AN APPROACH UTILIZING MULTIPLE VIEWPOINTS TO IDENTIFY FRAUDULENT ACTIVITY IN COMMERCIAL TRANSACTIONS INVOLVING MULTIPLE PARTIES

E. Murali*1, T. Nikitha*2, C. Keerthana*3, P. Muni Puspharaj*4, A. Madhan Mohan Reddy*5, G. Ganesh*6

*1Associate Professor, Siddartha Institute Of Science And Technology, Puttur, India.
*2,3,4,5,6Scholars, Siddartha Institute Of Science And Technology, Puttur, India.

DOI : https://www.doi.org/10.56726/IRJMETS53114

ABSTRACT

Identifying and preventing deceitful transactions on e-commerce platforms has persistently been the prime objective of transaction security mechanisms. However, owing to the covert nature of e-commerce, apprehending perpetrators solely based on historical order data is an arduous undertaking. Numerous studies have endeavored to develop technologies to thwart frauds, yet they have neglected to consider the dynamic conduct of users from multiple perspectives, leading to inefficient detection of fraudulent behaviors. To address this issue, this project proposes a novel fraud detection approach that amalgamates machine-learning and process mining models to monitor user behaviors in real-time. Firstly, we establish a process model concerning the B2C e-commerce platform, incorporating the detection of user behaviors. Secondly, a method for analyzing anomalies that can extract salient features from event logs is presented. Subsequently, we feed the extracted features to a Support Vector Machine (SVM) based classification model, which can detect fraudulent behaviors. Through experiments, we demonstrate the efficacy of our method in capturing dynamic fraudulent behaviors in e-commerce systems.

Keywords: Fraud, Users, Ecommerce, Transactions.

I. INTRODUCTION

While the proliferation of e-commerce and the expansion of contemporary technologies present enhanced prospects for online enterprises, novel security threats have emerged over the past few years. Reports indicate that the substantial escalation in the number of online fraud incidents incurs colossal financial losses amounting to billions of dollars globally on an annual basis. The dynamic and decentralized nature of the Internet has rendered anti-fraud systems indispensable to ensure the security of online transactions[7]. Existing fraud detection systems, which primarily concentrate on identifying abnormal user behaviors, still exhibit vulnerabilities when mitigating emerging security threats. A crucial issue in existing fraud detection systems is their lack of efficient process management during the trading process. The imperfect monitoring function is one of the key issues that necessitate attention. The detection perspective is usually inadequate due to the lack of process capture for the existing work[8].

Process mining possesses the capability to identify a substantial number of anomalous transactions, which traditional methodologies are unable to discern. The emerging process mining approach has been postulated as an appropriate solution to mitigate against fraud by incorporating internal affairs. For instance, conformance checks have been applied to monitor the process of melanoma patients. An alignment and replay technique has been employed to verify the conformance of the electronic medical record log with the hospital workflow model. Research has focused on monitoring and evaluating the sequence of processes occurring in the historical medical event log by establishing corresponding training and testing models for conformance checking. Tools such as ProM, Disco, and Heuristic Miner are extensively utilized for conformance checking. Process mining can be an efficient approach for fraud detection[9].

Particularly, it is crucial to be dynamic and multi-perspective when detecting fraudulent user behaviors. Process mining aids in comparing the actual data against the standard model to identify outliers. Despite existing progress in fraud detection, it is still necessary to develop hybrid learning methods to improve the accuracy of detection. To promote the understanding and development of process mining for anomaly detection, a method
of multi-perspective anomaly detection is proposed that goes beyond the perspective of control flow, including time and resources. Previous research on using process mining to detect fraudulent transactions demonstrated that process mining is capable of detecting fraudulent transactions, and it can effectively prevent audit fraud at a much earlier stage due to the continuous monitoring nature of event logs[6].

To address this, we propose a process-based method, where user behaviors are recorded and analyzed in real-time, and historical data is transformed into controllable data. Additionally, we incorporate a multi-perspective detection of abnormal behaviors[6]. This project combines the advantages of process mining and machine learning models by introducing a hybrid method to solve the anomaly detection in data flows, which provides information about each action embedded in a control flow model[10].

II. LITERATURE SURVEY

In [1], In the contemporary era, illicit activities pertaining to online financial transactions have become increasingly intricate and transcended geographical boundaries, resulting in substantial financial detriments for both parties, customers and organizations alike. Numerous techniques have been proposed for fraud prevention and detection in the online realm. However, notwithstanding their shared objective of identifying and combating fraudulent online transactions, each of these techniques is accompanied by its unique characteristics, advantages, and limitations. Within this context, this paper reviews the existing research conducted in fraud detection with the aim of identifying the employed algorithms and analyzing each of these algorithms based on specific criteria. To analyze the research studies in the field of fraud detection, the systematic quantitative literature review methodology was applied. Based on the machine-learning algorithms most frequently invoked in scientific articles and their characteristics, a hierarchical typology is constructed. Consequently, our paper highlights, in a novel manner, the most suitable techniques for detecting fraud by combining three selection criteria: accuracy, coverage, and costs.

In [2], With the proliferation of online shopping, transaction fraud has witnessed a grave escalation. Consequently, the study on fraud detection has garnered significant interest and importance. An essential approach to detecting fraud is the extraction of user behavior profiles (BPs) based on their historical transaction records, followed by the verification of whether an incoming transaction is fraudulent or not in view of their BPs. Markov chain models have gained popularity in representing users' BPs, which is effective for those users whose transaction behaviors exhibit relative stability. However, with the advancement and widespread adoption of online shopping, users have been afforded greater convenience for Internet-based consumption, thereby diversifying their transaction behaviors. Therefore, Markov chain models are ill-suited for the representation of these behaviors. In this paper, propose a logical graph of BP (LGBP), which is a total order-based model to represent the logical relation of attributes of transaction records. Based on LGBP and users' transaction records, we can compute a path-based transition probability from one attribute to another. Concurrently, define an information entropy-based diversity coefficient to characterize the diversity of a user's transaction behaviors. Additionally, we define a state transition probability matrix to capture the temporal features of a user's transactions.

In [3], Credit card fraud identification emerges as a pivotal area of study in the contemporary mobile payment era. Enhancing the performance of a fraud detection model while maintaining its stability poses formidable challenges, as users' payment patterns and criminals' fraudulent behaviors are often subject to change. In this article, we concentrate on acquiring profound feature representations of legitimate and fraudulent transactions from the perspective of the loss function of a deep neural network. Our objective is to attain superior separability and discrimination of features, thereby improving the performance of our fraud detection model and preserving its stability. We propose a novel loss function, termed the full center loss (FCL), which considers both distances and angles among features, and thus can comprehensively supervise the deep representation learning process. We conduct extensive experiments on two large datasets of credit card transactions, one private and the other public, to demonstrate the detection performance of our model by comparing FCL with other state-of-the-art loss functions. The results illustrate that FCL outperforms its counterparts.

In [4], Unauthorized utilization of mobile transactions through identity usurpation or credit card theft to illicitly acquire monetary funds constitutes mobile payment fraud. The rapid proliferation of smartphones and online
transaction services has fueled the escalation of mobile payment fraud as a burgeoning concern. In the practical realm, a highly accurate process for detecting mobile payment fraud is imperative, as financial fraud precipitates monetary losses. Consequently, our approach proposes a comprehensive process for identifying mobile payment fraud based on machine learning, employing supervised and unsupervised methods to detect fraud and process substantial volumes of financial data. Furthermore, our approach incorporates a sampling process and feature selection process to facilitate swift processing of large transaction data volumes and achieve high accuracy in mobile payment fraud detection. The F-measure and ROC curve are utilized to validate our proposed model.

In [5], Data mining and process mining offer solutions for fraud identification. Nevertheless, the automated methodologies grounded in historical data still necessitate enhancement. In this regard, we propose a hybrid approach that amalgamates association rule learning and process mining. In this case, process mining scrutinizes the event log. Through expert verification, the itemset of the association rule learning is employed to generate positive and negative rules, which are applied for compliance checking towards the testing dataset. The resultant findings demonstrate that the hybrid method exhibits a lower false discovery rate and furnishes higher accuracy in comparison to the process-mining technique.

III. PROPOSED SYSTEM

Perpetrators of fraud often dynamically adapt their behavioral patterns to circumvent existing fraud detection methodologies. In the realm of online credit card fraud detection, Support Vector Machines (SVMs) can classify user behaviors under intricate scenarios and deliver reliable outcomes. Numerous researchers leverage the advantage of amalgamating multiple detection techniques for comprehensive fraud identification. For instance, a method combining supervised and unsupervised learning has been employed, with a focus on payment fraud applications. The majority of machine learning-based approaches utilize historical data to analyze fraudulent transactions. However, they have not placed sufficient emphasis on the transactional process flow and dynamic user behaviors. The second category of fraud detection methods employs process mining, concentrating on extracting knowledge from existing event logs in information systems for the purpose of monitoring and enhancing the operational process within business IT infrastructure. Process mining specializes in comparing the event log with an established model to further detect, locate, and interpret the deviation between the established model and the actual event log.

Fraudsters often change their behavioral pattern dynamically to overcome existing fraud detection methods. In online credit card fraud detection, SVM can classify user behaviors under complex scenarios and deliver reliable results. Many researchers take the advantage of combining multiple detection methods for comprehensive fraud detection. For example, focusing on payment fraud applications, a method by combining supervised and unsupervised learning.

The proposed system combines the advantages of process mining and machine learning models by introducing a hybrid method to solve the anomaly detection in data flows, which provides information about each action embedded in a control flow model. By modeling and analyzing the business process of the e-commerce system, this method can dynamically detect changes in user behaviors, transaction processes, and noncompliance situations, and comprehensively analyze and identify fraudulent transactions from multiple perspectives.

Important contributions of this project are listed as follows:

1) A conformance checking method based on process mining is applied in the field of ecommerce transactions to capture the abnormalities.

2) A user behavior detection method is proposed to perform comprehensive anomaly detection based on Petri nets.

3) An SVM model is developed by embedding a multi perspective process mining into machine learning methods to automatically classify fraudulent behaviors.

Algorithms:

1. Naive Bayes
   • Type: Supervised learning, classification
• Concept: A probabilistic classifier based on Bayes' Theorem. It calculates the probability of an email being spam (or any other class) based on the probabilities of individual words appearing in spam emails.
• Strengths: Simple, efficient, fast for large datasets, effective for text classification (spam filtering, sentiment analysis).
• Weaknesses: Assumes independence of features (which may not always hold true), can be sensitive to rare features.

2. Support Vector Machine (SVM):
• Type: Supervised learning, classification
• Concept: Finds a hyperplane that best separates data points of different classes with the maximum margin.
• Strengths: Effective for high-dimensional data, good for small datasets, can handle non-linear data using kernel functions.
• Weaknesses: Can be computationally expensive for large datasets, difficulty in interpreting the model, may require careful parameter tuning.

3. Logistic Regression:
• Type: Supervised learning, classification (also used for regression)
• Concept: Models the relationship between features and a binary class label using a sigmoid function. Outputs the probability of a data point belonging to a particular class.
• Strengths: Simple to understand and interpret, works well with linear data, efficient for large datasets.
• Weaknesses: Limited to binary classification problems by default (can be extended to multi-class with one-vs-rest approach), may not perform well for non-linear data.

4. Decision Tree Classifier:
• Type: Supervised learning, classification
• Concept: Creates a tree-like model where each internal node represents a feature test, and each leaf node holds the class label. Data is classified by traversing the tree based on feature values.
• Strengths: Easy to interpret, can handle both categorical and numerical features, works well with missing data.
• Weaknesses: Prone to overfitting if not pruned, can be sensitive to small changes in the data.

5. Extra Trees Classifier:
• Type: Ensemble learning, classification (an extension of Random Forest)
• Concept: Similar to a random forest, but builds multiple decision trees using random feature selection at each split point. Improves accuracy and reduces overfitting.
• Strengths: Robust to overfitting, handles mixed data types, good feature importance estimation.
• Weaknesses: Can be less interpretable compared to single decision trees, may require more computational resources for training.

In essence, these algorithms offer various approaches to classification tasks in machine learning, each with its own strengths and weaknesses. The choice of algorithm depends on the specific problem, data characteristics, and desired outcome.

Accuracy:

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive Bayes</td>
<td>46.9348</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>50.0</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>49.2</td>
</tr>
<tr>
<td>Decision Tree Classifier</td>
<td>47.7011</td>
</tr>
<tr>
<td>Extra Trees Classifier</td>
<td>48.36</td>
</tr>
</tbody>
</table>

These tabular values changes based on the user transactions.
The Architectural diagram illustrates the general process of building a predictive model using machine learning techniques,

Data Collection from Different Sources: The process begins with collecting data from various sources, represented by multiple data storage symbols.

Data Cleansing: The collected data often contains noise, inconsistencies, or missing values. The data cleansing step, depicted by the funnel-like symbol, involves preprocessing the data to remove any irrelevant or noisy information, ensuring the data is clean and ready for model training.

Clean Data set + Applying Algorithm: The clean data set is combined with the chosen algorithms, in this case, the Support Vector Machine (SVM) algorithm, represented by the gear symbols.

Model: The puzzle piece symbol represents the model obtained by applying the SVM algorithm to the clean data set. SVM is a supervised learning algorithm that constructs a hyperplane or a set of hyperplanes in a high-dimensional space, which can be used for classification or regression tasks.

Model Evaluation: The model is then evaluated, as indicated by the icon with lines and dots, to assess its performance and make any necessary adjustments or optimizations.

Prediction: Finally, the evaluated model is used for making predictions, represented by the icon with dots connected by a line, which could be used for tasks such as classifying new data points or predicting future outcomes.

IV. IMPLEMENTATION

he Support Vector Machine (SVM) algorithm to identify fraudulent activity in commercial transactions involving multiple parties, the typical steps involved would be as follows:

1. Data Collection: Gather data related to commercial transactions, including information about the parties involved, transaction details, historical transaction records, and any other relevant features that could help distinguish between legitimate and fraudulent transactions.
2. Data Preprocessing: Clean and preprocess the collected data by handling missing values, removing noise or irrelevant features, and converting the data into a format suitable for the SVM algorithm. This may involve techniques such as feature scaling, encoding categorical variables, and data normalization.

3. Feature Engineering: Identify and extract relevant features from the preprocessed data that can effectively capture the characteristics of fraudulent and legitimate transactions. This may involve domain knowledge, statistical analysis, or automated feature selection techniques.

4. Training Data Preparation: Split the preprocessed data into training and testing sets. The training set will be used to train the SVM model, while the testing set will be used to evaluate the model’s performance.

5. Model Training: Train the SVM algorithm on the training data set. The SVM algorithm will construct a hyperplane or a set of hyperplanes that best separates the fraudulent and legitimate transactions in the feature space. The choice of kernel function (e.g., linear, polynomial, radial basis function) and the tuning of hyperparameters (e.g., regularization parameter, kernel parameters) will influence the model’s performance.

6. Model Evaluation: Evaluate the trained SVM model using the testing data set. Common evaluation metrics for fraud detection include accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (ROC-AUC). This step helps assess the model’s ability to correctly identify fraudulent and legitimate transactions.

7. Model Optimization: Based on the evaluation results, optimize the SVM model by adjusting hyperparameters, trying different kernel functions, or incorporating additional features. This iterative process can improve the model’s performance and generalization ability.

8. Model Deployment: Once satisfied with the model’s performance, deploy the trained SVM model for identifying fraudulent activity in real-time or batch commercial transactions involving multiple parties.

V. RESULT AND DISCUSSION

Each dataset comprises 82 anomaly detection features, and each feature characterizes a value of 0 or 1, where 1 signifies abnormality, and 0 represents normality. The statistical detection indicators of F1-score and AUC of our proposed SVM-based fraud detection model were obtained based on 10-fold cross-validation. The F1-score under the fusion of control flow and data flow data is higher than when only one type of data is considered, indicating that when the data of control flow and data flow are considered comprehensively, better user anomaly detection is achieved.

To further validate the fraud detection effects of our model under the three aforementioned cases, we consider various performance indicators under 50 rounds of tests and calculate their average values. The perspectives under the case of integrating data flow and control flow features illustrate that these two kinds of characteristic data only consider one aspect of the user’s anomaly. After learning the two types of data through the machine-learning model, the information of the two aspects is fully utilized to comprehensively detect user anomalies with better effect. SVM aims to find the hyperplane that best separates the data points of different classes while maximizing the margin.

The SVM formulation involves solving the optimization problem:

\[
\begin{align*}
\min_{w,b} & \quad \frac{1}{2} ||w||^2 + C \sum_{i=1}^{N} \xi_i \\
\text{subject to:} & \quad y_i(w \cdot x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0, \text{ for } i = 1, \ldots, N
\end{align*}
\]

where:

- \( w \) is the weight vector.
- \( b \) is the bias term.
- \( C \) is the regularization parameter.
- \( y_i \) is the class label (+1 for legitimate transactions, -1 for fraudulent transactions).
- \( x_i \) is the feature vector for the i-th transaction.
\( \xi \) are slack variables.
\( N \) is the number of samples.

Once the SVM model is trained, predictions can be made for new transactions by computing the decision function:

\[
f(x) = w \cdot x + b
\]

where \( x \) is the feature vector of the new transaction.

If \( f(x) \geq 0 \), the transaction is classified as legitimate; otherwise, it is classified as fraudulent.

In summary, when compared with considering only one perspective of information, our proposed model exhibits higher F1-score and AUC indicators. The detection effect of abnormal e-commerce users is better in our model. Therefore, our proposed method can detect abnormal ecommerce users more comprehensively. Additionally, compared with the related deep learning methods for fraud detection in e-commerce, our methodology can depict the transaction process and structures, and it is interpretable.

VI. CONCLUSION

This paper proposes a hybrid approach to identify fraudulent transactions by integrating formal process modeling and dynamic user behavior analysis. The e-commerce transaction process was analyzed from five major perspectives and an SVM model was created to perform fraudulent transaction detection. Our extensive experiments demonstrated that the proposed method can effectively capture fraudulent transactions and behaviors. The overall performance of our proposed multi-perspective detection method surpassed the single-perspective detection approach. As future work, related deep learning techniques and model checking methods would be incorporated into the proposed framework to achieve higher accuracy. Additionally, incorporating more time-based features into the behavior patterns to enhance the precision of risk identification is a future endeavor.

VII. REFERENCES


