EFFICIENT FEATURE AWARE MACHINE LEARNING MODEL FOR DETECTING FRAUDULENT TRANSACTION IN STREAMING ENVIRONMENT

Arati Shahapurkar, Dr. Sunil F. Rodd*

[1] Dept of CSE, KLS GOGTE Institute of Technology, Belagavi
India

* Corresponding Author: gitarti20@gmail.com

Abstract: The emergence of the internet and social streaming environment makes users circulate private multimedia data with similar people across global. Taking benefits from these applications; it becomes much easier for the spammer to attack digitally. As a result, an effective intrusion detection system is required. In this paper, an efficient feature extraction and selection method addressing class imbalance problems for detecting fraudulent links in a streaming environment is presented. Here an improved feature-aware machine learning-based classification algorithm for detecting fraudulent transactions in a streaming environment is presented. The results are compared over the existing supervised classification methodologies for validating the proposed methodology using standard datasets with serious class imbalance issues.

Keywords: Class Imbalance, Concept Drift, Deep Learning, Ensemble Learning, Intrusion Detection System, Machine Learning, Social Network.

1. INTRODUCTION

With the growth of technology, the contemporary attack aims to impact availability, integrity, and confidentiality continuously. As a result, the prerequisite to designing an effective intrusion detection system for preventing a wide range of attacks. Generally, the intrusion detection system is classified into two classes such as misuse and anomaly detection. The anomaly detection mechanism is generally used for detecting malicious links by establishing deviation in normal pattern; thus, prerequisite minimal false-positive rate. On contrary, misuse detection is used for differentiating between normal conditions with respect to malicious conditions. They are effective in identifying known attacks; however, perform miserably bad for detecting unknown attacks. Thus, it is important to design intrusion detection mechanisms with a high detection rate. Recently, machine learning (ML) has been employed for detecting misuse and anomaly detections. The job of the ML-based intrusion detection model must not only detect whether a transaction link is malicious or not but at the same time must tell what type of attack impacting the system. Nonetheless, in the real-time streaming environment, only a few transaction links are malicious and the rest of the transaction links are normally leading to class imbalance problems. As a result, it makes it difficult to aching a higher detection rate and keep false positives minimal. It showed building an ensemble classifier by combining multiple classifiers for building a predictive model achieves better performance. Additionally, network traffic dimensions and various attack type makes additional challenges such as high time and computational complexities. In addressing this, an effective feature selection strategy plays a very important role in building an effective intrusion detection system, by selecting semantic features with high importance and eliminating unnecessary features with minimal importance. It addressed both feature reduction and class imbalance issues together. However, the cross-validation scheme adopted fails to obtain a good number of useful features; thus, affecting the overall performance. In addressing the research problem this paper presents efficient feature extraction and selection method addressing the feature imbalance problem for detecting fraud in a streaming environment. Here feature aware XGBoost (FA-XGB) algorithm is designed. The FA-XGB algorithm employs an effective cross-validation scheme for building a predictive model. The research significance is given below:

- The FA-XGB employ an effective cross-validation scheme for selecting meaningful feature during the training of the predictive model.
- The FA-XGB-based attack detection model achieves much better accuracy, recall, and F-measure performance in comparison with the existing ensemble-based classification model.

2. LITERATURE SURVEY

In this section, a survey of various existing methods, the advantages of the methods, and the challenges faced by each of these methods for the detection of fraudulent online transactions using machine learning methodologies have been discussed.

2.1. Class Imbalance and Concept Drift Problems

The change in data from time to time makes the machine learning algorithms or methods unable to predict or detect any kind of fraudulent activities in the online environment. The problem of not being able to predict or detect these activities is called concept drift. Due to the problem of the concept drift many methodologies have been developed to resolve the issue of the detection and prediction and the challenges of the concept drift. Many researchers have attained a result that if the drift problem is addressed in the concept of drift, then the prediction can be done easily without facing any challenges. Many models have been used to resolve the problem of concept drift. Hence in [1], they have reviewed 130 papers related to the concept drift problems and their challenges. They have also mentioned the various methods used by each research and their frameworks for a better understanding of how the concept drift problem works and its limitations. They have used reviewed the 10 most used datasets which have been used to evaluate their models. In [2], they have proposed a machine learning algorithm using the concept drift approach which detects the fake website using the URL as an input to the model which detects whether the website is fake or not. The model has been trained and tested for accuracy using various machine learning classifiers like the Gradient Boosted Decision Tree, Random Forest, and neural networks. In [3], to resolve the problem of the cyber-attacks in the business enterprises using the concept drift-based approach a model has been proposed. In this model, the main objective of the model is to first detect the incoming attacks, locate the given attack and classify the concept drift instantaneously without using the storage.

Due to the problem of the drift in the concept drift, the data imbalance problem arises. The imbalanced data has a major challenge in the detection and prediction using a machine learning model as each sample of each of the classes may vary from the original sample value. Hence for this problem, in [4], they have proposed a model, TSCS, using the class imbalance and the concept drift approach. This model uses the feature selection method for the classification of the imbalanced data and provides a window for the adaptation of the drift detection method. In [5], they have discussed the problems faced by the concept drift approaches in the multi-class imbalanced data streams. They have also developed a detection technique using the restricted Boltzmann machine which can detect the changes in the sample of the multi-class imbalanced data streams. In [6], similar to [5], they have also found that there is a problem in the concept drift approaches in the multi-class imbalanced samples. To face this challenge, they have proposed a model using the AUC to classify the multi-class imbalanced data in the online environment. In [7], They have proposed a model, RE-DI, for the concept drift and class imbalance challenges. This model can detect the problem of the drift in the concept drift and resolves the issue of class imbalance.

In the real-time application, both the concept drift approaches and the data imbalance methods have been in current use for the detection and prediction of various kinds of attacks. One of the examples is twitter spam detection. In Twitter, various kinds of business enterprises use various kinds of advertisements to attract the users and the users end up in various kinds of scams. To resolve the problem of Twitter spam, in [8], they have introduced a Twitter spam detection technique using the class imbalance and concept drift method. In this model, they have used millions of o tweets to train the model using the Lfun method. After using the Lfun method the spammed tweets and the unlabeled tweets are classified according to the class imbalance method. This increases the accuracy of the detection of spam in Twitter.

2.2. Machine Learning based Fraudulent Online Transaction Detection Models

There have been many ML models which are currently being used for the detection of various fraudulent online transactions in credit card companies. In [9], they have proposed a model using machine learning algorithms that will detect whether the transaction made is legitimate or fraudulent. They have used five machine learning algorithms that preprocess the data and classify them according to their performance. The dataset used for the evaluation of the results was imbalanced data, they have used the values of kappa statistics. In [10], they have examined various classification methods which use the public dataset to train their model and to determine whether the transaction being processed is fraudulent or is legitimate. This research also explains which machine learning algorithm is best for classification when there is an imbalanced dataset. They achieved a result that using the SVM algorithm helps to identify the detection of the fraud payments and also has a higher performance rate. In [11], they have used the gradient boosting tree method (GBT) to detect the transaction of the credit card in real-time when there is a drift problem. This model is trained daily to identify the different problems faced by the model when there occurs a drift problem. In [12], they have developed a CatBoost method using the existing machine learning algorithm for the detection of fraud transactions. This method uses feature extraction in which only the required features are extracted for the classification. This method uses memory compression which helps to speed up the detection of online transactions. They have used the public dataset, IEEE-CIS for the evaluation of the results. In [13], they have evaluated the real-time transactions using the machine learning algorithms and classified them into legitimate or fraudulent transactions. In [14], they have proposed a method using neural networks based on unsupervised learning methods. In [15], they have surveyed different deep learning methods which have been used for the classification of online transaction fraud detection. They have also surveyed the various datasets which have been used for the classification. In [16], they have analyzed the different machine learning algorithms like logistic regression, SVM, and quadratic discriminant analysis which use the feature extraction method for the classification of the credit card fraud detection.

In [17], they have used the IDS method instead of the machine learning model to maintain the integrity and to protect the confidential data from getting in hands of the malicious user. In this model, they have used the feature selection method to acquire all the required data and discard the data which is not required by the model. This classification helps to increase the performance of the model. In this model an algorithm called the CFS-BA has been proposed to reduce the size of the dataset, then they have introduced the C4.5 model which ensembles the machine learning algorithms like the RF and Forest PA. For the detection of the attack, they have used the voting technique. In [18], they have developed an optimized XGBoost method for the class imbalance datasets without any of the resampling methods. They have used the Randomized Search CV hyperparameter optimization method to find the optimal parameters for the XGBoost. The sampling methods used in the model help to attain higher performance for the detection of fraudulent credit card transactions. However, the major factor affecting the existing model is very less important is given feature selection, in addressing the research problem an improved cross-validation scheme is modeled for the XGBoost algorithm in the next section.

3. PROPOSED METHODOLOGY

This section introduces an improved machine learning-based malicious transaction detection mechanism for the streaming environment. Here an improved XGBoost algorithm namely Feature aware XGB (FA-XGB) is modeled. The FA-XGB model employs an improved cross-validation mechanism for selecting useful features and creating an ensemble classification model. The proposed model will aid in achieving high detection accuracy with less false positive considering for detection of different malicious transaction links in the online streaming environment under a highly imbalanced environment.

3.1. Standard XGBoost Model

The standard XGBoost algorithm is an enhancement of distributed gradient boosting algorithm where multiple tress classifiers are combined for improving classification performance. Therefore, for detecting malicious transaction links in the streaming environment such as Twitter, the standard XGBoost algorithm trains dataset with o samples using multiple classifiers is obtained as follows

$$\widehat{\mathrm{A}}_{\mathrm{j}} = \mathrm{H}(\mathrm{Z}_{\mathrm{j}}) = \sum_{\mathrm{m}=1}^{\mathrm{M}} \mathrm{h}_{\mathrm{m}}(\mathrm{Z}_{\mathrm{j}}), \qquad \mathrm{h}_{\mathrm{m}} \in \alpha \tag{1}$$

where Z_{j} defines the j^{th}sample within training sample considered K represent the size of tree utilized to perform classification to identify whether a transaction link is malicious or not in streaming environment dataset, $\widehat{\mathrm{A}}_{\mathrm{j}}$ represent the output of the classification model considering definite dimension size, k^{th} dimension defines probabilities of it belongs to k^{th} class, during the classification process, and α represent decision tree (DT) sets which are obtained as follows

$$\alpha = \{h(z) = y_{u(z)}\} \tag{2}$$

where $h(z)$ defines different trees which must satisfy the leaf weight y and optimization parameter u. The standard XGB model is designed through minimization of loss parameter defined through the below equation

$$N(H) = \sum_{k} n(\hat{a}_k, a_k) + \sum_{m} \beta(h_l) \tag{3}$$

where,

$$\beta(h_l) = \delta V + \mu \|y\|^2 \tag{4}$$

In Eq. (3), the first parameter $n(\hat{a}_k, a_k)$ describes the loss operation between original and output predicted and the second parameter $\beta(h_l)$ describes the penalty parameter for optimizing the model; V defines the size of the leaf of the corresponding tree, δ and μ define the parameter sets utilized for controlling computation overhead. Here a weighted loss operation is utilized for training streaming data z whose labels are represented as n, the negative log probability loss operation is computed as follows

$$n(\hat{a}_k, a_k) = -\sum_{l} a(l) \log \hat{a}\,(n) = -\log \hat{a}\,(n) \tag{5}$$

where $a(l)$ defines l^{th} dimension of a, $\hat{a}(n)$ defines the l^{th} dimension of output $\hat{a}$. Alongside, the optimization process of loss operation is done through an iterative process (i.e., considering u^{th} iteration) to minimize the loss as follows

$$N^K = \sum_{k=1}^{p} n\left(\hat{a}_k^{(p-1)} + h_p(z_k), a_k\right) + \beta(h_v) \tag{6}$$

3.2. Class Imbalance Aware XGBoost Model

The proposed modified XGB model computes h_p which assures minimum loss through the greedy process as described below

$$\begin{aligned} N^p &\cong \sum_{k=1}^{p} \left[n\left(\hat{a}_k^{(P-1)} + a_k\right) + i_k h_k(z_k) + \frac{1}{2} j_k h_p^2(z_k) \right] + \beta\left(h_p\right) \\ &\propto \sum_{k=1}^{p} \left[i_k h_k(z_k) + \frac{1}{2} j_k \left[h_p(z_k)^2\right] \right] + \beta\left(h_p\right) \end{aligned} \tag{7}$$

where h_j and j_k defines first and second-order gradient of $n\left(\hat{a}_k^{(P-1)} + a_k\right)$, respectively; thus, the tree h_p are obtained through minimizing Eq. (7). The weighted loss operation for malicious transaction link identification in the streaming environment is computed as follows

$$\mathcal{N}_y = -\sum_{k=1}^{q} (\alpha a_k \log(\hat{a}_k) + (1 - a_k) \log(1 - \hat{a}_k)) \tag{8}$$

where α defines the bias parameter describing feature imbalance. The optimization process of addressing feature imbalance issues is done through effective cross-validation presented in the subsection.

3.3. Effective Cross-Validation Technique

Here we employ cross-validation (CV) mechanism for selecting useful feature sets in optimizing the predictive model described in Eq. (7). Here we select a predictive model that reduces validation error. Most of the standard attack detection models have employed K-fold CV scheme for optimizing output. In K-fold CV, the dataset is divided randomly into K subset of identical size; then $K - 1$ subset is used for building a predictive model and the leftover subset is used for predicting errors in the model. Finally, the K combination of predicted error is average for obtaining cross-validation errors. Later, a grid of l suitable values is generated for establishing ideal optimizing parameters in minimizing CV errors. Finally, the model with minimum cross-validation error is selected; However, such model results in poor classification outcomes when data is highly imbalanced and also fails to establish the correlation among features for different attacks or between attack and normal transaction link.

Here we model the effective cross-validation (ECV) technique to build a predictive model that minimizes prediction error considering feature importance. The proposed model employs cross-validation with two layers. In the first layer, feature subsets are chosen as the main features. In the second layer, the main subset feature selected from the first layer is used for building the final predictive model. The standard cross-validation model by constructing multiple sets of K folds rather than constructing single K-fold sets; the single fold cross-validation error is obtained using the following equation

$$CV(\sigma) = \frac{1}{M}\sum_{k=1}^{K}\sum_{j\in G_{-k}} P\left(b_j, \hat{g}_{\sigma}^{-k(j)}(y_j,\sigma)\right) \tag{9}$$

The improved version i.e., an effective cross-validation error is obtained using the following equation

$$CV(\sigma) = \frac{1}{SM}\sum_{s=1}^{S}\sum_{k=1}^{K}\sum_{j\in G_{-k}} P\left(b_j, \hat{g}_{\sigma}^{-k(j)}(y_j,\sigma)\right) \tag{10}$$

Then, the optimization parameter for selecting the optimal value $\hat{\sigma}$ is obtained using the following equation

$$\hat{\sigma} = \underset{\sigma\in\{\sigma_1,\dots,\sigma_l\}}{\arg\min}\ CV_s(\sigma) \tag{11}$$

In the above equations, $P(\cdot)$ represent loss function, $\hat{g}_{\sigma}^{-k(j)}(\cdot)$ represent a function for estimating coefficients, and M describes training data size. In ECV, the predictive model is built by estimating multiple optimization parameters. Here multilayer of K-folds cross-validations are created; the layer size is set according to the optimization parameter considered; i.e., if the parameter is optimized in layer 1, the parameter value is fixed and given to layer 2 for estimating additional optimization parameters.

4. RESULTS AND DISCUSSION

Here the performance of the Feature Aware XGBoost (FA-XGB) model is compared with the other existing models such as ensemble-based [17], Statistical-based [8], and XGB-based [19]. Accuracy, Recall, and F-measure are performance metrics used for validating the proposed FA-XGB-based malicious transaction model over the existing model. The proposed and other existing model is implemented using the python framework and the system used is i5 Intel processor with 12 GB RAM. The dataset used for the experiment is given below.

4.1. Dataset Description

Here experiment is conducted widely used streaming environment dataset [8], [19]. The Dataset comprises 12 feature sets that are described in Table I. A total of 10,000 malicious transaction link data is collected per day and the data similarly has been collected for 10 days. According to the real-world scenarios, only 5% of data is discarded as it is of no use. The dataset is composed of 95000 normal data and the remaining 5000 are malicious URLs; thus, the dataset has some serious imbalance issues.

Table I. Dataset Considered for Experimental Analysis

Feature Number	*Feature Named*
F1	*Account Age*
F2	*No_follower*
F3	*No_following*
F4	*No_userfavourites*
F5	*No_lists*
F6	*No_tweets*
F7	*No_retweets*
F8	*No_hashtags*
F9	*No_usermention*
F10	*No_URLs*
F11	*No_char*
F12	*No_digits*

4.2. Performance Metric

Here the performance of the FAXGB and existing systems such as ensemble-based, XGB, and FA-XGB is evaluated. The experiments are carried out for studying the impact of data imbalance affecting the overall classification accuracy of different models. The performance is evaluated in terms of accuracy, recall, and F-measure. The ROC confusion matrix is shown in Table II

Table II. ROC Confusion Matrix

	Malicious URL	*Normal URL*
Predicted Malicious URL	*True Positive (TP)*	*False Positive (FP)*
Predicted Normal URL	*False Negative (FN)*	*True Negative (TN)*

The accuracy performance is calculated as follows

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \qquad (12)$$

The recall performance is calculated as follows

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

The F-measure performance is calculated as follows

$$F - measure = \frac{2 * Precision * Recall}{Precision * Recall} \tag{14}$$

4.3. Accuracy

This section checks the accuracy of the proposed model with the other existing other models. The accuracy of the model is calculated using Eq. (12). The accuracy achieved by the FA-XGB algorithm over the existing system such as ensemble-based and XGBoost is shown in Fig. 1. From the figure, it can be seen that the model FA-XGB gives a better accuracy rate than the other algorithm used in the existing model. The ensemble-based algorithm and XGB-based show an accuracy of 98.43% and 98.85%, respectively. The proposed FA-XGB provides an accuracy of 99.684%. The higher accuracy of the FA-XGB-based malicious transaction link identification model is very effective in addressing serious class imbalance issues.

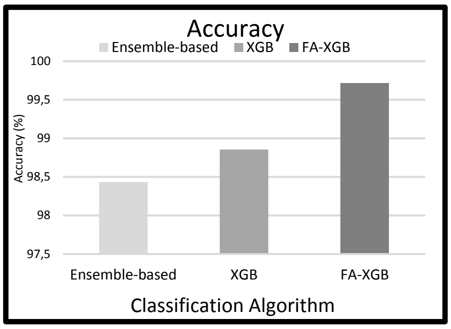

Fig. 1. Accuracy performance.

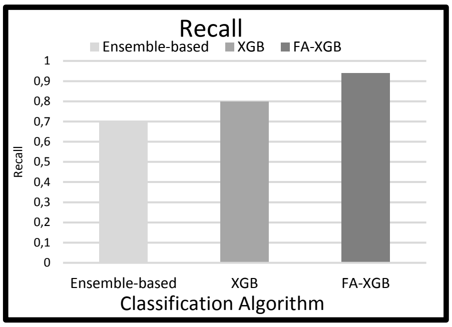

Fig. 2. Recall Performance.

4.4. Recall

This section checks the recall performance of the proposed model with the other existing other models. The recall performance of the model is calculated using Eq. (13). The recall performance achieved by the FA-XGB algorithm over the existing system such as ensemble-based and XGBoost-based is shown in Fig. 2. The ensemble-based algorithm and XGB-based show a recall performance of 0.69 and 0.79, respectively. The proposed FA-XGB provides a recall performance of 0.938. The higher recall performance of the FA-XGB-based malicious transaction link identification model is very effective in addressing serious class imbalance issues.

4.5. F-measure

This section checks the F-measure performance of the proposed model with the other existing other models. The F-measure performance of the model is calculated using Eq. (14). The F-measure performance achieved by the FA-XGB algorithm over the existing system such as ensemble-based and XGBoost is shown in Fig. 3. The ensemble-based algorithm and XGB-based show an F-measure performance of 0.803 and 0.867, respectively. The proposed FA-XGB provides an F-measure performance of 0.974. The higher F-measure performance of the FA-XGB-based malicious transaction link identification model is very effective in addressing serious class imbalance issues.

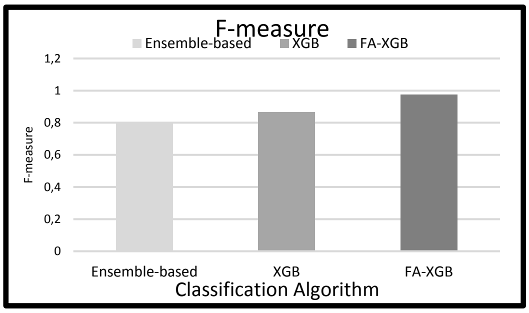

Fig. 3. F-measure Performance.

5. CONCLUSION

This paper first identified the problem of malicious transaction links in the streaming environment; then presented Feature Aware XGBoost for addressing class imbalance issues. The FA-XGB encompasses an improved cross-validation mechanism that selects the right kind of feature to achieve good classification performance even when data is imbalanced. The experiment is conducted using a streaming environment such as Twitter which has some serious data imbalance issues. The experiment outcome shows the FA-XGB achieves an accuracy improvement of 0.86% and 1.28% over XGB-based and ensemble-based attack detection methods. Then, FA-XGB achieves a recall performance improvement of 14% and 24.2% over the XGB-based and ensemble-based attack detection methods. Similarly, the FA-XGB achieves an F-measure performance improvement of 10.6% and 17.1% over the XGB-based and ensemble-based attack detection methods. The result shows the effectiveness of FA-XGB in addressing serious class imbalance issues in the streaming environment. Though the class imbalance issue is addressed through an improved machine learning model; however, the attack varies over time. As a result, affect the detection rate. Thus, it is important to address the drift problem in the streaming environment.

REFERENCES

[1] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama and G. Zhang. Learning under Concept Drift: A Review, *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 12, Dec. 2019, doi: 10.1109/TKDE.2018.2876857, pp. 2346-2363.

[2] S. Singhal, U. Chawla and R. Shorey. Machine Learning & Concept Drift based Approach for Malicious Website Detection, *2020 International Conference on Communication Systems & NetworkS (COMSNETS)*, 2020, doi: 10.1109/COMSNETS48256.2020.9027485, pp. 582-585.

[3] N. Liu, J. Huang and L. Cui. A Framework for Online Process Concept Drift Detection from Event Streams, *2018 IEEE International Conference on Services Computing (SCC)*, 2018, doi: 10.1109/SCC.2018.00021, pp. 105-112.

[4] Y. Sun, Y. Sun and H. Dai. Two-Stage Cost-Sensitive Learning for Data Streams With Concept Drift and Class Imbalance, *IEEE Access*, vol. 8, 2020, doi: 10.1109/ACCESS.2020.3031603, pp. 191942-191955.

[5] Ł. Korycki and B. Krawczyk. Concept Drift Detection from Multi-Class Imbalanced Data Streams, *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, 2021, doi: 10.1109/ICDE51399.2021.00097, pp. 1068-1079.

[6] S. Wang and L. L. Minku. AUC Estimation and Concept Drift Detection for Imbalanced Data Streams with Multiple Classes, *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, doi: 10.1109/IJCNN48605.2020.9207377, pp. 1-8.

[7] H. Zhang, W. Liu, S. Wang, J. Shan and Q. Liu. Resample-Based Ensemble Framework for Drifting Imbalanced Data Streams, *IEEE Access*, vol. 7, 2019, doi: 10.1109/ACCESS.2019.2914725, pp. 65103-65115.

[8] 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, April 2017, doi: 10.1109/TIFS.2016.2621888, pp. 914-925.

[9] K. AbdulSattar and M. Hammad. Fraudulent Transaction Detection in FinTech using Machine Learning Algorithms, *2020 International Conference on Innovation and Intelligence for Informatics, Computing and Technologies (3ICT)*, 2020, doi: 10.1109/3ICT51146.2020.9312025, pp. 1-6.

[10] R. Banerjee, G. Bourla, S. Chen, M. Kashyap and S. Purohit. Comparative Analysis of Machine Learning Algorithms through Credit Card Fraud Detection, *2018 IEEE MIT Undergraduate Research Technology Conference (URTC)*, 2018, doi: 10.1109/URTC45901.2018.9244782, pp. 1-4.

[11] B. Bayram, B. Köroğlu and M. Gönen. Improving Fraud Detection and Concept Drift Adaptation in Credit Card Transactions Using Incremental Gradient Boosting Trees, *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2020, doi: 10.1109/ICMLA51294.2020.00091, pp. 545-550.

[12] Y. Chen and X. Han. CatBoost for Fraud Detection in Financial Transactions, *2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, 2021, doi: 10.1109/ICCECE51280.2021.9342475, pp. 176-179.

[13] A. Thennakoon, C. Bhagyani, S. Premadasa, S. Mihiranga and N. Kuruwitaarachchi. Real-time Credit Card Fraud Detection Using Machine Learning, *2019 9th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, 2019, doi: 10.1109/CONFLUENCE.2019.8776942, pp. 488-493.

[14] A. K. Rai and R. K. Dwivedi. Fraud Detection in Credit Card Data using Unsupervised Machine Learning Based Scheme, *2020 International Conference on Electronics and Sustainable Communication Systems (ICESC)*, 2020, doi: 10.1109/ICESC48915.2020.9155615, pp. 421-426.

[15] Kanika and J. Singla. A Survey of Deep Learning based Online Transactions Fraud Detection Systems, *2020 International Conference on Intelligent Engineering and Management (ICIEM)*, 2020, doi: 10.1109/ICIEM48762.2020.9160200, pp. 130-136.

[16] P. Naveen and B. Diwan. Relative Analysis of ML Algorithm QDA, LR and SVM for Credit Card Fraud Detection Dataset, *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, 2020, doi: 10.1109/I-SMAC49090.2020.9243602, pp. 976-981.

[17] Yuyang Zhou, Guang Cheng, Shanqing Jiang, Mian Dai. Building an efficient intrusion detection system based on feature selection and ensemble classifier, *Computer Networks*, Volume 174, 2020, 107247, ISSN 1389-1286, https://doi.org/10.1016/j.comnet.2020.107247.

[18] C. V. Priscilla and D. P. Prabha. Influence of Optimizing XGBoost to handle Class Imbalance in Credit Card Fraud Detection, *2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 2020, doi: 10.1109/ICSSIT48917.2020.9214206, pp. 1309-1315.

[19] X. Wang, Q. Kang, J. An and M. Zhou. Drifted Twitter Spam Classification Using Multiscale Detection Test on K-L Divergence, *IEEE Access*, vol. 7, 2019, doi: 10.1109/ACCESS.2019.2932018, pp. 108384-108394.

Information about the authors:

Arati Shahapurkar was born in Karnataka, India, in 1978. She received the B.E. degree in Computer Science & engineering from the Visveswaraya Technological University, in the year 2005 and MTech in Computer Science and engineering from the Gogte Institute of Technology (GIT) Belagavi India, in 2011.

Dr. Sunil F. Rodd was born in Karnataka, India, in 1968. He received the B.E. degree in Computer Science & engineering from KREC(NITK) Surathkal from Mangalore University in 1990. and MTech in Computer Science and engineering from the IIT Bombay in 1995 and Ph.D. from Graphic Era University, Graphic Era Deemed University in 2014.

Manuscript received on 30 April 2022
Revised version received on 05 June 2022