In-Context Learning for Smarter Fraud Detection in Remote Flea Market Transactions

Hyunwoo Kim
Danggeun Pay Inc.
Seoul, Republic of Korea
peter.kim@daangnpay.com

Hyunmyoung Oh
Danggeun Pay Inc.
Seoul, Republic of Korea
hammer@daangnpay.com

Sunghyon Kyeong*
Danggeun Pay Inc.
Seoul, Republic of Korea
devyn@daangnpay.com

Abstract

this part is not ready yet.

CCS Concepts

• **Computing methodologies** → **Machine learning**; **Natural language processing**; • **Information systems** → **Enterprise information systems**.

Keywords

In-context learning, context-based fraud detection, fraud detection, remote secondhand transactions

ACM Reference Format:

Hyunwoo Kim, Hyunmyoung Oh, and Sunghyon Kyeong. 2025. In-Context Learning for Smarter Fraud Detection in Remote Flea Market Transactions. In *6th ACM International Conference on AI in Finance (ICAIF '25), November 15–18, 2025, Singapore.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3604237.3626838

1 Introduction

The global market for secondhand goods has been steadily expanding, driven in large part by the rise of online platforms that facilitate non-face-to-face peer-to-peer transactions. Prominent marketplaces in this domain include Facebook Marketplace (worldwide), Danggeun Market—also known as Karrot—operating in regions such as North America, Korea, and Japan, and Mercari, which is widely used in Japan. These platforms promote the reuse of goods, contributing to environmental sustainability, and attract a growing user base who are motivated by shared social and ecological values.

To ensure secure and convenient transactions for the majority of well-intentioned users, platform providers have implemented protective measures, including escrow-based financial services. Nevertheless, the anonymity and remote nature of these platforms are frequently exploited by malicious actors. For instance, some fraudulent sellers post items at unusually low prices and fail to deliver the products, engaging in what is commonly referred to as merchant fraud. In response, platforms invest substantial effort into detecting and sanctioning such fraudulent activities.

Traditional fraud detection systems have largely relied on rule-based approaches or supervised machine learning (ML) models trained on historical transaction data. Despite ongoing model retraining enabled by MLops platforms, these approaches exhibit clear limitations in rapidly evolving environments like secondhand marketplaces, where contextual factors heavily influence transaction dynamics. In particular, the early detection of novel fraud schemes remains structurally constrained.

Recently, Large Language Models (LLMs) have emerged as a promising alternative to address these limitations. In the context of secondhand trading, LLMs can effectively analyze unstructured textual data—such as listing titles, product descriptions, and seller profiles—to identify suspicious language patterns or detect fraud strategies that resemble previously known cases. Unlike traditional models that rely solely on structured features, LLMs excel at natural language understanding and can capture subtle linguistic cues, inconsistencies in phrasing, and tone variations that might otherwise elude human analysts.

Furthermore, LLMs possess the ability to cross-reference contextual information across multiple transactions. For example, the repeated use of similar phrases, emojis, or urgent language across listings from different user accounts may be linked to a single fraud actor. This capability is significantly enhanced through In-Context Learning (ICL), which enables LLMs to perform fraud detection tasks with minimal examples and without requiring explicit model fine-tuning. ICL is particularly advantageous in scenarios where large-scale labeled datasets are unavailable and where fraud tactics evolve rapidly.

Moreover, the increasing sophistication of fraudsters—who now leverage generative AI to craft convincing phishing messages, fabricate identities, and produce deepfake documents—underscores the urgency for platforms to deploy equally advanced AI-based defense mechanisms. This technological arms race necessitates the adoption of LLM-powered, intelligent fraud detection frameworks.

In this study, we propose a novel fraud detection approach tailored to non-face-to-face secondhand trading environments, leveraging LLM-based In-Context Learning. Specifically, we extract salient features from previously confirmed fraud cases using LLMs to analyze unstructured elements such as listing titles and seller profiles. These features are then compared against ongoing transactions to assess their likelihood of being fraudulent. Finally, we evaluate the effectiveness of our approach in comparison with traditional machine learning-based detection methods to determine its potential performance gains in real-world settings.

*Corresponding author

ICAIF '25, Singapore

ACM ISBN 979-8-4007-0240-2/23/11
https://doi.org/10.1145/3604237.3626838

2 Related Works

2.1 ML-Based Fraud Detection

A wide range of studies in both industry and academia have sought to advance techniques for financial fraud detection. One line of research focuses on representing transaction histories between bank accounts as graphs, enabling the development of graph-based fraud detection models that significantly outperform traditional baselines in terms of F1 score performance [5, 11].

Simultaneously, increasing attention has been paid to fraud in peer-to-peer transactions within online marketplaces, where financial transactions often accompany interpersonal exchanges. A prominent example is merchant fraud, in which a scammer lists trending products at unusually low prices, receives payment, and fails to deliver the goods. This type of fraud is especially prevalent in remote secondhand platforms. Some studies have addressed this issue by analyzing fraudulent seller accounts and building machine learning-based detection models using features derived from transaction histories and product listings [4, 10].

2.2 LLM-Based Fraud Detection

With the advent of large language models (LLMs), researchers and practitioners have actively explored their potential for financial fraud detection. Traditional methods—such as logistic regression, random forests, and neural networks—have long been applied to detect fraud (e.g., in credit card transactions), but these models face limitations when dealing with highly imbalanced datasets and evolving fraud patterns [12].

Recent studies suggest that Transformer-based LLMs are better suited for capturing long-range dependencies and subtle correlations in transaction data, leading to improved detection performance [3, 7]. For example, Yu et al. (2024) demonstrate that Transformer-based models outperform conventional machine learning approaches in terms of accuracy and are particularly effective at identifying rare fraudulent cases [8, 12]. The pretraining of LLMs on vast corpora enables them to form a form of commonsense understanding of sequences, which can be further enhanced through retrieval-augmented generation (RAG) methods to boost detection capabilities [9].

Moreover, LLMs have proven useful in processing unstructured data alongside structured transactional features. Butler (2025) highlights that LLMs can detect fraud-indicative language and anomalies in textual sources such as transaction notes, emails, and chat logs. This capacity allows them to surface social engineering attempts or abnormal phrasing in online interactions—types of fraud that often evade detection by traditional rule-based or statistical systems.

2.3 In-Context Learning for Fraud Pattern Recognition

In-context learning (ICL) has emerged as a powerful paradigm that enables LLMs to perform tasks without explicit fine-tuning. Introduced by Brown et al. (2020) with the release of GPT-3, ICL allows a model to generalize to new tasks using only a prompt containing a few labeled examples [2]. This characteristic makes ICL particularly well-suited for fraud detection scenarios, where only a small number of examples of emerging fraud types may be available.

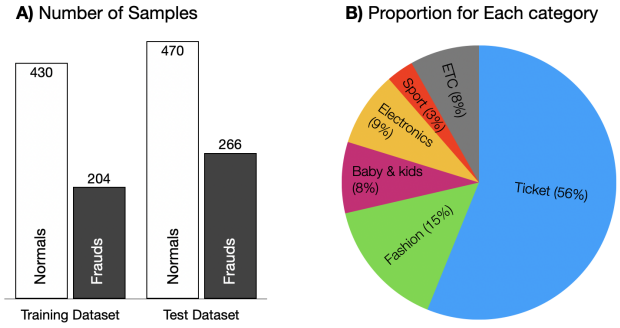

Figure 1: Overview of the sampled dataset with fraud label distribution

Through ICL, LLMs can implicitly learn patterns from a few in-context examples and adapt to new fraud types in real time. Liu et al. (2024) apply this concept to graph-based anomaly detection, using a handful of normal nodes as context to identify outliers in unseen graphs without additional training [6]. Similarly, Bhattacharya et al. (2025) propose a system that converts structured transaction features (e.g., amount, location, device information) into natural language descriptions and feeds them into an LLM along with a few labeled examples, enabling accurate classification of novel transactions as fraudulent or legitimate [1].

3 Datasets

This study leverages proprietary real-world transaction data provided by Danggeun Pay Inc., a financial technology company that operates the official payment infrastructure for Danggeun Market Inc.—a widely used local community platform in South Korea. The platform supports a variety of services, including secondhand goods trading, real estate listings, part-time job postings, and more. Within this ecosystem, Danggeun Pay facilitates peer-to-peer (P2P) payments, enabling the collection of fine-grained transactional records that are particularly rich in behavioral signals relevant to fraud detection.

The dataset comprises transaction-level records labeled as either fraudulent or normal. Each transaction is augmented with accompanying listing metadata as well as detailed behavioral features extracted from the seller's historical activity. The dataset was curated for the express purpose of facilitating machine learning research on fraud detection in P2P commerce and offers a comprehensive foundation for studying behavioral patterns in online trust-mediated environments.

3.1 Training and Test Split

The dataset includes transactions conducted over a two-month period, from April to May 2025. To ensure balanced model training and fair evaluation across fraud classes, stratified sampling based on ground-truth fraud labels was applied. As shown in Fig. 1A, a total of 1,370 transactions were selected, with a class distribution

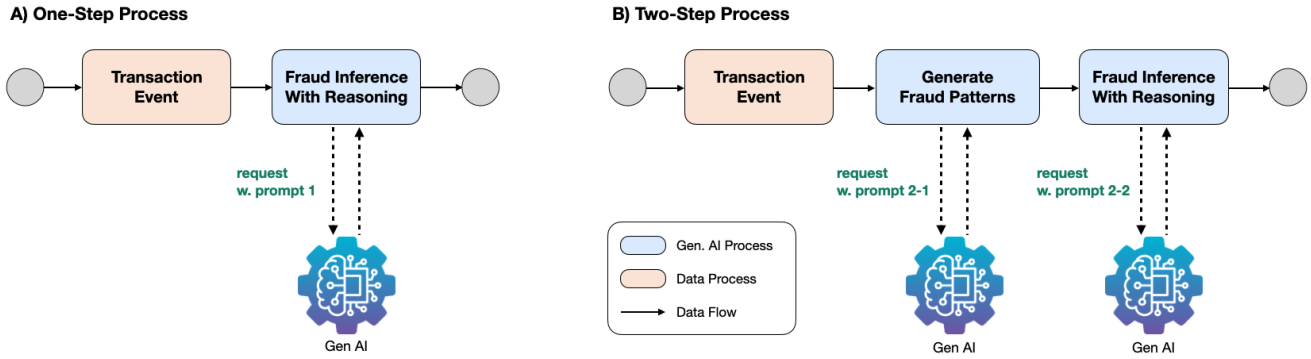

Figure 2: Overview of the experimental processes. A) One-step inference approach. B) Two-step inference approach.

ratio of approximately 1:2 (fraudulent to normal transactions), providing a sufficiently diverse dataset for evaluating fraud detection performance.

The dataset was temporally partitioned based on the transaction date. Transactions that occurred in April 2025 were designated as the training dataset (430 legitimate cases; 204 fraudulent cases), while those from May 2025 were allocated to the test dataset (470 legitimate cases; 266 fraudulent cases). The training dataset was not primarily used for direct model training but rather for extracting recent fraud cases during the experimental process. The test dataset was utilized to evaluate the performance of the proposed fraud detection methods.

3.2 Category-Wise Distribution

The sampled transactions span seven major product categories, each exhibiting distinct fraud risk profiles as shown in Fig. 1B. Table 1 summarizes the distribution of fraudulent and legitimate transactions across these categories.

Notably, categories with high liquidity and resale value—such as tickets—exhibit a disproportionately high rate of fraudulent activity. This heterogeneity underscores the importance of incorporating category-specific behavioral patterns into fraud detection models.

3.3 Feature Overview

Each transaction instance in the dataset is represented by a set of features grouped into four key dimensions:

- **Listing Metadata**: This feature includes the listing title (`title`), listed price (`price`), and product category (`category`).
- **Transaction Details**: This feature captures transaction timestamp (`tx_dttm`) and transaction amount (`tx_amt`).
- **Seller Profile**: This feature includes demographic and account-level attributes such as seller age (`seller_age`) and account tenure in days (`seller_account_tenure`).
- **Recent Seller Activity**: This feature summarizes behavioral signals over a 24-hour window preceding the listing. This includes the number of prior transactions (`recent_tx_cnt`), cumulative transaction amount (`recent_tx_amt_sum`), and number of unique counterparties (`recent_unique_buyers`).

These feature groups collectively capture both static attributes and dynamic behavioral cues, facilitating a comprehensive analysis of user behavior for fraud detection.

Category	Fraud	Normal
Tickets	305	464
Fashion & Miscellaneous	25	184
Baby & Kids	78	37
Electronics	37	84
Sports	12	32
Others	13	99

Table 1: Category-wise distribution of the dataset by fraud label

1. Role
You are a fraud detection expert working at an remote flea market.

2. Evaluation Flow
Step 1.
Step 2.
...

3. Current Transaction
{{current_transaction}}

4. High-risk Indicators
- Mismatch between seller's age and item category
- ...

4. Recent Fraud Examples
{{recent_fraud_examples}}

5. Output Format
Your response must match the following format exactly:

```json
{
    "fraud": "Y" or "N",
    "reasoning": "Reason 1 (in Korean) || Reason 2 (in Korean) || Reason 3 (in Korean)"
}
```

Figure 3: Partial description for one-step approach. The red text enclosed in double curly brackets denotes dynamic input parameters that were updated at each LLM invocation.

4 Experiments

This study investigates whether a large language model (LLM) can accurately assess the likelihood of fraud in financial transactions, given rich contextual information specific to peer-to-peer remote secondhand marketplaces. To this end, we designed a series of experiments aimed at evaluating the LLM's ability to make context-aware inferences about the legitimacy of each transaction event.

Two primary inference approaches were explored as illustrated in Fig. 2. In the one-step inference method, prompts were constructed to directly assess whether a given transaction was fraudulent. Each prompt embedded comprehensive contextual information about the target transaction along with recent fraud examples, enabling the LLM to leverage both transaction-specific features and patterns from past fraud cases.

In the two-step inference method, the fraud detection process was decomposed into two stages. In the first stage, the LLM was prompted with recent fraud examples to identify distinct fraud clusters and extract representative features for each group. In the second stage, these extracted fraud patterns were combined with the full contextual information of the target transaction in a new prompt, allowing the LLM to determine whether the transaction aligned with any known fraud patterns.

4.1 Proprietary LLM

We conducted experiments to evaluate whether state-of-the-art proprietary LLMs can assess the likelihood of fraud by leveraging in-context learning, using real-world examples of recent fraudulent transactions along with rich contextual information surrounding the current transaction. Specifically, we investigated the extent to which each model could comprehend the nuances of financial fraud scenarios and accurately infer fraud likelihood in a context-sensitive manner. Three leading proprietary LLMs were selected for this evaluation: OpenAI's GPT-4.1, Google's Gemini 2.5 Flash, and Anthropic's Claude Sonnet 4.0. These models were tested under identical prompt structures and inference conditions to ensure a fair comparison of their reasoning capabilities and context sensitivity in the domain of peer-to-peer financial fraud detection.

4.2 One-Step Fraud Inference

In the one-step fraud inference approach, the prompt was constructed by inserting comprehensive contextual information about the transaction under evaluation, recent fraud examples, and a description of high-risk fraud indicators. The large language model (LLM) was then prompted to determine whether or not the transaction under evaluation is fraud. The prompt format used in the one-step fraud inference experiment is illustrated in Fig. 3. In this figure, the red text enclosed in double curly brackets denotes dynamic input parameters that were updated at each LLM invocation.

The `{{current_transaction}}` slot was filled with full contextual information about the transaction, incorporating all features described in Section 3.3. The `{{recent_fraud_examples}}` slot included listing attributes and seller profile information from past transactions that had been labeled as fraudulent.

To investigate the impact of the number of few-shot examples on inference performance, we varied the number of recent fraud examples (N) inserted into the prompt. Specifically, for each target transaction, we extracted the most recent N fraud cases that occurred within the 24-hour period preceding it.

The experiments were independently conducted for each of the three proprietary LLMs—GPT-4.1, Gemini 2.5 Flash, and Claude Sonnet 4.0—by repeatedly applying the same protocol across varying values of N (10, 30, 50, 70, 90). This ensured a consistent and comparative evaluation of the one-step inference performance across all models under different prompt sizes.

4.3 Two-Step Fraud Inference

In the two-step fraud inference approach, the evaluation of a target transaction's likelihood of being fraudulent was conducted through a two-stage invocation of the LLM (Fig. 2B). In the first step, the LLM was prompted with recent fraud examples to perform fraud clustering and extract representative features for each identified cluster (see the descriptive LLM prompt in Fig. 4A). The fraud features generated in this step—derived in a data-driven manner—were subsequently used as input for the second LLM invocation.

In the second step, the LLM was presented with a new prompt (Fig. 4B) that incorporated both the fraud patterns extracted in the first step and the full contextual information of the target transaction as described in Section 3.3. Based on this prompt, the LLM was tasked with assessing whether the given transaction exhibited fraudulent characteristics or not.

As in the one-step inference approach, we investigated how the number of few-shot examples influenced both the quality of feature extraction and the performance of fraud inference. For each transaction, we selected the most recent N fraud cases that occurred within a 24-hour window prior to the transaction.

The same two-step inference procedure was independently applied to all three proprietary LLMs—GPT-4.1, Gemini 2.5 Flash, and Claude Sonnet 4.0—repeatedly across different values of N (10, 30, 50, 70, and 90), enabling a model-by-model comparison under identical experimental conditions.

4.4 Evaluation Metrics

To comprehensively assess the performance of the proposed fraud inference approaches, we report three widely adopted evaluation metrics: precision, recall, and F1-score. These metrics are particularly important in real-world fraud detection scenarios, where both false positives (misclassifying a legitimate transaction as fraudulent) and false negatives (failing to detect an actual fraud) entail significant practical consequences. Precision quantifies the proportion of transactions predicted as fraudulent that are indeed fraudulent. It reflects the model's effectiveness in minimizing false alarms and is particularly valuable when the cost of incorrectly flagging legitimate users is high.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \tag{1}$$

Recall measures the proportion of actual fraudulent transactions that are correctly identified by the model. It captures the model's sensitivity to fraud cases and is especially critical when the cost of missing fraudulent behavior is high.

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \tag{2}$$

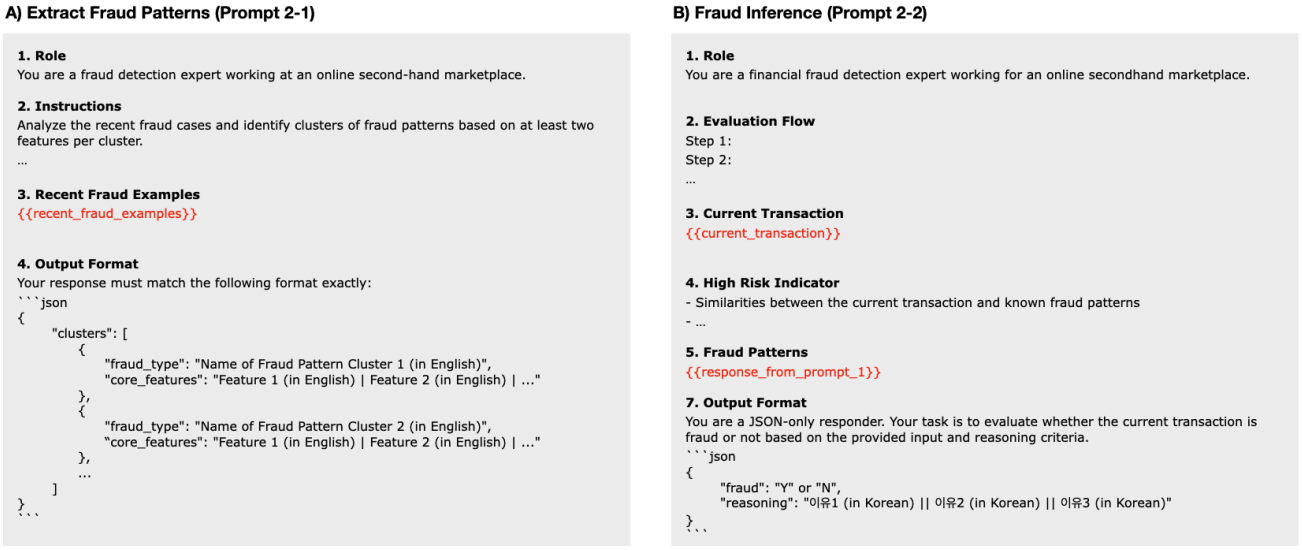

Figure 4: Illustrative description for Two-step prompts. A) First step prompt to cluster frauds and extract features. B) Second step prompt to determine whether the current transaction is fraudulent. The red text enclosed in double curly brackets denotes dynamic input parameters that were updated at each LLM invocation.

F1-score is the harmonic mean of precision and recall, providing a balanced metric that accounts for both types of classification errors. It is particularly suitable when neither precision nor recall can be compromised, as is often the case in fraud detection systems.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

5 Experimental Results

this part is not ready yet.

5.1 Performance of fine-tuning models

this part is not ready yet.

6 Conclusions

this part is not ready yet.

References

[1] Indranil Bhattacharya and Ana Mickovic. 2024. Accounting fraud detection using contextual language learning. *International Journal of Accounting Information Systems* 53 (2024), 100682. doi:10.1016/j.accinf.2024.100682

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL] https://arxiv.org/abs/2005.14165

[3] Zhengyu Chen, Jixie Ge, Heshen Zhan, Siteng Huang, and Donglin Wang. 2021. Pareto Self-Supervised Training for Few-Shot Learning. arXiv:2104.07841 [cs.CV] https://arxiv.org/abs/2104.07841

[4] Fahim Hasan, Sourov Kumar Mondal, Md. Rayhan Kabir, Md Abdullah Al Mamun, Nur Salman Rahman, and Md. Sagar Hossen. 2022. E-commerce Merchant Fraud Detection using Machine Learning Approach. In *2022 7th International Conference on Communication and Electronics Systems (ICCES)*. 1123–1127. doi:10.1109/ICCES54183.2022.9835868

[5] Junhong Lin, Xiaojie Guo, Yada Zhu, Samuel Mitchell, Erik Altman, and Julian Shun. 2024. FraudGT: A Simple, Effective, and Efficient Graph Transformer for Financial Fraud Detection. In *Proceedings of the 5th ACM International Conference on AI in Finance* (Brooklyn, NY, USA) *(ICAIF '24)*. Association for Computing Machinery, New York, NY, USA, 292–300. doi:10.1145/3677052.3698648

[6] Shuo Liu, Di Yao, Lanting Fang, Zhetao Li, Wenbin Li, Kaiyu Feng, XiaoWen Ji, and Jingping Bi. 2024. AnomalyLLM: Few-shot Anomaly Edge Detection for Dynamic Graphs using Large Language Models. arXiv:2405.07626 [cs.LG] https://arxiv.org/abs/2405.07626

[7] Xiaopeng Liu, Yan Liu, Meng Zhang, Xianzhong Chen, and Jiangyun Li. 2019. Improving Stockline Detection of Radar Sensor Array Systems in Blast Furnaces Using a Novel Encoder–Decoder Architecture. *Sensors* 19, 16 (2019), 3470. doi:10.3390/s19163470

[8] Weimin Lyu, Songzhu Zheng, Lu Pang, Haibin Ling, and Chao Chen. 2023. Attention-Enhancing Backdoor Attacks Against BERT-based Models. arXiv:2310.14480 [cs.LG] https://arxiv.org/abs/2310.14480

[9] Anubha Pandey. 2024. Retrieval Augmented Fraud Detection. In *Proceedings of the 5th ACM International Conference on AI in Finance* (Brooklyn, NY, USA) *(ICAIF '24)*. Association for Computing Machinery, New York, NY, USA, 328–335. doi:10.1145/3677052.3698692

[10] Shini Renjith. 2018. Detection of Fraudulent Sellers in Online Marketplaces using Support Vector Machine Approach. *International Journal of Engineering Trends and Technology (IJETT)* 57, 1 (March 2018), 48–53. doi:10.14445/22315381/IJETT-V57P210

[11] Yeeun Yoo, Jinho Shin, and Sunghyon Kyeong. 2023. Medicare Fraud Detection Using Graph Analysis: A Comparative Study of Machine Learning and Graph Neural Networks. *IEEE Access* 11 (2023), 88278–88294. doi:10.1109/ACCESS.2023.3305962

[12] Chang Yu, Yongshun Xu, Jin Cao, Ye Zhang, Yixin Jin, and Mengran Zhu. 2024. Credit Card Fraud Detection Using Advanced Transformer Model. In *2024 IEEE International Conference on Metaverse Computing, Networking, and Applications (MetaCom)*. 343–350. doi:10.1109/MetaCom62920.2024.00064