Contents lists available at GrowingScience

International Journal of Data and Network Science

homepage: www.GrowingScience.com/ijds

EFN-SMOTE: An effective oversampling technique for credit card fraud detection by utilizing noise filtering and fuzzy c-means clustering

Hadeel Ahmad[a,b*], Bassam Kasasbeh[c], Balqees AL-Dabaybah[d] and Enas Rawashdeh[e]

[a]*Computer Science, Applied Science Private University, Jordan*
[b]*MEU Research Unit, Middle East University, Jordan*
[c]*Data Science Artificial Intelligence, Al Hussein Technical University, Jordan*
[d]*Computer Science, Luminus Technical University College, Jordan*
[e]*Management Information Systems, Albalqa' Applied University, Jordan*

C H R O N I C L E

Article history:
Received: February 24, 2023
Received in revised format: April 12, 2023
Accepted: June 7, 2023
Available online: June 7, 2023

Keywords:
Oversampling technique
Credit card fraud detection
Unbalanced dataset
Fuzzy C-means (FCM)
SMOTE

A B S T R A C T

Credit card fraud poses a significant challenge for both consumers and organizations worldwide, particularly with the increasing reliance on credit cards for financial transactions. Therefore, it is crucial to establish effective mechanisms to detect credit card fraud. However, the uneven distribution of instances between the two classes in the credit card dataset hinders traditional machine learning techniques, as they tend to prioritize the majority class, leading to inaccurate fraud predictions. To address this issue, this paper focuses on the use of the Elbow Fuzzy Noise Filtering SMOTE (EFN-SMOTE) technique, an oversampling approach, to handle unbalanced data. EFN-SMOTE partitions the dataset into multiple clusters using the Elbow method, applies noise filtering to each cluster, and then employs SMOTE to synthesize new minority instances based on the nearest majority instance to each minority instance, thereby improving the model's ability to perceive the decision boundary. EFN-SMOTE's performance was evaluated using an Artificial Neural Network model with four hidden layers, resulting in significant improvements in classification performance, achieving an accuracy of 0.999, precision of 0.998, sensitivity of 0.999, specificity of 0.998, F-measure of 0.999, and G-Mean of 0.999.

1. Introduction

The use of credit cards has significantly expanded in recent times, especially with the development of technology and the emergence of various applications in multiple industries, including online payment solutions. This increase in cashless payment systems has led to a rise in the chances of fraudulent activity by unauthorized individuals, ultimately resulting in the proliferation of credit card fraud (Torgo et al., 2013; Khader et al., 2021). Fraudsters are developing new strategies to find loopholes, leading to an increase in fraudulent transactions. To address this issue, it is necessary to detect and predict fraudulent transactions using Machine Learning (ML) techniques. However, one of the significant challenges in identifying fraudulent activity is the unbalanced dataset. Unbalanced data occurs when classes are not equally represented in a dataset, resulting in a skewed distribution due to significant variance in data. This can adversely affect the quality of data and classification outcomes (Prasetiyo et al., 2021; Ahmad et al., 2022). Standard machine learning models trained on unbalanced class distribution ratios show satisfactory performance with respect to the majority class but often fail to identify fraudulent transactions (Mishra & Ghorpade, 2018; Tran & Dang, 2021a). Furthermore, evaluation metrics such as accuracy are not suitable for

* Corresponding author.
E-mail address: h_ahmad@asu.edu.jo (H. Ahmad)

ISSN 2561-8156 (Online) - ISSN 2561-8148 (Print)

doi: 10.5267/j.ijdns.2023.6.003

unbalanced classes, as they neglect the minority class and train the model based on the dominant class. Therefore, implementing re-sampling techniques to obtain a balanced dataset before training models using ML algorithms is crucial.

There are numerous techniques used to address the issue of class unbalance in a dataset. Oversampling and undersampling are two common approaches used to tackle class imbalance (Ramentol et al., 2012). Oversampling aims to increase instances from the minority class to achieve a specific balance ratio, but this may lead to overfitting. On the other hand, undersampling involves removing instances from the majority class, which could result in loss of important data. In our work, we deal with data unbalance by oversampling the minority class. Our goal is to determine the most effective oversampling strategy to bring the minority class to a level comparable to the majority class. To achieve this, we employ Synthetic Minority Over-Sampling Technique (SMOTE), which generates new synthetic minority instances with different features to increase the number of minor class instances and avoid overfitting (Chawla et al., 2002).

This research proposes a model that tackles the problems of class unbalance and overfitting in a dataset, making it a significant contribution. The framework presented in this study enhances the quality of the dataset by incorporating a sampling method that increases the precision and effectiveness of the detection process. Our approach involves clustering the dataset using Fuzzy C-means (FCM), which eliminates irrelevant noisy characteristics of the dataset. Then, SMOTE is applied to generate new instances with similar features, while maintaining the consistency and integrity of data features. Thus, our strategy effectively addresses overfitting and improves the overall dataset quality.

2. Related works

Imbalanced datasets pose a major challenge in data science as they lead to reduced performance of machine learning algorithms and a bias towards the majority class. Consequently, these algorithms perform well on the majority class but poorly on the minority class, even though the latter may hold more valuable information (Japkowicz, 2000). To address this issue, researchers have developed solutions that can be categorized into three primary areas: data-level methods, algorithm-level methods, and Ensemble Learning-based methods (Lebichot et al., 2020; Barua et al., 2014). In the field of machine learning, data-level methods have been classified into three main categories: under-sampling, oversampling, and hybrid-sampling. The objective of under-sampling is to create a balanced dataset by removing instances of the majority class from the training set (Galar et al., 2012). In contrast, oversampling aims to balance the data by increasing the number of minority class instances, typically through the generation of synthetic data Santoso et al. (2019). Under-sampling methods are commonly used to handle imbalanced datasets, including Random Under-sampling (RUS) which randomly removes instances of the majority class to balance the dataset (Batista et al., 2004). Under-sampling Based on Clustering (SBC) is another technique that divides the dataset into clusters and selects representative data from each cluster to improve the classification accuracy of the minority class (Yen & Lee, 2006; Laurikkala, 2001). Additional under-sampling techniques have been developed, including One-sided Selection (OSS) (Kubat et al., 1997), Distance- based Under-sampling (DUS) (Li et al., 2013), and Selection Based on Similarity (SBS) (Ahmad et al., 2023).

Oversampling techniques aim to increase the representation of the minority class by duplicating or synthesizing current samples (Ramentol et al., 2012). The simplest oversampling technique is random oversampling, which involves randomly replicating minority class samples, but it may lead to overfitting (Tantithamthavorn et al., 2020). To address this issue, the Synthetic Minority Over-sampling Technique (SMOTE) was introduced by Chawla et al. (2002). SMOTE generates synthetic samples from the minority class by creating linear combinations of two similar minority samples (Badic et al., 2022). SMOTE has been widely used to address imbalanced datasets, along with other resampling techniques such as ADASYN Tran and Dang (2021b) and the resampling methods proposed in Hordri et al. (2018). In the study by Zou (2021), both undersampling (using RUS) and oversampling (using ROS and SMOTE) were applied to manage imbalanced datasets in credit card fraud detection using artificial intelligence.

Several recent studies have demonstrated that while SMOTE is a powerful and effective oversampling technique, its performance can be degraded due to its extension with noisy and borderline samples (Batista et al., 2004; Verbiest et al., 2014; Sáez et al., 2015; Yi et al., 2022). This is due to the inherent limitation of SMOTE-based methods, known as blind oversampling, which leads to this problem (Yi et al., 2022). Consequently, change-direction and filtering-based methods are employed to address this issue. Change-direction methods are utilized to indirectly manage noisy and borderline samples. These methods aim to direct the generation of synthetic samples by SMOTE towards specific regions of the input space. Many SMOTE techniques fall under this category of improvements, such as Borderline- SMOTE, which generates additional synthetic boundary samples by finding the k-nearest neighbors in the majority class Han et al. (2005). Another method called Safe-Level-SMOTE uses kNN to identify safe areas and generates more synthetic samples in those regions Bunkhumpornpat et al. (2009). The K-means SMOTE clustering algorithm divides the dataset into k regions and generates more synthetic data in high-density regions Douzas et al. (2018). Adaptive- SMOTE uses inner and danger subsets to generate more synthetic samples at the class center to prevent an expansion of the class boundary and to strengthen the original dataset distributional characteristics (Pan et al., 2020). Filtering-based approaches, such as Tomek Link (TL), Edited Nearest Neighbors (ENN), and Iterative-Partitioning Filter (IPF), are strategies that combine SMOTE with error detection techniques. These techniques are specifically used to identify noisy and borderline samples. After these instances are identified, they are removed from the

dataset. An example of a filtering-based approach is SMOTE-ENN, which works by identifying samples as noisy if their kNN is classified incorrectly (Batista et al., 2004). Another technique is SMOTE- IPF, which uses an ensemble classifier to identify potentially misclassified samples by employing global error detection (Sáez et al., 2015).

A recent improvement on SMOTE is ASN-SMOTE, proposed by Yi et al. (2022). This version applies a noise filtering mechanism and an adaptive neighbor selection mechanism before synthesizing minority instances. The noise filtering mechanism filters out minority samples whose nearest neighbor is a majority sample. The adaptive neighbor selection mechanism only oversamples a minority neighbor that is closer to the closest majority sample. These two approaches prevent the generation of synthetic minority cases in the majority class region and make the decision boundary more general.

In this paper, we introduce a novel approach for oversampling that utilizes a noise filtering mechanism and an adaptive neighbor selection mechanism. The Elbow Method is employed to cluster the dataset into the optimal number of clusters. This technique is applied to a highly unbalanced credit card fraud detection dataset to distinguish between fraudulent and legitimate transactions (Dal Pozzolo et al., 2015).

3. Methodology

To address the issue of class imbalance, we introduce the EFN- SMOTE framework as a solution for balancing the dataset. The methodology behind this study is outlined in this section, with the EFN-SMOTE framework depicted in Fig. 1.

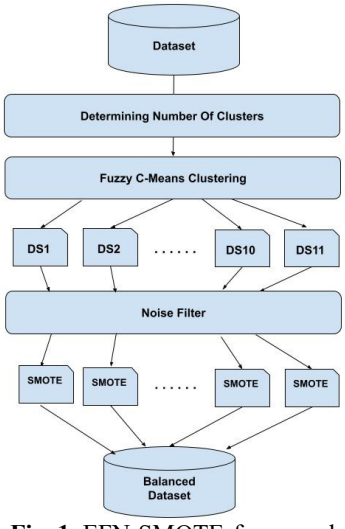


Fig. 1. EFN-SMOTE framework

3.1 Data selection

The confidentiality and privacy concerns associated with credit card fraud have limited the availability of actual credit card datasets. Therefore, we utilized a dataset obtained from Kaggle Dal Pozzolo et al. (2015). This dataset comprises 284807 transactions made by European cardholders in September 2013. It consists of 284315 normal transactions and 492 fraudulent transactions, making the minor class constitute only 0.172% of all transactions. Therefore, this dataset is highly imbalanced, and it poses a significant challenge for developing effective fraud detection models.

3.2 Fuzzy C-means

To ensure the reliability of the transactional features, we utilized Fuzzy C-Means to cluster the dataset and group the instances based on their similarities. Fuzzy C-Means is a machine learning algorithm that partitions the dataset into N clusters, where N denotes the number of clusters generated. The primary goal of clustering is to group the transactions into clusters, where each cluster comprises transactions with similar data features, and transactions belonging to different clusters are dissimilar as much as possible (Lei et al., 2018). The subsequent section outlines the approach adopted in this study to determine the optimal number of clusters.

3.3 Determining the optimal number of clusters

Determining the optimal number of clusters in partitioning a dataset is crucial. In this study, we employed the Elbow method, a heuristic technique widely used to identify the optimal number of clusters (Nainggolan et al., 2019). This method involves plotting the explained variation against the number of clusters and selecting the number of clusters based on the curve's

"elbow", where the rate of explained variation drastically decreases. The Elbow method computes the total within-cluster sum of squares (WCSS) for each cluster (k) using Eq. (1), where ci denotes cluster i, N_c is the number of clusters, x_{ci} is the cluster centroid, and x denotes sample mean. The pseudo-code for the Elbow method is presented in Algorithm 1.

$$WCSS = \sum_{i=1}^{N_c} \sum_{x \in c_i} d(x, \bar{x}_{ci})^2 \quad (1)$$

Algorithm 1 Elbow method

procedure Find the optimal number of clusters
  $k \leftarrow 2$
  $N \leftarrow n$ //which n is the max number of cluster
  while k <N **do**
    Calculate WCSS For each k
  $k \leftarrow k + 1$
  end while
  Plot the curve of WCSS according to the number of clusters k. Find location of band (knee)
end procedure

To ensure the reliability of our results, we conducted 100 runs of the Elbow method on the Fraud detection dataset (Dal Pozzolo et al., 2015). The optimal number of clusters appeared most frequently was recorded in Table 1. The Elbow method selected 12 clusters 34 times, which was the highest among all other iterations. As a result, we concluded that the optimal number of clusters for the dataset is 12, represented as DS1, DS2, DS3, ..., DS12. The implementation of the Elbow method and the identification of the optimal number of clusters are illustrated in Fig. 2.

Table 1
Number of iterations for each number of clusters by Elbow method

Number of clusters	10	11	12	13	14	15	16
Number of iterations	14	11	34	20	18	4	4

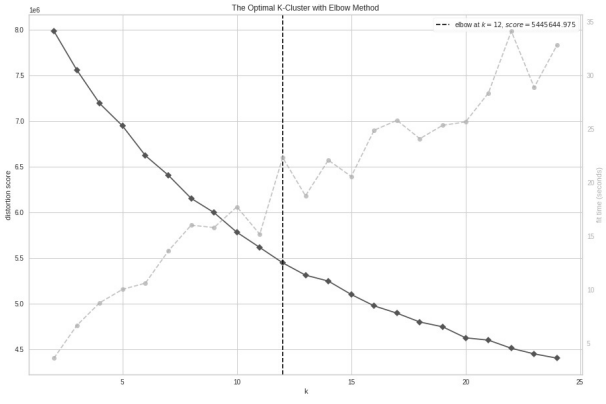


Fig. 2. Finding optimal number of clusters with Elbow method

3.4 Noise Filtering

In this study, we applied a noise filtering process (NFP) per each cluster, which is an essential step in improving the dataset's quality for machine learning applications. NFP aims to remove irrelevant and noisy features to ensure accurate results. We adopted the NFP algorithm proposed in Yi et al. (2022). This method employs a k-nearest neighbors (kNN) algorithm to reduce the generation of synthetic minority instances in the majority class. This algorithm identifies and filters out noise and minority instances located near the decision boundary by evaluating the nearest instance of each minority instance. This method not only targets the minority instances but also focuses on the decision border minority instances in the proximity of the decision boundary.

3.5 Dataset Resampling

In this study, we employed SMOTE, an oversampling technique, to increase the number of instances in the minority class. SMOTE is an algorithm that generates new synthetic instances in the minor class by duplicating the existing instances based on their distance to their nearest neighbor (Tran & Dang, 2021a). SMOTE generates new artificial instances using K-Nearest Neighbor (KNN) (Hordri et al., 2018), which generates new instances with similar features instead of replicating the existing ones.

After balancing the dataset, it was divided into a training set (80%) and a testing set (20%), then we trained the models using ANN.

4. Results and discussion

In this section, we applied the proposed EFN-SMOTE technique to the unbalanced Credit Card Fraud Detection dataset (Dal Pozzolo et al., 2015). We evaluated the classification performance of artificial neural network (ANN) algorithms with varying numbers of hidden layers, ranging from one to five. The proposed technique was compared with existing methods such as ASN- SMOTE (Yi et al., 2022) and the original SMOTE (Chawla et al., 2002). We provide a detailed description of the experimental setup, including the performance metrics used and the construction of Multi-Layer Perceptron (MLP) Neural Network Models. Finally, we present the results and discuss their implications.

4.1 Experimental setup

The study was conducted on a cloud-based platform called Google Collab- oratory (Colab) notebook, equipped with GPU hardware accelerators and various libraries, including Scikit-Learn, matplotlib, sklearn, and pandas. The Colab notebook is a Jupyter-based environment that is free and requires no setup, providing a convenient working environment with access to well-known machine learning libraries Bisong (2019). The experimental setup utilized an Intel Core i-7 3.0 GHz processor with 8.0 GB of RAM.

4.2 Performance Metrics

The performance evaluation of the proposed EFN-SMOTE model was con- ducted by comparing its classification accuracy with that of the MLP-ANN model using various performance metrics. These metrics included accuracy (Acc), sensitivity (Sen), specificity (Spe), precision (P), F-measure (F), and geometric mean (G-Mean), which were obtained from the confusion matrix data. The four basic elements of the confusion matrix are true positive (TP), true negative (TN), false positive (FP), and false negative (FN). To assess the effectiveness of the model in dealing with highly unbalanced datasets, we performed a comprehensive analysis of all relevant performance parameters for a typical classification task. The equations representing Acc, Sen, Spe, P, F, and G-Mean are represented by Eqs. (2-7).

$$\text{Accuracy (Acc)} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$\text{Sensitivity (Sen)} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{Specificity (Spe)} = \frac{TN}{TN + FP} \tag{4}$$

$$\text{Precision (P)} = \frac{TP}{TP + FP} \tag{5}$$

$$F\text{-}measure\ (F) = \frac{2 \times Sensitivity \times Precision}{Sensitivity + Precision} \tag{6}$$

$$G - Mean = \sqrt{\frac{TP}{TP+FN} x \frac{TN}{TN + FP}} \tag{7}$$

4.3 Building MLP Models

The Artificial Neural Network (ANN) is a machine learning algorithm that mimics the learning and reasoning ability of the human brain by processing and analyzing large datasets (Ileberi et al., 2022; Alkhalili et al., 2021). The Multilayer Perceptron (MLP) is a type of neural network architecture that comprises an input layer, one or more hidden layers, and an output layer Kasasbeh et al. (2022). The number of neurons in the input and output layers is determined by the number of input and output variables, respectively. The MLP utilizes interconnected layers of perceptrons that identify linearly separable features of the input data. These perceptrons produce outputs that are combined to generate the final output of the network (Kasasbeh et al., 2022; Jiang et al., 2021). To construct a MLP neural network, the initial step was to establish the appropriate number of layers and neurons in each layer. The number of hidden neurons was determined by applying Eq. (8).

$$\text{Nc} = \frac{N_p + N_0}{2} \tag{8}$$

where Nc is the number of neuron nodes in the current hidden layer, Np is the number of neuron nodes in the previous layer, and No is the number of neuron nodes in the output layer. Moreover, the number of hidden layers was determined following the approach proposed in Kasasbeh et al. (2022), which suggests that the number of nodes in the last hidden layer should be greater than the number of nodes in the output layer.

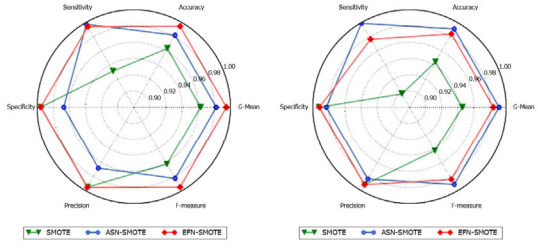
In this study, we evaluated the performance of MLP models with up to four hidden layers, using both a linear activation function (Identity) and a non-linear activation function (Tanh). We conducted experiments on 69 MLP models to test the proposed EFN-SMOTE method and compare it with the original SMOTE (Chawla et al., 2002) and ASN-SMOTE (Yi et al., 2022), as discussed in the following section. Table 2 summarizes the number of MLP models used for this evaluation.

Table 2
Number of MLP models in each layer

Number of hidden layers	1	2	3	4
Number of MLP models	3	9	27	30

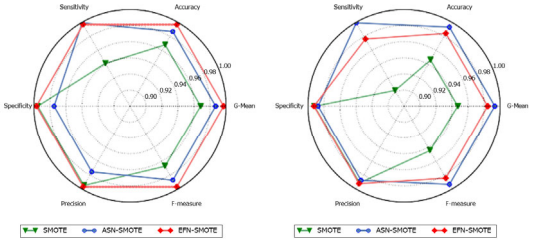
4.4 Results and discussion

This section presents the experimental results of applying EFN-SMOTE to the dataset Dal Pozzolo et al. (2015) and comparing it with ASN-SMOTE (Yi et al., 2022) and the original SMOTE (Chawla et al., 2002). Fig. 3 shows the results for the best ANN framework with one hidden layer, where EFN-SMOTE demonstrated superior accuracy compared to the other models. Figure 3a shows the results using Tanh activation function, where EFN-SMOTE achieved significantly higher accuracy than ASN-SMOTE and original SMOTE with accuracy values of 0.995, 0.982, and 0.963, respectively. The precision ratio for fraud cases confirmed this, which was 0.996, 0.968, and 0.995 for EFN-SMOTE, ASN-SMOTE, and original SMOTE, respectively. In terms of sensitivity ratio (representing the performance measurement to detect the fraud cases), EFN- SMOTE and ASN-SMOTE exhibited convergence. As for the specificity ratio (representing the performance measurement to detect normal cases), EFN- SMOTE and original SMOTE showed convergence. Moreover, F-measure and G-means were significantly higher in EFN-SMOTE compared to ASN-SMOTE and SMOTE. This can be attributed to the noise filtering process performed on each cluster individually, which prevented the synthesis of new minority instances in the majority class region. It is also worth noting that EFN- SMOTE exhibited superior performance because noise filtering within each cluster was more precise than when applied to all data.



(a) Tanh activation Function (b) Identity activation Function

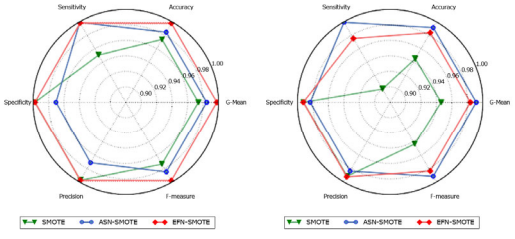
Fig. 3. ANN with one hidden Layer



(a) Tanh activation Function (b) Identity activation Function

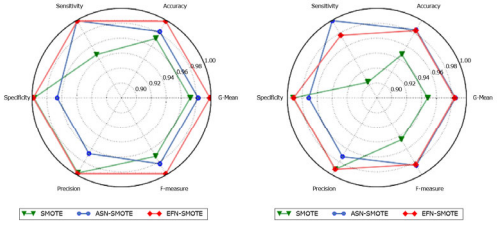
Fig. 4. ANN with two hidden Layers

In contrast, the use of the linear activation function called Identity function resulted in the superiority of the ASN-SMOTE over the EFN-SMOTE, as shown in Figure 3b, by approximately the same proportion as the superiority of the EFN-SMOTE over the ASN-SMOTE when using the Tanh function. As noted by Sharma (2020), the activation function in an ANN allows for non-linear transformation of inputs, enabling it to learn and perform more complex tasks. Our experimental results confirmed this, as the use of the Tanh function clearly outperformed the Identity function. Fig. 4 to Fig. 6 depict the performance comparison of the models with two, three, four, and five hidden layers, respectively, as previously described. The results demonstrate an increase in accuracy with each additional hidden layer. For example, using the non-linear Tanh function, the neural network with one hidden layer achieved an accuracy of 0.995. The accuracy improved to 0.997 in the neural network with two hidden layers, and then continued to rise to 0.998 and 0.999 in three and four hidden layers, respectively. However, the accuracy results slightly decreased to 0.998 when using five hidden layers.



(a) Tanh activation Function (b) Identity activation Function

Fig. 5. ANN with three hidden Layers



(a) Tanh activation Function (b) Identity activation Function

Fig. 6. ANN with four hidden Layers

5. Conclusion

This study proposes a novel oversampling technique, EFN-SMOTE, which utilizes a new SMOTE framework to enhance the classification performance of credit card fraud detection. The proposed method involves partitioning the data into clusters using the Fuzzy C-means algorithm and the Elbow method. This approach incorporates noise filtering and adaptive neighbor selection mechanisms for each cluster to remove any minority instance in proximity to the majority class. In the synthetic oversampling process, the method selects the minority neighbors closest to the nearest majority instance. The experimental results indicate that the EFN-SMOTE technique outperforms existing techniques by 1.0%- 5.5% in classification performance, as evaluated by ANN. This demonstrates the effectiveness of EFN-SMOTE in improving the classification accuracy of imbalanced data and its potential in enhancing machine learning algorithms for classifying data.

References

Ahmad, H., Kasasbeh, B., Aldabaybah, B., & Rawashdeh, E. (2023). Class balancing framework for credit card fraud detection based on clustering and similarity-based selection (SBS). *International Journal of Information Technology*, *15*(1), 325-333.

Ahmad, S., Jha, S., Alam, A., Yaseen, M., & Abdeljaber, H. A. (2022). A Novel AI-Based Stock Market Prediction Using Machine Learning Algorithm. *Scientific Programming*, *2022*.

Alkhalili, M., Qutqut, M. H., & Almasalha, F. (2021). Investigation of applying machine learning for watch-list filtering in anti-money laundering. *IEEE Access*, *9*, 18481-18496.

Badic, B., Da-Ano, R., Poirot, K., Jaouen, V., Magnin, B., Gagnière, J., ... & Visvikis, D. (2022). Prediction of recurrence after surgery in colorectal cancer patients using radiomics from diagnostic contrast-enhanced computed tomography: a two-center study. *European Radiology*, *32*(1), 405-414.

Barua, S., Islam, M. M., Yao, X., & Murase, K. (2012). MWMOTE--majority weighted minority oversampling technique for imbalanced data set learning. *IEEE Transactions on knowledge and data engineering*, *26*(2), 405-425.

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, *6*(1), 20-29.

Bisong, E. (2019). *Building machine learning and deep learning models on Google cloud platform* (pp. 59-64). Berkeley, CA: Apress.

Bunkhumpornpat, C., Sinapiromsaran, K., & Lursinsap, C. (2009). Safe-level-smote: Safe-level-synthetic minority oversampling technique for handling the class imbalanced problem. In *Advances in Knowledge Discovery and Data Mining: 13th Pacific-Asia Conference, PAKDD 2009 Bangkok, Thailand, April 27-30, 2009 Proceedings 13* (pp. 475-482). Springer Berlin Heidelberg.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.

Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015, December). Calibrating probability with undersampling for unbalanced classification. In *2015 IEEE symposium series on computational intelligence* (pp. 159-166). IEEE.

Douzas, G., Bacao, F., & Last, F. (2018). Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. *Information Sciences*, *465*, 1-20.

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2011). A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *42*(4), 463-484.

Han, H., Wang, W. Y., & Mao, B. H. (2005). Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In *Advances in Intelligent Computing: International Conference on Intelligent Computing, ICIC 2005, Hefei, China, August 23-26, 2005, Proceedings, Part I 1* (pp. 878-887). Springer Berlin Heidelberg.

Hordri, N. F., Yuhaniz, S. S., Azmi, N. F. M., & Shamsuddin, S. M. (2018). Handling class imbalance in credit card fraud using resampling methods. *Int. J. Adv. Comput. Sci. Appl*, *9*(11), 390-396.

Ileberi, E., Sun, Y., & Wang, Z. (2022). A machine learning based credit card fraud detection using the GA algorithm for feature selection. *Journal of Big Data*, *9*(1), 1-17.

Japkowicz, N. (2000, June). The class imbalance problem: Significance and strategies. In *Proc. of the Int'l Conf. on artificial intelligence* (Vol. 56, pp. 111-117).

Jiang, Y., Li, C., Sun, L., Guo, D., Zhang, Y., & Wang, W. (2021). A deep learning algorithm for multi-source data fusion to predict water quality of urban sewer networks. *Journal of Cleaner Production*, *318*, 128533.

Kasasbeh, B., Aldabaybah, B., & Ahmad, H. (2022). Multilayer perceptron artificial neural networks-based model for credit card fraud detection. *Indonesian Journal of Electrical Engineering and Computer Science*, *26*(1), 362-373.

Khader, M., Karam, M., & Fares, H. (2021). Cybersecurity Awareness Framework for Academia. *Information*, 12(10), 417.

Kubat, M., & Matwin, S. (1997, July). Addressing the curse of imbalanced training sets: one-sided selection. In *Icml* (Vol. 97, No. 1, p. 179).

Laurikkala, J. (2001). Improving identification of difficult small classes by balancing class distribution. In *Artificial Intelligence in Medicine: 8th Conference on Artificial Intelligence in Medicine in Europe, AIME 2001 Cascais, Portugal, July 1–4, 2001, Proceedings 8* (pp. 63-66). Springer Berlin Heidelberg.

Lebichot, B., Le Borgne, Y. A., He-Guelton, L., Oblé, F., & Bontempi, G. (2020). Deep-learning domain adaptation techniques for credit cards fraud detection. In *Recent Advances in Big Data and Deep Learning: Proceedings of the INNS Big Data and Deep Learning Conference INNSBDDL2019, held at Sestri Levante, Genova, Italy 16-18 April 2019* (pp. 78-88). Springer International Publishing.

Lei, T., Jia, X., Zhang, Y., He, L., Meng, H., & Nandi, A. K. (2018). Significantly fast and robust fuzzy c-means clustering algorithm based on morphological reconstruction and membership filtering. *IEEE Transactions on Fuzzy Systems*, *26*(5), 3027-3041.

Li, H., Zou, P., Wang, X., & Xia, R. (2013). A new combination sampling method for imbalanced data. In *Proceedings of 2013 Chinese Intelligent Automation Conference: Intelligent Information Processing* (pp. 547-554). Springer Berlin Heidelberg.

Mishra, A., & Ghorpade, C. (2018, February). 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)* (pp. 1-5). IEEE.

Nainggolan, R., Perangin-angin, R., Simarmata, E., & Tarigan, A. F. (2019, November). Improved the performance of the K-means cluster using the sum of squared error (SSE) optimized by using the Elbow method. In *Journal of Physics: Conference Series* (Vol. 1361, No. 1, p. 012015). IOP Publishing.

Pan, T., Zhao, J., Wu, W., & Yang, J. (2020). Learning imbalanced datasets based on SMOTE and Gaussian distribution. *Information* Sciences, 512, 1214-1233.

Prasetiyo, B., Muslim, M. A., & Baroroh, N. (2021, June). Evaluation performance recall and F2 score of credit card fraud detection unbalanced dataset using SMOTE oversampling technique. In *Journal of Physics: Conference Series* (Vol. 1918, No. 4, p. 042002). IOP Publishing.

Ramentol, E., Caballero, Y., Bello, R., & Herrera, F. (2012). Smote-rs b*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using smote and rough sets theory. *Knowledge and information systems*, *33*, 245-265.

Sáez, J. A., Luengo, J., Stefanowski, J., & Herrera, F. (2015). SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering. *Information Sciences*, *291*, 184-203.

Santoso, N., Wibowo, W., & Himawati, H. (2019). Integration of synthetic minority oversampling technique for imbalanced class. *Indonesian Journal of Electrical and Engineering Computation Sciences*, *13*(1), 102-108.

Sharma, O. (2020). A novel activation function in convolutional neural network for image classification in deep learning. In *Data Science and Analytics: 5th International Conference on Recent Developments in Science, Engineering and Technology, REDSET 2019, Gurugram, India, November 15–16, 2019, Revised Selected Papers, Part I 5* (pp. 120-130). Springer Singapore.

Tantithamthavorn, C., Hassan, A. E., & Matsumoto, K. (2018). The impact of class rebalancing techniques on the performance and interpretation of defect prediction models. *IEEE Transactions on Software Engineering*, *46*(11), 1200-1219.

Torgo, L., Ribeiro, R. P., Pfahringer, B., & Branco, P. (2013). Smote for regression. In *Progress in Artificial Intelligence: 16th Portuguese Conference on Artificial Intelligence, EPIA 2013, Angra do Heroísmo, Azores, Portugal, September 9-12, 2013. Proceedings 16* (pp. 378-389). Springer Berlin Heidelberg.

Tran, T. C., & Dang, T. K. (2021, January). Machine learning for prediction of imbalanced data: Credit fraud detection. In *2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM)* (pp. 1-7). IEEE.

Verbiest, N., Ramentol, E., Cornelis, C., & Herrera, F. (2014). Preprocessing noisy imbalanced datasets using SMOTE enhanced with fuzzy rough prototype selection. *Applied Soft Computing*, *22*, 511-517.

Yen, S. J., & Lee, Y. S. (2006). Under-sampling approaches for improving prediction of the minority class in an imbalanced dataset. In *Intelligent Control and Automation: International Conference on Intelligent Computing, ICIC 2006 Kunming, China, August 16–19, 2006* (pp. 731-740). Springer Berlin Heidelberg.

Yi, X., Xu, Y., Hu, Q., Krishnamoorthy, S., Li, W., & Tang, Z. (2022). ASN-SMOTE: a synthetic minority oversampling method with adaptive qualified synthesizer selection. *Complex & Intelligent Systems*, *8*(3), 2247-2272.

Zou, H. (2021, February). Analysis of best sampling strategy in credit card fraud detection using machine learning. In *2021 6th International Conference on Intelligent Information Technology* (pp. 40-44).