



IJIRCCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 10, Issue 5, May 2022

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 8.165



9940 572 462



6381 907 438



ijircce@gmail.com



www.ijircce.com

E-Commerce Transaction Fraud Detection through Ensemble Learning

Vijaykrishna Hansraj Yadav, Prashant Sanjay Vaishnav, Sourabh Ramesh Sonekar, Pratiksha

Shivaji Shinde, Prof. Vikas Maral

Dept. of Computer, KJCOEMR, Pune, India

Dept. of Computer, KJCOEMR, Pune, India

Dept. of Computer, KJCOEMR, Pune, India

Dept. of Computer, KJCOEMR, Pune, India

Dept. of Computer, KJCOEMR, Pune, India

ABSTRACT: The accelerated expansion of Fintech industry, internet based transactions utilizing alternative financing techniques have indeed been intimately connected with people's everyday routines. Whereas the instant payment technique makes life easier for customers, it also increases the risk of internet trade fraud. Individuals are hard to identify because of the hidden and various fraud methods, which result in losses for both customers and institutions. Many fraudsters have devised methods to abuse customers and thieve the credit card details in order to use that for unlawful transactions since the advent of credit card payments and contactless transactions. Nearly every day, this results in a large number of fraudulent transactions. This leads to a lot of losses that can be extremely problematic, therefore this research article defines an effective approach for the identification of fraudulent transactions through the implementation of deep learning methodologies. The prescribed approach utilizes Linear Clustering, Entropy Estimation, and Hypergraph formation along with Deep Belief Networks and Decision Making to achieve transaction fraud detection. The experimental evaluation has indicated that the approach achieves much better accuracy than the conventional approaches for transaction fraud detection.

KEYWORDS: Deep Belief Network, Information gain, Hypergraph, neo4j, transaction Fraud.

I. INTRODUCTION

Digital transactions have expanded at an extraordinary rate in the recent decade. The total amount of transactions is higher compared to a year ago, and the number of transfers has climbed dramatically. Similarly, the number of electronic payments in India has expanded markedly in previous years. Online banking and mobile payments have provided billions of people across the globe access to financial services. They've also given customers, businesses, and financial institutions substantial benefits, including that of the capacity to grow, cheaper operational expenses, ease of use, accessibility, and improved productivity.

But at the other hand, fraudulent techniques have swiftly evolved to take advantage of the growing fast-paced online payment environment. Traditionally, detecting fraud and financial crime relied on a plethora of laws and focuses on being objective, such as maximal transaction limits, to recognize suspicious behavior. Fraudsters have worked out how to get around rigid restrictions, therefore such administrative and rule-based methods have failed in recent years. Contradictory to latest industry figures, global financial misconduct is anticipated to increase dramatically. Embezzlement accounts for a significant number of the unlawful charges in this area, contributing for a significant proportion of the frauds.

Avoiding and recuperating from fraud has also become much more expensive. Fraud strategies have developed dramatically in response to the surge in fraudulent activities. As online payments have become more common, card fraud has declined considerably in recent years. Consequently, the risk of online transaction fraud has increased significantly. Electronic fraudulent financial cases surged drastically as a result of the worldwide prevalence and the subsequent significant growth in digital commerce quantities.

Due to the rapid progress of digital transactions, credit card transactions have become one of the most common methods of payment in recent years. Nonetheless, the development of sensitive electronic financial transactions has

resulted in unprecedented types of financial fraud. Scammers routinely gather information and evidence from consumers in order to effectively carry out illegal operations in a short period of time. As a consequence, banks should use a range of methods in the actual world as well as on the web to improve credit card fraud detection and protect consumers' privacy.

Such fraud instances are becoming increasingly complex, which is a major concern because present techniques to fraud detection and classification are inadequate. The stationary circumstances are readily detected and overcome by these fraudsters, making it incredibly difficult to pinpoint. Preventative measures are rapidly developed, and these approaches might go undiscovered for lengthy periods of time, permitting fraudsters to continue their nefarious activities. As a result, for the objective of detecting fraudulent transactions, an effective and successful technique is required. For the objective of detecting transaction fraud, the model of machine learning approaches has been among the foremost important.

The technique outlined in this paper is based on the linear clustering approach. The linearly clustered financial data, as well as the retrieved characteristics, are used to detect and prevent fraud. For the hypergraph to be created, entropy estimate and frequent itemset mining are used. The hypergraph created using Neo4j has shown to be among the most effective in determining the relationship between various traits and qualities. Due of the large and complicated composition of the graphs, this study delivers great understanding but can be difficult to assess. As a consequence, machine learning techniques are a powerful tool for extracting insightful data and detecting transactional fraud, which would be efficiently categorized utilizing decision making to reach the required outcomes.

In this research article related works are mentioned in the section 2. The proposed technique is deeply narrated in the section 3. The experimental evaluation is performed in section 4 and whereas section 5 concludes this research article with the scope for future enhancement.

II. LITERATURE SURVEY

Rongrong Jing [1] represents the fact that now the online community has expanded the simplicity and efficiency of residing for a sizable sub-section of general public. This has been extremely beneficial in the framework of banking services provided via the digital platform. These enable efficient transactions via the internet paradigm, hence enhancing the quality of life for its consumers. This is why the majority of people have been using the internet platform to enhance their lifestyles and enable transactions online. The internet platform allows fraudsters to realize their aims of fraudulent transactions, which could be a significant occurrence. To address incomplete data and label imbalances, the authors presented a data integrity enhancement technique to facilitate better credit card fraud identification.

PradheepanRaghavan [2] explains that because the initiation of electronic shopping and card payments have become the standard practice, that there's been an increasing trend in the median transactions which are being accomplished utilizing these technologies. Because of the increased amount of transactions, fraudsters and other cybercriminals have been drawn to these forums to carry out their malicious intentions. Because of the absence of comprehension and the cunning of the fraudsters, the frequency of fraudulent operations on such platforms has skyrocketed. The scam has no trends since it is dynamic, making it extremely difficult to discover and recognize. As a result, an effective technique for detecting fraud based on machine learning and deep learning is required.

Na Ruan [3] describes the pervasiveness of fraud in multiple sectors of the society. These really are persons with malevolent purpose who are extremely driven to perform illegal acts in order to get their motivation. There have already been successful strategies created to address the rising number of fraudulent transactions that already have plagued the banking system. However, fraudsters have become more resourceful and have found strategies to avoid the surveillance initiatives undertaken by these organizations. As a result, the authors of this publication have recommended using data mining on Call Detail Recordings to detect fraud done cooperatively while protecting privacy.

Aastha Bhardwaj [4] illustrates how the prevalence of fraudulent purchases has grown in current history. Among the most widely employed measures for reducing these activities and making the banking industry safer and dependable has been the recognition of fraud in banking transactions. There have been several ways attempted to aid in the identification of fraudulent purchases, however the majority of them have been inadequate in terms of apparent reliability and transactional evaluation. The researchers of this study have proposed an effective technique for accomplishing qualitatively assessment of the financial accounts and identifying fraud.

Ruoyu Wang [5] specifies that there's been a substantial quantity of fraud that is getting publicly perpetrated online where it has been continuing to increase in capacity day after day. This is because online transactions have given thieves and anyone with malicious intentions a great deal of influence in performing these nefarious transactions. There are various alternative ways for the objective of identifying and preventing fraudulent transactions that banking firms have a tough time assessing. Because criminals have gotten more proficient in circumventing the traditional identification strategy, studies in this field have recommended the use of quantitative recognition for the aim of collaborative fraud prevention and detection.

Gabriel Castaneda [6] explains that there's been an elevated priority toward these prevention and treatment due to the troublesome circumstances which have been experienced by the wider population. With the epidemic nearing, many people have turned to a lifestyle change. This also implies that the quantity of fraud cases in this area, which can be highly damaging, has increased. Medical fraud is among the most challenging events to recognize and prevent, and the researchers have provided a new technique that accomplishes fraud detection on huge data utilizing the maxout computational model for fraud recognition to give a resolution to this dilemma.

Xiaoguo Wang [7] discusses the tremendous rise in the volume of people using the internet platform to achieve an appropriate enhancement in the bank transaction methodology. The growth in the quantity of customers, as well as the significant surge in the frequency of operations, has garnered a significant number of scammers and persons with malicious intentions. As a result, a considerable quantity of malfeasance is perpetrated on digital sites, creating a troublesome scenario that might result in large-scale liabilities. The developers of this technique have developed a novel technique to combat fraud by combining K means clustering with the Markov Model.

AbdollahEshghi [8] articulates the reality that the percentage of suspicious purchases are continuing to increase substantially over the preceding few years. This may be ascribed to the massive growth in the volume of payments in recent years as a result of the usage of the digital platform for numerous financial and other economic operations. This has resulted in a significant increase in the number of payments, shifting the emphasis of criminals. These illicit operations have posed a significant risk to the site's stability, prompting a variety of measures aimed to recognize suspicious transactions that have been inadequate or ambiguous in their implementations. The authors of this paper suggest a unique mix of semi-supervised and supervised algorithms to enhance this approach of fraud detection.

ErenKurshan [9] elaborates that it is a surge in the portion of online payments or the mobile banking that have already been continuing to increase in prominence across the globe. This surge in utilization might be ascribed to the simplicity of operation and accessibility that these platforms provide. The enhanced qualities have proven beneficial in attaining a significant transaction volume on this portal nearly every day, with new customers being introduced on a regular basis. There seems to be a noticeable increment in the quantity of various forgeries, schemes, and other illegal actions occurring online, which must be curtailed in order to improve trustworthiness. As a result, the authors recommend using graph computation in conjunction with machine learning to confirm the prevalence of corruption.

BayuNurPambudi [10] explains that there's a steadily increasing reported incidents of adoption of internet financing options for the intention of attaining the money transfers. In compared to physical or cash operations, the amount of transactions completed on the web platform has grown. This advancement has not come devoid consequences, as there has been a spike in the volume of fraudsters and criminals on the web platform, leading to a dramatic surge in the frequency of scams as well as other criminal enterprises such as laundering money. As a result, the researchers of this journal have advocated the usage of an improved SVM or Support Vector Machine to identify financial fraud successfully.

MasoudErfani [11] describes that there has been a surge in the percentage of residents that incorporate the online system for the goal of accomplishing their economic and other financial transactions. This is because the internet banking procedure is incredibly simple and it can be completed with just a few keystrokes. Because of this ease, the bulk of transactions are now conducted over the web, establishing a new opportunity for thieves to fulfill their malicious goals. This is extremely troublesome since the fraud may result in enormous quantities of money being lost and costing a big number of people and organizations a great deal of money. As a result, the authors have presented a comprehensive explanation of support vector data for the goal of detecting bank fraud.

III. PROPOSED METHODOLOGY

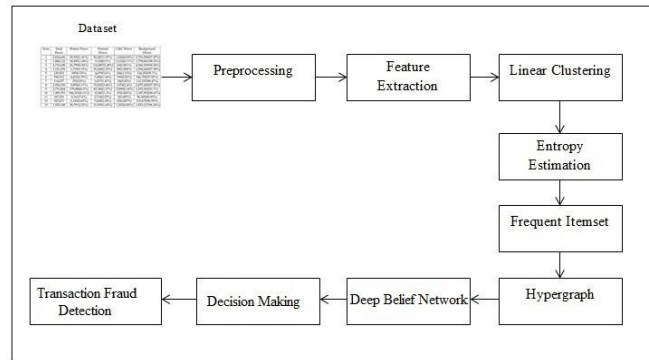


Figure 1: System Overview

The proposed model for E-commerce Fraud detection system is represented in the figure 1. The steps that are involved in the process are properly narrated below.

Step 1: Dataset Preprocessing and Feature Extraction – In the first step of the prescribed approach, where an E-Commerce transaction dataset has been downloaded from the URL - <https://www.kaggle.com/vbinh002/fraudecommerce/data>.

The downloaded dataset contains a number of attributes such as user_id, signup_time, purchase_time, purchase_value, device_id, source, browser, sex, age, ip_address and class. This dataset is stored in a workbook and then it is fed to the This data is saved in a worksheet before being input into the aforementioned fraud detection algorithm. The dataset is written into the double dimension list because it is supplied into the system, which will then be submitted to the preprocessing operation.

Major properties are chosen during preprocessing, while the others are either vomited or utilised afterwards. The source, browser, and class attributes are discarded in this phase, while the remainder are positioned in a suitable list. Three qualities from any of this preprocessed set are regarded key traits that play an essential part in the fraud detection process in E-Commerce transactions. The algorithm picks three characteristics in three categories for this purpose: signup time, purchase time, and purchase value. For this purpose system selects three attributes like signup_time, purchase_time, purchase_value, signup_time and in three different lists.

Step 2: Linear Clustering and Entropy Estimation – After the pre-processed list is completed, the information from the list is partitioned into several clusters. It is easier to pick a group of data around which to apply a neural network after constructing clusters.

The preprocessed list is partitioned into five equal index ranges throughout the Linear clustering phase. Subsequently a single cluster is produced for every one of the range indexes, which is then included to the resulting linear cluster list. This procedure is described in the algorithm 1 below.

ALGORITHM 1: Cluster Formation

```

//Input : Preprocessed List PL
//Output:Linear Cluster List CL
1: Start
2: Index=∅ [Index List]
3: DIV=PL size / N [N= Number of Cluster]
4: begin=0, end=0
5: for i=0 to N
6:   Range=∅ [Range List]
7:   Range[0]=begin
8:   end=begin+DIV
    
```

```

9:   Range[1]=end
10:  Index= Index+Range
11:  begin=end
12:  end for
13:  for i=0 to Size of Index
14:      TL=∅ [Temp List]
15:      R= Index[i]
16:      MIN= R[0]
17:      MAX=R[1]
18  for j=MIN to MAX
19:      TL= TL + PL[j]
20:  end for
21:  CL= CL+ TL
22:  end for
23:  return CL
24: Stop
    
```

The suggested method incorporates a methodology to calculate E-commerce fraud across all groups. Based on this procedure, every cluster is counted as P if it has the identical amount of rows with almost the same registration and purchasing period. The cluster dimensions is then calculated as S. Then, using Shannon information gain on every cluster, an entropy value is calculated, which shows the dispersion of the proportion of rows that subsequently meet this technique, as shown in Equation 1.

$$IG = -\frac{P}{S} \log \frac{P}{S} - \frac{(S-P)}{S} \log \frac{(S-P)}{S} \text{ -----(1)}$$

Where

IG = Information Gain of the cluster

The Shannon information gain formula yields an information gain factor for the cluster that ranges from zero to 1. Any number approaching 1 reflects the cluster's relevance in terms of the fraud identification procedure. A double dimension list is used to hold every cluster index and its associated gain value. This list is arranged in descending order to determine the best top clusters, which will ultimately comprise certain useful information for the fraud detection procedure.

Step 3: Frequent itemset and Hypergraph estimation – The data from the top clusters is then combined into a single list. Then, from such a single list, all activities with the same registration and buy times are detected, and the associated purchase amount is collected in a list. The hash set operation is then used to the purchase price list to obtain the distinctive purchase price in a separate list, which is referred to as the frequent item list.

The acquired frequent item list is used to execute a parameter foresight assessment in a single input collection. To calculate the purchase value, the registration and purchase times are compared to see if they are comparable. This transaction value is then compared with the frequent pattern list's property to create a hyper graph objects with user ID as well as purchase value as vertices and edge String as "purchase value." This hyper graph is saved in a sophisticated graph database, such as neo4j, and may be seen in a browser.

Step 4: Deep Belief network – The DBN method is provided the resulting frequent item list as an input list. In which the hidden and output layers for the properties User ID and purchase value are evaluated for every transaction rows in the frequently item list. This is accomplished by using the same quantities for the randomized weights W1,W2,W3,W4,W5,W6,W7,W8,B1,B2 as the goal values. B1 and B2 are the bias levels that are employed to keep the neurons stable. The output layers are then approximated employing Equations 2 and 3 for Hidden layer as well as Activation function. The retrieved output layers are combined with the goal values to create a new prediction list that will be used to identify fraud.



$$X = (AT1 * W1) + (AT2 * W2) + B1 \quad (2)$$

$$H_{LV} = \frac{1}{(1 + \exp(-X))} \quad (3)$$

Where attribute AT1 and AT2 are the USERID and Purchase Value respectively. Then the activation function called the sigmoid function is given by Equation 3 of the neural network. H_{LV} – Indicates the hidden layer value.

Step 5: Decision Making – To identify E-Commerce fraud IDs, the resulting fraud detection probability list from previous phase is fed into the Decision Classification process. DBN’s fraud recognition probability list is used to compute the upper and lower limits of purchase values in this procedure. Subsequently, employing these lowest and highest values, a distance separating them is calculated as a dividend, which is then reduced by division through 5 to provide the quotient Q. VERY LOW, LOW, MEDIUM, HIGH, and VERY HIGH are the five decision crisp values generated by this Quotient.

To retrieve the USER ID, the acquired decision crisp values are compared to the fraud detection likelihood list for their boundary. And then these USER IDs are classified into different cluster with respect to the decision crisp ranges. These classified clusters eventually indicates the different level of fraud right from the VERY LOW to VERY HIGH range, which is displayed on an interactive user interface.

IV. RESULTS AND DISCUSSIONS

The presented approach for establishing a precise e-commerce fraud recognition framework has been established employing the NetBeans IDE as well as the Java programming language. The suggested approach was created on a system with an Intel Core i5 CPU, 500GB of hard disk drive capacity, and 4GB of RAM. The MySQL database service is used to manage the database.

Substantial testing was done to determine the suggested methodology's performance indicators. The RMSE and MSE performance measures, which may systematically represent the efficiency of the suggested strategy, were employed to evaluate the validity of the proposed strategy. The performance metrics were thoroughly examined in order to demonstrate that the e-commerce fraud recognition method relying on Deep Belief Networks as well as Decision Classification described in this study was properly implemented.

Performance Evaluation based on Root Mean Square Error

The Root Mean Square Error (RMSE) is used to assess the predictive performance of the E-commerce Fraud detection system. The error value amongst the two interrelated and consistent variables is calculated using the RMSE method. The achieved fraudulent transactions as well as the expected amount of fraudulent transactions are two associated and continuous elements in the developed framework. Equation 4 below can be used to calculate this.

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (x_{1,i} - x_{2,i})^2}{n}} \quad (4)$$

Where,

\sum - Summation

$(x_1 - x_2)^2$ - Differences Squared for the summation in between the expected number of fraudulent transactions and the achieved number of fraudulent transactions

n - Number of samples or Trails

An indepth realization of the performance measures is realized using the RMSE metric, and the outcomes are populated in table 1 below.

Experiment No.	Rows	Expected no of Transactions	Achieved no of Transactions	MSE
1	1000	51	38	169
2	2000	109	84	625
3	3000	159	121	1444
4	4000	210	163	2209
5	5000	273	210	3969

Table 1: Mean Square Error measurement

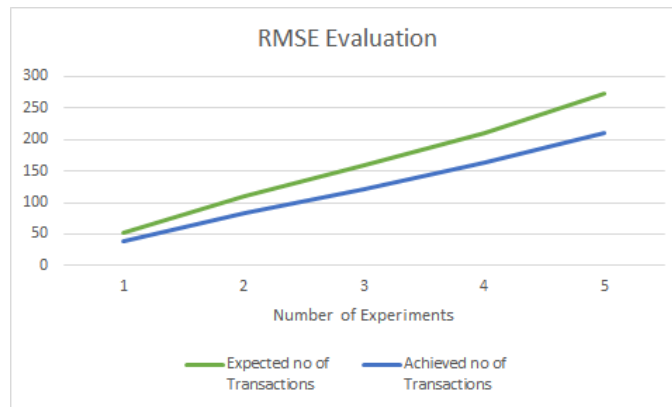


Figure 2: Comparison of MSE in between Expected number of Fraudulent Transactions V/s Achieved number Fraudulent Transactions

The mean square error rate among the number of achieved fraudulent transactions and the number of expected fraudulent transactions for more than just a set of five experiments is depicted in Table 1 and the graph represented in Figure 2. The initial experiment had 1000 records, and following attempts faced a 1000-entry increase consecutively. The average RMSE of the experimental outcomes is 41.02. For the identification of fraudulent transactions, the realized RMSE values are computed. A system is regarded acceptable if the RMSE value achieved is less than 50. The RMSE number obtained shows that the performance obtained for the initial installation of such a mechanism was immensely effective and remarkable.

V. CONCLUSION AND FUTURE SCOPE

The presented approach for the detection of transaction fraud has been elaborated in this research article. The methodology takes the transaction dataset as an input into the system. The Dataset is initially pre-processed to achieve the effective removal or the missing data, irrelevant values and incomplete data that might cause unnecessary wastage of computational power. The pre-processed dataset is then subjected to feature extraction where the relevant features or attributes are selected for use in the next steps of the approach. The pre-processed dataset with the features is provided to the Linear Clustering approach for the purpose of clustering the data. The achieved clusters are provided to the entropy estimation module to calculate the information gain through the Shannon information gain. The next step extracts the frequent itemset using the output from the earlier steps. These itemsets are then used for the purpose of generating the hypergraph through neo4j. The hypergraph along with the other features are then provided to the Deep Belief Network for effective identification of fraud. The Decision Making approach classifies the output from the DBN to achieve effective transaction fraud detection. The experimental evaluations have been useful in proving the effectivity of the transaction fraud detection approach.

The future research approach can be done in the direction of achieving the transaction fraud detection in a real time scenario. The approach can also be converted into an API for easier integration into existing banking systems.

REFERENCES

- [1] R. Jing et al., "Improving the Data Quality for Credit Card Fraud Detection," 2020 IEEE International Conference on Intelligence and Security Informatics (ISI), 2020, pp. 1-6, doi: 10.1109/ISI49825.2020.9280510.
- [2] P. Raghavan and N. E. Gayar, "Fraud Detection using Machine Learning and Deep Learning," 2019 International Conference on Computational Intelligence and Knowledge Economy (ICCIKE), 2019, pp. 334-339, doi: 10.1109/ICCIKE47802.2019.9004231.
- [3] N. Ruan, Z. Wei and J. Liu, "Cooperative Fraud Detection Model With Privacy-Preserving in Real CDR Datasets," in IEEE Access, vol. 7, pp. 115261-115272, 2019, doi: 10.1109/ACCESS.2019.2935759.
- [4] A. Bhardwaj and R. Gupta, "Qualitative analysis of financial statements for fraud detection," 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), 2018, pp. 318-320, doi: 10.1109/ICACCCN.2018.8748478.
- [5] R. Wang et al., "Statistical Detection Of Collective Data Fraud," 2020 IEEE International Conference on Multimedia and Expo (ICME), 2020, pp. 1-6, doi: 10.1109/ICME46284.2020.9102889.
- [6] G. Castaneda, P. Morris and T. M. Khoshgoftaar, "Maxout Neural Network for Big Data Medical Fraud Detection," 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService), 2019, pp. 357-362, doi: 10.1109/BigDataService.2019.00064.
- [7] X. Wang, H. Wu and Z. Yi, "Research on Bank Anti-Fraud Model Based on K-Means and Hidden Markov Model," 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC), 2018, pp. 780-784, doi: 10.1109/ICIVC.2018.8492795.
- [8] A. Eshghi and M. Kargari, "Introducing a Method for Combining Supervised and Semi-Supervised Methods in Fraud Detection," 2019 15th Iran International Industrial Engineering Conference (IIIEC), 2019, pp. 23-30, doi: 10.1109/IIIEC.2019.8720642.
- [9] E. Kurshan, H. Shen and H. Yu, "Financial Crime & Fraud Detection Using Graph Computing: Application Considerations & Outlook," 2020 Second International Conference on Transdisciplinary AI (TransAI), 2020, pp. 125-130, doi: 10.1109/TransAI49837.2020.00029.
- [10] B. N. Pambudi, I. Hidayah and S. Fauziati, "Improving Money Laundering Detection Using Optimized Support Vector Machine," 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), 2019, pp. 273-278, doi: 10.1109/ISRITI48646.2019.9034655.
- [11] M. Erfani, F. Shoeleh and A. A. Ghorbani, "Financial Fraud Detection using Deep Support Vector Data Description," 2020 IEEE International Conference on Big Data (Big Data), 2020, pp. 2274-2282, doi: 10.1109/BigData50022.2020.9378256.



INNO  **SPACE**
SJIF Scientific Journal Impact Factor

Impact Factor: 8.165

doi[®]
cross **ref**

ISSN INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 **9940 572 462**  **6381 907 438**  **ijircce@gmail.com**



www.ijircce.com

Scan to save the contact details