E-commerce Merchant Fraud Detection using Machine Learning Approach

Conference Paper · June 2022

DOI: 10.1109/ICCES54183.2022.9835868

6 authors, including:

Fahim Hasan
Daffodil International University
6 PUBLICATIONS **5** CITATIONS
SEE PROFILE

Sourov Kumar Mondal
Daffodil International University
3 PUBLICATIONS **4** CITATIONS
SEE PROFILE

Md Rayhan Kabir
University of Alberta
16 PUBLICATIONS **55** CITATIONS
SEE PROFILE

Md Abdullah Al Mamun
Jain University
5 PUBLICATIONS **27** CITATIONS
SEE PROFILE

Proceedings of the 7th International Conference on Communication and Electronics Systems (ICCES 2022)
DVD Part Number: CFP22AWO-DVD; ISBN: 978-1-6654-9633-9

E-commerce Merchant Fraud Detection using Machine Learning Approach

Fahim Hasan
Computer Science and engineering
Daffodil International University
Dhaka, Bangladesh
fahim15-1556@diu.edu.bd

Sourov Kumar Mondal
Computer Science and engineering
Daffodil International University
Dhaka, Bangladesh
sourov15-1688@diu.edu.bd

Md. Rayhan Kabir
Computer Science and engineering
Daffodil International University
Dhaka, Bangladesh
rayhan15-1557@diu.edu.bd

Md Abdullah al Mamun
Computer Science and engineering
Daffodil International University
Dhaka, Bangladesh
abdullah15-1549@diu.edu.bd

Nur Salman Rahman
Computer Science and engineering
Daffodil International University
Dhaka, Bangladesh
nur15-1531@diu.edu.bd

Md. Sagar Hossen
Informatics Engineering
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
sagar15-1504@diu.edu.bd

***Abstract*—At present, e-commerce has become a global phenomenon. With the great achievement of ecommerce, many are cruel Promotional services are also increasing: with the aim of growing sales, spiteful marketers try to improve their target spectators by improving the outcomes of an illegal search using false travel, shopping, etc. In this report, we read about the problem of deception in major commerce platforms. First, we want to list the merchant fraud, the names of those who have previously committed fraud in the business will be marked on the list. And will train machines using machine learning approach. So that, if a merchant id is given in the system, it can detect whether the id is fraud or not. Our lesson here paper is predictable to hut light on the defense in contradiction of e-commerce fraud of active commerce platforms. In this research report, we proposed a machine learning model to analyze and identify merchant fraud. As a machine learning model, we choose the Random forests, decision tree and logistic regression algorithm for our model.**

Keywords— Random Forest, Decision tree, Confusion matrix, Logistic regression, Ip-bound, Machine learning (ML), Merchant registration date, Internet banking, Cash on delivery

I. INTRODUCTION

E-commerce trade has now received a huge response around the world. The customer can buy these products through his credit card transaction sitting at home [23]. E-commerce is benefiting not only buyers, but also sellers. Merchants are able to present their products to customers with the help of e-commerce and their sales are increasing rapidly. But sellers are at risk because of some unscrupulous traders they are losing their respect. Unscrupulous traders are cheating buyers by not giving them the product at the right time or by taking money from them and didn't give the product to them. Unscrupulous traders show customers a good product, but when they deliver, the quality of that product is much worse. We can call such traders as Fraud Merchants. All of this is very risky for an e-commerce site. The e-commerce site is given priority over the customer so that the customer received the product correctly. So, it is very important and necessary for us to detect such merchant fraud. Merchant fraud has become a major impediment to the growth of e-commerce and has a significant impact on the economy. As a result, fraud detection is critical and important. Fraud detection is the technique of detecting a cardholder's deal action in instruction to control whether an incoming deal is being performed for the benefit of others. Method detection and anomaly detection are the two types of method finding. To evaluate whether an incoming operation is duplicitous or not, misuse detection and categorization approaches are used. Typically, such a strategy needs information of surviving types of scam in order to make models by learning diverse for patterns. Fraud detection is the process of constructing a profile of a cardholder's normal deal conduct founded on his or her historical transaction data [17]. This article is about the anomalous method. The normal is trained using random forest, and B has your features. Random forest is a classification algorithm that uses the votes of all based classifiers to determine categorization. This paper's main contribution can be summarized as follows. Random forests are used to train the regular fraud behaviour features in order to solve the fraud detection challenge. We also use Logistic regression to describe data and to explain the relationship between one dependent binary variable and one or more nominal ordinal independent variables. Conclusions drawn from the results of the tests would be useful for future study.

II. RRLATED WORK

A. Graph primarily based Fraud Detection:

In recent times, many frauds detection programs have sought to identify and prevent fraudulent activities that are related to online advertising. In [3],[5],[11], Tian et al modelled a programmatic advertisement fraud detection technique based on pattern recognition involving crowd behaviour. Similarly, Li et al projected the use of relational analysis for remotely detecting false click activity observed in mobile applications [4]. The following graph illustrates sample interactions between users and advertisers during an advertising campaign for order placement. Van Ital served as a spread engineer and was dominant in the algorithms market. He leveraged the propagation algorithmic program to check the effect of network data for future tax fraud detection, [5], [25]. A. Tseng explored a graph-based dishonourable telephony detection technique and projected it to detect telephone numbers with fraud tags automatically, [6]. They employed a weighted HITS algorithmic program to find out the trusted price of a signal and engineered 2

bipartite charts to represent behaviour among users (telecommunication conduct). Hu et al. developed an online mobile publicity fraud detection scheme that exploits techniques from nonnegative matrix factorization, WPROP and Random LDA coupled with the concept of bipartite matchings; Label Propagation Method is taken into account where Facebook's approach has been modified by introducing mutual information which involves weighted edges within random bipartite graphs instead [11].

B. Combination Model-based mostly Fraud Detection:

Combination models Leverage the mix of constantly applied mathematics deliveries to perfect the traditional illustrations and nonstandard instances [12], [14], [13]. Patriarch et al. assumed that the traditional and anomalies are each produced from Gaussian deliveries, while the anomalies have a bigger variance [12]. Byers et al. used a Page 4 ©Daffodil International University Poisson combination model to characterize the traditional information and then notice irregularities that doesn't fit in to any of the learned mock-ups [14].

III. PROPOSED METHOD

Our research is on to predict merchant fraud and use machine learning techniques to find fraud. As here mentioned, that we use machine learning so we need a dataset to learn our system. After learning these data systems can give a result so first, we need to process data. In our work we use Google collaborator environment Python and some libraries. Our training data set includes some missing or invalid entries which are represented by Nan. After prepossessing our data then we input our data in a machine learning algorithm. Applying our choosing algorithm then we see our representation data.

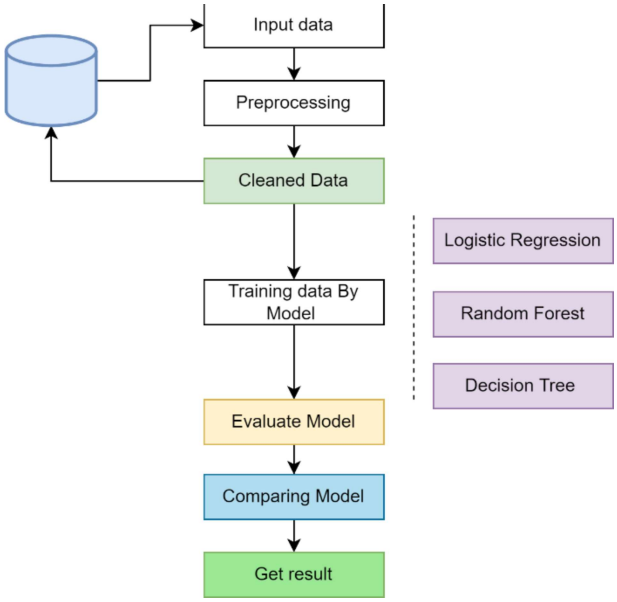

Fig 1. Block Diagram of Proposed Method

IV. EXPERIMENT ANALYSIS

By applying the AI method [Fig 1] in the informational collection, we got the yield result. This exploration works is finished by utilizing machine learning calculation. Here we use decision tree, confusion matrix, and random forest and the research work done with merchant and customers information and there are 16 attributes. The data divided into two data set one is training set and the other is test data set. First applying algorithms into the training data set and training dataset trained. At that point utilizing preparing informational collection the calculations applied to the test Informational collection. That is the test informational collections works Following with the preparation informational collections and figure out which calculation Discover the best exactness. Making a specialist Framework that attempts to help distinguish misrepresentation shipper. In view of it is a manmade brain power-based inability recognizing framework, so it is very well maybe distinguished extortion trader. For this [Fig 2] offices individual can recognize misrepresentation vendor of their crosscut in before stage.

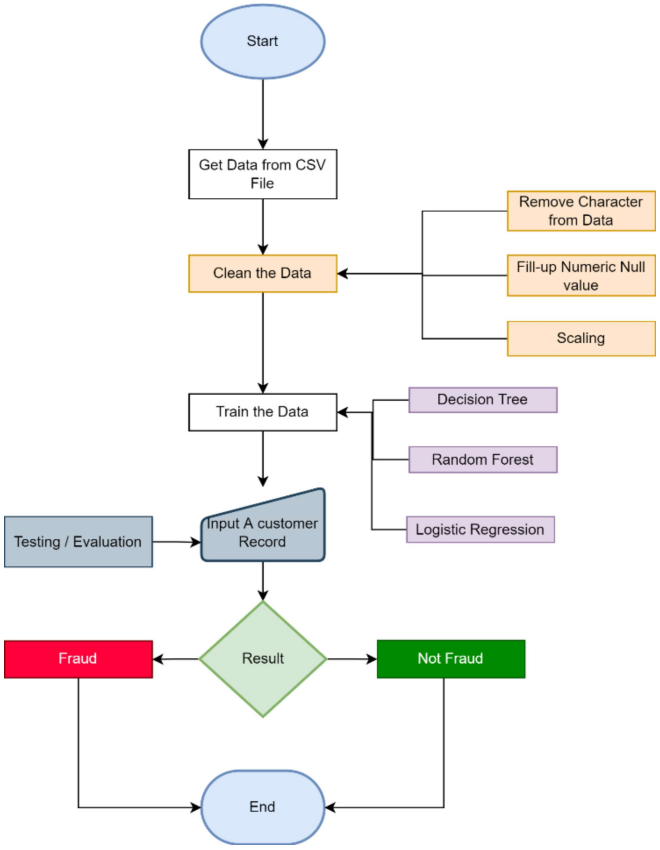

Fig 2. Block Diagram of Experiment Analysis.

A. Aalgorithm

Step 1: Start the methodology.
Step 2: Input information from database
Step 3: Data Cleaning: remove character, Null value fill up and Scaling
Step 4: Training method of data with models.
Step 5: Testing the recorded data.
if Fraud data is highly match it is Fraud result is 1
else if a Not Fraud result is 0
Step 6: End Methodology.

B. Dataset Description

At first, we have a tendency to pick some data so we have a tendency to get coaching data from there and set them into a

coaching knowledge set. In step with this coaching, knowledge takes some testing knowledge. Between that succeed in our expected outcome of the result. We've sorted our coaching knowledge set through direct search in an exceedingly range of places. Coaching knowledge, we've got to use our selecting formula. We have customer behavior data so we can easily predict that person is fraud or not. if you have authentic credit or debit card as your phone number or identification number then we can find the person or the purchase is authentic. So, from this data we can predict the fraud person for further purchase. We have Ecommerce Provider ID, Merchant ID, Merchant Registration Date, Registered Device ID, Gender, Age, IP Address parameter so we can detect by Ip address also.

C. Data Preprocessing

We need to find some organized info that we have a tendency to use to find our traditional outcome. Initially, we have a tendency to work to preprocess our preparation informational index. From getting a ready informational index we have a tendency to discover cleanup dataset. We gather info from the IEEE process Intelligence Society. Likewise, we have a tendency to contact varied on-line enterprises whose area unit break away at this region to find extortion shippers and consumers. Here we are cleared character where we need only numerical data by python library ord(chr).

D. Model Description

Down examined the lion's share category to coordinate with a variety of occasions of minority category all at once for the model to not add up towards the bigger half category (non-fraudulent). Converted the all-out factors into numeric utilizing hot coding. In provision regression l1 regularization is employed for highlight determination, irregular terra firm utilizes bootstrapping to settle onset of things all directly, and XGBoost limits blunder iteratively to assemble a solid classifier addicted to the frail classifier. Due to the dimensions of the dataset and restricted calculation power, the boundary tuning for the XGBoost model is finished physically addicted to determined conjectures and native space suggestions. With additional calculation power, the boundary calibration ought to be potential consequently.

1) Random Forest

Arbitrary Forest [Fig 1] forms numerous choice trees and unions them along to encourage an extra right and stable expectation. It very well may be utilized for both characterization and replace issues, which structure most of current AI framework. RF total numerous choice trees to restrict over fitting just as mistake because of predisposition and in this manner yield helpful outcomes. A random Forest characterization model is an assortment of grouping tree indicators. {h (x, k), k=1, 2...., T} RF is an assortment of Decision trees. It gives precise expectations to numerous sorts of uses. It can quantify the significance of each element regarding the preparation dataset. On the off chance that we utilize numerous trees in our woodland, in the long run the entirety of our element will be included. It is an administered learning model with related learning algorithmic principles. What's more it additional examinations information for arrangement and relapse investigation.

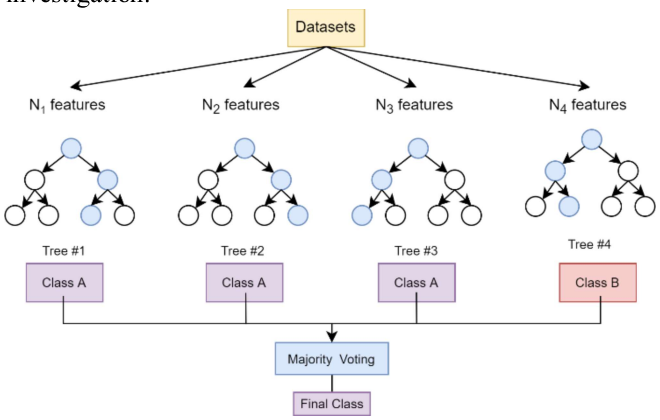

Fig 3. Random Forest.

2) Decision Tree

Decision Tree [Fig 4] is the one in all the supervised formulas. It desires a piece of information set for making a tree. Specifically, the knowledge set demand area unit given here. The attributes of a tuple unit tested against the selection tree. From the inspiration, a path is traced to a leaf node that contains the prediction for that tuple. In a call tree, our goal is to predict a choice or category from an information set. Therefore, first we've got to style our knowledge set as a coaching dataset wherever it's some attribute and a category. Category depends upon this totally different attribute. Establish first we've got to calculate the Entropy of this class. Then for every attribute, we've got to search out the data then Entropy for every attribute. Finally, after we work out Attribute Entropy from category Entropy, we'll get our Gain (class or decision).

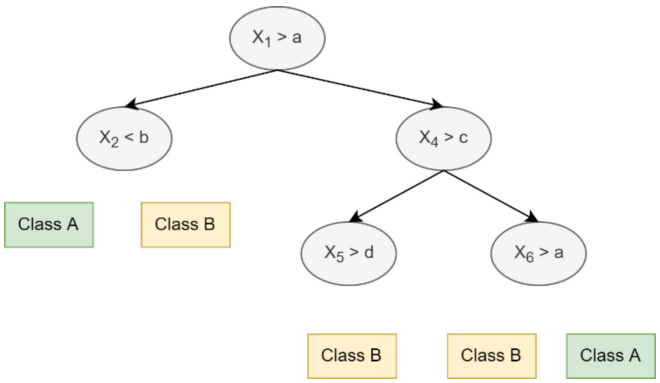

Fig 4. Deccision Tree

3) Logistic Regression

Logarithmic modification on the result variable permits America to demonstrate a non-straight relationship during a direct manner. This is often the condition used in supply Regression. Calculated relapse is used to amass probabilities proportionally inside the sight of quiet one logical variable. The system [Fig 5] is extremely like varied straight relapse, with the exemption that the reaction variable is binomial. The end result is that the result of each issue on the possibilities proportion of the detected occasion of interest. Straight Regression is AN AI calculation hooked

in to direct knowledge. It does a relapse task. Models of relapse a subjective expected esteem based on free variables. The assignment to predict a dependent variable price (y) visible of a specific free issue is played out via direct relapse (x).

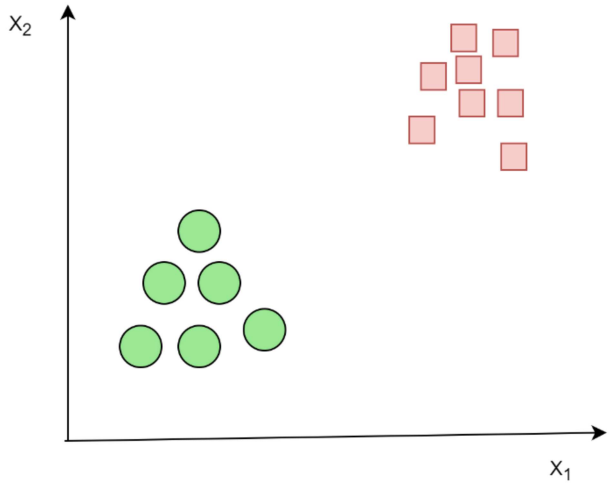

Fig 5. Logistic Regression.

V. Result And Analysis

In the modern science will be automation by machine learning in this circumstance we prepare a regular life problem solution that is use in our daily life. If your system is not intelligence then your business will be ruin. because in the future fraud will be clever so you have to more secure your business, for this study we find and analysis algorithm for finding fraud transaction. In here our algorithm perform very much. We got 84% accuracy [Fig 6] from RF. And another algorithm is worked much better

TABLE I. Results

Algorithm	sensitivity	specificity	precision	Accuracy
Decision tree	0.8314	0.7361	0.7061	0.7822
Random Forest	0.8236	0.8703	0.8781	0.8455
Logistic regression	0.7513	0.7093	0.6798	0.7138

TABLE II. Comparative Study

Previous Method	Proposed Method
There are lot of writer showed in their technology how to detect fraud.[3]	In our proposed system we showed that how to detect and how to solution this fraud problem
C.-H. Park analysis and identify key factor effect on fraud ecommerce system [6]	We are show in this paper with keyword and some identifying solution
Duman, E written in his paper detecting credit card fraud.	We are detecting whole ecommerce fraud and its solution.

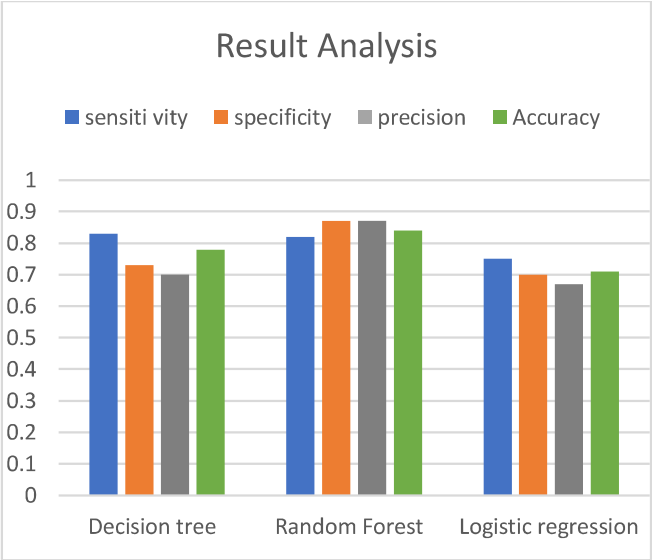

Fig 6. Result and analysis

VI. Avantage

In this new era, we are concern about our data. Data is the power of all hackers and fraud people. So, in daily life we cannot concentrate to our data. Every day we use credit card for shopping or any billing method. So, any fraud people can use our data from cloud [26], [27]. But if ecommerce uses new technology of machine learning, then they can pretend from fraud people. In this paper we are discussing about that. Machine learning has the power to detect fraud transaction and decline the transaction. Even our method is working so much accurate better than any previous model. This is the high time to save our business and save the ecommerce

VII. Future Work

The research work will help the e-commerce industry and customers to predict e-commerce fraud detection. On the basis of attributes in different categories, the percentage of frauds can be determined. If the percentage of fraud is high that means risk of e-commerce fraud is high then it is possible to take immediate step to overcome the problem. The various algorithms presented in this paper can be expanded to the online learning approach of machine learning in the future. They can be investigated in both offline (collecting data) and real-time scenarios for obtaining better results with reasonable accuracy. The demonstration of online learning will detect fraud cases in real time with the minimum time required for processing. Here, this research work is done by classification technique. In future we will work to find fraud customers and develop a system that will manage a software it easier and we will work to analyze the data more deeply. Finally, from this research work we can perform the test data set to real word and it help the e-commerce authority who prevent this system earlier

VIII. Conclusion

This paper has inspected the exhibition of random forest models. A real dataset on MasterCard exchanges is used in our trial. Though random forest acquires nice outcomes on very little set info, their square measure still

has some problems like unbalanced info. Our future work can zero in on braving these problems. And Sensitivity, Specificity, precision and accuracy are used to evaluate the performance of the proposed system. The accuracy for the Decision tree, logistic regression and random forest classifier is 0.8236, 0.8703, 0.8781 and 0.8455 respectively. By comparing all the three methods, found that the random forest classifier is better than the logistic regression and decision tree. The calculation of random forest itself needs to be improved. As an illustration, the democratic system expects that each one among base classifiers consumes eq, but a number of them may well be the next priority than others. On these lines, we tend to in addition plan to create some improvement for this calculation.

REFERENCES

[1] Random Forest for Credit Card Fraud Detection,Shiyang Xuan, Guanjun Liu, Zhenchuan Li,2018 IEEE

[2] Alae Chouiekha, EL Hassane Ibn EL Hajb "The First International Conference On Intelligent Computing in Data Sciences ConvNets for Fraud Detection analysis"10.1016/j.procs.2018.01.107

[3] Ruinan Zhang,Fanglan Zheng and Wei Min "Sequential Behavioral Data Processing Using Deep Learning and the Markov Transition Field in Online Fraud Detection"KDD, 2018, London, UK

[4] Evandro Caldeira, Gabriel Brandao˜and Adriano C. M. Pereira "Characterizing and Preventing chargebacks in next generation Web payments services" 978-1-4673-4794-5/12/$31.00 c 2012 IEEE

[5] Amin Ullah,Khan Muhammad,Killichbek Haydarov and Ijaz Ul Haq "One-Shot Learning for Surveillance Anomaly Recognition using Siamese 3D CNN",Conference Paper · July 2020

[6] C.-H. Park and Y.-G. Kim,"Identifying key factors affecting consumer purchase behavior in an online shopping context," International Journal of Retail & Distribution Management, 2003.

[7] T. Tian, J. Zhu, F. Xia, X. Zhuang, and T. Zhang, "Crowd fraud detection in internet advertising," in ACM WWW 2015

[8] V. Van Vlasselaer, T. Eliassi-Rad, L. Akoglu, M. Snoeck, and B. Baesens, "Gotcha! network-based fraud detection for social security fraud," Management Science, 2016.

[9] Soheil Jamshidi, Mahmoud Reza Hashemi "An Data Enrichment Scheme for Fraud Detection UsingSocialNetwork Analysis"2012 IEEE.

[10] Duman, E., and Ozcelik, M. H. (2011). Detecting credit card fraud by genetic algorithm and scatter search. Expert Systems with Applications, 38(10), 13057-13063

[11] S. Parvinder and M. Singh, "Fraud Detection by Monitoring Customer Behavior and Activities", International Journal of Computer Applications, vol. 111, no. 11, pp. 23-32, 2015

[12] E. Caldeira, G. Brandao and A. C. M. Pereira, "Fraud Analysis and Prevention in e-Commerce Transactions", 9th Latin American Web Congress, Ouro Preto, pp. 42-49, 2014

[13] E. Caldeira, G. Brandão, H. Campos and A. Pereira, "Characterizing and Evaluating Fraud in Electronic Transactions", Eighth Latin American Web Congress, Cartagena de Indias, pp. 115-122, 2012

[14] J. S. R. Jang, "ANFIS: adaptive-network-based fuzzy inference system", IEEE Transactions on Systems, Man, and Cybernetics, vol. 23, no. 3, pp. 665-685, 1993

[15] J. S. R. Jang and C.-T. Sun, "Neuro-fuzzy modeling and control", in Proceedings of the IEEE, vol. 83, no. 3, pp. 378-406, 1995

[16] T. K. Behera and S. Panigrahi, "Credit Card Fraud Detection: A Hybrid Approach Using Fuzzy Clustering & Neural Network", Second International Conference on Advances in Computing and Communication Engineering, Dehradun, pp. 494-499, 2015

[17] A. Srivastava, A. Kundu, S. Sural and A. Majumdar, "Credit Card Fraud Detection Using Hidden Markov Model", in IEEE Transactions on Dependable and Secure Computing, vol. 5, no. 1, pp. 37-48, 2008

[18] J. J.-S. Roger, C.-T. Sun, and E. Mizutani, "Neuro-fuzzy and soft computing: A computational approach to learning and machine intelligence", 1997

[19] W. Sultani, C. Chen, and M. Shah, "Real-world anomaly detection in surveillance videos," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6479- 6488

[20] Askari, S.M.S., Hussain, M.A.: Credit card fraud detection using fuzzy ID3. In: Proceeding -IEEE International Conference on Computing Communication and Automation ICCCA 2017, January 2017, pp. 446–452 (2017)

[21] Kumari, P., Mishra, S.P.: Analysis of credit card fraud detection using fusion classifiers. In:Behera, H., Nayak, J., Naik, B., Abraham, A. (eds.) Computational Intelligence in Data Mining, vol. 711, pp. 111–122. Springer, Singapore (2019)

[22] Akila, S., Srinivasulu Reddy, U.: Cost-sensitive Risk Induced Bayesian Inference Bagging (RIBIB) for credit card fraud detection. J. Comput. Sci. 27, 247–254 (2018)

[23] Lakshmi, S., Kavila, S.D.: Machine learning for credit card fraud detection system. Int. J. Appl. Eng. Res. 13(24), 16819–16824 (2018)

[24] Su, C.-H., et al.: A ensemble machine learning based system for merchant credit risk detection in merchant MCC misuse. J. Data Sci. 17(1) (2019)

[25] Wang, S., Liu, C., Gao, X., Qu, H., Xu, W.: Session-based fraud detection in online e-commerce transactions using recurrent neural networks. In: Altun, Y., et al. (eds.) Machine Learning and Knowledge Discovery in Databases. Lecture Notes in Computer Science, pp. 241–252. Springer, Cham (2017)

[26] V. Mareeswari and G. Gunasekaran, "Prevention of credit card fraud detection based on HSVM," in IEEE International Conference on Information Communication and Embedded Systems, 2016

[27] Mittal, P. (2020) '[CREDIT CARD FRAUD DETECTION SYSTEM]', (May), pp. 0–27. doi: 10.13140/RG.2.2.28192.81924

[28] Li, C. et al. (2021) 'Application of Credit Card Fraud Detection Based on CS - SVM', 11(1). doi: 10.18178/ijmlc.2021.11.1.1011.

[29] Krishnan, M. Soumya, and E. R. Vimina. "E-commerce Logistic Route Optimization Deciphered Through Meta-Heuristic Algorithms by Solving TSP." In International Conference on Communication, Computing and Electronics Systems: Proceedings of ICCCES 2020, vol. 733, p. 419. Springer Nature, 2021.

[30] Kumar, E.Rajesh, A. Aravind, E. Jotheeswar Raghava, and K. Abhinay. "Decision Making Among Online Product in E-Commerce Websites." In Inventive Computation and Information Technologies, pp. 529-536. Springer, Singapore, 2021