No Pie in The Sky: The Digital Currency Fraud Website Detection

Haoran Ou[1], Yongyan Guo[1], Chaoyi Huang[1], Zhiying Zhao[1], Wenbo Guo[1], Yong Fang[1], and Cheng Huang[1,*]

School of Cyber Science and Engineering, Sichuan University, Chengdu, China
Corresponding author: Cheng Huang
opcodesec@gmail.com

Abstract. In recent years, digital currencies based on blockchain technology are growing rapidly. Therefore, many criminal cases related to digital currency also took place. One of the most common ways is to induce victims to invest. As a result criminals can obtain a large number of profits through fraud. Cybercriminals usually design the layout of digital currency fraud websites to be similar to normal digital currency websites. Use some words related to blockchain, digital currency, and project white papers to confuse victims to invest. Once the victims have invested a lot of money, they cannot use digital currency to cash out. Digital currency is also difficult to track due to its anonymity. In this paper, we classified and summarized the existing methods of identifying digital currency scams. At the same time, we collected 2,489 domain names of fraudulent websites in the digital currency ecosystem and conducted statistical analysis from the four aspects of website text, domain names, rankings, and digital currency transaction information. We proposed a method to detect the website based on domain name registration time, website ranking, digital currency exchange rate, and other characteristics. We use the random forest algorithm as a classifier. The experimental results show that the proposed detection system can achieve an accuracy of 0.97 and a recall rate of 0.95. Finally, the case study results show that the system gets better detection and accuracy than other security products.

Keywords: Blockchain · Digital currency fraud website · Ponzi scheme · Phishing.

1 Introduction

Since the first Bitcoin emerged in 2009, with development of the blockchain technology and the digital economy ecosystem, digital currencies have seen explosive growth. In addition to Bitcoin, thousands of digital currencies appear from time to time. As of the end of 2018 [1], there are more than 2,000 different digital currencies. Their total market value of up to 100 billion US dollars, higher than the GDP of 127 countries (as of 2019) [2]. As an indispensable trading platform for the ecosystem, hundreds of digital currency exchanges are emerging to facilitate

transactions between digital assets and traditional legal tender or other digital assets.

However, various MLM currency scams under the guise of block-chain [3] are also increasing. These digital currency fraud websites use high rebates as a gimmick to attract everyone to participate and absorb membership dues to collect money. In the end, the scam was exposed due to the severance of the capital chain [4]. The general public lacks professional network security knowledge, and they are often deceived by the advanced technical guise of these websites and various lofty backgrounds to invest in and finally cause serious economic losses [5]. The existence of these coins seriously threatens the safety of people's property and hinders the normal ecological development of blockchain technology.

How to identify digital currency fraud sites and prevent fraud attacks is a hot spot. [6] The blockchain community has begun to pay attention to fraudulent websites in the digital currency ecosystem. Several open-source databases (such as Crypto Scam DB and Etherscam DB) have collected this type of malicious domain names and their related URLs. The scammers [7] take advantage of these domain names and addresses to defraud victims and raise funds to obtain economic benefits.

The current research in this area is mainly on the detection of Ponzi schemes based on digital currencies, and the scope of detection is often limited to Bitcoin. It is detected by analyzing the characteristics of the smart Ponzi scheme [8], extracting smart contracts [9], and analyzing the abnormal transaction behavior of Bitcoin [10]. On the other hand, there are also studies on phishing websites and phishing accounts related to Bitcoin [11]. However, at this stage, there is still a lack of research on the use of automated technology to identify and classify digital currency websites as to whether they are fraudulent.

In response to the above problems, this paper proposes a method for identifying digital currency fraud websites based on machine learning. we first need to screen out the effective features from the extracted website data. Then we use the classification model, to realize the automatic recognition of the website.

The contribution of this article as follows:

- We propose a detection method for digital currency fraud sites. The detection of digital currency fraud sites is still rare. We have analyzed and categorized the existing detection methods for fraudulent digital currencies. Most of the current research is about whether there are abnormal transactions in Bitcoin or other digital currencies or the detection of Ponzi schemes related to digital currencies.

- We extract effective features to improve the accuracy of website detection. After statistical analysis and related literature review, we extracted text, domain names, website rankings, and mainstream exchanges from digital currency fraud sites. Using random forest algorithms, we established a digital currency fraud site detection model and realized the detection of digital currency fraud sites. And accurate classification of normal.

– Compared with the test results of Tencent website security center, our test results are more accurate. The experimental data set sources of this article are (https: //cryptoscamdb.org/) and (https://www.badbitcoin.org/ thebadlist/index). Based on this data set, the recall rate of the detection model is 0.95, and the accuracy is 0.97. At the same time, we conducted comparative experiments and feature importance analysis. Finally, there was a case study carried out to verify the accuracy of the classification results.

The rest of this article is structured as follows: Section 2 reviews the background and related literature on digital currency fraud sites. The third section discusses the data source and data extraction process used. Section 4 presents the classification results of machine learning algorithms and model performance evaluation. The fifth section ends with a discussion.

2 Background and Related Work

2.1 Digital currency

Digital currency is a digital asset that uses cryptography to ensure its creation security and transaction security. The first and most famous digital currency, Bitcoin, was released in 2009 [12]. So far, there are more than 2500 different digital currencies. With the rise of digital currencies in 2017, people pay more and more attention to digital currency exchanges to obtain or trade digital currencies.

Digital currency exchange [9] is a market where users can buy and sell digital currency. Many of them only provide trading services between digital currencies, while a few provide fiat currencies (such as U.S. dollars or euros) for digital currency transactions. Similar to the stock market [13], people obtain benefits because of changes in digital currency prices. There are three types of digital currencies: centralized exchanges (CEX) managed by companies or organizations, decentralized exchanges (DEX) that provide automated processes for peer-to-peer transactions, and hybrid transactions that combine the two.

2.2 Ponzi scheme

Ponzi scheme [14] refers to a means of investment fraud in the financial field. Many illegal pyramid schemes use Ponzi schemes to collect money. The essence of the Ponzi scheme is to pay the investors of the next round as proceeds to the investors of the previous round, and so on, involving more investors and funds. But investors and funds are limited. When investors and funds are unsustainable, the entire scam will immediately collapse.

Ponzi schemes [15] generally have common characteristics such as low risk, high return, and pyramid-like investor structure.

Bitcoin is currently used as the payment infrastructure for Ponzi schemes [16]. These are financial frauds disguised as high-profit investment projects: in fact, the Ponzi scheme only uses funds invested by new users who join the program to repay users, so when there are no longer new investments, it will collapse.

A large number of victims have realized that these websites are fraudulent and illegal in many countries, but Bitcoin-based Ponzi schemes are still spreading on the Internet [17]. A recent study investigated posts on bitcointalk.org (a popular Bitcoin discussion forum), and the results showed that there were more than 1,800 Ponzi schemes from June 2011 to November 2016. [18] Due to the lack of a data set of Bitcoin-related Ponzi addresses, it is very difficult to measure the economic impact of Bitcoin-based Ponzi schemes. Conservative estimates from September 2013 to September 2014, Bitcoin-based Ponzi schemes have raised more than 7 million U.S. dollars. [19]

The current research on digital currency fraud includes two categories, smart Ponzi schemes, and phishing.

2.3 Smart Ponzi schemes detection

Weili Chen et al. [14] obtained 200 smart Ponzi schemes by manually checking more than 3000 open-source smart contracts on the Etalum platform. Two characteristics are extracted from the transaction history and operation code of the smart contract. Finally, a classification model of the smart Ponzi scheme is proposed.

Marie Vasek et al. [8] studied the supply-demand relationship of Bitcoin-based Ponzi schemes. Daniel Liebau et al. [9] defined scams and used empirical data to evaluate the number of cases that met this definition to establish a digital currency world. Massimo Bartoletti et al. [1] conducted a comprehensive investigation on Ponzi scams and analyzed their behavior and impact from different perspectives.

Shen Meng et al. [20] took two types of abnormal trading behaviors, airdrop candy, and greedy capital injections, as typical representatives, and designed the two types of abnormal trading behavior judgment rules, and then abstracted the abnormal trading pattern diagram. On this basis, they use the subgraph matching technology to realize the recognition algorithm of Bitcoin's abnormal transaction behavior.

2.4 Phishing scam detection

Most of the existing methods of analyzing Bitcoin scams require the manual or semi-manual collection of websites related to digital currency scams on the Internet [21]. Then the researchers can use automated tools to analyze. Quantify the impact of the scam by examining the related transactions on the blockchain. Ross Phillips et al. [22] analyzed open-source blockchain-based website data. They applied DBSCAN clustering technology to the content of fraudulent websites. The result shows that the types of digital currency fraudulent websites are corresponding with prepaid and phishing fraud.

Xiongfeng Guo et al. [23] proposes a method of phishing account detection system based on blockchain transactions and uses Ethereum as an example to

verify its effectiveness. Specifically, they propose a graph-based cascading feature extraction method based on transaction records and a GBM-based double-sampling set algorithm to establish a recognition model. A large number of experiments show that the algorithm can effectively identify phishing scams

The existing detection methods (mentioned in section 2.3 and section 2.4) for information fraud websites do not detect digital currency fraud websites, but can only detect traditional phishing websites. Or it is only possible to classify the open-source digital currency fraud websites through the clustering algorithm. It cannot detect whether the website is a normal website or a digital currency fraud website.

Based on the research of open-source digital currency fraud websites, our paper designs a classification model of digital currency fraud websites based on machine learning algorithms by analyzing and extracting effective features such as text and domain names in the website. The model realize the two classifications of normal websites and digital currency fraud websites.

3 Methodology

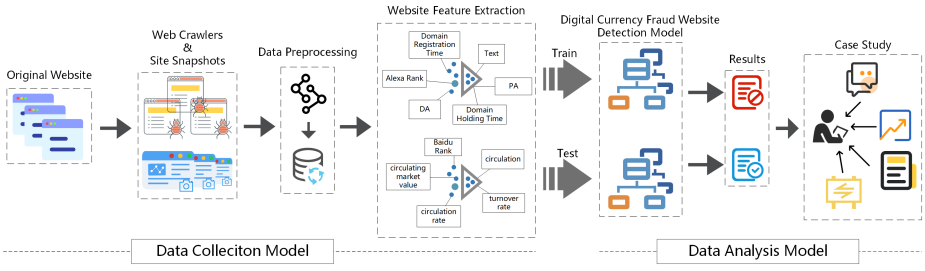

Fig. 1. The digital currency fraud website detection system

The framework of the digital currency fraud website detection system is shown in Figure 1. Data sources include normal websites and digital currency fraud websites. For more information about it, see section 3.1. We use web crawlers to collect the website's content, and for inaccessible websites, we use snapshots to collect related information. Next, we filter the website to remove unqualified websites and translate text. After data preprocessing operations, we obtain blacklist and whitelist data that meet the requirements. After analysis, we selected features such as text, search engines, website rankings (such as Alexa rankings), domain names, and mainstream exchanges to generate feature vectors as input to the detection model. Random forest (RF) is selected as the classifier, and support vector machine (SVM), naive Bayes (NB), and K-nearest neighbor algorithm (KNN) algorithms are comparative experiments. Use accuracy and precision indicators to evaluate the performance of the algorithm. Finally, for

the detection results, we conduct case study and verify the validity and accuracy of our detection results by consulting authoritative digital currency forums, exchange platforms, news media, and official announcements about the relevant comments and reports of this type of digital currency.

3.1 Data collection and preprocessing

Data collection It is divided into whitelist and blacklist data collection. The source of the whitelist data is the current mainstream digital currency websites and the Alexa's top 5000 websites. The blacklist includes digital currencies that are not on the mainstream trading platform and the publicly maintained blacklist list CryptoScamDB [24]. It is an open-source dataset website that stores more than 6,500 known fraud records on many chains such as Ethereum, Bitcoin, XRP, NEO, etc., which can be used to track malicious URLs and associate them with them Address. The entire website is open source, and all data sets and documents are available on GitHub. Among the collected websites are live websites and expired websites. We first traverse the secondary webpages belonging to the surviving website. After that, it needs to collect information such as pictures, text, DOM tree, homepage screenshots, and external URL links in the secondary webpages and the homepage. If we have no access to the webpage, we collect text and DOM tree information of these websites from screenshots.

Data preprocessing First of all, we translate the text of the collected websites. There is very little Chinese websites. According to the characteristics of the data set, we translate all other languages into English. As a result, the model can realize the detection of any language type website. Next, clean the text and delete all non-ASCII characters in the data. Finally, we obtain 3508 whitelisted websites with more than 12,000 pages and 2498 blacklisted websites with more than 12,000 pages.

3.2 Website feature extraction

Text feature Text information is an important part of the website and is widely used to identify phishing websites [25–27]. Among them, Adebowale et al. [25] proposed an intelligent phishing detection and protection scheme based on comprehensive features such as images, frames, and text. Digital currency scam website recognition also considers text features, while integrating other features for more accurate detection.

We use the Baidu translation interface to translate the source text information in the data set into English, delete punctuation marks and stop words, and perform word frequency statistics on the websites in the blacklist and whitelist. The top 300 words in the whitelisted word frequency statistics are removed from the top 300 words in the blacklisted word frequency statistics, which are used as the text features of the whitelisted website. Meanwhile, the words in the top 300 of the blacklisted word frequency statistics are removed from the top 300

of the whitelisted word frequency statistics and used as the text features of the blacklisted website.

Domain feature The domain is also one of the important features to effectively identify a digital currency scam website. Li Xiaodong et al. [28] proposed a multi-dimensional feature of the malicious website detection method. They incorporated the registration-level feature data into the study. Ross et al. [22] applied it to detect digital currency scam websites for the first time. The domain names selected in this paper have domain name registration time and domain name holding time.

Digital currency scam websites often have a relatively new registration time, and the domain name expires quickly, and the domain name is held for a short time. Statistics shows that 93% of normal website domain names were registered earlier than 2019, while only 45% of digital currency scam websites registered earlier than 2019. 41% of normal website domain names expire later than 2021, but only 18% of digital currency scam websites expire later than 2021. Finally, 71% of normal websites that hold domain names for more than 4 years, while only 18% of digital currency scam websites have more than 4 years.

Website ranking feature The website ranking feature is a side reflection of the popularity and authority of the website [18]. We can collect evaluation data through the ranking and inclusion situation provided by well-known companies and research on phishing website identification. Hu et al. [29] used publicly available website ranking data to build a classifier based on the machine learning algorithm. The website ranking features selected in the model include Alexa rank (AR), Baidu index (BD), Domain Authority value (DV), and Page Authority value (PV).

Alexa website ranking analyzes website visits to determine the website's popularity and give the website's world ranking [30]. It is the current more authoritative website visits evaluation index; Baidu inclusion means that the website is crawled by Baidu search engine, which can be passed The keywords are searched on Baidu. The more included, the higher the website weight and the more website traffic. Domain Authority value (DV) is an important index to measure the authority of a website. It has an effect on the authority of the whole site. Page Authority value (PV) can evaluate the authority of a page, and it affects the weight of a single page. According to statistics, 92.6% of digital currency scam websites are not in the previously collected Alexa top one million, and only 31% of normal websites are not in the previously collected Alexa top one million. Meanwhile, we counted the number of times that the domestic search engine Baidu included 75% of digital currency scam websites that were not included by Baidu, and only 15% of normal websites were not included by Baidu.

Mainstream exchange feature The characteristics of mainstream exchanges are unique to the identification of digital currency scam websites. Digital currency scams use digital currency as a gimmick to lure victims. Regular digital

currencies can inevitably be queried in mainstream exchanges, so we select the trading platform [31], circulation market value, circulation, circulation rate, and turnover rate as the characteristics of the digital currency listed on the market mainstream exchanges.

The trading platform on which coin is listed has not entered the comprehensive ranking of global exchanges and has not participated in the ranking; at the same time, the coin cannot inquire about the characteristics of the circulating market value, circulation, circulation rate, supply and turnover rate. The calculation formula for the feature is as follows:

$$Circulating\ market\ value = Circulation * Currencyprice \tag{1}$$

$$Circulation\ rate\ (CR) = (\frac{Total\ circulation}{Maximum\ supply}) * 100\% \tag{2}$$

The turnover rate is also called "turnover rate", which refers to the frequency of changing hands in the market within a certain period. It is an indicator reflecting the strength of liquidity. The 24H turnover rate calculation formula is as follows:

$$Turnover\ (TO) = \frac{24H\ Turnover}{Circulating\ market\ value} * 100\% \tag{3}$$

3.3 Digital currency fraud website detection model

The detection model of digital currency fraud websites proposed in this paper is implemented based on the random forest algorithm.

Random forest (RF) [32] is an ensemble learning algorithm, which is composed of a large number of decision trees aggregation, compared with a single decision tree, in order to reduce the variance of the experimental results. By aggregating the predictions of decision trees, a new prediction result is obtained. In the regression problem, the most direct and common process of random forest is to average the prediction results of a single decision tree, and use voting to determine the final prediction result, that is, the new prediction result of the T decision trees with the most classification The result is decided.

Now suppose that a fixed training data set D is composed of n observation results. a random forest algorithm model with T decision trees derives prediction rules based on the data set D. In an ideal situation, these prediction rules are estimated based on an independent test data set D_{test}, which consists of n_{test} test observations.

The true value of the i-th observation in the test data set $(i = 1, ..., n_test)$ is represented by y_i. In regression, in the case of binary classification, it is represented as a value of 0 or 1. The predicted value $(t = 1, ..., T)$ output by the decision tree t is represented as $haty_{it}$, and $\hat{y}_i$ is used to represent the predicted value of the entire random forest output. In the case of regression, the calculation formula of $\hat{y}_i$ is as follows:

$$\hat{y}_i = \frac{1}{T}\sum_{t=1}^{T}\hat{y}_{it} \tag{4}$$

In the case of classification, the value of i is usually obtained by majority voting. For binary classification, it is equivalent to calculating the same average value as regression, and the calculation formula adopted is as follows:

$$\hat{p}_i = \frac{1}{T}\sum_{t=1}^{T} I(\hat{y}_{it} = 1) \tag{5}$$

Use i to represent the probability, and finally derive the calculation formula of i as:

$$\hat{y}_i = \begin{cases} 1, & \hat{p}_i > 0 \\ 0, & \text{others} \end{cases} \tag{6}$$

The random forest algorithm has the advantages of fast detection speed, high detection accuracy, good anti-noise ability, not easy to overfit, and few hyperparameters. At the same time, the random forest algorithm can get the feature importance ranking, can process discrete data, and does not need to normalize the data set. It can be well applied to the data set of our paper, and it can help analyze which feature has the greatest impact on the detection result. Based on the above reasons, we finally chose the random forest algorithm to apply to the digital currency fraud website detection model.

4 Experiment

4.1 Dataset

The experimental data set consists of two parts, the normal website (whitelist) and the digital currency fraud website (blacklist). Whitelist is composed of the current mainstream digital currency official website and the top 5000 websites in Alexa. Blacklist is composed of digital currency websites which are not included in the mainstream trading platform and a publicly maintained list of fraudulent websites. -After screening, the data set used in the experiment has 3508 whitelist data and 2498 blacklist data. Each piece of data includes the website's text, domain name, search engine indexing, website ranking, mainstream exchanges, and other characteristics.

4.2 Experimental environment and evaluation metrics

To evaluate the detection model, we conducted experiments using a Ubuntu server with a 4-core 3.2 GHz Intel Core i7-8700 processor, 6GB GeForce GTX 1070 graphics processing unit (GPU), and 16GB memory.

To evaluate the performance of the model, the following indicators are used:

True Positive (TP). The model correctly predicts that the digital currency fraud website is a digital currency fraud website.

True Negative (TN). The model correctly predicts a normal website as a normal website.

False positive (FP). The model incorrectly predicts a normal website as a digital currency fraud website.

False Negative (FN). The model incorrectly predicted negative instances, that is, the model predicts that the digital currency fraud website is a normal website.

Accuracy (AC). The percentage of correctly classified records relative to the total records. If false positives and false negatives have similar costs, accuracy will be best:

$$AC = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

Precision (P). The percentage of predicted digital currency fraud sites to actual digital currency fraud sites:

$$P = \frac{TP}{TP + FP} \times 100\% \tag{8}$$

Recall rate (R). The ratio of the total number of correctly classified digital currency fraud websites to the total number of positive records. A high recall rate means that the class is correctly identified with a small amount of FN:

$$R = \frac{TP}{TP + FN} \times 100\% \tag{9}$$

F1-Score. This is the harmonic average (percentile) of precision and recall. It is a value near the smaller value of precision or recall. Provides a more realistic way to use precision and recall to evaluate the accuracy of the test. If the false positive and false negative values are very different, the F1 value works best:

$$F1 = \frac{2 \cdot P \cdot R}{P + R} \times 100\% \tag{10}$$

4.3 Experimental settings

The ten-fold cross-validation method is used to test the performance of the model. Divide the data set into ten parts and take turns using 9 parts as training data and 1 part as test data for testing. Each test will get the corresponding accuracy and other indicators. The average value of indicators such as the accuracy of the results of 10 times is used as an evaluation of the performance of the model.

The experiment uses a random forest classification model based on scikit-learn, and the input of the model is a feature matrix containing 16 normalized features. In the stage of experimental preprocessing, use 90% data for training and 10% data for verification, and repeat the same experiment 10 times in total.

In terms of parameter settings, follow the principles of availability. Set the number of decision trees in the random forest to 10, and set the maximum number of features allowed for a single decision tree to $\sqrt{N}$(where N is the total number of features). The maximum depth set to None, the minimum number of samples required for the subdivision of internal nodes set to 2, and whether to use sampling with replacement when building the decision tree set to True;

In terms of experimental evaluation, use sklearn.metrics to calculate evaluation indicators such as the accuracy, accuracy, recall, and F1 value of the experimental results. Three machine learning algorithms are used, Support Vector Machine (SVM), Naive Bayes (NB), and K-Nearest Neighbor Algorithm (KNN) as the as the comparative experiments.

4.4 Results and discussion

Results summary For digital currency fraud websites, we take precision, recall, and F1 test model's performance. Then we analyze the website instances to prove the effectiveness of the model.

Table 1. Comparison of four algorithms evaluation

Method	Accuracy	Precision	Recall	F1
NB	0.46	0.45	0.99	0.62
SVM	0.90	0.95	0.83	0.88
KNN	0.94	0.93	0.96	0.95
RF	**0.97**	**0.98**	**0.95**	**0.96**

The experimental results are shown in Table 1. The random forest model used in this article has the highest accuracy and the most superior performance.Its accuracy is 0.97, precision is 0.98, recall is 0.95, F1-score is 0.96. Support vector machine and K-nearest neighbor algorithm can also achieve relatively good classification results, but the accuracy is lower than random forest. The accuracy of the support vector machine algorithm is 0.90, and the accuracy of the K-nearest neighbor algorithm is 0.92. The Naive Bayes algorithm has the worst performance. Its accuracy is only 0.46, precision is 0.45, recall is 0.99, and F1-score is 0.62.

ROC curve is shown in the figure 2. The horizontal axis represents the specificity of the false positive rate (FPR), which divides the proportion of all negative examples in the instance to all negative examples. The vertical axis represents the true positive rate(TPR), also recall rate. The different solid lines in the figure represent the ROC curves of different machine learning algorithms. Each point on the line corresponds to a threshold.

The larger the FPR, the more actual negative cases in the predicted positive cases.The larger the TPR, the more actual positive cases in the predicted positive cases. Ideal target: TPR=1, FPR=0, that is, the point (0,1) in the figure, so the

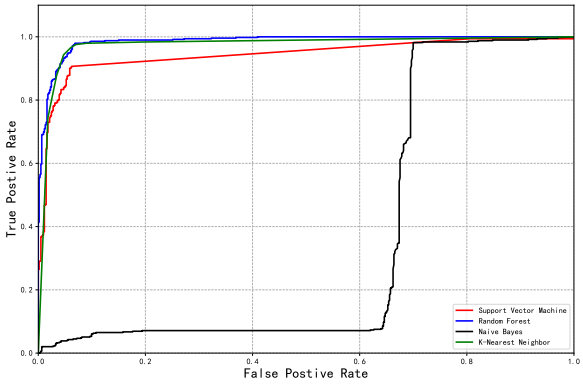

Fig. 2. ROC of four algorithms

closer the ROC curve is to the point (0,1), the more it deviates from the 45-degree diagonal, the better the algorithm performance.

Therefore, we can intuitively see from the figure that RF performance is the best, TPR is 0.96, and FPR is 0.06. SVM and KNN algorithms are second. The worst is NB, TPR is 0.99, FPR is 0.93.

Compared with existing detection methods Tencent Security [33] is a leading brand in Internet security. Its URL Security Center can block malicious URLs and identify phishing websites within seconds. The website type detection that can be implemented by the website security center includes information fraud websites, such as fake investment and wealth management websites and fake brokerage websites. These two types of websites are similar to the digital currency fraud websites mentioned in this article. Therefore, we submit the blacklist and whitelist data sets used in this article to the Tencent Security Website Testing Center for testing and compare the statistical detection results with the detection model of this article.

The results of the comparative experiment are shown in Table 2. In the selected sample set, for this kind of fraudulent website based on digital currency, the Tencent Website Security Inspection Center's detection results of the website are divided into three categories, normal websites, risky websites, and unknown websites. We classify the detection results as normal websites and unknown websites as positive examples, and risk websites as negative examples to calculate the accuracy, precision, recall and F1 score of the detection results. The detection accuracy rate is 0.55, the precision rate is 0.55, the recall rate is 0.97, and the F1 score is 0.70. The experimental results show that the random forest algorithm and extracted website features used in this article can make the detection effect of digital currency fraud websites better.

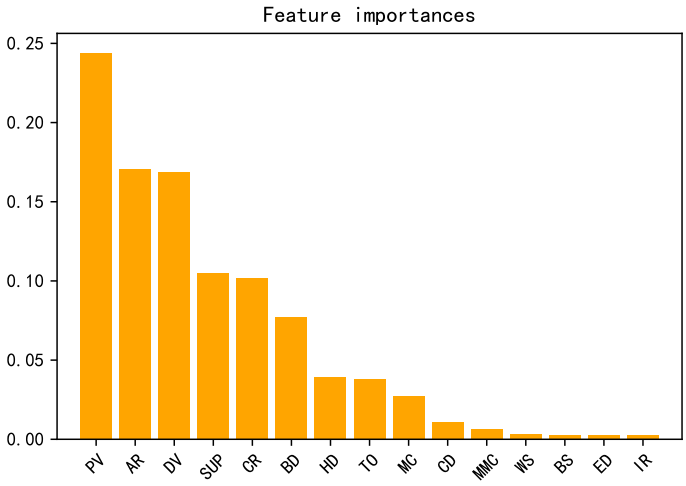

Fig. 3. Feature importance

Table 2. Compared with Tencent Security URL Security Center

Method	Accuracy	Precision	Recall	F1
Tencent	0.55	0.55	0.97	0.70
RF	**0.97**	**0.98**	**0.95**	**0.96**

Feature importance Figure 3 shows the input feature importance score of the digital currency fraud website detection model.

The idea of using the random forest [34] for feature importance evaluation is to observe the contribution of each feature on each tree in the random forest, and then take an average value, and finally compare these values.

We use the method of evaluating with the *Gini* index here. [35] Variable importance measures are represented by VIM, and the *Gini* index is represented by GI. Assuming there are m features $X_1, X_2, X_3..., X_C$, now we need to calculate the *Gini* index score of each feature X_j, $VIM_j^{(Gini)}$, that is, the average change in the impurity of node splitting of the j-th feature in all decision trees in RF.

The formula for calculating the *Gini* index is

$$GI_m = \sum_{k=1}^{|K|} \sum_{k' \neq k} p_{mk} p_{mk'} = 1 - \sum_{k=1}^{|K|} p_{mk}^2 \tag{11}$$

K indicates that there are K categories, and p_{mk} indicates the proportion of category k in node m. Intuitively, it is the probability that we randomly select two samples from node m, and their category labels are inconsistent.

The importance of feature X_j at node m, that is, the change in $Gini$ index before and after the branch of node m is as follows:

$$VIM_j^{(Gini)} = GI_m - GI_l - GI_r \tag{12}$$

Where GI_l and GI_r respectively represent the $Gini$ indices of the two new nodes after branching;

If the set of nodes that feature X_j appears in decision tree i is M, then the importance of X_j in the i-th tree is:

$$VIM_j^{(Gini)} = \sum_{i=1}^{n} VIM_{ij}^{(Gini)} \tag{13}$$

Finally, normalize all the obtained importance scores:

$$VIM_j = \frac{VIM_j}{\sum_{i=1}^{c} VIM_i} \tag{14}$$

The full name of the features are as follows:

1) *WhiteSet (WS).* Text data set of whitelist website.
2) *BlackSet (BS).* Text data set of blacklist website.
3) *Alexa Rank (AR).* Website's Alexa ranking.
4) *Baidu Index (BD).* Website being indexed by Baidu.
5) *Domain Authority Value (DV).* Domain Authority Value (DV): It is an important index to measure the authority of a website.
6) *Page Authority Value (PV).* It can evaluate the authority of a page, and it affects the weight of a single page.
7) *Supply (SUP).* Supply of digital currency.
8) *Circulation Rate (CR).* Circulation rate of digital currency.
9) *Turnover (TO).* It refers tothe frequency of changing hands in the market within a certainperiod.
10) *Mainstream Markets Count (MMC).* The number included by mainstream trading platforms.
11) *Markets Count (MC).* The number included by all of the trading platforms.
12) *Creation Date (CD).* Whether the domain name registration time is earlier than 2019, if yes, set it to 0, otherwise it is 1.
13) *Expiration Date (ED).* Whether the domain name expiration time is later than 2021, yes, set it to 0, otherwise it is 1.
14) *Handle Date (HD).* Whether the domain name has been held for more than 4 years, if yes, set it to 0, otherwise it is 1.

It can be found that the two features Page Authority Value(PV) and Alexa Rank(AR) have the greatest impact on the model classification results. Among them, the feature importance score of PV is 0.228, and AR is 0.213. BS has the least influence on the model, and the feature importance score is only 0.003.

4.5 Case study

Our model identified a total of 2,498 digital currency fraud sites. We tested these sites in well-known secure URL detection center, Tencent [33]. It can detect phishing fraud, information fraud, false advertising, spam and other websites. We found that this detection center has difficulty determining the security of these digital currency fraud sites. According to the test results, more than 90% showed that no risk has been found or the safety is unknown.

To verify the accuracy of the identified websites, we conducted manual case analysis on 2,498 digital currency fraud websites detected by the model. By analyzing the website's text and image information, domain name features, website rankings, and token transaction information, we can make a relatively accurate judgment about whether the website is a digital currency fraud site. Take one of the digital currency fraud websites as an example, the specific analysis is as follows.

Taking the website (http://51-mdd.com/) as an example, we first checked the domain name information of the website. The registration time of the domain name is May 27, 2020, and the expiration time of the domain name is May 27, 2021. The domain name is only held for one year. It is in line with the late registration time and short holding time of our statistics on digital currency fraud websites. Next, query the ranking of the website. Check the Alexa rank of the website, not in the Alexa top one billion previously collected. The result can reflect that the number of page views and user coverage of the site is low. The website was not indexed by Baidu, while most of the websites can be inquired about the ranking and inclusion information on these professional website analysis agencies, such as Alexa and Baidu. Then we inquire about the transaction information of the digital currency on the major digital currency trading platforms. Basic transaction information such as the circulating market value, circulating quantity, circulation rate, and turnover rate of the digital currency cannot be queried at all.

Finally, analyze the content of the official website. The website provides two interfaces in Chinese and English. It is composed of the introduction, ecological construction, project functions, plans, and tokens. The website provides links to promotional videos and white papers, but they are not accessible at all. The content promoted by the website is camouflaged using blockchains, such as distributed operating systems, super security, cross-chain wallet, secure transaction, decentralization, open-source, and other words. The propaganda of the project used magnificent but very false words such as all-mankind, sustainability, the universe, and the world's top 500. Regarding the plan of the project, there is nothing to achieve. The content related to project profit is similar to the Ponzi scheme, low investment, high return, low risk, etc.

Based on the above characteristics, we can completely determine that the website is a digital currency fraud website. Similar analysis methods are used for the other websites, which have all or most of the above features. The case study confirms the validity of the features selected in this paper and the accuracy of the classification.

5 Conclusion

The purpose of this article is to analyze the characteristics of the website and realize the accurate classification of digital currency fraud websites and normal websites. In order to solve the above problems, we collected the top 5000 websites in Alexa and mainstream digital currency websites as whitelists. At the same time, we collected digital currency websites that were not disclosed on mainstream trading platforms and fraudulent websites that were publicly maintained as blacklists. The collected websites are filtered and the text is translated and cleaned, so the model can realize the classification of websites in any language type. After statistics and effectiveness analysis, we selected text, domain names, website rankings, etc. as features. Random forest algorithm is used as the website classification model, and the performance of the model is tested through ten-fold cross-validation. The accuracy rate is 0.97, and the recall rate is 0.95. This identification method helps to detect and classify digital currency fraud sites in a timely manner.

The classification of digital currency fraud websites based on website characteristics proposed in this article is a new insight. But there are still shortcomings. In future work, we will continue to research the following aspects: (1) Mainstream blockchain transaction methods (2) Mainstream blockchain traceability algorithm (3) Mainstream digital currency value tracking.

Acknowledgment

This research is funded by the National Natural Science Foundation of China (U20B2045, No.61902265), Sichuan Science and Technology Program (No.2020YFG0076).

References

1. Bartoletti, M., Carta, S., Cimoli, T., Saia, R.: Dissecting ponzi schemes on ethereum: identification, analysis, and impact. Future Generation Computer Systems **102**, 259–277 (2020)
2. Tang, C., Chen, S., Fan, L., Xu, L., Liu, Y., Tang, Z., Dou, L.: A large-scale empirical study on industrial fake apps. In: 2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP). pp. 183–192. IEEE (2019)
3. Zhao, J.L., Fan, S., Yan, J.: Overview of business innovations and research opportunities in blockchain and introduction to the special issue (2016)
4. Huang, D.Y., Aliapoulios, M.M., Li, V.G., Invernizzi, L., Bursztein, E., McRoberts, K., Levin, J., Levchenko, K., Snoeren, A.C., McCoy, D.: Tracking ransomware end-to-end. In: 2018 IEEE Symposium on Security and Privacy (SP). pp. 618–631. IEEE (2018)
5. Alrwais, S., Yuan, K., Alowaisheq, E., Li, Z., Wang, X.: Understanding the dark side of domain parking. In: 23rd {USENIX} Security Symposium ({USENIX} Security 14). pp. 207–222 (2014)

6. Agten, P., Joosen, W., Piessens, F., Nikiforakis, N.: Seven months' worth of mistakes: A longitudinal study of typosquatting abuse. In: Proceedings of the 22nd Network and Distributed System Security Symposium (NDSS 2015). Internet Society (2015)
7. McCorry, P., Möser, M., Ali, S.T.: Why preventing a cryptocurrency exchange heist isn't good enough. In: Cambridge International Workshop on Security Protocols. pp. 225–233. Springer (2018)
8. Vasek, M., Moore, T.: Analyzing the bitcoin ponzi scheme ecosystem. In: International Conference on Financial Cryptography and Data Security. pp. 101–112. Springer (2018)
9. Liebau, D., Schueffel, P.: Cryptocurrencies & initial coin offerings: are they scams?-an empirical study. The Journal of The British Blockchain Association **2**(1), 7749 (2019)
10. Kiffer, L., Levin, D., Mislove, A.: Analyzing ethereum's contract topology. In: Proceedings of the Internet Measurement Conference 2018. pp. 494–499 (2018)
11. Holub, A., O'Connor, J.: Coinhoarder: Tracking a ukrainian bitcoin phishing ring dns style. In: 2018 APWG Symposium on Electronic Crime Research (eCrime). pp. 1–5. IEEE (2018)
12. Chen, W., Guo, X., Chen, Z., Zheng, Z., Lu, Y.: Phishing scam detection on ethereum: Towards financial security for blockchain ecosystem. In: International Joint Conferences on Artificial Intelligence Organization. pp. 4506–4512
13. Dam, T., Klausner, L.D., Buhov, D., Schrittwieser, S.: Large-scale analysis of pop-up scam on typosquatting urls. In: Proceedings of the 14th International Conference on Availability, Reliability and Security. pp. 1–9 (2019)
14. Chen, W., Zheng, Z., Ngai, E.C.H., Zheng, P., Zhou, Y.: Exploiting blockchain data to detect smart ponzi schemes on ethereum. IEEE Access **7**, 37575–37586 (2019)
15. Bartoletti, M., Pes, B., Serusi, S.: Data mining for detecting bitcoin ponzi schemes. In: 2018 Crypto Valley Conference on Blockchain Technology (CVCBT) (2018)
16. Chen, W., Zhang, T., Chen, Z., Zheng, Z., Lu, Y.: Traveling the token world: A graph analysis of ethereum erc20 token ecosystem. In: Proceedings of The Web Conference 2020. pp. 1411–1421 (2020)
17. Torres, C.F., Steichen, M., et al.: The art of the scam: Demystifying honeypots in ethereum smart contracts. In: 28th {USENIX} Security Symposium ({USENIX} Security 19). pp. 1591–1607 (2019)
18. Bian, S., Deng, Z., Li, F., Monroe, W., Shi, P., Sun, Z., Wu, W., Wang, S., Wang, W.Y., Yuan, A., et al.: Icorating: A deep-learning system for scam ico identification. arXiv preprint arXiv:1803.03670 (2018)
19. Boshmaf, Y., Elvitigala, C., Al Jawaheri, H., Wijesekera, P., Al Sabah, M.: Investigating mmm ponzi scheme on bitcoin. In: Proceedings of the 15th ACM Asia Conference on Computer and Communications Security. pp. 519–530 (2020)
20. SHEN Meng, SANG An-Qi, Z.L.H., SUN Run-Geng, Z.C.: Abnormal transaction behavior recognition based on motivation analysis in blockchain digital currency **44**(01), 193–208 (2021)
21. Xu, J., Livshits, B.: The anatomy of a cryptocurrency pump-and-dump scheme. In: 28th {USENIX} Security Symposium ({USENIX} Security 19). pp. 1609–1625 (2019)
22. Phillips, R., Wilder, H.: Tracing cryptocurrency scams: Clustering replicated advance-fee and phishing websites. In: 2020 IEEE International Conference on Blockchain and Cryptocurrency (ICBC). pp. 1–8. IEEE (2020)

23. Chen, W., Guo, X., Chen, Z., Zheng, Z., Lu, Y.: Phishing scam detection on ethereum: Towards financial security for blockchain ecosystem. In: Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence IJCAI-PRICAI-20 (2020)
24. cryptoscamdb. `https://cryptoscamdb.org/`, 2021
25. Adebowale, M.A., Lwin, K.T., Sanchez, E., Hossain, M.A.: Intelligent web-phishing detection and protection scheme using integrated features of images, frames and text. Expert Systems with Applications **115**, 300–313 (2019)
26. Ray, K.S., Kusshwaha, R.: Detection of malicious urls using deep learning approach. In: The" Essence" of Network Security: An End-to-End Panorama, pp. 189–212. Springer (2021)
27. Verma, P., Goyal, A., Gigras, Y.: Email phishing: text classification using natural language processing. Computer Science and Information Technologies **1**(1), 1–12 (2020)
28. TIAN Shuang-Zhu, C.Y., YAN Zhi-Wei, L.X.D.: Illegitimate website detection based on multi-dimensional features. Computer Systems and Applications pp. 207–211 (2017)
29. Hu, Z., Chiong, R., Pranata, I., Susilo, W., Bao, Y.: Identifying malicious web domains using machine learning techniques with online credibility and performance data. In: 2016 IEEE Congress on Evolutionary Computation (CEC). pp. 5186–5194. IEEE (2016)
30. Vasek, M., Moore, T.: There's no free lunch, even using bitcoin: Tracking the popularity and profits of virtual currency scams. In: International conference on financial cryptography and data security. pp. 44–61. Springer (2015)
31. Xia, P., Wang, H., Zhang, B., Ji, R., Gao, B., Wu, L., Luo, X., Xu, G.: Characterizing cryptocurrency exchange scams. Computers & Security **98**, 101993 (2020)
32. Probst, P., Boulesteix, A.L.: To tune or not to tune the number of trees in random forest? Journal of Machine Learning Research **18** (2017)
33. Tencent security url security center. `https://urlsec.qq.com/index.html`, 2021
34. Genuer, R., Poggi, J.M., Tuleau-Malot, C.: Vsurf: Variable selection using random forests. Pattern Recognition Letters **31**(14), 2225–2236 (2016)
35. Raschka, S.: Python machine learning. Packt publishing ltd (2015)