

Class imbalance aware drift identification model for detecting diverse attack in streaming environment

Arati Shahpurkar, Rudragoud Patil, Kiran K. Tangod

Department of Computer Science and Engineering, K. L. S. Gogte Institute of Technology,
Affiliated to Visveswarya Technological University, Belagavi, India

Article Info

Article history:

Received Jul 14, 2023

Revised Oct 15, 2023

Accepted Nov 23, 2023

Keywords:

Class imbalance

Cross validation

Drift identification

Fraudulent transaction

NADS-RA

Xtreme gradient boosting

ABSTRACT

Detecting fraudulent transactions in a streaming environment presents several challenges including the large volume of data, the need for real-time detection, and the potential for data drift. To address these challenges a robust model is needed that utilizes machine learning techniques to classify transactions in real-time. Hence, this paper proposes a model for detecting fraudulent transactions in a streaming environment using xtream gradient boost (XGBoost), cross-validation and class imbalance aware drift identification (CIADI) model. The performance of the proposed method is evaluated using datasets named credit card and Network Security Laboratory (NSL-KDD) dataset. The results demonstrate that the model can effectively detect fraudulent transactions with high accuracy, recall, and F-measure. The results show that the proposed CIADI model attained 95.63% for the credit card dataset which is higher accuracy in comparison to the generative-adversarial networks (GAN), network-anomaly-detection scheme-based on feature-representation and data-augmentation (NADS-RA) and feature-aware XGBoost (FA-XGB). Further the proposed CIADI model attained 98.5% for the NSL-KDD dataset which is higher accuracy in comparison to the NADS-RA, stacked-nonsymmetric deep-autoencoder (sNDAE) and convolutional neural-network (CNN). This study suggests that the proposed method can be an effective model for detecting fraudulent transactions in streaming environments.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Arati Shahpurkar

Department of Computer Science and Engineering, K. L. S. Gogte Institute of Technology

Affiliated to Visveswarya Technological University, Belagavi, Karnataka, India

Email: asshahpurkar@git.edu

1. INTRODUCTION

With the increasing use of digital transactions, the risk of fraudulent activities has also grown significantly. Fraudulent transactions not only cause financial loss but also damage the reputation and trust of businesses and financial institutions [1]. To address this issue, there is a need for effective and efficient fraud detection systems that can quickly identify and prevent fraudulent activities. Traditional approaches to fraud detection often involve analyzing historical data in batch mode, which can be time-consuming and may not be able to detect new types of fraud in real-time [2]. In recent years, there has been a growing interest in detecting fraudulent transactions in real-time through the use of streaming data analytics [3]. A streaming environment allows for continuous processing of data as it is generated, which enables quick detection and response to fraudulent activities. However, detecting fraud in a streaming environment poses unique challenges, such as dealing with high data volume and velocity, identifying changing patterns of fraud over time, and handling concept drift [4]–[6].

A fraudulent transaction typically involves a sequence of events, which can vary depending on the specific type of fraud being perpetrated [7]. However, there are some common patterns that can be observed in many fraudulent transactions. The first step in a fraudulent transaction is often the identification of a potential victim, such as a business or an individual with valuable assets or information. The fraudster may gather information about the victim through various means, such as social engineering [8] or hacking [9]. Once a victim has been identified, the fraudster typically initiates a contact with them through a variety of means, such as email [10], phone [11], or social media [12]. The next step involves the fraudster tricking the victim into providing sensitive information or performing an action that benefits the fraudster, such as transferring funds or installing malicious content [13]. This can be done through various means, such as phishing emails [14] or fake websites [15]. After the fraudulent transaction is initiated, the fraudster often attempts to cover their tracks and avoid detection. Hence, detecting and preventing fraudulent transactions often requires a combination of technical measures, such as fraud detection systems and security protocols, as well as awareness and education of individuals and organizations about the risks and warning signs of fraud [16]. An example has been given in Figure 1. The Figure 1 illustrates a common sequence of events in a fraudulent transaction, highlighting the various techniques and tactics used by cybercriminals to infiltrate and deceive users. Hence, in this work, to address these challenges, a robust approach is needed that utilizes machine learning techniques, to classify transactions in real-time. Hence, in this paper, we propose a method for detecting fraudulent transactions in a streaming environment using xtreme gradient boosting (XGBoost), cross-validation and class imbalance aware drift identification (CIADI) model.

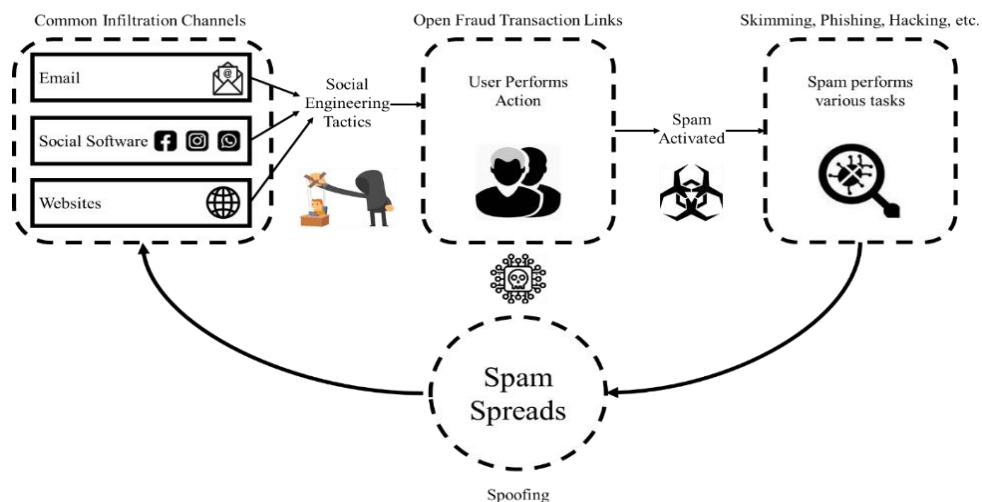


Figure 1. Common sequence of events in a fraudulent transaction

2. LITERATURE SURVEY

Zhou *et al.* [17], explain that datasets with high dimensions make intrusion detection system (IDS) classification difficult because of the existence of a great deal of unrelated information. They have also pointed towards the fact that several assaults are capable of being recognized with just one classification algorithm, as well as that most current algorithms were developed using out-of-date information, limiting their adaptability to evolving threats. Correlation feature selection bat algorithm (CFS-BA) is a technique for IDS that makes use of feature-selection (FS) and ensemble approaches as a solution to the issues mentioned above. The given technique selects the optimal subset by considering the relationship between the characteristics introduced in the first phase of the reduction of dimensionality. Following this, they introduced an ensemble approach that takes into account the strengths of the random forest (RF), the C4.5, as well as the forest by penalizing-attributes (F-PA). Finally, this technique uses a voting method for integrating the fundamental learner's distribution of probabilities for attack identification. Yotsawat *et al.* [18] explained, almost all earlier efforts on scoring credit employed the concept of an ensemble classification to deal with data imbalances. In this approach, resampling methods were utilized to generate a large number of training subgroups from which several base classifiers could be constructed. Nevertheless, this approach has several drawbacks that diminish its categorization effectiveness. These include model-overfitting, loss of data, and high computational costs. In order to develop a cost-sensitive-neural network-based scoring credit approach, they present a new ensemble technique called cost-sensitive-neural-network-ensemble (CS-NNE).

The proposed approach allows multi-base-neural-networks to accommodate imbalanced classes by making adjustments regarding the weighting of various classes based on the initial training information. This technique allows for the problem-free generation of a large number of distinct baseline classifiers. For the evaluation of this method, they have used credit card dataset. The results show that the proposed method addresses the class imbalance and drift issues efficiently. Fiore *et al.* [19], proposed a generative adversarial network (GAN) for detecting fraud in the credit cards. They have used the GAN as these networks provide flexibility, learnability and this network is a powerful deep-learning method. The GAN model addresses the class imbalance problem. The results show that the proposed model has attained higher accuracy in comparison to the existing models for the credit card dataset.

Liu *et al.* [20], they have proposed network anomaly detection scheme representation and augmentation (NADS-RA) model for detecting network anomalies by utilizing the feature representing method and data augmentation method in the dataset. In this model, they have presented a Re-circulation pixel-permutation technique for attaining all the features. Further, they have presented an image-based augmentation technique which will generate the data augmentation of the input. They have addressed the class imbalance issue in this work. They have used credit card and Network Security Laboratory (NSL-KDD) dataset for evaluating their model and comparing it with other methods. The results show that the proposed NADS-RA model has attained the highest performance when compared with the existing models. Shahapurkar and Rodd [21], proposed model called feature aware-XGBoost (FA-XGB) which detects the fraudulent transaction. In this work they have proposed a machine learning model, XGBoost, which addresses the class imbalance issues. In this work they have used the cross-validation technique for constructing their predictive model. In this work, they have used the Twitter spam dataset for evaluating their model and comparing it with the existing models. The results show that by addressing the class imbalance issue in the dataset, the performance of the model can be increased. Ni *et al.* [22], proposed a spiral-oversampling-balancing technique (SOBT) for the detection of the fraud in the credit cards. They have used a feature boosting method in this work. Further, in this work they have also modelled a multi-factor synchronous-embedding technique (MSET) which enhances the selection of the features as well as for the improvement of the decision-making capability. The results show that the proposed model has attained better performance in comparison to the existing methods.

3. PROPOSED MODEL

3.1. Architecture

The proposed CIADI methodology's architecture is described in this section. The suggested architecture is depicted in Figure 2. There are five steps to the proposed CIADI procedure. First input is taken into account, then data is preprocessed and cleaned. After this initial step the data is split in half for technique learning and the other half for validation. Learned classes can be either binary or multi-class based on the input data. Binary and multi-class classifications are used iteratively to build the CIADI method to attack detection. The second step involves optimizing the test set. Once optimization is complete, the procedure moves on to the next step. The third step is to evaluate the CIADI classifier's accuracy using the modified cross-validation model given below in light of the generated classifiers and the testing sets. Adjustments to the CIADI models parameters can be made using the suggested optimization framework if performance drops below expectations. To succeed it must realize its full potential. In the final step, we evaluate the CIADI classifier's accuracy with respect to the threshold (drift indicator) and determine the optimal value for the classifier's hyperparameter. If the CIADI classifier's accuracy is below the set threshold the algorithm will iterate back to step two and update the classifier. If a classifier has greater accuracy its effectiveness can be evaluated by how well it works.

3.2. XGBoost

The XGBoost algorithm has been introduced in this section. The XGBoost model is a means of improving upon a decision-tree (DT) model by combining boosting with a gradient-descent method [23]. When training a new data set the XGBoost model is periodically refreshed with the results of previous DT iterations. The XGBoost model's minimum objective function value can be calculated using (1).

$$\text{Obj} = -\frac{1}{2} \sum_{j=1}^t \frac{G_j^2}{H_j + \lambda} + \gamma t \quad (1)$$

Where, G and H represent the total cost function of the first and second-order gradients. t represents the leaves present in the DT. λ and γ are used for denoting the penalty coefficients. Since the XGBoost model is built from the DT iteration results accumulating all of the DT is necessary for precision.

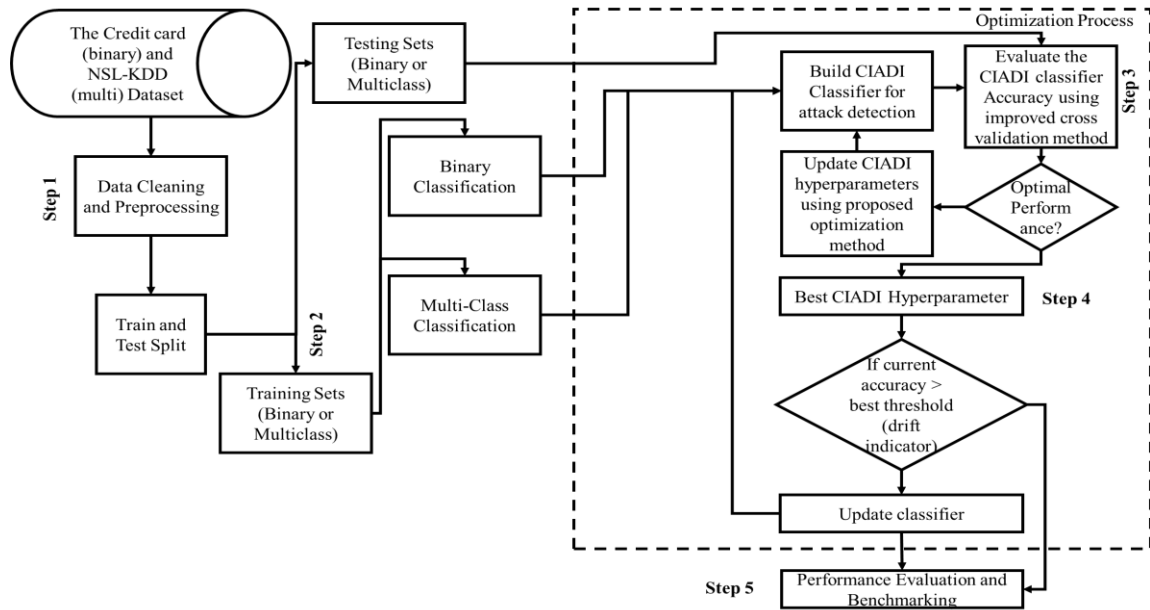


Figure 2. Architecture of the proposed CIADI model

3.3. Cross validation

In this section, the cross-validation (CV) model has been presented. To optimize the feature-imbalance in the dataset, the CV is utilized. The CV model is constructed by utilizing the different groups of K – folds. For the evaluation of an individual K – fold which has CV, the (2) is utilized.

$$CV(\sigma) = \frac{1}{M} \sum_{k=1}^K \sum_{j \in G_{-k}} P(b_j, \hat{g}_{\sigma}^{-k(j)}(y_j, \sigma)) \tag{2}$$

By evaluating of an individual K – fold which has CV, the model fails to attain higher accuracy as the dataset may have imbalanced values. To decrease the error during the CV [21], has presented a CV which has two layers. These two layers consist of important features attained from the dataset as well as the important features which have been chosen by utilizing the previous layer. Both the layers are utilized for the construction of the predictive model. The two-layer CV is done using (3).

$$CV(\sigma) = \frac{1}{SM} \sum_{s=1}^S \sum_{k=1}^K \sum_{j \in G_{-k}} P(b_j, \hat{g}_{\sigma}^{-k(j)}(y_j, \sigma)) \tag{3}$$

Using the (3), to optimize the parameters of the CV and to select the best value for the parameters of CV, the (4) is used.

$$\hat{\sigma} = \underset{\sigma \in \{\sigma_1, \dots, \sigma_l\}}{\text{arg min}} CV_s(\sigma) \tag{4}$$

In (3), the parameter for the gradient loss is represented using P(·). The size of training is represented using M. $\hat{g}_{\sigma}^{-k(j)}(\cdot)$ is utilized for the evaluation of the coefficients. By utilizing the standard XGBoost model and CV model, one can attain better results, but the drift issues won't be addressed. Hence in this work we address the drift issues by proposing the CIADI model. The CIADI model has been presented in the next section.

3.4. CIADI model

An online model or algorithm designed to handle concept drift must be able to quickly adjust to any novel concept whenever changes are introduced into the data stream while maintaining stability under constant conditions. Consider a situation where a given classifier in the given ensemble model fails to correctly classify the sample from the incoming streaming data, then its weight is reduced using the variable β for all the p timestamps. Hence, to address the issue of concept drift in the streamed data, the CIADI model

generates and removes the classifier periodically. In a scenario where the value (weight) of the classifier is below a certain level, then the CIADI model removes the classifier. Moreover, the proposed CIADI model generates a classifier whenever the ensemble classifier fails to classify the sample from the incoming streaming data. Hence, the successful performance of CIADI as an ensemble technique can be attributed towards the dynamic modulation for assigning the classifier values (weights). The strategy for adjusting the values (weights) takes into account two factors in particular. The first factor is the time factor. Whenever a classifier which is very far from the given timestamp, that classifier is assigned less value (weight). The second factor is the concept-drift factor. The classifier which is constructed by using the outdated concept (old classifiers) will be given less value (weight) whenever the concept-drift occurs, with the rate of value (weight) loss increasing over time. An identical concept is built within the Lfun method [24], wherein the time-factor is combined alongside the error to determine the importance for every classifier. In Algorithm 1, the pseudocode for the CIADI model has been given. The proposed algorithm addresses the drift issue in the dataset. The proposed algorithm has attained better performance which has been discussed in the next section.

Algorithm 1. Pseudocode for the CIADI model

```

Input  Data stream  $t: \mathcal{D} = \{x_i \in \mathcal{X}, y_i \in \mathcal{Y}\}, i = 1, \dots, N$ , number of classes  $c$ , the threshold for deleting
       individual classifiers  $\theta$ , factor for decreasing weights  $\beta$ , period between
       classifier removal, creation, and weight update  $p$ .

1:   $m \leftarrow 1$ ;
2:   $w_m \leftarrow 1$ ;
3:   $H_m \leftarrow \text{CreateClassifier}$ ;
4:   $\mathcal{H} \leftarrow \{H_m\}$ 
5:  for  $i \leftarrow 1$  to  $N$  do
6:      for  $j \leftarrow 1$  to  $m$  do
7:          if  $H_j(x_i) \neq y_i$  and  $i \bmod p = 0$  then
8:               $w_j \leftarrow \beta w_j$ ;
9:          end if
10:     end for
11:     Predict  $x_i$  by the ensemble classifier:
        $\bar{y}_i \leftarrow \text{sign}(\sum_{j=1}^m w_j H_j(x_i))$ ;
12:     if  $i \bmod p = 0$  then
13:         Normalize classifier weight:
                $w \leftarrow w / \sum_j w_j$ ;
14:         Remove classifiers with a weight less than  $\theta$ :
                $\mathcal{H} \leftarrow \mathcal{H} \setminus \{H_j | w_j < \theta\}$ ;
15:         if  $\bar{y}_i \neq y_i$  then
16:              $m \leftarrow m + 1$ ;
17:              $H_m \leftarrow \text{CreateClassifier}$ ;
18:              $\mathcal{H} \leftarrow \mathcal{H} \cup H_m$ ;
19:              $w_m \leftarrow 1$ ;
20:         end if
21:     end if
22:     for  $j \leftarrow 1$  to  $m$  do
23:          $H_j \leftarrow \text{UpdateClassifier}(H_j, x_i, y_i)$ ;
24:     end for
25: end for
Output Ensemble Classifier set  $\mathcal{H}$ , the weight of individual classifiers  $w$ .

```

4. RESULTS AND DISCUSSIONS

In this section, the results for the proposed CIADI model have been evaluated by evaluating it with the benchmarked credit card dataset and NSL-KDD dataset. Further, the proposed CIADI model has been compared with the existing works. The configuration used for the experimentation is as follows: Windows 10, 16GB RAM, 250 GB ROM, Intel i7 processor. For the coding, Python has been used and the Anaconda 3 has been used for running the code. For evaluating the proposed model and comparing it with the existing models, the performance evaluation metrics has been given in the next section. The evaluation has been done using credit card dataset [25] and NSL-KDD dataset [26].

4.1. Performance evaluation metrics

In this section, the focus is on presenting the performance metrics employed for assessing the CIADI model, alongside comparisons with other existing models. The evaluation criteria encompass metrics such as accuracy, recall, and F-measure, which are quantified through the utilization of (5) to (7)

respectively. These metrics serve as essential benchmarks for gauging the effectiveness and efficiency of the proposed CIADI model in relation to its counterparts in the study.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \tag{5}$$

$$\text{Recall} = \frac{TP}{TP+FN} \tag{6}$$

$$F - \text{measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

4.2. Credit card dataset

The proposed CIADI model is contrasted with GAN [19], NADS-RA [20] and FA-XGB [21]. The results for accuracy, recall and F-measure have been given in Figures 3 and 4 respectively. The proposed CIADI model performs better than GAN, NADS-RA, and FA-XGB by a margin of 6.317%, 16.2558%, and 0.86764%, respectively for accuracy performance. To improve accuracy the suggested CIADI model acknowledges the dataset's class imbalance. Also, the results demonstrate that the GAN approach has achieved a relatively low level of recall when contrasted with the remaining models. The NADS-RA performs better in comparison to the GAN model but does not attain the best recall in comparison to the FA-XGB and CIADI. The FA-XGB attains better recall in comparison to the GAN and NADS-RA model. The CIADI model and FA-XGB have attained similar recall. The proposed CIADI method is contrasted with GAN [19], NADS-RA [20] and FA-XGB [21]. Furthermore, according to the findings the F-measure achieved by the GAN approach is the lowest of all the models tested. The NADS-RA performs better in comparison to the GAN model but does not attain the best F-measure in comparison to the FA-XGB and CIADI. The FA-XGB attains better recall in comparison to the GAN and NADS-RA model. When compared to other methods the suggested CIADI method achieves an improved F-measure.

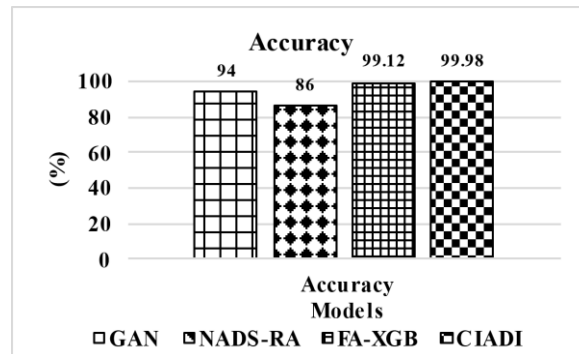


Figure 3. Accuracy for credit card dataset

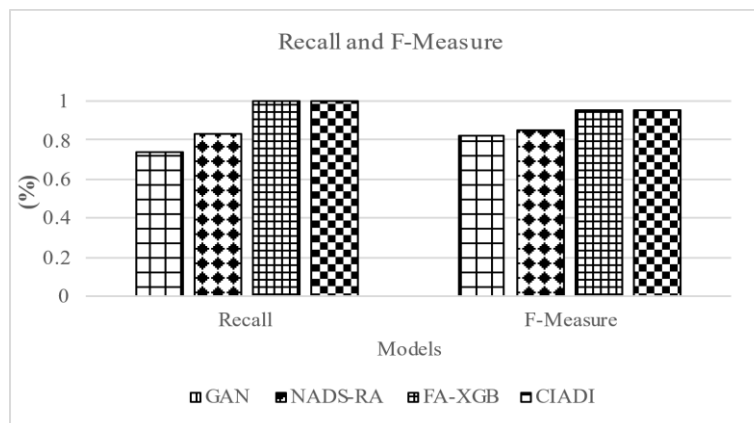


Figure 4. Recall and F-measure for credit card dataset

4.3. NSL-KDD dataset

NADS-RA [20], sNDAE [20], and CNN [20] have all been evaluated to the suggested CIADI model. The results for accuracy, recall and F-measure have been given in Figures 5 and 6 respectively. The outcomes demonstrate that the suggested CIADI model performed better than NADS-RA, sNDAE, and CNN by 2.3908%, 15.3396%, and 8.1229%, for accuracy performance respectively. The proposed CIADI model is contrasted with NADS-RA [20], sNDAE [20] and CNN [20]. From findings, it appears to indicate that NADS-RA has higher recall than the sNDAE as well as the CNN model, but lower recall than the suggested CIADI model. While the sNDAE outperforms the CNN in the recall, it lags below the NADS-RA as well as the CIADI model in these respects. The CNN fails to attain better recall in comparison to all the existing models and proposed model. The proposed CIADI model attains the highest recall in comparison to all the existing models. Furthermore, the NADS-RA offers a greater F-measure than the CNN model but a lower F-measure than the sNDAE and CIADI model. The sNDAE model shows better F-measure in comparison to the NADS-RA and CNN model but fails to attain higher F-measure in comparison to the proposed CIADI model. The CNN model attains the least F-measure when compared with NADS-RA, sNDAE, and proposed CIADI. The proposed CIADI model has attained the highest F-measure.

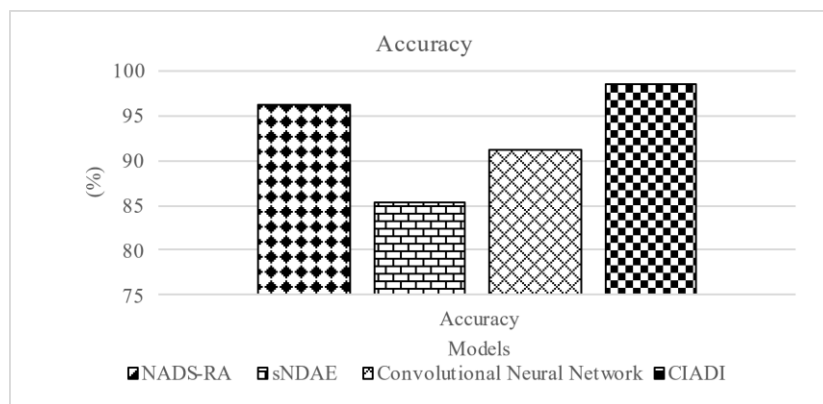


Figure 5. Accuracy for NSL-KDD dataset

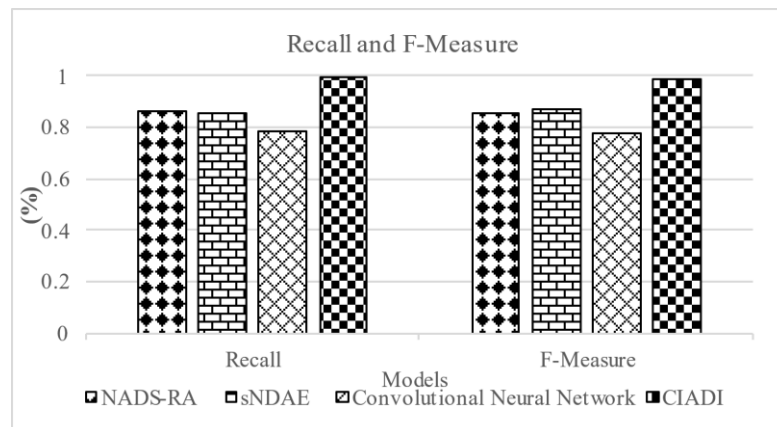


Figure 6. Recall and F-measure for NSL-KDD dataset

4.4. Discussion

Here we discuss the degree of accuracy achieved using the credit card dataset as well as the NSL-KDD dataset. Table 1 displays the results of the comparison analysis. Given the outcomes we can deduce that just NADS-RA has employed both datasets in the assessment of their approach. The majority of existing models either use the NSL-KDD dataset or the credit card dataset. The suggested method has been tested as well as the results compared against the NSL-KDD as well as credit card datasets. As can be seen from the findings the suggested CIADI method performs better than all other methods on the two datasets.

Table 1. Comparitive study

Models	Accuracy attained for credit card dataset	Accuracy attained for the NSL-KDD dataset
GAN [19]	94%	-
NADS-RA [20]	86%	96.2%
sNDAE [20]	-	85.4%
CNN [20]	-	91.1%
CIADI (Proposed)	95.63%	98.5%

5. CONCLUSION

Detecting fraudulent transactions in a streaming environment presents several challenges, including the large volume of data, the need for real-time detection, and the potential for data drift. To address these challenges a robust approach is needed that utilizes machine learning techniques to classify transactions in real-time. Hence in this work a model has been presented to address these challenges. In this work first the XGBoost model has been presented which further utilizes the cross-validation model to attain higher accuracy. Further for addressing the drift in the given sample an algorithm called CIADI has been presented. For evaluating the proposed CIADI model two standard datasets credit card and NSL-KDD dataset have been used. The performance metrics accuracy, recall and f-measure have been used for the evaluation. The results show that the proposed CIADI model attained 95.63% for the credit card dataset which is higher accuracy in comparison to the existing models. Further the proposed CIADI model attained 98.5% for the NSL-KDD dataset which is higher accuracy in comparison to the existing models. This work suggests that the proposed model can be an effective for detecting fraudulent transactions in streaming environments. For the future work the proposed CIADI model can be used to evaluate other datasets such as KDD99, DARPA1998, and UNSW-NB15.




REFERENCES

- [1] H. van Driel, "Financial fraud, scandals, and regulation: a conceptual framework and literature review," *Business History*, vol. 61, no. 8, pp. 1259–1299, Oct. 2019, doi: 10.1080/00076791.2018.1519026.
- [2] M. Aschi, S. Bonura, N. Masi, D. Messina, and D. Profeta, "Cybersecurity and fraud detection in financial transactions," in *Big Data and Artificial Intelligence in Digital Finance*, Springer International Publishing, 2022, pp. 269–278.
- [3] R. Bin Sulaiman, V. Schetinin, and P. Sant, "Review of machine learning approach on credit card fraud detection," *Human-Centric Intelligent Systems*, vol. 2, no. 1–2, pp. 55–68, May 2022, doi: 10.1007/s44230-022-00004-0.
- [4] M. Karimian and H. Beigy, "Concept drift handling: A domain adaptation perspective," *Expert Systems with Applications*, vol. 224, p. 119946, Aug. 2023, doi: 10.1016/j.eswa.2023.119946.
- [5] H. Mehmood, P. Kostakos, M. Cortes, T. Anagnostopoulos, S. Pirttikangas, and E. Gilman, "Concept drift adaptation techniques in distributed environment for real-world data streams," *Smart Cities*, vol. 4, no. 1, pp. 349–371, Mar. 2021, doi: 10.3390/smartcities4010021.
- [6] S. Thudumu, P. Branch, J. Jin, and J. Singh, "A comprehensive survey of anomaly detection techniques for high dimensional big data," *Journal of Big Data*, vol. 7, no. 42, Jul. 2020, doi: 10.1186/s40537-020-00320-x.
- [7] W. Hilal, S. A. Gadsden, and J. Yawney, "Financial fraud: a review of anomaly detection techniques and recent advances," *Expert Systems with Applications*, vol. 193, p. 116429, May 2022, doi: 10.1016/j.eswa.2021.116429.
- [8] F. Salahdine and N. Kaabouch, "Social engineering attacks: A survey," *Future Internet*, vol. 11, no. 4, p. 89, Apr. 2019, doi: 10.3390/FII11040089.
- [9] A. Mishra and C. Ghorpade, "Credit card fraud detection on the skewed data using various classification and ensemble techniques," in *2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science, SCEECS 2018*, Feb. 2018, pp. 1–5, doi: 10.1109/SCEECS.2018.8546939.
- [10] R. Valecha, P. Mandaokar, and H. Raghav Rao, "Phishing email detection using persuasion cues," *IEEE Transactions on Dependable and Secure Computing*, vol. 19, no. 2, pp. 747–756, 2022, doi: 10.1109/TDSC.2021.3118931.
- [11] S. E. Ayeb, B. Hemery, F. Jeanne, and E. Cherrier, "Community detection for mobile money fraud detection," in *2020 7th International Conference on Social Network Analysis, Management and Security, SNAMS 2020*, Dec. 2020, pp. 1–6, doi: 10.1109/SNAMS52053.2020.9336578.
- [12] F. Khan, R. Alturki, G. Srivastava, F. Gazzawe, S. T. U. Shah, and S. Mastorakis, "Explainable detection of fake news on social media using pyramidal co-attention network," *IEEE Transactions on Computational Social Systems*, pp. 1–10, 2022, doi: 10.1109/tcss.2022.3207993.
- [13] D. Myalil, M. A. Rajan, M. Apte, and S. Lodha, "Robust collaborative fraudulent transaction detection using federated learning," in *Proceedings - 20th IEEE International Conference on Machine Learning and Applications, ICMLA 2021*, Dec. 2021, pp. 373–378, doi: 10.1109/ICMLA52953.2021.00064.
- [14] E. D. Frauenstein and S. Flowerday, "Susceptibility to phishing on social network sites: A personality information processing model," *Computers and Security*, vol. 94, p. 101862, Jul. 2020, doi: 10.1016/j.cose.2020.101862.
- [15] S. Kodate, R. Chiba, S. Kimura, and N. Masuda, "Detecting problematic transactions in a consumer-to-consumer e-commerce network," *Applied Network Science*, vol. 5, no. 90, Nov. 2020, doi: 10.1007/s41109-020-00330-x.
- [16] G. J. Priya and S. Saradha, "Fraud detection and prevention using machine learning algorithms: A review," in *Proceedings of the 7th International Conference on Electrical Energy Systems, ICEES 2021*, Feb. 2021, pp. 564–568, doi: 10.1109/ICEES51510.2021.9383631.
- [17] Y. Zhou, G. Cheng, S. Jiang, and M. Dai, "Building an efficient intrusion detection system based on feature selection and ensemble classifier," *Computer Networks*, vol. 174, p. 107247, Jun. 2020, doi: 10.1016/j.comnet.2020.107247.




- [18] W. Yotsawat, P. Wattuya, and A. Srivihok, "A novel method for credit scoring based on cost-sensitive neural network ensemble," *IEEE Access*, vol. 9, pp. 78521–78537, 2021, doi: 10.1109/ACCESS.2021.3083490.
- [19] U. Fiore, A. De Santis, F. Perla, P. Zanetti, and F. Palmieri, "Using generative adversarial networks for improving classification effectiveness in credit card fraud detection," *Information Sciences*, vol. 479, pp. 448–455, Apr. 2019, doi: 10.1016/j.ins.2017.12.030.
- [20] X. Liu *et al.*, "NADS-RA: network anomaly detection scheme based on feature representation and data augmentation," *IEEE Access*, vol. 8, pp. 214781–214800, 2020, doi: 10.1109/ACCESS.2020.3040510.
- [21] A. Shahapurkar and S. F. Rodd, "Efficient feature aware machine learning model for detecting fraudulent transaction in streaming environment," *International Journal on Information Technologies and Security*, vol. 14, no. 3, pp. 3–14, 2022.
- [22] L. Ni, J. Li, H. Xu, X. Wang, and J. Zhang, "Fraud feature boosting mechanism and spiral oversampling balancing technique for credit card fraud detection," *IEEE Transactions on Computational Social Systems*, pp. 1–16, 2023, doi: 10.1109/TCSS.2023.3242149.
- [23] K. Chandrashekar and A. T. Narayanreddy, "An ensemble feature optimization for an effective heart disease prediction model," *International Journal of Intelligent Engineering and Systems*, vol. 16, no. 2, pp. 517–525, Feb. 2023, doi: 10.22266/ijies2023.0430.42.
- [24] C. Chen, Y. Wang, J. Zhang, Y. Xiang, W. Zhou, and G. Min, "Statistical features-based real-time detection of drifted Twitter spam," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 4, pp. 914–925, Apr. 2017, doi: 10.1109/TIFS.2016.2621888.
- [25] "Credit card fraud detection," *Machine Learning Group - ULB*, 2018. <https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud> (accessed Apr. 25, 2023).
- [26] "NSL-KDD dataset," *UNB, University of New Brunswick*, 1999. <https://www.unb.ca/cic/datasets/nsl.html> (accessed Apr. 25, 2023).

BIOGRAPHIES OF AUTHORS






Arati Shahapurkar    currently working as an Assistant Professor, at Department of Computer Science Engineering, Karnatak Law Society's (KLS) Gogte Institute of Technology, Belagavi. She has 16 years of Teaching Experience at KLS Gogte Institute of Technology, Karnataka. She has published over 10 papers in international journals, and conferences of high repute. Her subjects of interest include machine learning, big data management, and network security. She can be contacted at email: asshahapurkar@git.edu.



Dr. Rudragoud Patil    currently working as an Associate Professor, at Department of CSE, KLS Gogte Institute of Technology, Belagavi. He has 12 years of Teaching Experience at Professional Institutes across Karnataka. He published over 13 papers in international journals, book chapters, and conferences of high repute. His subjects of interest include cloud computing, distributed computing, machine learning, and network security. He can be contacted at email: rspatil@git.edu.



Dr. Kiran K. Tangod    presently holding the position of Professor and Department Head in Information Science and Engineering at KLS Gogte Institute of Technology in Belagavi, Dr. Kiran K. Tangod boasts an impressive teaching tenure of 22 years within various esteemed educational institutions throughout Karnataka. His scholarly contributions encompass more than 13 publications in distinguished international journals, book chapters, and esteemed conferences. His academic pursuits are particularly focused on the realm of data mining. He can be contacted at email: kirankt@git.edu.