International Journal of Engineering Inventions
e-ISSN: 2278-7461, p-ISSN: 2319-6491
Volume 11, Issue 12 [December. 2022] PP: 97-103

Fraud Detection in Credit Card using Data Mining Classification approach: A review

Mohd Towqeer Ul Haq, Navdeep Kumar Chopra
Department of Computer Science & Engineering
Chandigarh Engineering College, Landran, Mohali
Punjab, India

Abstract—

Due to “developments in technology and network communication, the adoption of transactions using online methods in daily life has drastically expanded over the previous decade. Credit card fraud by system abuse resemble the theft or misuse of user credit card information for his personal gain without the cardholder's consent. To detect such types of fraud, it is mandatory to investigate the usage behavior of users in previous operations. Credit card fraud is the unauthorized use of a credit card or the unauthorized obtaining of sensitive credit card information. Due to the convenience, simplicity, and easy to use of the online action system, many new users joining large number of participants in such a program. Many ML algorithms and DM strategies are used for obtaining CCF. Data mining (DM) involves a fundamental technique that deepens data as opposed to simple data and information. Indeed, data mining is an important component of the acquisition process. The credit card companies (CC) provide their consumers with a variety of cards. The research compares the performance of Decision Tree algo, Logistic Regression algo and Random Forest algo in detecting the fraud in credit card. The work is implemented using Python as language and uses three different ML classification techniques. The accuracy score, f1-score, precision, and recall score parameters are used to measure the algorithm’s performance.

Keywords: Credit card fraud, detection and classification, techniques

I. INTRODUCTION

Fraud in financial method is a rising problem in today’s world with far-reaching implications for the financial industries dealing with financial transactions, corporations, and governments.Fraud is defined as an illegal act carried out with the intent to gain financial gain. The growing reliance on various internet technologies led to a significant growth in credit card transactions. As credit card transactions replace cash as the primary mode of payment for both online and offline purchases, credit card fraud is rapidly increasing.However, financial institutions have concentrated on the most recent accounting systems to combat credit card fraud. To distinguish the fraud in credit card system data mining tools are playing an important role. Data mining is a class of machine-learning algos that can analyze and extract non-trivial patterns from data. Because of its effectiveness, data mining is a popular method of combating fraud. Methods for data mining are well-defined processes that use data as an input and produce patterns as their result. In other words, data mining is used to evaluate massive huge data and extract actionable information that can be interacted with. for future reference. Once we've identified the best model for the data, we can use it to forecast future events by categorizing the data. The supplied data is examined using a suitable model to determine whether or not it indicates any fraudulent actions. Machine Learning and data mining are both becoming increasingly important areas of focus for many commercial enterprises, including the banking industry. It is the procedure of evaluating facts from various angles and distilling it into helpful knowledge. DM assists banks in detecting patterns in groups and discovering previously unknown relationships in data. It can be utilized in credit risk management and fraud detection. In the field of credit risk management, banks provide credit cards to customers after checking various data. Even if banks are cautious when issuing credit cards, there remains a potential that users will fail on the credit card. Data mining and machine learning approaches can assist in distinguishing between clients who pick up a credit card immediately and those who do not. It also predicts when a consumer will become delinquent and whether issuing a credit card to a specific customer would result in a negative credit card.

CREDIT CARD FRAUD

It states to the illegal use of credit card or theft of information from the owner of the card. The fraud trick apps and behaviors are associated with two types of fraud. When app fraud happens, the perpetrators either ask the bank to give a new card or give it to businesses that use false information.In order to commit duplicate fraud, a person can submit numerous applications with identical or almost identical descriptions (named identity fraud). Theft/loss of cards, fake cards, mail theft and "current card holder does not exist" fraud are the four main subcategories of behavioral fraud.A credit card is stolen or obtained when a lost or stolen card fraud happens. Fraud using mails occurs when a cheater receives personal information of a user from a bank in the mail before it passes to its original card holder. There is no mention of Fake & Card Holders Fraud & credit card descriptions. Previously, remote communications could be conducted utilizing card information sent via mail, phone, or internet. Second, based on card data, fraudulent cards are manufactured.

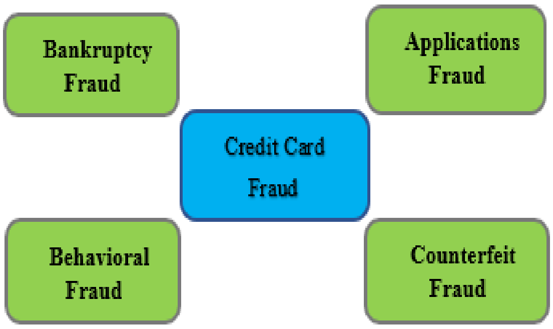

Fig 1: Frauds in CC

TYPES OF FRAUD

Bankruptcy Fraud: It is the first type of fraud that occurs when a user while knowing that there is no money on the card uses his own credit card and the bank is required to pay by sending a bill to the address.
Applications Fraud: It is the second type of fraud occurs when an application form for a credit card is sent to the bank with false information.
Behavioral Fraud: Third, while purchasing an item online, the card information of any other credit card is submitted without being in knowledge of the owner.

Theft Fraud/ Counterfeit Fraud: This type of fraud encompasses taking a credit card and using it by posing as the card's owner until the bank block the card.

DIFFICULITIES IN CREDIT CARD FRAUD DETECTION:

- Unbalanced data- Information from CCFD is not balanced. This suggests that all transactions using credit cards are fraudulent. Fraudulent transactions are quite challenging to identify.
- Various misclassifications with varying significance- Dissimilar diversification errors have dissimilar significance in the fraud detection process. A typical abortion transaction is not fraudulent. If you commit the first error, classification will be given more consideration.
- Data that cross through- The frequency of fraudulent transactions can be high, although they frequently occur (false positive), and fraudulent transactions can sometimes look legitimate (false negative). As a result, fraud detection technologies are critical to procurement with a low rate of false positives and fake bids.
- Fraud detection costs- The system that is detecting the fraud must calculate for both the cost of detecting fraud as well as the cost of preventing it. For instance, avoiding obtaining money while stopping a few dollars' worth of fraudulent transactions.

CLASSIFICATION

It is the most popular data mining technique, that uses classification method that involves the use of a set of pre-classified samples that helps to build a model that categorizes information from a population on a larger scale. Furthermore, this retrieves important data information. An example is ‘Gmail’. They utilize an algorithm to determine if the email received is spam or a legitimate email. This classifier-training technique works out the set of parameters required for correct discrimination using pre-classified samples. An algorithmic software is used to encodes these given parameters into a model known as a classifier model. It is heavily reliant on the classifier.

II. LITERATURE REVIEW

Shiyang Xuan et al: They proposed random forest and CART-based random forest algorithms. They train the behavior aspects of regular and abnormal transactions using distinct random forest algos, and the algorithms differ on the basis of classifications and performances. They implemented the algorithms on a data-set derived from a Chinese e-commerce company. In which the fraud transaction ratio in subgroups ranges from 1:1 to 10:1. Consequently, the accuracy of the random tree algo based random forest is 91.96%, with respect to the accuracy of the CART-based random forest, which is 96.7%. Because the data that was used was from the B2C dataset, hence various issues arose like imbalanced data. This result in enhancement of the algorithm. [Random Forest for Detecting Credit Card Fraud].

Kuldeep Randhawa et al. - For CCFD, they installed twelve machine learning algorithms ranging from a neural network to various deep learning algorithms. They are analyzing the benchmark and real-world dataset performance. Furthermore, the majority voting methods and AdaBoost are used to build the hybrid models. As explained in the accompanying paper, single and hybrid models exist. They had given the results for both parameters (Benchmark and real-world data-sets) using their twelve selected algorithms, which are Random Forest, Decision Tree, Nave Bayes, Gradient Boosted Tree, Linear Regression, Decision Stump, Random Tree, Neural Network, Deep Learning, SVM, Logistic Regression and Multilayer Perceptron. When all these algorithms were combined with majority voting methods and AdaBoost using benchmark data, the Random Forest algorithm achieved the highest accuracy i.e. 95% and sensitivity i.e. 91%. The accuracy rate remains above 90% when tested with real-world data including 30% noise in the dataset. MCC (Mathews Correlation Coefficient) is a common measure of model performance, and the top MCC score with majority vote is 0.823, whereas 0.942 with 30% noise included to the dataset.

Krishna Modi et al- Previous transaction data from clients is being analyzed to determine cost behavior. A fraudulent transaction is one that deviates from the available cost pattern. Banks and credit card firms use a variety of data mining techniques, such as fuzzy clustering, hidden Markov models, decision perspectives, rule-based mining, NN, and hybrid methods, to identify fraud.Based on previous customer actions, both techniques are utilised to evaluate prevalent usage patterns. This study compares different strategies for identifying fraud.

III. OBJECTIVES

The following are the study's primary objectives:

1. Examine the advantages and disadvantages of the several credit card fraud detection methods.

2. Using data classification, create a better credit card fraud detection technique.

3. Simulate and compare the proposed scheme's findings with existing credit card fraud detection approaches.

RESEARCH GAPS

Authors-Aisha Mohammad Fayyomi, Derar Eleyan, Amina Eleyan(2021) -The primary objective of the research is to create algorithms that credit card issuers may use to more quickly, cheaply, and accurately identify fraudulent transactions.

Authors-Dr.V. Thiagarasu & M. Sathyapriya (2019)

This proposed system by the authors (credit card expenditure model) performs improved in terms of lowering the false alarms rate, which may be accomplished by studying the association between transactions noticeable by real fraud and transactions suspected of fraud.

Authors- Janaki K, Keerthana S., B.V. Harshitha, Harshitha Y.V, Ramyashree K.(2019) - The proposed method will aid in the detection and prevention of fraudulent transactions and activities, hence lowering the unit of drop in the economic industry.

Author-Kuldeep Randhawa(2018)- The experimental findings of this study demonstrate that the majority voting approach may accurately identify credit card fraud.

Authors-S.Vimala, K.C Sharmili (2017)-In this Study they conclude that using a decision tree with a Hidden Markov model is the finest technique to detect fraud.

Authors- Ong Shu Yee, Saravanan Sagadevan et.al. (2015)- Five Bayesian classifiers are used to test classification metrics. The interpretation was conducted using two types of data-sets: the first was fake data-set that expresses the features of credit card data, and the second was a modified data-set that uses data standardization and Principal component Analysis techniques.

Authors- Suman, Mitali Bansal (2014)- Various strategies can be used to detect fraud. The method uses unsupervised learning approach, with networks that are taught to detect fraud.

IV. PROPOSED SYSTEM AND METHODOLOGY

The key goal of this work is to develop a system for detecting credit card theft in order to inform individuals about this type of fraud.This approach eliminates the opportunity for fraudsters to make numerous transactions on a stolen or forged card before the card holder becomes aware of the illegal behavior. Then, this model is employed to ascertain whether a transaction is fraudulent or not. The main area of this study is to eliminate erroneous fraud classifications while detecting 100% of fraudulent transactions.

DATASET: The dataset used in Kaggle was publicly available credit card transaction-related to the detection of fraud during September 2013 in Europe. This data set contains 2, 84, 807 incidents divided into two categories: fake and legitimate. There are 492 incidents of fraud and 2,84,315 instances of real.

DATA PREPROCESSING: Data preprocessing is a technique that turns provided raw data into a clean data collection.Any time unprocessable raw data is collected, particular procedures are used to reduce the data to a manageable clean data collection.This strategy is used prior to performing iterative analysis. Data Preprocessing is a series of steps that includes Cleaning, Integration, Transformation, and Reduction. Because of the presence of erroneous data, data preprocessing is required (missing data).

Inaccurate Data- There are numerous reasons for missing data, including data that is not continuously collected, a mistake in data entry, technical issues, and so on.

Inconsistent Data- The occurrence of inconsistencies is caused by factors such as data duplication, human data entry, having errors in data or code, and so on.

DIVIDE THE DATASET INTO TEST AND TRAINING DATA: The data-set is separated into two parts: training data and test data. From the given data set 70% is being trained, while the remaining 30% is being tested. Some supervised ML methods are used. Algos used are Logistic Regression, Random Forest, and Decision Trees.

LOGISTIC REGRESSION: For classification tasks, logistic regression is utilized. This approach is straightforward for binary and multivariate classification tasks. Binomial has only two possible values (0 or 1). Multinomial has three different types, none of which are ordered, but Ordinal is in the ordered group (very poor, poor, good, very good)

DECISION TREE:A decision perspective is a visual depiction of the most likely outcome of an option under advantageous conditions.Starting with the root node, the decision view is divided into distinct areas and connected to further nodes. The leaf node is the terminating up node of a decision tree.The leaf node at each node is a class of labels, and the decision view at each node is an experiment linked by a branch that displays its findings. Decision views in a complex topic are usually simplified using this strategic strategy of differentiation and decision-making.

Steps Working of Decision Tree

In the first phase, the dataset is used to work from the root node forward. Find the dataset's top feature in the second phase using attribute selection metrics.The last node is referred to as a leaf node when nodes cannot be categorised. Based on the labels, the root node is then separated into a decision node and a single leaf node.At some point, the node separates into two leaves (accepted and rejected).

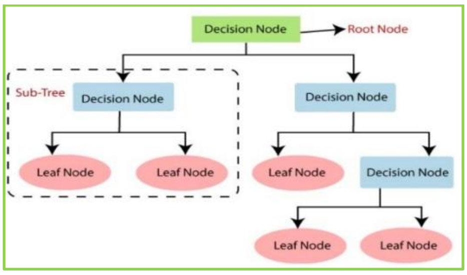

Fig 2: Working process of Decision Tree

RANDOM FOREST

This classifier algorithm generates decision trees in a subset for the given data, gather, condense, and combine their information to decide the predictive outcomes of the entire dataset. As an alternative of depending on a single decision tree, the RF takes each tree's projections and predict the decisive output based on the majority vote of forecasts. By utilizing multiple trees in the forest, it increases accuracy and removes the over-fitting issue. It predicts output with high accuracy and scales well with huge datasets. Additionally, it can continue to be accurate even when a sizable amount of data is lacking.

RF WORKING STEPS

These steps are depicted in Fig 3 below; in the first step, select (K) at random from the drill set as data points. Second, build the DT that is linked to the selected data points (Subsets). Then, enter the digit (N) for the number of decision trees you want to build. Then repeat Steps 1 and 2.Determine the new data points that each decision tree predicted, then allocate the new data points to the group that received the most votes.Usingfollowing scenario, explain how RF works: Assume you have a dataset including photos of several fruits. As a result, this dataset will be passed to the RF classifier.A piece of the dataset is supplied to each decision tree to process. The Random Forest classifier envisages the result based on most outcomes whenever a new data-point arises.

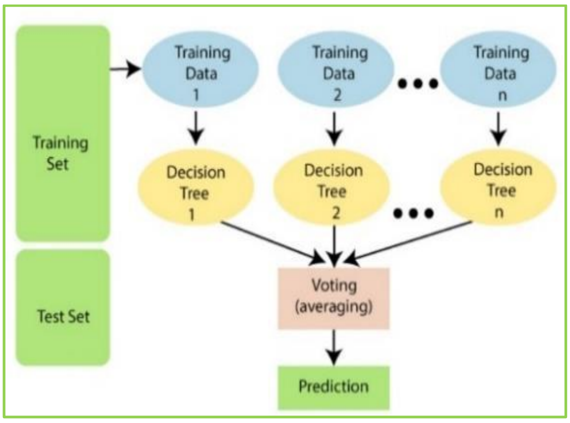

Fig 3: General structure working of the RF

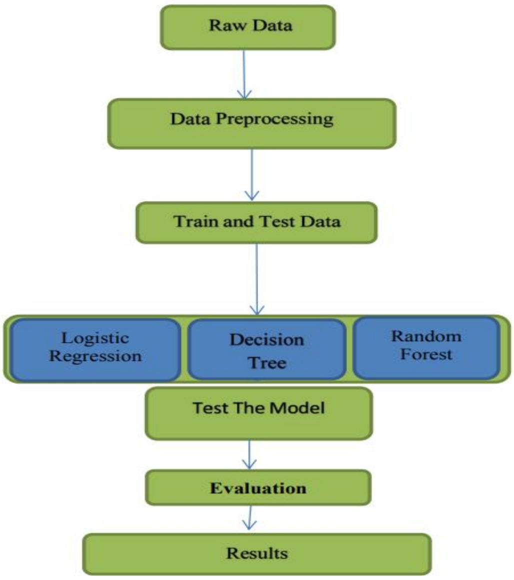

Fig 4: Flowchart of the proposed framework

PROPOSED ALGORITHM

1. Import the necessary items
2. Examine the Dataset.
3. Exploratory data analysis, such as locating duplicate values and null values,
4. Choosing the columns for Features (X) and Target (Y)
5. The dataset will be divided into test and train data.
6. Create the model by training it.
7. Model testing, or model prediction
8. System evaluation, including accuracy score, F1-score, etc.

TABLE I: Advantages And Disadvantages Of Various Ml-Techniques

Techniques	Advantages	Disadvantages
Logistic Regression	Training that is incredibly effective and easier to implement and interpret.	Logistic regression is based on linear decision surface; hence it cannot resolve the non-linear issue.
Decision Trees	It can be quite helpful in finding solutions to decision-making and action issues. a high degree of adaptability that helps in weighing all possible answers to an issue. The necessity for data cleansing is limited.	This method is challenging since it has numerous layers. It might have too many issues for the RF algorithm to be fully understood. The DR's mathematical complexity can rise.
Random Forest	Both organization and relapse tasks can be accomplished by RF. It can handle big data files with expanding dimensions. It encourages model thoroughness and guards against over-fitting issues.	Although RF can be used for both organization and relapse functions, Regression activities may not benefit more from its use.

EVALUATION CRITERIA

There are several parameters, including the accuracy score, classification report, F1-score, confusion matrix, etc., to evaluate the outcomes of the classification algorithms.

1) **Accuracy** – It is the percentage of accurate predictions to all input samples.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

2) **Confusion Matrix** -It is the matrix that is often used to describe how a classification model(a "classifier") performed on a set of test data whose true values were known.

3) **Precision (Specificity)-**It is the ratio of true positive results to those that the classifier expected to be true positive.

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

4) **Recall (Sensitivity**)- It is determined as the sum of all relevant samples divided by the quantity of precisely positive results (all samples that should have been identified as positive).

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

5) **F1 score** - The Harmonic Mean between recall and precision is known as the F1 Score. F1 Score's range is [0, 1].

$$\text{F1} = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

V. RESULTS

Three machine learning algorithms were employed in this study to identify credit card system fraud. 30% of the dataset is used for testing, while the remaining 70% is used to train the algorithms. The performance of these four techniques is assessed using accuracy, F1-score, precision, and recall score.

Table II. Summary of Evaluation criteria

Algorithm	'Accuracy'	'F1-score'	'Precision'	'Recall'
Logistic-Regression	0.99926	0.78351	0.81720	0.75248
Decision Tree	0.99947	0.84536	0.88172	0.81188
Random Forest	0.99958	0.87368	0.93258	0.82178

As displayed in Table II the DT,LR, and RF algos all have excellent accuracy scores. The accuracy is, correspondingly, 0.99926, 0.99947, and 0.99958. However, when we include the other 2 factors, it becomes evident that the Random Forest classifiers outperform all of the aforementioned classifiers and accurately, precisely, and consistently anticipate fraudulent transactions. The best algorithm for detecting credit card fraud, according to our comparison of the three algorithms, is the Random Forest classifier.

CONCLUSION

This study examined credit card system fraud was detected using data mining classification techniques including ML techniques like DT, LR and RF. The proposed system's performance was assessed using accuracy, F1-score, confusion matrix, sensitivity, and specificity. Although the accuracy of each method was excellent, it was discovered that the Random Forest classifier outclassed Logistic Regression and even Decision Tree when other assessment factors were taken into consideration.

REFERENCES

[1]. M Sathyapriya, Dr.V. Thiagarasu, "A Cluster Based Approach for Credit Card Fraud Detection System using HMM with the Implementation of Big Data Technology", International Journal of Applied Engineering Research ISSN Volume- 14, Number 2(2019)

[2]. Aswathy MS, LIJI Sameul, "Survey on Credit Card Fraud Detection", International Research Journal of Engineering and Technology (IRJET), Volume: 05 Issue: 11, Nov 2018.

[3]. Agrawal A.N Dermala N., "Credit card fraud detection using SVM and Reduction of false alarms," International Journal of Innovations in Engineering and Technology (IJIET), vol. 7, no. 2, pp. 176-182, 2016.

[4]. Phua C., Lee V., Smith, Gayler K.R., "A comprehensive survey of data mining-based fraud detection research", arXiv preprint arXiv:1009.6119, 2010.

[5]. Stojanovic A., Aouada D., Ottersten B Bahnsen A.C., "Cost-sensitive credit card fraud detection using Bayes minimum risk," in 12th International Conference on Machine Learning and Applications (ICMLA) 2013.

[6]. Carneiro E.M., Dias L.A.V., Da Cunha A.M., Mialaret L.F.S., "Cluster analysis and artificial neural networks: A case study in credit card fraud detection", in 12th International Conference on Information Technology-New Generations, pp.122- 126, 2015.

[7]. S. Aghili and P. Zavarsky K. T. Hafiz, "The use of predictive analytics technology to detect credit card fraud in Canada," in 11th Iberian Conference on Information Systems and Technologies (CISTI), pp. 1-6, 2016

[8]. Bansal M Sonepat H.C.E., "Survey Paper on Credit Card Fraud Detection," International Journal of Advanced Research in Computer Engineering & Technology, vol. 3, no. 3, pp. 827-832, 2014.

[9]. S., Tuyls, K., Vanschoenwinkel, B. and Manderick, "Credit card fraud detection using Bayesian and neural networks," in Proceedings of the 1st international naiso congress on neuro-fuzzy technologies, pp. 261-270, 2002.

[10]. Kuldeep Randhawa, Chu Kiong Loo, Manjeevan Seera, Chee Peng Lim and Asoke K. Nandi, "Credit card fraud detection using AdaBoost and majority voting," IEEE Access, vol. 6, pp. 14277-14284, 2018.

[11]. A. Roy and J. Sun and R. Mahoney and L. Alonzi and S. Adams and P. Beling, "Deep learning detecting fraud in credit card transactions," in Systems and Information Engineering Design Symposium (SIEDS), pp. 129-134, 2018.

[12]. Guanjun Liu, Zhenchuan Li, Lutao Zheng, Shuo Wang and Changjun Jiang Shiyang Xuan, "Random Forest for Credit Card Fraud Detection," in IEEE 15th International Conference On Networking, Sensing and Control (ICNSC), pp.1-6, 2018.

[13]. Zarrabi, H. Kazemi, "Using deep networks for fraud detection in the credit card transaction," IEEE 4th International Conference In Knowledge-Based Engineering and Innovation (KBEI), pp. 0630-0633, 2017.

[14]. John O., Adebayo O. Adetunmbi, and Samuel A. Oluwadaren Awoyemi, "Credit card fraud detection using machine learning techniques: A comparative analysis."International Conference on Computing Networking and Informatics (ICCNI), pp. 1-9, 2017.

[15]. S. Dutta, A. K. Gupta and N. Narayan, "Identity Crime Detection Using Data Mining, "3rd International Conference on Computational Intelligence and Networks (CINE), Odisha, pp. 1-5, 2017.

[16]. K. Modi and R. Dayma, "Review on fraud detection methods in credit card transactions, "International Conference on Intelligent Computing and Control (I2C2), Coimbatore, pp. 1-5, 2017.87

[17]. D. Pojee, S. Zulphekari, F. Rarh, and V. Shah, "Secure and quick NFC payment with data mining and intelligent fraud detection, "2nd International Conference on Communication and Electronics Systems (ICCES), Coimbatore, pp. 148-152, 2017.

[18]. D. S. Sisodia, N. K. Reddy and S. Bhandari, "Performance evaluation of class balancing techniques for credit card fraud detection," IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI), Chennai, pp. 2747-2752, 2017.

[19]. L. Vergara, A. Salazar, J. Belda, G. Safont, S. Moral and S. Iglesias, "Signal processing on graphs for improving automatic credit card fraud detection," International Carnahan Conference on Security Technology (ICCST), Madrid, pp. 1- 6, 2017.

[20]. Credit card fraud detection: A realistic modeling and a novel learning strategy. IEEE Transactions on Neural Networks and Learning Systems, 29(8):3784–3797, August 2018.

[21]. J. O. Awoyemi, A. O. Adetunmbi, and S. A. Oluwadare. Credit card fraud detection using machine learning techniques: A comparative analysis. In 2017 International Conference on Computing Networking and Informatics (ICCNI), pages 1–9, 2017.

[22]. M. Azhan, M. Ahmad, and M. S. Jafri. Metoo: Sentiment analysis using neural networks (grand challenge). In 2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM), pages 476–480, 2020.