

# Bitcoin Illegal Transaction Detection Model Based on Time Step and Ensemble Learning

Zelai Yang and Xudong Li<sup>(✉)</sup>

College of Software, Nankai University, Tianjin 300000, China  
leexudong@nankai.edu.cn

**Abstract.** In recent years, bitcoin, as the leading digital currency, has grown in value. However, due to the anonymity of the bitcoin system, it is convenient for people to carry out illegal activities on it. This will lead to irreversible loss to investors' rights and interests, resulting in significant economic loss. Thus, detecting and combating illegal transactions becomes crucial. This research is dedicated to solving the problem of detecting illegal transactions in open-source bitcoin transaction datasets. We propose a bitcoin illegal transaction detection model based on time step and ensemble learning. The model focuses on the time-sensitive nature of illegal behaviour in reality by grouping the dataset at the time step. Moreover, the model leverages oversampling techniques in the ensemble learning stage to improve the recall. Experimental results indicate that the proposed model can better capture the prevalent illegal patterns of the bitcoin system on different time steps. Results also show that this model can achieve high precision and recall (precision=0.99, recall=0.9) in the scope of elliptic dataset, thus improving the detection rate of illegal transactions.

**Keywords:** Bitcoin, Illegal transaction, Time step, Ensemble learning, Oversampling.

## 1 Introduction

Blockchain technology has garnered significant attention in academia and industry due to its decentralized, immutable, anonymous, and secure nature. Currently, the digital currency, represented by the Bitcoin, has been researched and utilized by an increasing number of individuals, thus leading to a rise in its value and recognition. However, the decentralized and anonymous features of blockchain, when compared with the traditional financial system, render transactions more challenging to trace, offering convenience for illegal activities. As a result, illegal transactions such as money laundering and phishing are commonplace, using Bitcoin as a cloak. Hence, detecting and combating illegal transactions becomes crucial. Although these illegal transactions represent a small fraction of the total, identifying them from the massive transactions not only helps to analyze the transactions and detect coin mixing on cryptocurrencies [14], but also provides strong support for the regulatory agencies to fight against illegal and criminal activities.

Several Researchers have publicly released the Bitcoin [17] and Ether [18] transaction datasets. In Bitcoin research, scholars have employed classical machine learning techniques such as Random Forest, XGBoost, and neural networks for data mining. Although Random Forest shows superior performance in precision and recall metrics, the highly imbalanced dataset, with illegal transactions being a minority, hampers the effectiveness of these algorithms, particularly in recall. Considering that the recall represents the percentage of detected illegal transactions to the total number of illegal transactions, the cost of misclassifying illegal money laundering transactions as legal ones is unacceptable in reality. Therefore, increasing the recall rate without decreasing the precision becomes a difficult problem to be solved.

Aiming at the problem of detecting illegal transactions on the open-source Bitcoin transaction dataset, this paper constructs a Bitcoin illegal transaction detection model based on time step and ensemble learning. The main contributions of this paper are as follows:

1. With reference to the reality that illegal behaviors are often time-sensitive, the model groups the dataset based on time steps to make it easier for the classifier in each stage to learn the mainstream illegal patterns in different time steps. Then, the model integrates the classifiers in each time step through ensemble learning to make the model achieve a better performance than the individual classifiers.
2. Considering the high cost of missing to detect illegal transactions in reality, the model we proposed alleviates the problem of imbalanced categories in the dataset by introducing oversampling at ensemble learning stage, which effectively improves the ability to identify illegal transactions.
3. It is experimentally verified that the model proposed in this paper shows high precision and recall on Elliptic dataset. In addition, we do comparative experiments on oversampling techniques to get the applicability of SMOTE and ROS when they used at different stages of the model.

## 2 Related Work

This section describes the work related to transaction classification in the Bitcoin domain through three aspects: datasets, conventional machine learning and deep learning models, and the applicability of supervised and unsupervised learning.

### 2.1 Work on the Dataset

In 2019, Weber et al. [17] contributed Elliptic, the largest publicly available transaction dataset on Bitcoin, to provide dataset support for subsequent illegal transaction detection work. The dataset consists of over 200k Bitcoin transaction data and 234k payment streams. Youssef Elmougy et al. [11] expand the Elliptic dataset and propose Elliptic++ by adding 17 features to each transaction and provides a dataset with 822k wallet addresses.

## 2.2 Machine Learning and Deep Learning Work

In bitcoin transaction classification task, the task can also be viewed as anomaly detection. This type of problem is often modeled and handled by machine learning and deep learning models. Weber et al. [17] use machine learning and GCN models to classify illegal transactions, while Johrha Alotibi et al. [5] explore the applicability of several machine learning methods and deep learning methods in anti-money laundering tasks. The results [5,17] show that random forests compared to other machine learning methods have better performance in this task. Ahmad Al Badawi et al. [7] use shallow neural networks and random forests to handle money laundering detection and get better accuracy on Elliptic dataset. [6] and [12] both leverage LGBM model for transaction classification, where the former classifies the data, while the latter forms a transaction history summary by extracting features from transactions, and then classifies the transaction accounts by LGBM model. Ismail Alarab et al. [3] explore the effectiveness of various ensemble learning and points out that ensemble learning outperforms the classical machine learning models.

Since Bitcoin naturally has the property of transferring in and out, if transactions are considered as nodes and money flows as edges, the transaction classification problem can be viewed as a graph node classification problem. [17] uses GCN model to classify illegal transactions on the Elliptic dataset, and the results are used as a benchmark by subsequent research. Ismail Alarab et al. [4] use an improved graph convolutional network model, which connects potential features from the outputs of graph convolutional and linear layers to predict illegal transactions. [15] discards the gradient-based optimizer and uses a differential optimization algorithm as the optimizer for the graph convolutional model, which reduced the training time. [16] proposes scalable graph convolutional neural networks and experimentally demonstrated the promise of graph deep learning in the field of anti-money laundering. In general, traditional machine learning models have superior performance, but we still maintains the expectation of improving graph convolutional networks and other deep learning models to migrate to this task.

## 2.3 Applicability of Supervised and Unsupervised Learning

Supervised learning and unsupervised learning are the two main machine learning paradigms in this domain. Supervised learning models learn from transaction labels to identify known illicit behaviors, and unsupervised learning models need to explore the features of Bitcoin transaction data and identify potential money laundering patterns.

Madhuparna Bhowmik et al. [8] compare the performance of various supervised machine learning techniques for the detection of illicit transaction patterns. Joana Lorenz et al. [13] use unsupervised learning models and active learning for illegal pattern detection. Experiments show that existing unsupervised learning methods are not sufficient to detect illicit patterns in bitcoin transaction datasets. Moreover, anomalies in the feature space do not indicate illegal behavior, partial synthetic data can be misleading for the model. [2] explores the impact of various sampling techniques on classification of bitcoin data and ethereum data in imbalanced dataset conditions. Among them, undersampling performs the best. This study agrees with Joana Lorenz's [13] experimental conclusion that oversampling technique performs synthesis of samples and some of the synthesized data is misleading, making the dataset distorted. Taking a different view,

Ana Isabel Canhoto et al. [9] argues that there is a lack of high-quality datasets and therefore supervised learning approaches have limited applicability. Therefore, the authors suggest using reinforcement machine learning, or unsupervised learning approach for money laundering detection.

### 3 Methodology

The overall illicit transaction detection model proposed in this paper consists of three steps. The data preprocessing part accomplishes data screening, data grouping and training/testing set partitioning. The next step is to train the classifier for each time step and select qualified classifiers. Finally, ensemble learning is used to integrate the classifiers obtained in the previous stage and obtain the classification results. The structure of the model proposed in this paper is shown in Fig. 1.

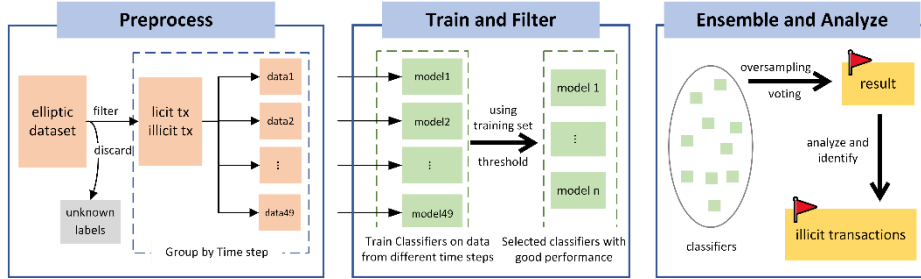


Fig. 1. The structure of our model.

#### 3.1 Preprocessing

In the preprocessing stage, considering that the classification model used subsequently is a supervised learning model, which requires real labels on the dataset, data with unknown labels are discarded. In addition, this paper draws on the reality that illegal behavior patterns are characterized by timelines. The popularity of a scam may spread rapidly over a certain period of time, attracting a large number of victims [10].

Therefore, we divide the dataset according to time step based on the popular scam trend, which makes it easier for the model to learn the popular illicit patterns in the Bitcoin system on different timelines. We first divide the dataset into training set and testing set according to a certain ratio, and subsequently divide the data in the training set according to 'Time step' feature values.

We denote *Elliptic* as the original transaction dataset, where one transaction is denoted as  $tx$ ,  $dataset$  represents a collection of transactions labeled as 'licit' or 'illicit',  $Timestep(tx)$  represents the values of the time step features of the transaction  $tx$ ,  $Label(tx)$  represents the label of the transaction  $tx$ , and  $data_i$  represents grouped data packets, then:

$$dataset = \{tx \mid Label(tx) = licit \vee Label(tx) = illicit, tx \in Elliptic\}$$

$$data_i = \{tx \mid Timestep(tx) = i, tx \in dataset\}$$

Since the Time-step's values range from 1 to 49, we can obtain data packets  $data_i$ , where  $i$  range from 1 to 49,  $i \in Z^+$ .

### 3.2 Training and Selecting Classifiers

Let  $classifier_i$  represent the classifier trained with  $data_i$ , then the data packets  $data_i$  can train  $n$  classifiers ( $n = 49$ ). Each classifier can be targeted to learn illegal trading patterns in the corresponding time step.

Since the Random Forest algorithm shows better performance on this task [5,17], this paper uses Random Forest(RF) as classifier in the experiments. RF classifies samples by constructing multiple decision trees and combining their results. A subset of features is randomly selected at each node for decision tree splitting. we adopt the RF as classifier for  $data_i$  on each time step  $i$ , and utilize  $data_i$  to train the RF classifiers. 49 random forest models can be trained and obtained in 49 time steps. Then, we filtered these 49 classifiers according to their performance on the training set, and set the threshold  $threshold$  to discard some random forest classifiers to ensure that all the random forests participating in the next ensemble stage have nice classification ability.

Denote the recall and precision of  $classifier_i$  on  $data_i$  as  $Recall(classifier_i)$  and  $Precision(classifier_i)$  respectively, the set of classifiers  $classifier_i$  that have been selected satisfies:

$$Recall(classifier_i) \geq threshold \wedge Precision(classifier_i) \geq threshold$$

### 3.3 Ensemble Learning and Oversampling

The purpose of this step is to integrate the classifiers obtained after training and filtering, so that the model obtains a better performance than the individual classifiers. We use ensemble learning(soft and hard voting methods) to do the combination of selected classifiers. The soft voting method votes based on the predicted probabilities of the classifier and the hard voting method votes based on the classification results of the classifier.

Oversampling is a method of dealing with unbalanced datasets, which contributes to category balance by increasing the number of minority class samples. Considering the imbalance in the dataset (low percentage of illegal transactions), oversampling was used to help the model focus on illegal transactions when conducting the voting method. Common oversampling methods are ROS and SMOTE. ROS copies the original minority class samples to increase the number of minority, while SMOTE interpolates between the original minority class samples and generates new synthetic samples.

## 4 Experiment

### 4.1 Dataset Description

The Elliptic dataset [1] is the classic bitcoin transaction dataset available in open source. The dataset consists of 203,769 node transactions and 234,355 directed edge

payments flows. Each transaction can be viewed as a node with 166 features, including 94 local features and 72 aggregated features. The nodes are labeled with three types (illicit, licit, and unknown). The description of the transaction is shown in the Table 1.

**Table 1.** Elliptic dataset description.

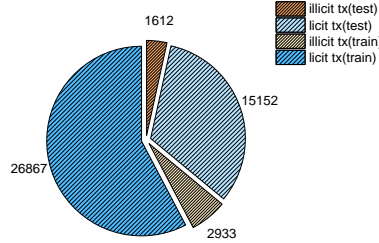
Transactions	Licit	Illicit	Unknowm	Total
Number	42019	4545	157205	203769
Proportion	21%	2%	77%	100%

In addition, the dataset contains 49 time steps, each representing a collection of transactions that appeared in the Bitcoin blockchain in less than three hours. The time step feature in each transaction represents an estimate of the time when the transaction is confirmed by the Bitcoin network.

#### 4.2 Data Preprocessing

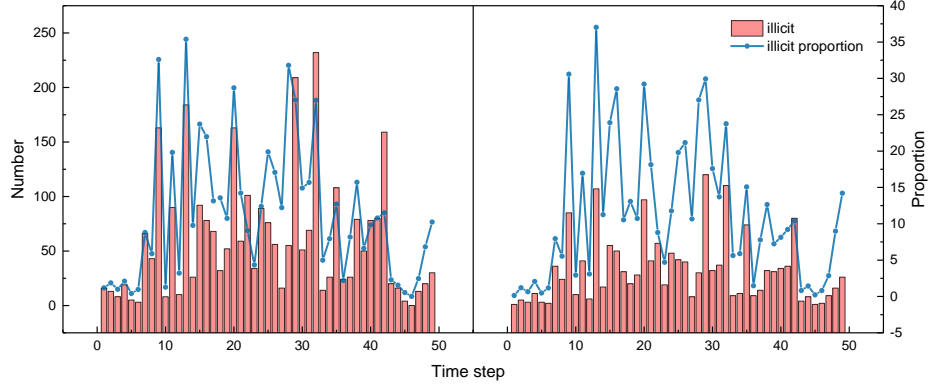
**Data Cleaning.** The classifier used for training with  $data_i$  in each time step is Random Forest. Since the supervised learning model requires true labels on the dataset, data with unknown labels are discarded for filtering. Thus we extracted 4,545 illegal transactions and 42,019 legal transactions to get a dataset containing 46,564 transactions.

From Fig. 2, it can be seen that labels in the dataset and testing set are extremely unbalanced. In the training set, there are 2,933 illegal transactions, accounting for 9.84% of the total transactions, and in the testing set, there are 1,612 illegal transactions, accounting for 9.62%.



**Fig. 2.** Detailed label distribution.

**Data Grouping.** In the original paper [17], the data with time step  $< 35$  is used as the training set, and the part with time step  $\geq 35$  is used as the testing set. We still keep the ratio of the number of transactions in the original paper. The illegal trade information of the divided training and test set at each time step is shown in Fig. 3.

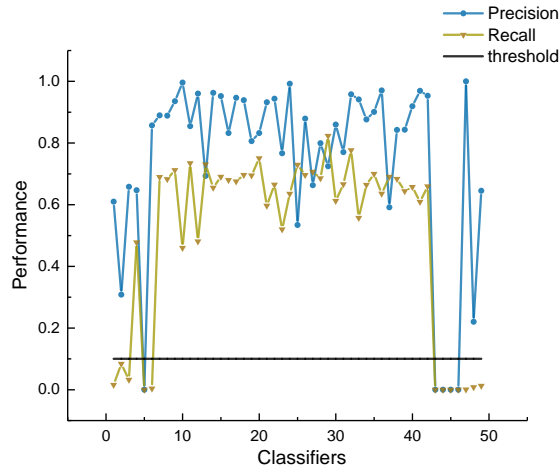


**Fig. 3.** The number and proportion of illicit transactions at each time step in the training set(left) and the testing set(right).

The figure reveals non-uniform percentages of illegal transactions at each time step. time step 13 and 46 exhibit the highest and lowest percentages of illegal transactions, accounting for 35.38% and 0%, respectively. Similarly, in the testing set, time step 13 and 1 exhibit the highest and lowest percentage, accounting for 37% and 0.13%, respectively.

### 4.3 Parameter Setting

We use Python and sklearn packages for model implementation in the experiment. RF is trained with  $data_i$ , where the parameter  $n\_estimators$  for RF is set to 50 (same as [17]) and the parameter  $threshold$  is set to 0.1.



**Fig. 4.** The performance of 49 RF classifiers on the training set.

The performance of RF classifiers on the training set is shown in Fig. 4. It is clear that 12 RF classifiers will be discarded and only 37 RF classifiers remain. The reason for

the poor performance is related to the input training data. It is difficult for the classifiers to learn useful information due to the small percentage of illegal transactions in the training data. For example,  $classifier_{46}$ , the percentage of illegal transactions in the training set at its corresponding time step is 0%, so it is understandable for  $classifier_{46}$  to perform poorly.

The oversampling is implemented using the imblearn library with parameters setting:  $sampling\_strategy=0.5$ ,  $random\_state=42$ . The parameter  $k\_neighbors$  is set to 3 for SMOTE; MLP uses Adam as the optimizer, with  $learning\_rate\_init$  set to 0.01,  $max\_iter$  set to 1500.

#### 4.4 Model Evaluation

The model performance is presented through four metrics: Precision, Recall, F1 and micro-F1. In this paper, we focus on the Precision and Recall values, where the former represents the accuracy of predicting positive instances, and the latter measures the model's ability to find out all illegal transactions. We use the transaction data in Elliptic as the dataset and regard the original paper [17] as the benchmark.

Following three research questions of interest, experiments are designed, and answers are given based on the results of the experiments.

- **Question1:** Does the model perform better than the original paper [17]?
- **Question2:** Is the time-step division of data effective for the task of detecting illicit transactions?
- **Question3:** Is the introduction of oversampling in ensemble learning stage effective?

## 5 Results and Analysis

This section designs experiments and gives experimental results based on the three problems in section 4.4.

Table 2 shows the performance of the model proposed in this paper when the data is grouped by time steps, and the overall model structure is shown in Fig. 1.

Table 3 demonstrates the performance of the model when the data is not grouped by time step, and the result of original paper is located in the first row of Table 3 as a benchmark. In this table, we do not use ensemble learning, and the oversampling step is put ahead into the classifier training phase.

#### Answer to Question1

From Table 2, it can be seen that ensemble learning using voting has higher Precision and Recall than XGBoost and MLP, and the model we proposed outperforms the original paper on all four metrics (see the first row of Table 3).

Both SMOTE and ROS improve the model's recall value, which indicates that the model's ability to find out all illegal transactions has improved. The recall value of the model using SMOTE and soft voting can reach 0.9, which is about 2% – 3% higher than that of the model without using oversampling, but its precision also decreases from 0.995 to 0.99. Considering that the cost of missing to detect the illicit transactions in



reality will be high, we argue that it is worthwhile to sacrifice a slight loss of precision to improve the recall value.

**Table 2.** The model performance when data is grouped by time steps(Using ensemble learning).

Oversampling	Ensemble Methods	Precision	Recall	F1	Micro-F1
\	soft voting	0.995	0.875	0.931	0.988
\	hard voting	0.995	0.873	0.93	0.987
\	XGBoost	0.96	0.841	0.897	0.981
\	MLP	0.947	0.846	0.894	0.98
SMOTE	soft voting	0.99	0.9	0.943	0.99
SMOTE	hard voting	0.99	0.899	0.942	0.989
SMOTE	XGBoost	0.928	0.848	0.886	0.979
SMOTE	MLP	0.897	0.839	0.867	0.975
ROS	soft voting	0.99	0.897	0.94	0.989
ROS	hard voting	0.99	0.895	0.94	0.989
ROS	XGBoost	0.926	0.851	0.887	0.979
ROS	MLP	0.9	0.848	0.871	0.976

### Answer to Question2

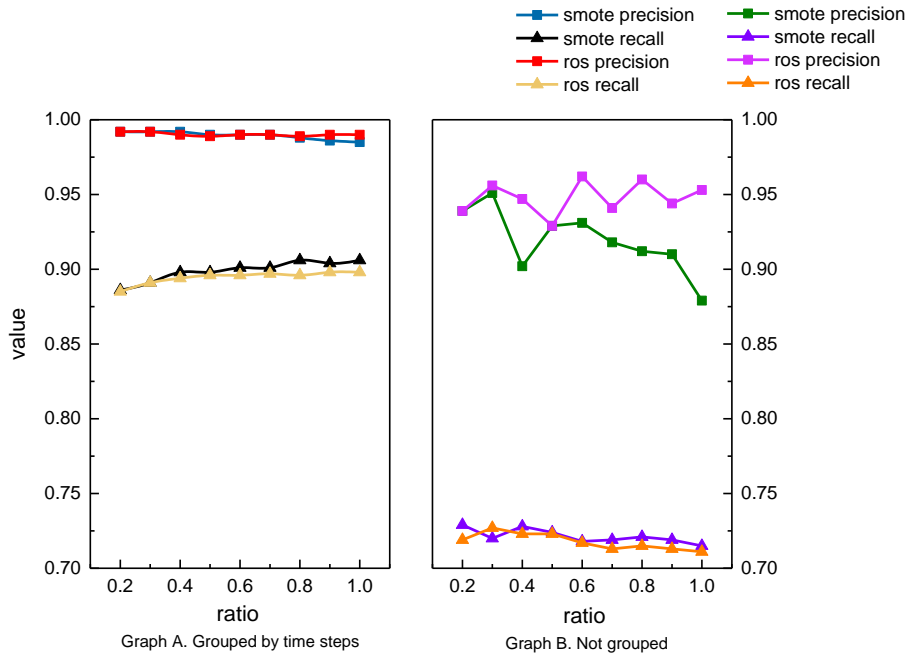
The models in Table 3 do not use ensemble learning and the oversampling step is put ahead into classifier training stage. It is easy to find that the models that used the oversampling step have higher recall than those that do not. Among them, model with XGBoost and oversampling achieves the highest recall of 0.73. In terms of precision, the original paper has the highest precision of 0.971, and the rest of the models have lower precision, which is all less than 0.95.

As a whole, the models involved in Table 3 are lower than those in Table 2 in four metrics. The step of grouping  $data_i$  according to the time step allows the model to better learn the patterns of illegal behaviors at different time steps, which enables the model to detect a fuller range of illegal behaviors along with high prediction. Therefore, the step is effective for the illegal transaction classification task. This not only demonstrates the effectiveness of time-step division in the illegal transaction classification task, but also emphasizes the importance of time as a feature in modeling illegal behavior detection.

### Answer to Question3

It is worth noting that, when models in Table 2 and our proposed model both use oversampling technique to obtain the improvement of recall, the precision of models in Table 3 decrease more drastically. This is due to the fact that when the oversampling method is used in the classifier training, SMOTE synthesizes the data of illegal transactions and ROS copies the data of illegal transactions. Since the data synthesized by SMOTE is likely to be distorted, it will mislead the model, whereas the transaction data copied by ROS is real, and thus has less impact on the model's precision. This is

consistent with the view expressed in [13] - "anomalies in the feature space do not indicate illegal behavior, and experiments performed on partially synthesized data can be misleading".



**Fig. 5.** The influence of sampling methods at different stages of the model

**Table 3.** The model performance when data is not grouped by time steps(Not using ensemble learning).

Oversampling	Classifier	Precision	Recall	F1	Micro-F1
\	RF	0.971	0.675	0.796	0.978
\	XGBoost	0.916	0.724	0.809	0.978
\	MLP	0.598	0.602	0.6	0.948
SMOTE	RF	0.895	0.724	0.8	0.977
SMOTE	XGBoost	0.943	0.727	0.821	0.979
SMOTE	MLP	0.508	0.642	0.567	0.936
ROS	RF	0.929	0.723	0.813	0.978
ROS	XGBoost	0.892	0.73	0.803	0.977
ROS	MLP	0.504	0.66	0.571	0.936

Fig. 5 demonstrates the impact of using oversampling at different stages, where the x-axis represents the ratio of the number of minority class to majority class after

oversampling. The precision and recall values when oversampling is used at the ensemble stage(Graph A) are greater than the case when it is used at the classifier training stage(Graph B). This is due to the fact that oversampling done during ensemble stage does not involve the synthesis of illegal transaction data, but rather the synthesis of the classification results of the random forest classifiers.

When the sampling ratio gradually increases, the precision value in Fig. 5 Graph A decreases very little, but the recall value tends to increase. In Graph B, regardless of using SMOTE or ROS, the recall value of the model fluctuates slightly. When ROS is used, there is no significant decrease in the precision value as the sampling ratio increases. Therefore, ROS performs better than SMOTE when oversampling is done during classifier training.

## 6 Conclusion

In this paper, we construct an illegal transaction detection model based on time-step and ensemble learning in bitcoin, which makes it easier for the classifier to learn the illegal transaction patterns on different timelines by grouping the data according to time steps. In addition, instead of classifier training phase, the model introduces oversampling in the ensemble learning phase, which avoids the synthesis of distorted transaction data. The experimental part evaluates the model using the Elliptic dataset and provides an in-depth analysis of three research questions of interest. The results show that our proposed model can improve the detection rate of illegal transactions while maintaining a high precision. This is crucial for detecting illegal transactions in the real world, which helps to deter illegal transactions, combat illegal criminal activities and reduce economic losses.

In addition, the experimental results also point out that when using SMOTE to synthesize the raw data, the model is misled by the distortion of the synthesized data because the model identifies the synthesized data as real data for training. If some cases have to use oversampling to increase the proportion of minority samples, try to use ROS, or refer to this model and move the oversampling phase to the ensemble learning phase.

## References

1. <https://www.elliptic.co>
2. Alarab, I., & Prakoonwit, S. (2022). Effect of data resampling on feature importance in imbalanced blockchain data: Comparison studies of resampling techniques. *Data Science and Management*, 5(2), 66-76.
3. Alarab, I., Prakoonwit, S., & Nacer, M. I. (2020, June). Comparative analysis using supervised learning methods for anti-money laundering in bitcoin. In *Proceedings of the 2020 5th international conference on machine learning technologies* (pp. 11-17).
4. Alarab, I., Prakoonwit, S., & Nacer, M. I. (2020, June). Competence of graph convolutional networks for anti-money laundering in bitcoin blockchain. In *Proceedings of the 2020 5th international conference on machine learning technologies* (pp. 23-27).

5. Alotibi, J., Almutanni, B., Alsubait, T., Alhakami, H., & Baz, A. (2022). Money Laundering Detection using Machine Learning and Deep Learning. *International Journal of Advanced Computer Science and Applications*, 13(10).
6. Aziz, R. M., Baluch, M. F., Patel, S., & Ganie, A. H. (2022). LGBM: a machine learning approach for Ethereum fraud detection. *International Journal of Information Technology*, 14(7), 3321-3331.
7. Badawi, A. A., & Al-Haija, Q. A. (2021, November). Detection of money laundering in bitcoin transactions. In *4th Smart Cities Symposium (SCS 2021)* (Vol. 2021, pp. 458-464). IET.
8. Bhowmik, M., Chandana, T. S. S., & Rudra, B. (2021, April). Comparative study of machine learning algorithms for fraud detection in blockchain. In *2021 5th international conference on computing methodologies and communication (ICCMC)* (pp. 539-541). IEEE.
9. Canhoto, A. I. (2021). Leveraging machine learning in the global fight against money laundering and terrorism financing: An affordances perspective. *Journal of business research*, 131, 441-452.
10. Cross, C. (2019). Online fraud. In *Oxford research encyclopedia of criminology and criminal justice*.
11. Elmougy, Y., & Liu, L. (2023, August). Demystifying Fraudulent Transactions and Illicit Nodes in the Bitcoin Network for Financial Forensics. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (pp. 3979-3990).
12. Lin, Y. J., Wu, P. W., Hsu, C. H., Tu, I. P., & Liao, S. W. (2019, May). An evaluation of bitcoin address classification based on transaction history summarization. In *2019 IEEE international conference on blockchain and cryptocurrency (ICBC)* (pp. 302-310). IEEE.
13. Lorenz, J., Silva, M. I., Aparício, D., Ascensão, J. T., Bizarro, P., & Ascensão, J. T. (2005). Machine learning methods to detect money laundering in the bitcoin blockchain in the presence of label scarcity (2020). Preprint arxiv.
14. Sun, X., Yang, T., & Hu, B. (2022). LSTM-TC: Bitcoin coin mixing detection method with a high recall. *Applied Intelligence*, 52(1), 780-793.
15. Tasharofi, S., & Taheri, H. (2021, March). DE-GCN: differential evolution as an optimization algorithm for graph convolutional networks. In *2021 26th International Computer Conference, Computer Society of Iran (CSICC)* (pp. 1-6). IEEE.
16. Weber, M., Chen, J., Suzumura, T., Pareja, A., Ma, T., Kanezashi, H., ... & Schardl, T. B. (2018). Scalable graph learning for anti-money laundering: A first look. arxiv preprint arxiv:1812.00076, 1-7.
17. Weber, M., Domeniconi, G., Chen, J., Weidele, D. K. I., Bellei, C., Robinson, T., & Leiserson, C. E. (2019). Anti-money laundering in bitcoin: Experimenting with graph convolutional networks for financial forensics. arxiv preprint arxiv:1908.02591.
18. Zheng, P., Zheng, Z., Wu, J., & Dai, H. N. (2020). Xblock-eth: Extracting and exploring blockchain data from ethereum. *IEEE Open Journal of the Computer Society*, 1, 95-106.