ISSN 2347 - 3983
Volume 8. No. 8, August 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter93882020.pdf
https://doi.org/10.30534/ijeter/2020/93882020

Recognizing Credit Card Fraud Transaction using Spending Behavior-based Transaction Features

Sihar Simbolon[1], **Yaya Heryadi**[2]
[1]BINUS Graduate ProgramMaster of Computer Science Jakarta, Indonesia, Sihar.simbolon@binus.ac.id
[2]Doctor of Computer Science Program BINUS Jakarta, Indonesia, Yayaheryadi@binus.edu

ABSTRACT

Flexibility and easiness for making financial transaction has continued to make credit card transaction a popular alternative financial payment especially during Corona-19 Pandemic that hamper almost worldwide. However, recognizing credit card fraudulent transaction is still an open problem despite many methods have been proposed. The main problems, among others, are imbalanced nature of the credit card transaction data and no agreed upon features to represent the transaction. This paper presents a novel approach to address such problem by exploiting card user spending behavior as the main feature to represent the transaction. in our experiment with Random forest we put max features parameter value 18 and 75 as number of trees in Random forest. While our experiment with OCSVM is by using 4 different kernels which are RBF, Polynomial, Linear and Sigmoid, and, we tuned nu and gamma parameter to get the optimal precision, recall and AUC for our Classifier. The experiment results found that One-Class SVM model achieve 0.984 AUC for training and 0.985 AUC for testing; whilst, Random Forest model achieves 0.991 AUC for training and 0.943 AUC for testing.

Key words: Anomaly Detection, Bagging, Machine Learning, One-Class SVM, Random Forest.

1. INTRODUCTION

Credit card usage for e-commerce transactions has been increasing. Flexibility and easiness to make financial transaction, especially during Corona-19 Pandemic in the past couple of months, has made credit card transaction become more popular than transaction using cash. In order to secure against malicious transactions, Visa, MasterCard and JCB has used two factor authentications techniques to make internet trans-action more secure by using 3 Domain Secure authentications (3D Secure). But in the application not all merchants and bank implement the 3D Secure authentication causes the e-commerce transaction not always secure. To address this issue most of Banks, have an integrated system to monitor the credit card transaction in order to recognize genuine and fraudulent transaction specially in e-commerce transaction. A lot work has been done to prevent and avoid Credit Card fraud, such as implementing sophisticated technology inside the physical credit card and using machine learning to monitor credit card usage behavior[1].

The large number of credit card transactions caused a huge size of data to be maintained by the bank. Whilst the data keep growing with a fast pace, some of those data may contains some fraudulent transactions. With such characteristics make the data cannot be effectively analyzed using conventional methods. As credit card transaction data can be categorized as a Big Data, data analysis should adopt new approach such as machine learning[2] which can facilitate analysis of large-scale data. One of the analysis is recognizing fraudulent transaction in a fast and accurate way.

Although many methods have been proposed to address the task of identifying credit card fraudulent transaction in situ, such problem has not been solved completely. The main challenges, among others, are dataset availability and no agreed upon features to represent credit card transaction. In addition, the imbalanced nature of the fraudulent transaction [3] make the classification model tends to bias to non-fraudulent transaction.

Despite many reports have been reported; to the best of our knowledge, little have been said about credit card fraudulent transaction recognition in Indonesia especially among Government owned Banks using real data. Therefore, this study aims to propose a novel method for recognizing credit card fraudulent transaction by focusing to transaction features associated with card user spending behavior such as frequency and amount spending [4] of each cardholder.

The remaining of this paper is organized as follows. Section 2 will describe related works. Section 3 will elaborate on research method. Section 4 will explain research results followed by conclusion in Section 5.

2. RELATED WORKS

2.1 Fraud Transaction Recognition Model

The problem of Fraud detection is to make segregation of transaction into two classes which are genuine and fraudulent

transaction[5]. Fraud detection system commonly recognize customer spending in physical merchant and customer behavior. In the recognition process, Fraud detection system use predicting algorithm to identify the classes. A transaction will be labeled as Fraudulent transaction if the system observes a deviation in the normal spending behavior. These are some research that have been used in Credit Card fraud detection.

Table 1: Credit Card Fraud Detection Techniques

Method	Contribution	Accuracy	FPR	Ref
OC-SVM	Optimal Kernel Parameter	96.60%	0.085	[6]
Logistic Regression	APATE (Anomaly Prevention using Advanced Transaction Exploration)	87.4%	0.243	[7]
Neural Network		87.9%	0.232	
Random Forest		93.2%	0.126	
KNN + Dynamic Random Forest	Combining KNN + DRF	97.47%	0.013	[8]

In thetable 1 above we can see that Supervised and Unsupervised Learning method in Machine Learning has been used in Credit Card Fraud detection.

From above research we know that machine learning approach can give high accuracy model to detect Credit Card Fraud and have a big challenge to analyze imbalance data since fraudulent transaction is very rare compare to genuine transactions.

2.2 One-Class Support Vector Machine

Support Vector Machine (SVM) was introduce by Vladimir N. Vapnik in 1999 [9]. The basic idea of SVM was to build a hyperplane that can covered most of either positive or negative class. The optimal hyperplane can be created by calculating distance between hyperplane and the maximum data class that been covered[10].

The idea of One Class Classification was originated by Moya in his research[11]. In term of anomaly detection one class classification can build a model by learning from only target examples[12].

SchÖlkopf[13] adopted SVM method to One Class Classification by searching for hyperplane with a maximum mar-gin between the region containing target data and the origin in feature space. The target data is again mapped to feature space via a kernel

Recent research of anomaly detection using actual financial transaction dataset from Indonesian bank was reported by Heryadi[14], They were implementing commonly used kernel function in OC-SVM that are Gaussian Radial Basis Function (RBF), Sigmoid Function and Polynomial Function to find higher performance of OC-SVM in financial transaction. The research conclude that Sigmoid and RBF kernel show high contribution on training, validation and testing accuracy of One-Class SVM model.

2.3 Random Forest Classifier

Random forest (RF) consists of many individual decision trees that operate as an ensemble classifier. Ensemble classifier is a method of combining multiple classifier. Each individual tree in RF has its own Characteristic and has low correlation to each other. Every tree will has its own prediction, than there will be a vote to choose one result that will become the model's prediction[15]. Bagging (Bootstrap and Aggregation) method is used as part of Random Forest Algorithm, where random forest steps are as follow:
(1)Create *n* random sampling data from input dataset where in creation of the sampling data there is a possibility of data duplication (bootstrap sampling). (2) From each *n* random sampling data use decision tree classifier to create n decision tree model. (3) Make prediction from each decision tree model and (4) From the n prediction result make a voting by choosing the majority prediction as the final output of the classifier model

The low correlation between each tree models in the RF algorithm is the virtue. uncorrelated models can produce ensemble predictions that are more accurate than any of the individual predictions. The reason for this virtue effect is that the trees protect each other from each individual error. This make each tree model can make right or wrong prediction, as a group the trees majority prediction result will be voted as the final prediction of RF.

Evaluation result from random forest can be determine without performing cross validation, this is because RF use bootstrap sampling method, there are some data that are not taken during bootstrap sampling step, this kind of data called out of bag data (OOB)[16]. This OOB is used by RF to make prediction and calculate the error. This error result from OOB is called as Out of Bag Error. Accuracy in RF can be measured as in (1)

$$Accuracy = 1 - Error\ Rate \quad (1)$$

Recent research of credit card detection using random forest is made by adisaputra[17] in their research they used SMOTE method to handle imbalance dataset that commonly happen in real dataset, the result from SMOTE is combine with Random

forest algorithm and has accuracy value of 95%.

3. RESEARCH METHOD

3.1 Fraud Transaction Recognition Model

In this research, the transaction log and fraudulent transaction samples are obtained from a National Bank in Indonesia under permission. In accordance to an agreement with the bank, the detail of the dataset attributes is disclosed. The dataset comprises of e-commerce transactions in the period of December 2019 until February 2020. The transaction dataset is highly imbalanced as the number of genuine transaction samples is: 14,447 (99.02 percent) and fraudulent transaction samples is: 143 (0.98 percent).

The Credit Card Transactions dataset is preprocessed by removing inconsistency transaction record, removing non ap-proved transaction and removing missing information. Since we still need the uniqueness of credit card number so we hashed the credit card number so it still can be used without publishing the sensitive information.

The raw data originally has 18 number of attributes, after preprocessing the remaining attributes are 13. Those attributes as shown in the Table 2.

Table 2: Original Financial Transaction Attribute

No	Variable Name	Description
1	Id_trx	Unique Transaction identifier
2	Trx_date	Date of transaction
3	Trx_time	Time of transaction
4	Merch_name	Merchant name of the Transaction
5	Merch_ctry	Country name of merchant location
6	trx_type	Type of transaction like retail or cash withdrawal
7	Trx_sts	Status of transaction declined or Approved
8	Pin_ind	Indicator weather transaction use pin or not
9	Acq_id	Information about financial institution that responsible for the merchant
10	Trx_pem	Information about the transaction source
11	Card_num	Credit Card number
12	Trx_mcc	Merchant Category Group Id
13	Trx_amt	Amount of each transaction

3.2 Transaction Spending Behavior

We are doing RFM (Recency, Frequency and Monetary) analysis[18] from preprocessing result. This is to examine Transaction Spending behavior of frequent pattern to Machine learning algorithm.

This RMF Analysis step by creating below information:

Table 3: Behavior Transaction Attribute

No	Variable Name	Description
1	Card_num	Hashed Card number
2	Trx_mcc	Merchant Category Group Id
3	Merch_ctry	Originating Country of the transaction
4	Hari	The Date of Maximum Transaction Frequency in daily basis
5	Minggu	The week number of maximum transaction Frequency in weekly basis
6	Bulan	The month number of maximum transaction Frequency in monthly basis
7	DfreqMax	The number of maximum transaction frequency in daily basis
8	DfreqMin	The number of minimum transaction frequency in daily basis
9	DfrqAvg	The number of average transaction frequency in daily basis
10	WfreqMax	The number of maximum transaction frequency in weekly basis
11	WfreqMin	The number of minimum transaction frequency in weekly basis
12	WfrqAvg	The number of average transaction frequency in weekly basis
13	MfreqMax	The number of maximum transaction frequency in monthly basis
14	MfreqMin	The number of minimum transaction frequency in monthly basis
15	MfrqAvg	The number of average transaction frequency in monthly basis
16	WSpndMax	The number of maximum transaction Spending in weekly basis
17	WSpndMin	The number of minimum transaction spending in weekly basis
18	WSpndAvg	The number of average transaction spending in weekly basis
19	MspndMax	The number of maximum transaction Spending in monthly basis
20	MspndMin	The number of minimum transaction spending in monthly basis
21	MspndAvg	The number of average transaction spending in monthly basis

In the Table 3, we get information about Credit card

transaction behavior for every Cardholder that doing transaction in various MCC and, we can get information of the originating transaction country

3.3 Experiment Design

In our research, nu and gamma parameter in OC-SVM will be explored to get the optimal value when training and testing using RMF Pattern dataset. And, we will explore 4 different Kernels which are Gaussian Radial Basis Function (RBF), Sigmoid, Linear and Polynomial. In anomaly detection during training we only feed genuine transaction sample to be predicted by the OC-SVM algorithm[19].

For random forest model we will exploring depth and weighting parameters using RMF Pattern Dataset. The experiment design in this paper as in Figure. 1

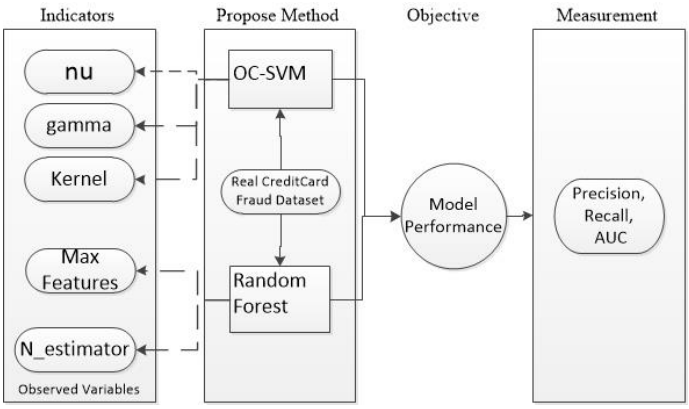

Figure 1. Experiment Design

3.4 Evaluation Criteria

Cross-validation used in this research is one-leave-out technique [20]where 80% of the dataset is used for training model and 20% of the dataset is used for testing. Performance of the tested classifier are as follows[21].

A. Precision

Precision is a metric that quantifies the number of correct positive predictions made[22]. In imbalanced classification Precision is the ratio between true positive / total predicted positive, as shown in (2).

$$Precision = \frac{TP}{TP+FP}. \quad (2)$$

B. Recall

Recall means that the test result will correctly Identify fraudulent transaction correctly. Recall is the ratio between true positive / total actual positive and called as sensitivity, as shown in (3)

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

C. ROC

Receiver Operating Characteristic (ROC) Curve. ROC Curve is two-dimensional curve that plots relation between the false positive rate (FPR) in horizontal axis and the true positive rate (TPR) in vertical axis. An optimal classifier model ideally has High TPR and low FPR, as shown in (4) and (5)

$$TPR = \frac{TP}{TP + TN} \quad (4)$$

$$FPR = \frac{FP}{FP + TN} \quad (5)$$

D. AUC

Area under the ROC Curve (AUC): is the area under ROC curve, it's a scalar value whose value between 0 and 1 that show expected performance of classifier. The statistical property of AUC is equivalent to the probability that the classifier will rank a randomly selected positive sample higher than a randomly selected negative sample. The closest AUC value to 1 will show the better classifier performance.

Where TF is transactions that both models predicted and actual are fraud transaction. TN: is transactions that both models predicted and actual are not fraud transaction. FP: is transactions that the model predicted as fraud transaction, but they are not fraud transaction. FN: is transactions that the model does not predicted as fraud transaction, but they are fraud transactions[23].

4. EXPERIMENT RESULT

OC-SVM have 5 well known Kernels, we train and testing our dataset by using RBF, Linear, Polynomial and Sigmoid kernel. Experiment using RBF and Linear Kernel we are tuning nu and gamma parameter until we found highest AUC. Experiment for Polynomial kernel we also tuning nu and gamma parameter and use degree of 3 to get the highest AUC. While for sigmoid kernel we are tuning nu and gamma value to get highest ROC value. We use scatter chart to present our experiment in OCSVM where colored box represents our training step and colored star represent our testing step.

Figure 2 shows our experiment with RBF Kernel where we experiment with 12 different value of nu and gamma until we found the optimal nu and gamma parameter for our dataset. The best AUC value with RBF kernel is generated by nu = 0.00818183 and gamma = 1e-08.

Figure 3 shows our experiment with polynomial kernel, in our dataset the Optimal nu parameter is 1e-05. And we found that the changes of nu parameter is not significantly give a better performance, otherwise combination of gamma and degree can increase the prediction result. In our experiment when we use higher gamma parameter than we need to also use higher degree parameter and vice versa.The best AUC value with

polynomial kernel is generated by degree = 7 and gamma = 1e-07.

Figure 4 shows our experiment with linear parameter where we use 112 different value of nu and 12 different value of gamma. From our experiment we found that gamma is not significantly change the performance of the classifier. And the best AUC value with polynomial kernel is generated by nu = 0.00438and gamma = 1e-08. Figure 5 show our experiment with Sigmoid Kernel where we experiment with 22 different value of nu and 12 different value of gamma parameter. And the best AUC value with sigmoid kernel is generated by nu = 0.00476and gamma = 7.27e-06.

Table 4: Matrix Result of OC-SVM

Class	Training		Testing	
	Precision	Recall	Precision	Recall
RBF	0.25	1.00	0.25	1.00
Poly	0.03	0.92	0.03	0.85
Linear	0.36	0.72	0.30	0.59
Sigmoid	0.12	0.81	0.12	0.81

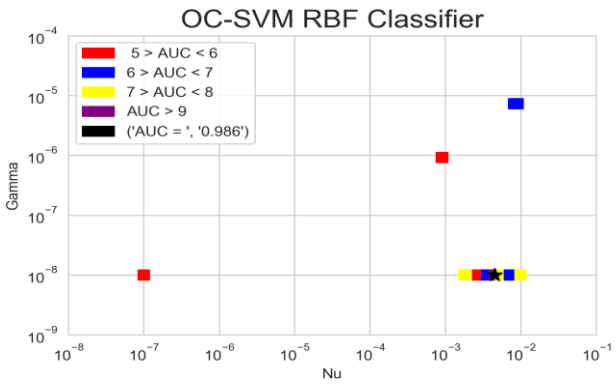

Figure 2: Experiment using RBF Kernel

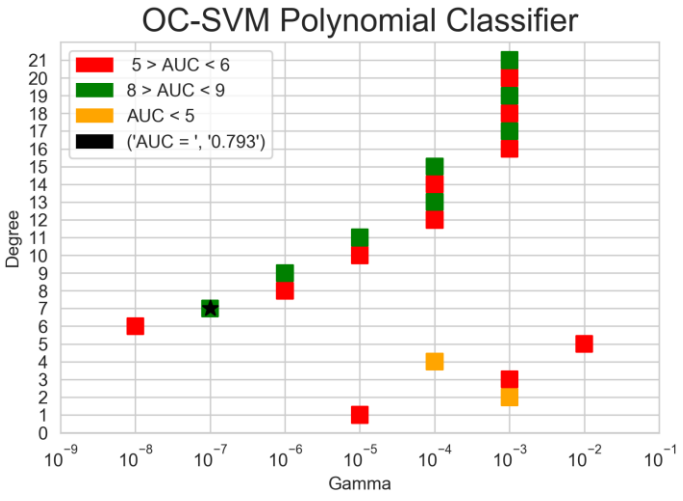

Figure 3: Experiment using RBF Kernel

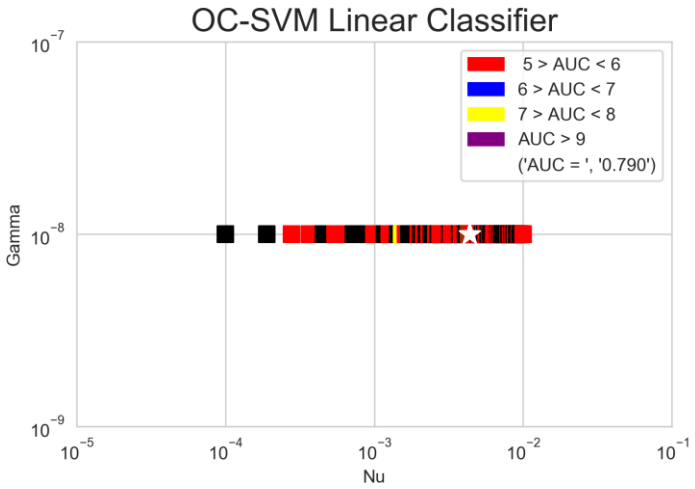

Figure 4: Experiment using Linear Kernel

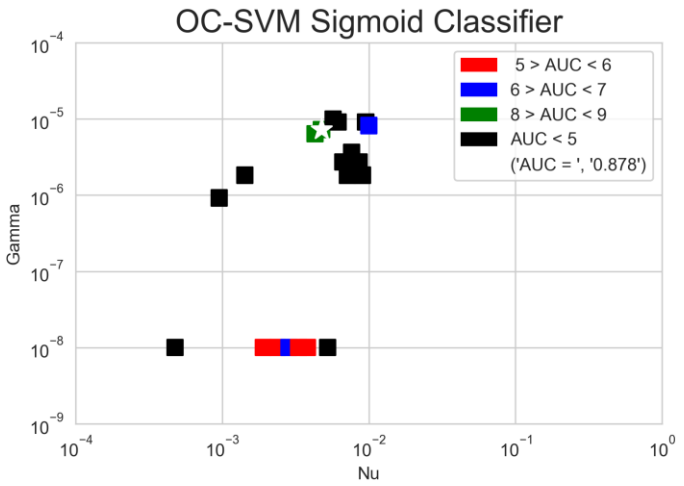

Figure 5: Experiment using Linear Kernel

Summary of our experiment in each kernel as described in the table 4 and the table 5. Recall value of 1 in table 4 tells that RBF Kernel manage to predict all the minority class, but it also missed predict a few of genuine transaction as fraudulent transaction as showed in precision value 0.25. Table 5 showing the Accuracy and AUC result during training and testing experiment.

Table 5: OC-SVM Experiment Result

Kernel	Training		Testing	
	Accuracy	AUC	Accuracy	AUC
RBF	98%	0.984	97%	0.985
Poly	83%	0.831	73%	0.792
Linear	98%	0.855	98%	0.789
Sigmoid	94%	0.874	94%	0.878

Random forest algorithm can be tuned by using the appropria te number of maximum feature and number of trees. Figure 6 show the experiment result when tuning number of maximum feature and number of trees we can get different AUC result. Figure 6 consist of two charts where the upper chart is the AU C result on training phase while the bottom chart is the AUC result for the testing phase.

Generally increasing max feature and add more trees can improve the performance of the model because with tree will have a higher number of features to be considered to make decision. And with higher number of trees can minimize the probability of wrong prediction winning the final voting of random forest.

Before we start using Random forest, we split the dataset 80 % for training that give us 11.556 genuine transaction and 116 fraudulent transactions, while 20% Dataset during testing that gives us 2.891 genuine transaction and 27 fraudulent transactions. Figure 7 shows the feature importance that been used in Random Forest Classifier. We can see that Merch_ctry (Country of merchant) have significant impact for fraudulent transaction recognition.

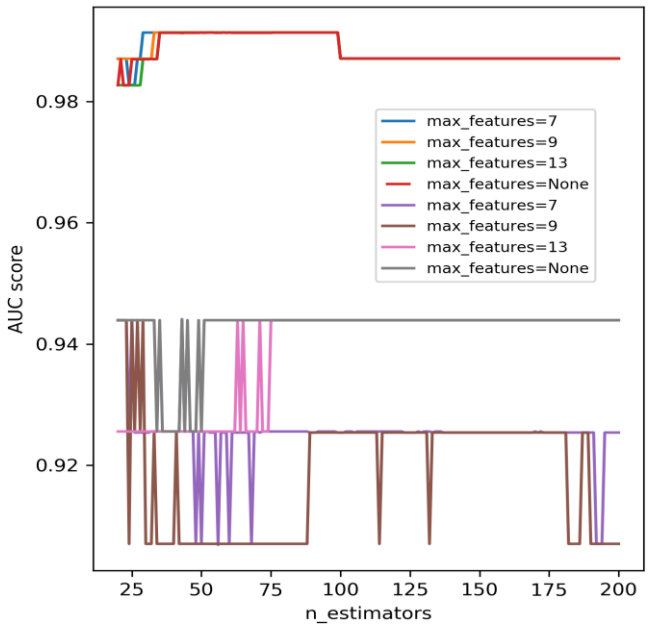

Figure 6: Random Forest AUC from max feature and number of trees

Table 6: Matrix Result of Random Forest

Class	Training		Testing	
	Precision	**Recall**	**Precision**	**Recall**
Genuine	1.00	1.00	1.00	1.00
Fraudulent	0.98	0.98	0.89	0.89

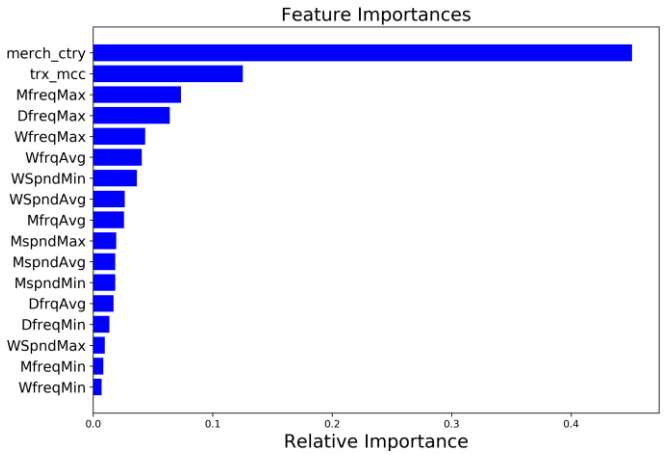

Figure 7: Feature importance of Random Forest

From our experiment we find that using all the feature as the maximum feature, with 75 number of trees and random state of 70 is the optimal parameter for our dataset. by using those parameters, we got training AUC 0.991 and testing AUC 0.943 while in the table 6 we can see the Precision and Recall result during training and testing.

5. CONCLUSION AND FUTURE WORKS

In this paper, we explore supervised and unsupervised learning model for fraudulent card transaction recognition. In initial feature the financial transaction has 13 features. After we calculate customer behavior, we get 21 transaction feature and from our preprocessing we decide to use only 18 features as our predictor for OCSVM and Random Forest Classifier.

Random forest Classifier performance AUC result on testing was 0.91 means that the classifier can maximize all the feature that has been given to predict genuine and fraudulent transaction well. Merchant country is the feature important for Random forest because most of the fraudulent transaction was performed outside Indonesia and the fraudulent transaction happen on merchant that represent by the merchant category code (MCC).

The result showed that the highest AUC was 0.985 that come from One Class SVM with using RBF Kernel, this high AUC result better than Sigmoid and Linear Kernels. This high result was also emphasis that Novelty Detection OCSVM have high performance when dealing with imbalance dataset.

Future works that comes for this experiment is getting more optimal parameter for Random forest from preprocessing step by using SMOTE or balancing the dataset using Balanced Random forest.

REFERENCES

[1] A. O. Adewumi and A. A. Akinyelu.**A survey of machine-learning and nature-inspired based credit**

card fraud detection techniques,*Int. J. Syst. Assur. Eng. Manag.*, 2016.

[2] S. Biswas, A. Ghosh, S. Chakraborty, S. Roy, and R. Bose.**Scope of sentiment analysis on news articles regarding stock market and GDP in struggling economic condition,***Int. J.*vol. 8, no. 7, 2020. https://doi.org/10.30534/ijeter/2020/117872020

[3] T. M. Mitchell.**Machine learning**. McGraw Hill, 1997.

[4] F. Carcillo, Y. Le Borgne, O. Caelen, and Y. Kessaci.**Combining unsupervised and supervised learning in credit card fraud detection,***Inf. Sci. (Ny).*, 2019.

[5] I. Sadgali, N. Sael, and F. Benabbou.**Detection of credit card fraud : state of art detection of credit card fraud : state of art,***Int. J. Comput. Sci. Netw. Secur.*, pp. 76–83, 2018.

[6] T. M. H. Tran, P. H., Tran, K. P., Huong, T. T., Heuchenne, C., HienTran, P. **Real time data-driven approaches for credit card fraud detection,** in *In Proceedings of the 2018 International Conference on E-Business and Applications*, 2018, pp. 6–9.

[7] V. Van Vlasselaer *et al.***APATE: A novel approach for automated credit card transaction fraud detection using network-based extensions,***Decis. Support Syst.*, vol. 75, pp. 38–48, 2015.

[8] S. Nami and M. Shajari.**Cost-sensitive payment card fraud detection based on dynamic random forest and k-nearest neighbors**, *Expert Syst. Appl.*, vol. 110, pp. 381–392, 2018.

[9] V. N. Vapnik.**An overview of statistical learning theory**,*IEEE Trans. neural networks*, vol. 10, no. 5, pp. 988–999, 1999.

[10] A. S. Nugroho, A. B. Witarto, and D. Handoko.**Support vector machine**, in *Proceeding Indones. Sci. Meeiting Cent. Japan.*, 2003.

[11] L. D. Moya, M. M., Koch, M. W., and Hostetler.**One-class classifier networks for target recognition applications**, *STIN*, vol. 93, p. 24043, 1993.

[12] V. L. C. B, M. Nicolau, and J. Mcdermott.O**ne-class classification for anomaly detection with kernel density estimation and genetic programming**, in *European Conference on Genetic Programming*, 2016, pp. 3–18.

[13] J. C. Schölkopf, B., Williamson, R. C., Smola, A. J., Shawe-Taylor, J., and Platt.**Support vector method for novelty detection**,*Adv. Neural Inf. Process. Syst.*, pp. 582–588, 2000.

[14] Y. Heryadi and Dandalina.**The effect of several kernel functions to financial transaction anomaly detection performance using one-calss svm**, in *2019 International Congress on Applied Information Technology (AIT)*, 2019, pp. 1–7.

[15] L. Breiman.**Bagging predictors**, *Mach. Learn.*, vol. 140, pp. 123–140, 1996. https://doi.org/10.1007/BF00058655

[16] F. Livingston.**Implementation of breiman ' s random forest machine learning algorithm**, *ECE591Q Mach. Learn. J. Pap.*, pp. 1–13, 2005.

[17] A. Saputra.**Fraud detection using machine learning in e-commerce**, *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 9, pp. 332–339, 2019.

[18] X. Zhang, Y. Han, W. Xu, and Q. Wang.**HOBA : A novel feature engineering methodology for credit card fraud detection with a deep learning architecture**, *Inf. Sci. (Ny).*, 2019.

[19] D. M. J. Tax and R. P. W. Duin.**Support vector domain description**, *Pattern Recognit. Lett.*, vol. 20, pp. 1191–1199, 1999.

[20] M. Bekkar, H. K. Djemaa, and T. A. Alitouche.E**valuation measures for models assessment over imbalanced data sets**, *J Inf Eng Appl*, vol. 3, no. 10, pp. 27–39, 2013.

[21] A. Al-qerem, M. A. L. Qerem, and I. Jebreen.I**dentifying driver behaviors in learning style classification**, *Int. J.*vol. 8, no. 6, 2020. https://doi.org/10.30534/ijeter/2020/73862020

[22] S. Smadi, M. Alauthman, O. Almomani, A. Saaidah, and F. Alzobi.**Application layer denial of services attack detection based on stacknet**, *Int. J. 3929*, pp. 2278–3091, 2020.

[23] Y. Heryadi, L. A. Wulandhari, and A. Bahtiar Saleh.**Recognizing debit card fraud transaction using CHAID and k-nearest neighbor: Indonesian bank case,** in *In 2016 11th International Conference on Knowledge, Information and Creativity Support Systems (KICSS)*, 2016, pp. 1–5.