2018 IEEE 16th Int. Conf. on Dependable, Autonomic & Secure Comp., 16th Int. Conf. on Pervasive Intelligence & Comp., 4th Int. Conf. on Big Data Intelligence & Comp., and 3rd Cyber Sci. & Tech. Cong.

Fraud Risk Monitoring System for E-Banking Transactions

Chaonian Guo*, Hao Wang†[1], Hong-Ning Dai‡, Shuhan Cheng§, Tongsen Wang¶
Division of Sci. and Tech., Fujian Rural Credit Union, Fuzhou, Fujian, China*§¶
Department of ICT and Natural Sci., Norwegian University of Sci. and Tech., Alesund, Norway, hawa@ntnu.no†
Faculty of Information Tech., Macau University of Sci. and Tech., Macau, China‡

***Abstract*—Fraudulent e-banking transactions have caused great economic loss every year. Thus, it is important for financial institutions to make the e-banking system more secure, and improve the fraud detection system. Researches for the fraud risk monitoring are mainly focused on score rules and data driven model. The score rule is based on expertise, which is vulnerable to new patterns of frauds. Data driven model is based on machine learning classifiers, and usually has to handle the imbalanced classification problem. In this paper, we propose a novel fraud risk monitoring system for e-banking transactions. Model of score rules for online real-time transactions and offline historical transactions are combined together for the fraud detection. Parallel big data framework: Kafka, Spark and MPP Gbase which integrated with a machine learning algorithm is presented to handle offline massive transaction logs. Experimental results show the effectiveness of our proposed scheme over a real massive dataset of e-banking transactions. This evaluation leads us to identify research gaps and challenges to consider in future research endeavors.**

***Keywords*—Fraud, risk monitoring, e-banking, big data, machine learning**

I. Introduction

The majority of modern commerce relies on e-banking and cashless payment systems, as the development of information technology and banking business. By offering e-banking services, traditional financial institutions seek to offer lower costs, improve consumer banking services, retain consumers and expand share of customer. E-banking services such as telephone bank, online bank and mobile bank, have provided great convenience for people's daily life. Particularly, the mobile bank, which is installed in smartphones, can check the accounts, initiate transactions, and confirm them.

Security is by far one of the major concerns of e-banking transactions. People using e-banking are worrying that intruders will get into their account and spend their money [1]. China's e-banking fraud case has entered a period of rapid growth, and the risk of e-banking transactions is rapidly increasing. According to the report [2], 29.17% of the phone frauds happened in finance and has caused a loss about CN¥ 91.5 billion in 2017. There are 1.8 million credit card transactions every day for Fujian Rural Credit Union (FRCU) in e-banking channels, among which there are 5 thousand fraudulent transactions. The e-banking business risks, such as low payment amount with large transaction volume, difficulties in virtual trading tracking etc., are also emerging. How to manage and control e-banking risks effectively, especially customer transaction risks, has become an important issue that regulators must pay attention to and banks must face.

Short Message Service (SMS) based *One-Time Passwords* (OTP) were introduced to counter phishing and other attacks against authentication and authorization of Internet service [3]. The prime example of SMS OTP is the mobile *Transaction Authorization Number* (mobile TAN or mTAN) that is used to authorize transactions for online bank and mobile bank services [4]. U shield is another authentication method of TAN for e-banking transactions [5]. Biometrics technology (such as fingerprint, iris and voice recognition) has accelerated at an immense pace for the authentication and identification of e-banking systems [6] [7] [8]. Researches show that most of the mobile bank applications of e-banking fail to preserve the integrity of their transactions [9] [10]. Exploits toward the mobile application in FinTech [11] start scene, show attackers could steal the customers' money because of a broad variety of partly severe security issues [12].

Its urgent for the banks to build an effective risk monitoring and management system for e-banking. The use of ontology makes the rule-based expert system more efficient for suspicious transactions detection of e-banking [13] [14]. Support vector machine was introduced to fight against credit card fraud, money laundering and mortgage fraud [15] [16]. GANs [17] was presented to deal with the problem of class imbalance in the application of supervised classification to detection of credit card fraud [18]. Big data and parallel computing technique were introduced to identify financial fraud [19] [20] [21].

When a fraudulent transaction happens, customers would suspect the security of the e-banking system after they lose the money. However, banks would question the identity of customer, which is a game between spear and shield. Researches for the fraud risk monitoring are mainly focused on the identity and security of e-banking terminals on the customer side. While, score rule expert system and data driven model machine learning system focus on the bank side. The score rule is based on expertise, which is vulnerable to new patterns of frauds. Data driven model is based on machine learning classifiers, and usually has to handle the imbalance

[1]Hao Wang is the corresponding author

978-1-5386-7518-2/18/$31.00 ©2018 IEEE
DOI 10.1109/DASC/PiCom/DataCom/CyberSciTec.2018.00030

issue [18] and deal with a big volume of data.

To address this problem, we propose a fraud risk monitoring system for e-banking transactions. The system is composed of two parts, online scoring subsystem based on expertise and offline machine learning subsystem based on big data. The online subsystem will generate the RAIB (Risk of Activity, Identity and Behavior of transaction) score of the transaction and give a risk level. In the offline subsystem, the big data framework for machine learning is introduced to handle historical transactions. Parallel random forest algorithm is proposed for the learning of fraudulent transactions, which have been show be particularly effective in fraud detection [22] [23] [24]. The major contributions of this paper are summarized as follows:

- We propose a novel design of fraud risk monitoring system for e-banking transactions;
- Score rule based on RAIB and data driven model based on parallel random forest algorithm over big data are designed together for the detection of fraudulent transactions.
- Big data framework KSMG: Kafka, Spark and MPP (Massively Parallel Processing) Gbase is presented to handle offline massive historical transactions for machine learning.

In section II, we introduce the background and evaluation of fraud detection for e-banking. Subsequently, in section III, we describe the proposed framework of fraud risk monitoring system. Section IV discusses the RAIB model, parallel random forest algorithm, KSMG framework and the monitoring algorithm for the fraud detection. Afterwards, experimental results for the parallel random forest on real dataset of e-banking transactions is given in section V. Section VI concludes the paper with an extensive research agenda.

II. Background

In this section, we briefly review fraud detection system of e-banking transaction and the evaluation approaches for fraud detection.

A. Fraud Risk Detection

Fraud risk detection in banks for e-banking transactions relies on multiple systems. Firstly, the identity of the customer and the device security where transaction happens should be checked. The transaction may come into the bank through different channels (terminals), which means various identity and security verification methods should be used. Secondly, the fraud risk level of transaction should be evaluated in the fraud monitoring system after identity check passed. Thirdly, a challenge could be given to the customer, or the bank could contact the customer for the realness validation of the transaction, if the fraud risk level is much high. Finally, if fraud is confirmed, the account in the banking core system of the fraudulent transaction should be froze. In this paper and hereafter, we mention the fraud detection in the fraud risk monitoring system.

There are various of fraud types in e-banking services, for a more detailed discussion of different fraud types we refer the reader to [25].

B. Fraud Risk Monitoring Evaluation

Receiver Operating Characteristic (ROC) and Area Under the ROC Curve (AUC) are usually used for the evaluation of fraud risk monitoring. ROC and AUC are extracted from a confusion matrix as show in Table I. From the confusion matrix, other statistics could be extracted:

- Sensitivity: $\frac{TP}{TP+FN}$, also called True Positive Rate/TPR
- False Positive Rate/FPR: $\frac{FP}{FP+TN}$

TABLE I
Classification confusion matrix

	Actual positive	Actual negative
Predicted positive	TP	FP
Predicted negative	FN	TN

III. Framework of Fraud Risk Monitoring System

In this section, we describe the overview of the fraud risk monitoring system. The transaction risk monitoring procedure and related operations are introduced.

A. System Overview

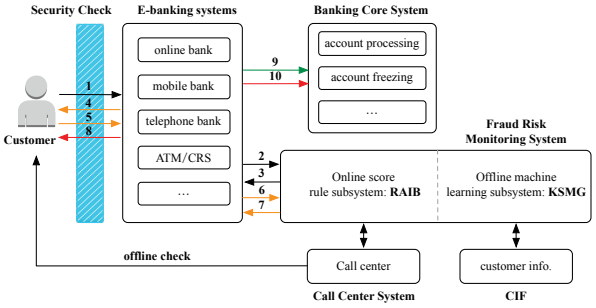

Fig. 1. Framework of the Fraud Risk Monitoring System

As shown in Fig. 1, the monitoring system is composed of online subsystem and offline subsystem. E-banking systems are those electronic channels: such as online bank, mobile bank, telephone bank and ATM/CRS (Automatic Teller Machine/Cash Recycling System). CIF is the customer information system that records key information for the customer (name, telephone number, gender, address, etc.). Banking core system holds all the account information and deal with daytime real-time transactions and nighttime bulk transactions from peripheral systems. Call center handles the inbound and outbound calls of customers.

The online subsystem calculates the risk score of the transaction based on the expertise model of the e-banking service. The risk score is generated based on RAIB, which composed of different weights of the three dimensions: identity, activity and behavior of the transaction. When a customer invokes

one transaction in any e-banking terminal, it will trigger the identity, activity and behavior check and hit corresponding rules.

The offline subsystem is a big data engine KSMG: Kafka[1], Spark[2] and MPP Gbase[3]. Kafka is a distributed publish-subscribe messaging queue system that is commonly used for log collection. It will retrieve transaction logs from all e-banking systems. Spark is a unified analytics engine for big data processing, with built-in modules for streaming, SQL, machine learning and graph processing. It classifies the transactions based on the random forest algorithm from the historical data, and store the transaction cold data to MPP Gbase. MPP Gbase is a parallel and scalable database for offline massive data processing.

B. The Main Risk Monitoring Procedure

When a transaction is invoked in any of the electronic channels by the customer, it is transferred to the risk monitoring system for risk detection. The following is the main procedure:

1) The customer invokes a transaction through any e-banking channel.
2) The transaction is passed through to the fraud risk monitoring system after security check.
3) The transaction RAIB score is calculated in online subsystem and is classified in offline subsystem, then the response is returned to the e-banking system with the next operation (continue normal transaction or challenge the suspicious transaction).
4) If suspicious transaction is detected, challenge would be given to the customer. The offline subsystem will get customer information from CIF system for the message construction of the challenge.
5) The customer responses to the challenge.
6) The response is transferred to the monitoring system and validated.
7) Return the validation result back to the e-banking channel.
8) If the customer fails with the challenge, the suspicious transaction is rejected (fraud).
9) The normal transaction or suspicious transaction with successful response for the challenge is continued.
10) Freeze the account of the fraudulent transaction.

IV. Design and Implementation of the System

We will analyze the RAIB model, the construction of parallel random forest classifier and the KSMG framework for the data processing of machine learning. And then introduce the fraud monitoring algorithm for e-banking transactions.

A. RAIB Risk Model

There are many systems in e-banking service, therefore the transaction risk model should be classified to different channels and scenarios. Overall, all the scenarios could be summarized to three dimensions: activity, identity and behavior.

Activity indicates the active information of the transaction, such as the transaction channel/terminal, transaction position, transaction frequency etc. Identity indicates the account status (active, inactive, frozen, etc.), the identity status of the customer (in black-list, grey-list or white-list), device status (trustable, fraudulent). Behavior indicates the normal behavior of the customer, such as the maximum transaction amount, the normal transaction time, the normal receiver of the transaction etc. For each transaction tx of e-banking, there are many rules under every dimension. And each rule has a subscore for fraud risk rating. We use $\overrightarrow{r_a}$, $\overrightarrow{r_i}$, $\overrightarrow{r_b}$ to denote rule tensors for transaction activity, identity and behavior. Then, use $\overrightarrow{s_a}$, $\overrightarrow{s_i}$, $\overrightarrow{s_b}$ to denote the subscore tensors for the rules. $S_{\mathcal{RAIB}}$ denotes the RAIB score of the transaction tx. All rules have initial weights $\overrightarrow{w_a}$, $\overrightarrow{w_i}$ and $\overrightarrow{w_b}$. Activity, identity and behavior have their initial weights, w_A, w_I and w_B. The RAIB score $S_{\mathcal{RAIB}}$ could be generated by equation (1).

$$S_{\mathcal{RAIB}} = w_A \sum \overrightarrow{p_a}\overrightarrow{w_a} + w_I \sum \overrightarrow{p_i}\overrightarrow{w_i} + w_B \sum \overrightarrow{p_b}\overrightarrow{w_b} \quad (1)$$

The rule engine is pre-configured by the e-banking service domain experts in the bank for the online subsystem. At the very beginning, many rules of transaction could be selected and tested for the RAIB rating model. As the historical transactions get progressively, the rules could be updated. Also, the weights of each rule could be trained by different machine learning algorithms in offline subsystem.

After the RAIB score $S_{\mathcal{RAIB}}$ is generated, it will be mapped to risk level. We divide the RAIB score to five levels: low, $medium$-low, $medium$, $medium$-$high$ and $high$. The higher the score level, the higher the risk of transaction fraud is. If the risk level is under the threshold, the transaction is authorized to go on. On the contrary, if the risk level above the threshold, the transaction will be transferred to offline KSMG subsystem for further classification.

B. Random Forest Classifier

1) Feature Engineering: There are many features in the transactions of e-banking. To build an effective classifier from massive data, we have to choose the most relevant features from initial set of features (raw features) for the generation of classifier, which is a feature engineering process. We denote the raw features f_r and f_e for the selected feature through engineering for classifier construction. The process could be represented as function E: $f_r \rightarrow f_e$. Alternative feature engineering techniques could be selected in research [26].

2) Parallel Random Forest Classifier: Random forest is using bootstrapping techniques to produce a large number of decision trees for the reduction of the over-fitting risk [27]. In random forest, the splitting process for an individual node is based on a randomly chosen subset of all features.

By introducing randomness into the feature selection for every decision node, the basic algorithm of random forest trains a set of decision tree separately from multiple bootstrap

[1]http://kafka.apache.org
[2]http://spark.apache.org
[3]http://www.gbase.cn/

sampling of the training set. Therefore, all leafs in each decision tree $tree_i$ represent one class, fraud or genuine. And then the most predicted class will be the representative for the tree. Finally, we can predict the outcome by taking the majority vote of the entire forest $\mathcal{F}$, $(tree_1,\ tree_2, ..., tree_n)$. We denote the random forest classifier as $\mathcal{C}_{rf}$.

Spark revolves around the concept of a resilient distributed dataset (RDD), which is a fault-tolerant collection of elements that can be operated on in parallel. Due to the independence of each decision tree in the random forest $\mathcal{F}$, the random forest classifier $\mathcal{C}_{rf}$ could be implemented in a distributed environment in order to train several decision trees in parallel. It leads us to implement the random forest using Spark parallelly, which improves the efficiency of the classifier generation.

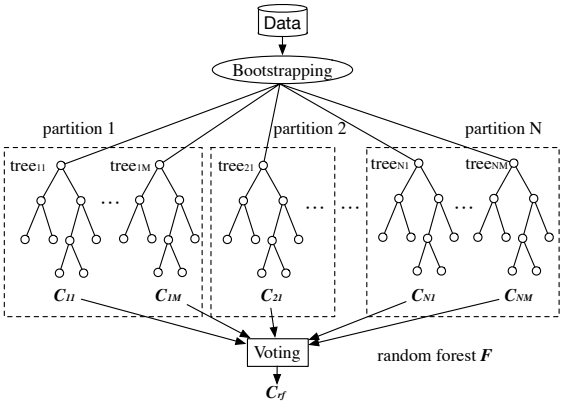

Fig. 2. The generation of the random forest in parallel

Let us denote the raw dataset as $\mathcal{DS}$. After feature engineering, the dataset would be $\mathcal{DS}_E$. For parallel computing, we divide $\mathcal{DS}_E$ into N partitions for decision trees generation. Random forest $\mathcal{F}$ is generated respectively in N partitions, as shown in Fig. 1. All the decision trees in each partition is collected to form the final classifier $\mathcal{C}_{rf}$. The parallel version of random forest could be shown in Algorithm 1.

Algorithm 1: Parallel random forest algorithm

Input: Historical transaction dataset $\mathcal{DS}$ for training
Output: Return classifier $\mathcal{C}_{rf}$

$nTree \leftarrow M$, number of trees in each partition
$nPartition \leftarrow N$, number of partitions
$\mathcal{DS}_E \leftarrow$ process feature engineering of $\mathcal{DS}$
$\{\mathcal{DS}_{E1}, ..., \mathcal{DS}_{EN}\} \leftarrow$ load $\mathcal{DS}_E$ in RDD for N partitions
$tArray \leftarrow \phi$
for $i \leftarrow 1,\ N$ **do**
    $trees \leftarrow$ build classifiers $trees$ using $\mathcal{DS}_{Ei}$
    $tArray \leftarrow$ append $trees$ to $tArray$
$\mathcal{C}_{rf} \leftarrow$ collect all classifiers in $tArray$
return $\mathcal{C}_{rf}$

C. KSMG Framework and Fraud Monitoring Algorithm

1) KSMG Framework: In the KSMG framework, Kafka collects the transaction log files periodically from all the e-banking systems, and aggregates the features together for the transactions. The raw features of the transactions are stored in MPP Gbase database for the query and machine learning. Then, Spark reads these massive data and processes the feature engineering, the construction of random forest. This process could be shown in Fig. 3.

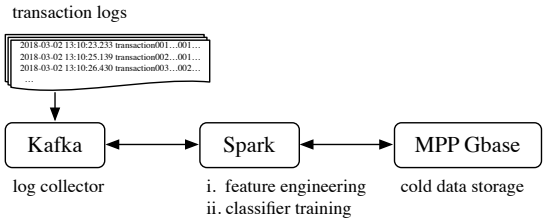

Fig. 3. Overview of the KSMG framework

2) Fraud Monitoring Algorithm: When a transaction tx enters into an e-banking system, it is transferred to fraud monitoring system. The online subsystem will calculate the RAIB score $S_{\mathcal{RAIB}}$ of the transaction, and map the $S_{\mathcal{RAIB}}$ to risk level. If the risk level exceeds the threshold of risk, the transactions will be classified in the KSMG offline subsystem. Finally, classification result and corresponding operation are returned to the e-banking system. The fraud monitoring procedure is showed in Algorithm 2.

Algorithm 2: Fraud monitoring algorithm

Input: transaction tx from e-banking system
Output: Return code c with operation op for the transaction tx

$S_{\mathcal{RAIB}} \leftarrow$ calculate the fraud risk score of tx using RAIB model in equation (1)
$l_{tx} \leftarrow$ map $S_{\mathcal{RAIB}}$ to the risk level
if $l_{tx} < threshold_{risk}$ **then**
    $op \leftarrow$ authorize tx
    return (0, op) ▶ Authorize transaction tx (no fraud)
else
    $f_e \leftarrow$ select important features f_e from f_r for tx
    $tx_{class} \leftarrow$ predict the class of tx via $\mathcal{C}_{rf}$
    if tx_{class} *is not fraud* **then**
        $op \leftarrow$ authorize tx
        return (0, op) ▶ Authorize transaction tx (no fraud)
    else
        $op \leftarrow$ challenge tx with further security and identity check
        return (1, op) ▶ Challenge the transaction tx (fraud)

V. Experiments and Evaluation Result

We evaluate the proposed parallel random forest algorithm based on credit card transactions of e-banking from European card holders in 2013, in which there are 284,807 transactions. And among all the transactions, there are 492 fraudulent ones, which is a critically imbalanced dataset.

The number of partitions $nPartition$ is set to 2, and the number of trees in each partitions $nTree$ is set to 50. Table II reveals the predicted outputs of confusion matrix under different ratios of training and validation datasets: 5:5, 7:3 and 8:2 respectively. Fig. 4 shows the corresponding ROC and

AUC curves under these three ratios of different thresholds. In the AUC curve, when threshold is 0.3, the 8:2 ratio get best performance, in which the AUC is 0.9463.

TABLE II
The parallel random forest confusion matrix outputs

	Actual (5:5)	Actual (7:3)	Actual (8:2)
Predicted	193 12 53 142146	125 7 26 85285	86 4 11 56861

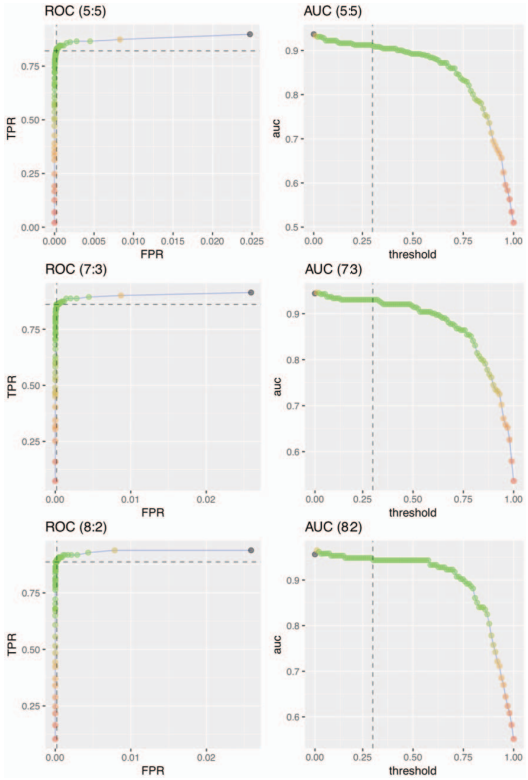

Fig. 4. ROC and AUC curves of different thresholds

VI. Conclusion and Future Work

Fraud risk monitoring is an important topic for e-banking service. It is necessary for banks to construct a fraud risk monitoring system for the e-banking transactions. A novel fraud monitoring system is introduced in this paper to handle this issue. We design the system to two parts, online fraud risk scoring subsystem and offline fraud prediction subsystem. In online subsystem, RAIB score model is introduced for the fraud monitoring, and in the offline system, the parallel random forest is designed for the fraud classifier training. We design KSMG framework for the offline data process of machine learning. Evaluation shows that the proposed fraud classification approach can reach excellent performance.

However, we have not done the oversampling and undersampling work for the imbalanced data of fraud monitoring, and have not introduced pruning strategies for the random forest. Currently, FRCU is working on the construction of the fraud risk monitoring system for e-banking transactions. Next step, we will continue this research with different sampling methods and pruning technique, also carry out other machine learning algorithm in fraud risk monitoring.

Acknowledgment

This work is partially funded by the Fujian Fumin Foundation.

References

[1] Sharma S.. A detail comparative study on e-banking VS traditional banking. International Journal of Advanced Research, 2, 302-307, 2016.
[2] 2017 Phone Fraud Situation Analysis Report. http://economy.china.com/list/11173296/20180328/32243592_all.html (accessed 19 April, 2018).
[3] 3rd Generation Partnership Project. 3GPP TS 23.040 - Technical realization of the Short Message Service (SMS). http://www.3gpp.org/ftp/Specs/html-info/23040.htm, September 2004.
[4] Mulliner C., Borgaonkar R., Stewin P., Seifert J.P.. SMS-Based One-Time Passwords: Attacks and Defense. In: Proceedings of the Conference on Detection of Intrusions and Malware & Vulnerability Assessment (DIMVA), 2013.
[5] Xu H.Y.. China's Internet Financial Risks and Countermeasures. International Conference on Financial Management, Education and Social Science (FMESS 2017), 2017.
[6] Akinyede R.O., Esese O.A.. Development of a Secure Mobile E-Banking System. International Journal of Computer (IJC), Vol 26, No 1, 2017.
[7] Gatali I.F., Lee K.Y., Park S.U., Kang J.. A qualitative study on adoption of biometrics technologies: Canadian banking industry. In: Proceedings of the 18th Annual International Conference on Electronic Commerce: e-Commerce in Smart connected World, 2016.
[8] Traynor P., McDaniel P., La Porta T.. Security for Telecommunications Networks. Springer, 2008.
[9] Reaves B., Scaife N., Bates A., Traynor P., Butler K. R.. Mo(bile) money, mo(bile) problems: analysis of branchless banking applications in the developing world. In: Proceedings of the 24th USENIX Security Symposium (USENIX Security 2015), USENIX Association, 2015.
[10] Haupert V., Müller T.. On App-based Matrix Code Authentication in Online Banking. Technical Report. Freidrich-Alexander-Universität ErlangenNürnberg, 2016.
[11] Schueffel P.. Taming the Beast: A Scientific Definition of Fintech. Journal of Innovation Management, 4(4) 32-54, 2017.
[12] Haupert V., Maier D., Müller T.. Paying the Price for Disruption: How a FinTech Allowed Account Takeover. In: Proceedings of the 1st Reversing and Offensive-oriented Trends Symposium, 2017.
[13] Rajput Q., Khan N. S., Larik A., Haider S.. Ontology Based Expert-System for Suspicious Transactions Detection, Canadian Center of Science and Education, Computer and Information Science; Vol. 7, No. 1, 2014.
[14] Leonard, K. J.. Detecting Credit Card Fraud Using Expert Systems. Computers and Industrial Engineering 25(1-4), 103-1, 1993.
[15] Quah J. T. S., Sriganesh M.. Real-time credit card fraud detection using computational intelligence, Expert Systems with Applications 35 (4), 1721–1732, 2008.
[16] Abdelhamid D., Soltani K., Ouassaf A.. Automatic Bank Fraud Detection Using Support Vector Machines. The International Conference on Computing Technology and Information Management (ICCTIM). Society of Digital Information and Wireless Communication, 2014.
[17] Goodfellow I. J., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A. C., Bengio Y.. Generative adversarial nets. In Proceedings of NIPS, pages 2672–2680, 2014.
[18] Fiore U., Santis A. D., Perla F., Zanetti P., Palmieri F.. Using generative adversarial networks for improving classification effectiveness in credit card fraud detection. Information Sciences, in press.
[19] Alexandre C., Balsa J.. A Multi-Agent System Based Approach to Fight Financial Fraud: An Application to Money Laundering. Preprints 2018, 2018010193.

[20] Melo-Acosta G. E., Duitama-Muñoz F., Arias-Londoño J. D.. Fraud Detection in Big Data using Supervised and Semi-supervised Learning Techniques. IEEE Colombian Conference on Communications and Computing (COLCOM), 2017.
[21] Hormozi H., Hormozi E.. Credit Cards Fraud Detection by Negative Selection Algorithm on Hadoop. In Proceedings of the 5th Conference on Information and Knowledge Technology (IKT), pp 40-43, 2013.
[22] Bhattacharyya S., Jha S., Tharakunnel K., Westland J. C.. Data mining for credit card fraud: A comparative study, Decision Support Systems, 50 (3), 602–613, 2011.
[23] Bahnsen A. C., Aouada D., Ottersten B.. Example-dependent cost-sensitive decision trees, Expert Systems with Applications, 42 (19), 6609–6619, 2015.
[24] Van Vlasselaer V., Bravo C., Caelen O., Eliassi-Rad T., Akoglu L., Snoeck M., Baesens B.. Apate: A novel approach for automated credit card transaction fraud detection using network-based extensions. Decision Support Systems, 75, 38–48, 2015.
[25] Baesens B., Van Vlasselaer V. V., Verbeke W.. Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection. John Wiley & Sons, 2015.
[26] Bahnsen A. C., Aouada D., Stojanovic A., Ottersten B.. Feature engieering strategies for credit card fraud detection. Expert Systems With Applications 51 134–142, 2016.
[27] Breiman L.. Random Forests, Machine learning, pp. 5–32, 2001.