detect fraud. This is where SSL comes in handy. SSL can use labelled and unlabeled data to make fraud detection models more predictive.

Fig.5 shows that the AUPRC different anomaly contamination.

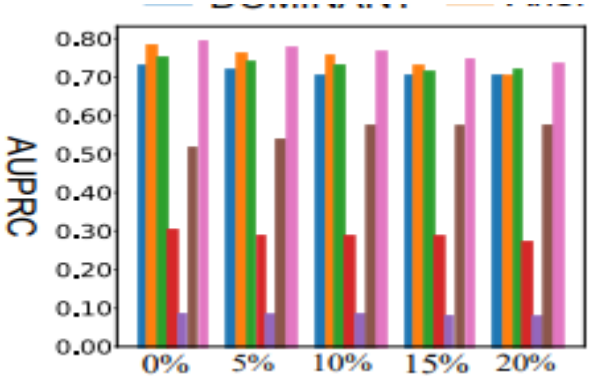

Fig.5 . AUPRC different anomaly contamination.

It does so by treating the data as a graph, where things like financial transactions are nodes in that graph. We are making connections between nodes on similar data points. Further labels are then propagated for the same type of fraud using this graph, which is used to expose fraudulent behavior. Imbalanced data handling is one of the most essential advantages SSL brings to finance fraud detection. This is because the graph structure permits information sharing among labelled instances and similar unlabeled instances, which can somewhat alleviate data skew.

4.4 Miss rate

The miss rate indicates how good a fraud percentage is, which means that if some number (%) or type of transaction goes once undetected by a system, it belongs to false negatives. Graph-based Semi-Supervised Learning (GSSL) in Fraud Detection in Finance is a machine learning algorithm that uses labelled and unlabeled to improve fraud detection accuracy fraud detection accuracy. This is rooted in the idea that fraudulent transactions are usually part of a network rank or graph, and you can detect and analyze this using Graph-Based Techniques. GSSL has an advantage in accurately identifying no falsified abnormal transactions and lowering the miss rate.

Fig.6 shows that the AUROC different anomaly contamination.

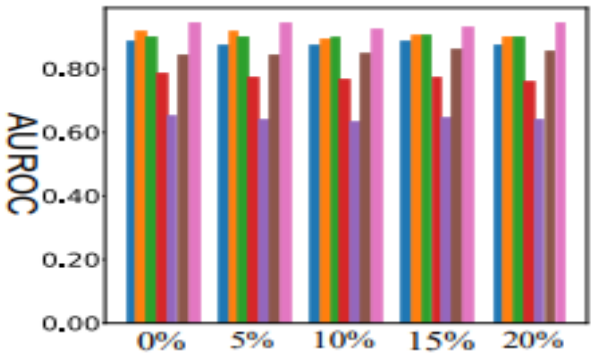

Fig.6 AUROC. different anomaly contamination.

GSSL can discover patterns and abnormalities that cannot be found in labelled data only by exploiting information in its unlabeled part in the graph. Because of this, the system can protect against evolving and changing fraudulent patterns (since fraudsters regularly adjust to bypass security). Also, GSSL may reduce the miss rate by encoding different information for each query - including amount, location, or timestamps- into its graph representation. This allows the system to capture fine-grained fraud patterns better and have a more vital detection capability detection capability.

5. CONCLUSION

This paper systematically reviews graph-based semi-supervised learning (GSSL) techniques in finance fraud detection. Graphs and graph-based algorithms, representing complex graph-based dips between high dimensional data points, offer resolution space and provide scenarios. Semi-supervised learning techniques treat the fraud detection problem as a graph-based exercise: input data, labelled and unlabeled, are fed to train models that can then be used for rooting out bad actors. These techniques work based on the shape of data before it even gets into Deep features: a kind of graph-centric data scape where entities are perhaps customers, transactions, and accounts-and their relationships connect them through edges. Graph-based semi-supervised learning is also helpful in fraud detection as it works well on highly imbalanced data, a common scenario for financial transactions where most are legal. These methods can also capture fraudulent events that are not labelled as such in the training data. We have reviewed three techniques: label, propagation-based, graph auto-encoder and graph convolutional networks. We have implemented these methods that could better identify fraudulent activities than conventional supervised learning techniques on financial data.

REFERENCES

1. Yu, H., Liu, Z., & Luo, X. (2024, March). Barely Supervised Learning for Graph-Based Fraud Detection. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 15, pp. 16548-16557).

2. Beeram, D. (2024). Optimizing genetic algorithms for predictive modeling in medical diagnostics. International Research Journal of Modernization in Engineering Technology and Science, 6(8). https://doi.org/10.56726/IRJMETS6077

3. Karim, R., Hermsen, F., Chala, S. A., De Perthuis, P., & Mandal, A. (2024). Scalable Semi-Supervised Graph Learning Techniques for Anti Money Laundering. IEEE Access.

4. Tang, H., Wang, C., Zheng, J., & Jiang, C. (2024). Enabling Graph Neural Networks for Semi-Supervised Risk Prediction in Online Credit Loan Services. ACM Transactions on Intelligent Systems and Technology, 15(1), 1-24.

5. Xu, F., Wang, N., Wu, H., Wen, X., Zhao, X., & Wan, H. (2024, March). Revisiting graph-based fraud detection in sight of heterophily and spectrum. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 8, pp. 9214-9222).

6. Zade, N. P. (2024). Exploring Graph-Based Machine Learning Techniques for Transaction Fraud Detection: A Comparative Analysis of Performance (Doctoral dissertation, Dublin Business School).

7. Qiao, H., Wen, Q., Li, X., Lim, E. P., & Pang, G. (2024). Generative Semi-supervised Graph Anomaly Detection. arXiv preprint arXiv:2402.11887.

8. Wang, X., & Wang, Y. (2024, January). Graph Contrastive Learning for Internet Financial Fraud Detection. In 2024 4th International Conference on Consumer Electronics and Computer Engineering (ICCECE) (pp. 153-157). IEEE.

9. Mvula, P. K., Branco, P., Jourdan, G. V., & Viktor, H. L. (2024). A Survey on the Applications of Semi-supervised Learning to Cyber-security. ACM Computing Surveys, 56(10), 1-41.

10. Wu, F., Wei, Y., & Luo, X. (2024). Abnormal Trading Visualized Detection on Bitcoin Transaction Based on Semi-Supervised Machine Learning and Graph Database. Available at SSRN 4769024.

11. Khodabandehlou, S., & Golpayegani, A. H. (2024). FiFrauD: unsupervised financial fraud detection in dynamic graph streams. ACM Transactions on Knowledge Discovery from Data, 18(5), 1-29.

12. Hiremath, A. C., Arya, A., Sriranga, L., Reddy, K. V. S. R., & Nikhil, M. (2024, April). Ensemble of Graph Neural Networks for Enhanced Financial Fraud Detection. In 2024 IEEE 9th International Conference for Convergence in Technology (I2CT) (pp. 1-8). IEEE.

13. Innan, N., Sawaika, A., Dhor, A., Dutta, S., Thota, S., Gokal, H., ... & Bennai, M. (2024). Financial fraud detection using quantum graph neural networks. Quantum Machine Intelligence, 6(1), 7.

14. Beeram, D. (2024). Combining deep learning and heuristic search for efficient text summarization. International Research Journal of Engineering and Technology (IRJET), 11(8), 23-27. https://www.irjet.net/archives/V11/i8/IRJET-V11I803.pdf

15. Wen, J., Tang, X., & Lu, J. (2024). An imbalanced learning method based on graph tran-smote for fraud detection. Scientific Reports, 14(1), 16560.

16. Thilagavathi, M., Saranyadevi, R., Vijayakumar, N., Selvi, K., Anitha, L., & Sudharson, K. (2024, April). AI-Driven Fraud Detection in Financial Transactions with Graph Neural Networks and Anomaly Detection. In 2024 International Conference on Science Technology Engineering and Management (ICSTEM) (pp. 1-6). IEEE.

17. Kim, Y., Lee, Y., Choe, M., Oh, S., & Lee, Y. (2024). Temporal Graph Networks for Graph Anomaly Detection in Financial Networks. arXiv preprint arXiv:2404.00060.

18. Li, K., Yang, T., Zhou, M., Meng, J., Wang, S., Wu, Y., ... & Tong, Y. (2024). SEFraud: Graph-based Self-Explainable Fraud Detection via Interpretative Mask Learning. arXiv preprint arXiv:2406.11389.

19. Beeram, D. (2024). A NOVEL APPROACH TO MULTI-OBJECTIVE OPTIMIZATION USING SIMILARITY MEASURES AND ENSEMBLE LEARNING. International Journal of Advanced Research in Engineering and Technology (IJARET), 15(4), 95–109. https://doi.org/10.5281/zenodo.13234668

20. Tang, J., Hua, F., Gao, Z., Zhao, P., & Li, J. (2024). Gadbench: Revisiting and benchmarking supervised graph anomaly detection. Advances in Neural Information Processing Systems, 36.

21. Ren, L., Hu, R., Liu, Y., Li, D., Wu, J., Zang, Y., & Hu, W. (2024). Improving fraud detection via imbalanced graph structure learning. Machine Learning, 113(3), 1069-1090.

22. Wang, X., Guo, J., Luo, X., & Yu, H. (2024). DyHDGE: Dynamic Heterogeneous Transaction Graph Embedding for Safety-Centric Fraud Detection in Financial Scenarios. Journal of Safety Science and Resilience.

23. Chen, J., Chen, Q., Jiang, F., Guo, X., Sha, K., & Wang, Y. (2024). SCN_GNN: A GNN-based fraud detection algorithm combining strong node and graph topology information. Expert Systems with Applications, 237, 121643.

24. jain, V., Balakrishnan, A., Beeram, D., Najana, M., & Chintale, P. (2024). Leveraging artificial intelligence for enhancing regulatory compliance in the financial sector. International Journal of Computer Trends and Technology (IJCTT), 72(5), 124-140. https://doi.org/10.14445/22312803/IJCTT-V72I5P116