Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Supplementary materials for

Min GAO, Shutong CHEN, Yangbo GAO, Zhenhua ZHANG, Yu CHEN, Yupeng LI, Qiongzan YE, Xin WANG, Yang CHEN, 2024. Detecting compromised accounts caused by phone number recycling on e-commerce platforms: taking Meituan as an example. *Front Inform Technol Electron Eng*, 25(8):1077-1095. https://doi.org/10.1631/FITEE.2300291

1 IEEE-CIS dataset information

The IEEE-CIS dataset (Mainali et al., 2022; Nti and Somanathan, 2024), available on Kaggle (`https://www.kaggle.com/`), is a representative real-world e-commerce transaction fraud dataset provided by a world's leading payment service company called Vesta (`https://www.vesta.io/`). This dataset contains transaction records with diverse features, ranging from device types to product features. These diverse features make it a suitable one for studying fraud detection and transaction behavior in financial platforms. This dataset could also contain other factors like identity theft (Zou et al., 2020; Wang and Zhu, 2022), another typical fraud type that could potentially contribute to compromised accounts. For example, e-commerce transaction fraud is a potential consequence of identity theft. Therefore, we examine our model on the IEEE-CIS dataset to detect compromised accounts via other factors like identity theft. Based on the dataset, we identify unique user IDs to represent individual user accounts. The transaction behavior sequences of each account have been extracted according to their associated user IDs. A compromised account has one or more fraudulent transactions, and is labeled as 1. A normal account has no fraudulent transactions, and is labeled as 0. With the above preprocessing strategy, we obtain 78 745 accounts. Among them, there are 8585 compromised accounts. We split the extracted transaction behavior sequences into two sub-sequences, which are used as inputs for our model. Because the temporal information for each account is relatively sparse, we simply set the lengths of the two sub-sequences to be 5. Each account has 356 statistical features after data preprocessing. Similarly, the training, validation, and test sets are randomly split with a ratio of 3:1:1.

2 Experimental setup for baseline methods

- LightGBM (Ke et al., 2017): A well-known implementation of the gradient boosting decision tree (GBDT) that is widely used in both industry and academia. In our experiments, we implement a LightGBM model with 500 decision trees for capturing statistical features.

- GBDT embedding + NN (Ke et al., 2019): We leverage a two-layer fully-connected neural network as a classifier based on the leaf embedding of a GBDT model trained on features with 500 decision trees.

- Long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997): A popular recurrent neural network (RNN) to process sequences. As a variant of the vanilla RNN (Keren and Schuller, 2016), it aims at tackling the gradient vanishing and explosion problem by introducing a memory cell controlled by gate functions to better characterize long-term dependencies. In our experiments, the LSTM model consists of two units for processing payment and login sequences, and a single-layer neural network for making the final prediction according to the concatenation of the final hidden states of two sequences.

- GRU (Cho et al., 2014): A popular variant of RNN and LSTM. GRU reduces the number of parameters and can converge faster than LSTM during training with a similar or lower loss (Cho et al., 2014). The

GRU model uses the same setting as the LSTM model except for replacing LSTM units with GRU units.

- DeepScan (Gong et al., 2018): A representative method for malicious account detection in location-based social networks. The method captures information from the spatiotemporal data, and it models user dynamic behaviors and statistical features. In our experiments, we follow the same setting as DeepScan. The payment and login sequences are modeled by a bidirectional LSTM network, and statistical features serve as the conventional features within DeepScan.

- Al-Qurishi et al. (2018)'s: An approach for detecting malicious behaviors in online social networks by integrating content, graph, and profile data. This method exploits principal component analysis (PCA) and random forest classifier to distinguish normal users from fake ones. In our experiments, there is no corresponding graph data. We take statistical features as the input of PCA, and a random forest classifier is used to predict whether an account is compromised or normal.

3 Performance of TSF for the IEEE-CIS dataset

3.1 Performance comparison

We present the results of all methods for the IEEE-CIS dataset in Table S1, and the corresponding precision–recall curves are illustrated in Fig. S1. First, one can observe that the proposed TSF outperforms all the baseline models for all metrics in Table S1. The result suggests that the proposed TSF can effectively identify the compromised accounts.

Second, the sequential models (i.e., the GRU and LSTM models) focus on learning the temporal relationships in users' behavioral sequences, while the ensemble models (i.e., LightGBM, Al-Qurishi et al.

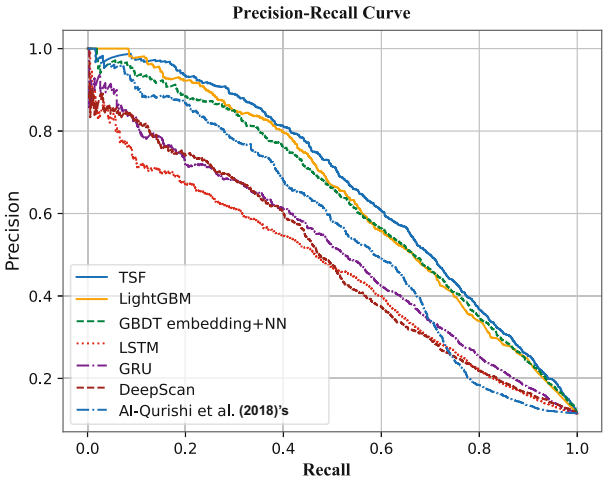

Fig. S1 Precision–recall curves of TSF and baselines for the IEEE-CIS dataset

Table S1 Performances of TSF and baselines for the IEEE-CIS dataset

Model	AUC	R@80%P	R@85%P	R@90%P
LightGBM (Ke et al., 2017)	0.8786	0.3966	0.2983	0.1486
GBDT embedding+NN (Ke et al., 2019)	0.8836	0.3508	0.1950	0.1829
GRU (Cho et al., 2014)	0.8308	0.1072	0.0696	0.0149
LSTM (Hochreiter and Schmidhuber, 1997)	0.8058	0.0640	0.0376	0.0149
DeepScan (Gong et al., 2018)	0.8100	0.1259	0.0564	0.0099
Al-Qurishi et al. (2018)'s	0.7968	0.2790	0.2121	0.1044
TSF	**0.8927**	**0.4133**	**0.3564**	**0.2878**

AUC: area under curve; R@T%P: Recall@T%Precision. T: 80, 85, or 90. The best results are in bold with $p < 0.001$

(2018)'s, and GBDT embedding + NN) use hand-crafted rules that may capture the anomalies in users' multiple operations to classify accounts. Different from the results of the Meituan dataset, according to Table S1, the ensemble models outperform the sequential models due to the sparse sequence patterns for accounts from the IEEE-CIS dataset. Although DeepScan leverages both sequential models and ensemble models, its performance suffers from poor feature fusion capabilities and inability to account for interactions among temporal sequences from the same account.

Overall, TSF outperforms other baselines with a 0.91%–8.69% AUC increase and a 1.67%–34.93% recall increase under three fixed thresholds of precision for the IEEE-CIS dataset. Furthermore, we conduct McNemar's test (McNemar, 1947) on the outcomes (in both Table S1 and Table S2), with the results ($p < 0.001$) confirming that our model is distinctly different from other methods.

3.2 Ablation study

To explore the importance of each module in our model, we perform ablation studies by removing one specific module at a time. Similarly, we evaluate the detection performance of the following variants for the IEEE-CIS data. (1) TSF (w/o TPE): TSF without the whole temporal pattern encoder, i.e., the statistical feature encoder in Fig. 4; (2) TSF (w/o SFE): TSF without the pre-trained GBDT or the single-layer neural network to transform leaf embedding, i.e., the temporal pattern encoder in Fig. 4; (3) TSF (w/o IA): TSF without the inter-attention mechanism to align payment behaviors with login behaviors; and (4) TSF (w/o SA): TSF without the self-attention mechanism but using concatenation to fuse information.

As shown in Table S2 and Fig. S2, the performance degradation indicates that all modules of our model contribute to the task of detecting compromised accounts for the IEEE-CIS dataset. The followings are the analyses of the ablation study:

Temporal pattern encoder. By removing the temporal pattern encoder, the AUC score drops by 1% to 0.8827, and the recall scores also decrease with the improvement of precision scores. Our results highlight the

Table S2 Results of the ablation study for the IEEE-CIS dataset

Model	AUC	R@80%P	R@85%P	R@90%P
TSF (w/o TPE)	0.8827 (−1.00%)	0.3408 (−7.25%)	0.2955 (−6.09%)	0.2039 (−8.39%)
TSF (w/o SFE)	0.8268 (−6.59%)	0.0912 (−32.21%)	0.0282 (−32.82%)	0.0099 (−27.79%)
TSF (w/o IA)	0.8891 (−0.36%)	0.3558 (−5.75%)	0.2912 (−6.52%)	0.1939 (−9.39%)
TSF (w/o SA)	0.8890 (−0.37%)	0.3541 (−5.92%)	0.2906 (−6.58%)	0.2088 (−7.90%)
TSF	**0.8927**	**0.4133**	**0.3564**	**0.2878**

AUC: area under curve; R@T%P: Recall@T%Precision. T: 80, 85, or 90. The values in parentheses indicate the ratio of metric value decrease compared with that of TSF. The best results are in bold with $p < 0.001$

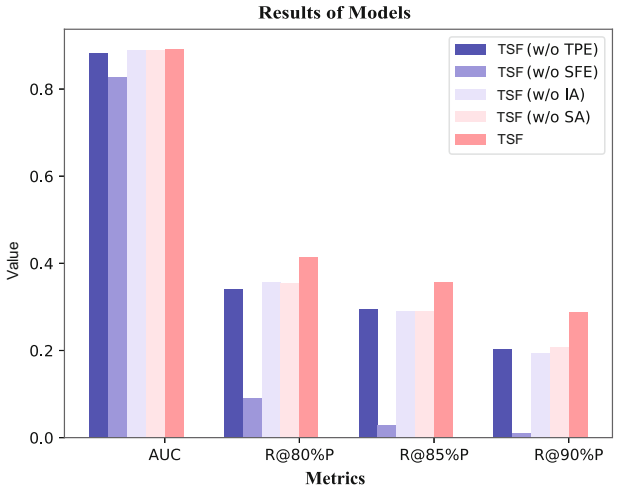

Fig. S2 Ablation study for the IEEE-CIS dataset

importance of capturing the temporal patterns. For the IEEE-CIS dataset, each account has little temporal information, resulting in a limited contribution from the temporal pattern encoder.

Statistical feature encoder. For the IEEE-CIS dataset, removing the statistical feature encoder has the greatest impact on model performance. The AUC score drops by 6.59%, and the recall score decreases by up to 32.82%. Additionally, the variant TSF (w/o SFE) model performs much worse than the variant TSF (w/o TPE) for both the AUC score and the recall scores under different precision settings. The result highlights the importance of specifying rules to capture significant behaviors. The technique of leaf embedding allows us to apply the knowledge from a pre-trained GBDT model to detect compromised accounts.

Inter-attention module. The performance of our model decreased slightly after we removed the inter-attention module, which captures the interaction between two different sequences. The result implies that the inter-attention module also contributes to the detection performance of TSF as well as that the correlation between two sequences is essential.

Self-attention module. In our experiments, the model with the self-attention module consistently outperforms TSF (w/o SA) with the recall improved by up to 7.90%. Notably, this module can interpret the detection results by demonstrating the significance of different representations to the final prediction.

4 Description of statistical features for the Meituan dataset

For the Meituan dataset, the statistical features are listed in Table S3. The features in the first four subsets characterize the corresponding behaviors, while those in "other features" describe the overall activity of the account. These features aim to capture the statistical features that can distinguish compromised accounts from normal ones. For example, the unusually high "total payment amount" in a short period suggests that the account may be undergoing abnormal transactions related to card fraud. Additionally, if "the card holder's name is not consistent with the real-name identity" when binding a new bank card, this indicates that the account's owner may have changed.

Table S3 Statistical features of each account used in the statistical feature encoder of TSF

Subset	Statistical feature
Payment features	• Number of payment records in 30 d/90 d/6 months • Number of failed payment records in 30 d/90 d • Total payment amount in 30 d/90 d/6 months • Average amount per payment in 30 d/90 d/6 months • Total amount of credit payments in 30 d/90 d/6 months • Average amount of credit payments in 30 d/90 d/6 months
Login features	• Number of logins via SMS in 90 d • Number of failed login attempts in 90 d • Number of logins via SMS on a new device in 90 d
Card binding features	• Number of card binding records • Is the cardholder's name consistent with the realname identity? • Number of failed card bindings in 30 d/90 d • Number of successful card bindings in 30 d/90 d
Payment password features	• Whether the account has attempts to retrieve payment password • Whether the account has ever successfully retrieved payment password
Other features	• Maximum length of silent periods • Total length of silent periods • Number of silent periods • Number of operations at midnight in 30 d/90 d

References

Al-Qurishi M, Hossain MS, Alrubaian M, et al., 2018. Leveraging analysis of user behavior to identify malicious activities in large-scale social networks. *IEEE Trans Ind Inform*, 14(2):799-813. https://doi.org/10.1109/TII.2017.2753202

Cho K, van Merriënboer B, Gulcehre C, et al., 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. Proc Conf on Empirical Methods in Natural Language Processing, p.1724-1734. https://doi.org/10.3115/v1/D14-1179

Gong QY, Chen Y, He XL, et al., 2018. DeepScan: exploiting deep learning for malicious account detection in location-based social networks. *IEEE Commun Mag*, 56(11):21-27. https://doi.org/10.1109/MCOM.2018.1700575

Hochreiter S, Schmidhuber J, 1997. Long short-term memory. *Neur Comput*, 9(8):1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Ke GL, Meng Q, Finley T, et al., 2017. LightGBM: a highly efficient gradient boosting decision tree. Proc 31st Int Conf on Neural Information Processing Systems, p.3149-3157.

Ke GL, Xu ZH, Zhang J, et al., 2019. DeepGBM: a deep learning framework distilled by GBDT for online prediction tasks. Proc 25th ACM SIGKDD Int Conf on Knowledge Discovery & Data Mining, p.384-394. https://doi.org/10.1145/3292500.3330858

Keren G, Schuller B, 2016. Convolutional RNN: an enhanced model for extracting features from sequential data. Proc Int Joint Conf on Neural Networks, p.3412-3419. https://doi.org/10.1109/IJCNN.2016.7727636

Mainali P, Psychoula I, Petitcolas FAP, 2022. ExMo: explainable AI model using inverse frequency decision rules. Proc 3rd Int Conf on Human-Computer Interaction, p.179-198. https://doi.org/10.1007/978-3-031-05643-7_12

McNemar Q, 1947. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153-157. https://doi.org/10.1007/BF02295996

Nti IK, Somanathan AR, 2024. A scalable RF-XGBoost framework for financial fraud mitigation. *IEEE Trans Comput Soc Syst*, 11(2):1556-1563. https://doi.org/10.1109/TCSS.2022.3209827

Wang C, Zhu HY, 2022. Representing fine-grained co-occurrences for behavior-based fraud detection in online payment services. *IEEE Trans Depend Secure Comput*, 19(1):301-315. https://doi.org/10.1109/TDSC.2020.2991872

Zou YX, Roundy K, Tamersoy A, et al., 2020. Examining the adoption and abandonment of security, privacy, and identity theft protection practices. Proc CHI Conf on Human Factors in Computing Systems, p.1-15. https://doi.org/10.1145/3313831.3376570