IPTEK The Journal of Technology and Science, 34(2), 2023 (e/pISSN:2088-2033, 0853-4098)

DOI: 10.12962/j20882033.v34i3.17183 | Received 15 June, 2023; Revised 15 June, 2023; Accepted 27 Juny, 2023

ORIGINAL RESEARCH

ENSEMBLE OVERSAMPLING FOR FINANCIAL FRAUD CLASSIFICATION OF IMBALANCED DATA

Moch Deny Pratama1 | Agus Budi Raharjo* | Diana Purwitasari

Department of Informatics, Institut Teknologi Sepuluh Nopember, Surabaya, , Indonesia

Correspondence

Email: agus.budi@its.ac.id

Present Address

Magister Tower, Jl Informatika No. 10,
Surabaya 60111, Indonesia

Abstract

Financial fraud classification cases such as credit card fraud and bitcoin fraud have highly imbalanced data problems that the oversampling data of fraud class is necessary. Financial transactions could have different attributes. In a credit card transaction, the attributes could represent a nominal amount, transaction period information, the status of deposits or other types like withdrawals or refunds, and more detailed information. In the financial transaction of bitcoin, the attributes could represent the number of nodes, transaction fee, output volume, and aggregated figures. The various characteristics of attributes in financial fraud data require an adaptable oversampling method so that the classification model can perform well. An Ensemble Oversampling method is proposed as a general context approach to handling financial fraud classification in credit cards and bitcoin. The proposed method combines generative with traditional approaches such as GAN, SMOTE, and ADASYN. In the classification step, Deep Learning algorithms such as CNN and LSTM are applied to provide better performance. The genetic algorithm is used to optimize Deep Learning hyperparameters. The evaluation was carried out by comparing four scenarios, i.e., without oversampling, using oversampling with GAN, SMOTE, ADASYN, original data, and Ensemble Oversampling. The combined oversampling of GAN and SMOTE with the CNN classifier model produces the highest evaluation score of all scenarios with an average F1-Score value of 0.995 and Kappa Statistics of 0.990. It shows that augmented data quality does affect prediction performance, and Ensemble Oversampling technique could be considered to improve classifier performance in financial fraud data.

KEYWORDS:

Financial Fraud Classification, Imbalanced Data, Ensemble Oversampling, Deep Learning.

1 | INTRODUCTION

Bank fraud, corporate fraud, insurance fraud, and cryptocurrency fraud are the four categories of financial fraud. Two types of financial fraud that are frequent, widely researched, and characterized by imbalanced data are credit card and bitcoin fraud[1]. Detection of fraud in credit cards is vital to identify fraudulent transactions that can result in significant losses[2]. On the other hand, bitcoin is a cryptocurrency that is still developing in the blockchain network[3]. Bitcoin fraud detection works by identifying losses and illegal transactions, such as ransomware attacks without verification on blockchain technology[4]. One of the approaches used in credit card fraud and bitcoin fraud classification with imbalanced datasets is using Machine Learning algorithms[5]. Learning classifiers such as Logistic Regression, Decision Tree, Support Vector Machine, and Artificial Neural Network are the best methods based on three performance tests, (Accuracy, Sensitivity, and Area Under Precision-Recall Curve (AUPRC))[6]. However, without oversampling, the good Accuracy could be caused by biased data.

Deep Learning techniques can be utilized in the context of financial fraud classification in addition to traditional Machine Learning algorithms[3] [6] [7]. The uses of the Synthetic Minority Oversampling Technique (SMOTE) for handling imbalanced data, Uniform Manifold Approximation and Projection (UMAP) to reduce the dimensionality of the data, and Long Short-term Memory Network (LSTM) as a Deep Learning algorithm showed good results with Accuracy and Recall values of 96.72% and 91.91%. LSTM has a feedback connection between hidden units, enabling linkages from long-term sequences that will be learned from transaction labels to make predictions based on historical transaction sequences in the past[8]. Another study that applied Homogeneity-oriented Behavior Analysis (HOBA) for feature generation and Convolutional Neural Network (CNN) as a Deep Learning algorithm in credit card fraud showed Accuracy, Precision, and Recall of 96.40%, 34.58%, and 69.90%, respectively[9]. This research proposed a feature engineering method without oversampling technique. Applying feature engineering does not guarantee unbiased results even after using the Deep Learning technique, indicated by a high score of Accuracy but low Recall and Precision scores. In the domain of cryptocurrency, Graph Convolutional Network (GCN) was employed in financial forensics and obtained an F1-Score of 70%, with Precision and Recall of 81% and 62%[5]. Other research reviews on GCN and LSTM showed that Deep Learning algorithms perform better than Logistic Regression, Random Forest, and Multilayer Perceptron[10] [11].

Selecting appropriate hyperparameters in Deep Learning can be challenging and affect performance. There are several hyperparameters that need to be optimized, such as filter values, number of units, number of kernels, dropout, optimizer, activation function, and number of epochs. The Genetic Algorithm (GA) is one of the algorithms that can be used to optimize hyperparameters and avoid overfitting bias[12]. GA was employed to optimize the CNN model during the training phase classification problem[13].

Financial fraud classification not only deals with model optimization but also with handling highly imbalanced data. There are three approaches to managing imbalanced data, i.e., oversampling, undersampling, and ensemble sampling[14]. Several studies have been conducted that applied oversampling technique. A previous study that applied an oversampling technique in the imbalanced data improved model prediction Accuracy, Precision, and Recall[15]. Another study that uses LSTM and ensemble classifiers showed that SMOTE increased the AUC score[16]. The undersampling method offers a simple concept and efficient computational complexity. Radial Undersampling, an enhanced method based on SMOTE algorithm, demonstrated a significant decrease in computation time while maintaining good performance[17]. Combining the two sampling techniques is called ensemble sampling[14]. Previous research reduced the false negative rate by using the ensemble random oversampling technique. The trained model with the proposed sampling method predicts more accurately from the fraud class[18].

There are two categories of oversampling based on the technique and technology used, namely traditional and generative approaches[19]. SMOTE and Adaptive Synthetic (ADASYN) are categorized as traditional approaches. On the other hand, Generative Adversarial Networks (GAN) and Variational Autoencoder (VAE) are two examples of generative approaches that use Neural Networks as their computational cores[20]. Compared to traditional approaches, the generative technique is generally superior in creating more realistic data augmentation samples because it requires less human intervention and is adaptive to future work applications on non-image data[21].

An Ensemble Oversampling method is proposed in this study by combining traditional and generative oversampling techniques to handle imbalanced data in financial fraud. This method combines traditional and generative approaches, such as SMOTE, ADASYN, and GAN. The purpose of ensemble oversampling is to take advantage of the strength of different methods to produce

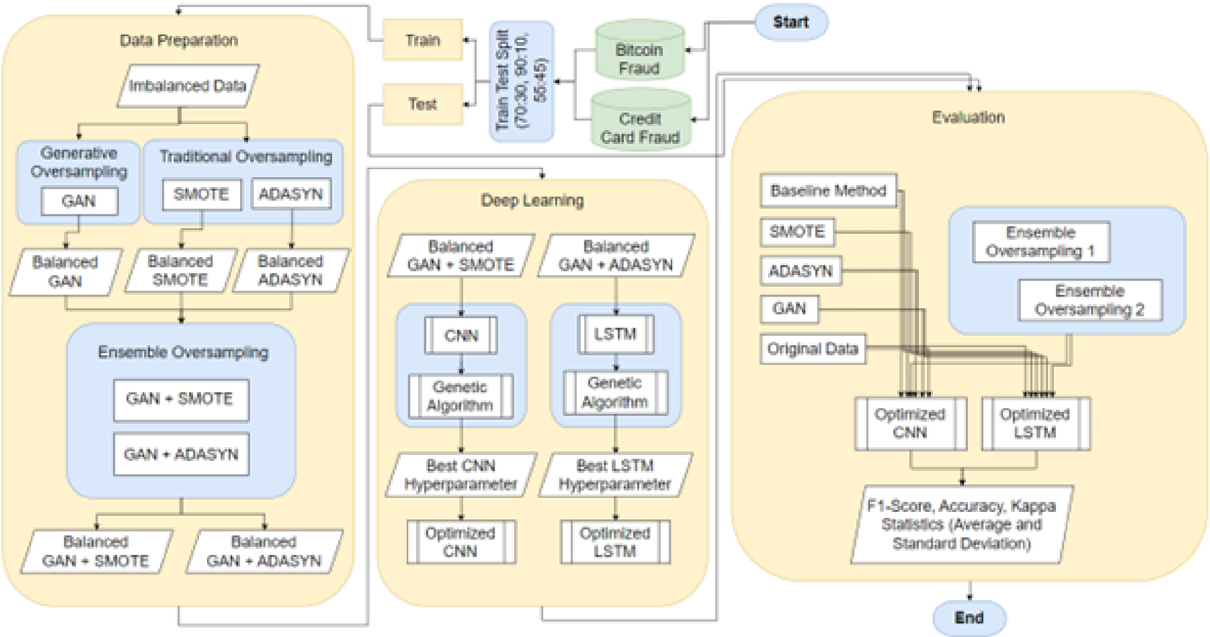

FIGURE 1 Research Methods.

more robust and diverse synthetic data. CNN and LSTM are applied to build fraud classification models. CNN extracts short-term temporal correlations and patterns of user transaction behavior, while LSTM captures long-term temporal activity and spots instances of fraud. Accuracy, F1-Score, and Kappa statistic are used to evaluate the performance of the prediction models as well as to observe the application of the ensemble oversampling method.

2 | PREVIOUS RESEARCHES

Theoretical foundations for the proposed research consist of oversampling techniques and Deep Learning algorithms described in Sections 2.

2.1 | Oversampling Techniques in Financial Fraud

Traditional oversampling such as SMOTE and ADASYN were widely used to handle imbalanced data[14] [15] [16]. On the other hand, with the rise of Deep Learning architecture, the Neural Network approach for oversampling, like Generative Oversampling, provides promising results in generating realistic data augmentation[18] [19] Figure 1 .

2.1.1 | Traditional Oversampling

One of the efficient oversampling techniques, SMOTE, duplicates instances of minority classes to produce new synthetic instances by considering the neighbor's sample in the minority class. SMOTE generates synthetic data in a random position near minority instances, and then the new data class is labeled based on their k nearest neighbors using Euclidean distance[22]. This oversampling algorithm has improved performance for imbalanced data[23].

ADASYN algorithm is the advancement of SMOTE. This method adaptively produces minority class samples by using a weighted distribution. When minority class samples are difficult to generate, additional synthetic data is produced. This method's effectiveness is demonstrated by enhanced data distribution-related training. This method adjusts to restrict classification judgments for complex samples while minimizing bias carried on by classification imbalance[22].

2.1.2 | Generative Oversampling

Generative Adversarial Networks (GAN) is a type of Deep Learning framework involving two main networks: the generator and the discriminator. They compete (adversarial) to produce augmented data[21]. The generator creates samples using a random noise vector from the latent space. The discriminator examines the erroneous of the created samples using input from actual training data[23]. GAN with Deep Learning architecture can be trained on the minority class data to generate synthetic samples close to the original data[20].

2.2 | Deep Learning in Financial Fraud

Deep Learning techniques, such as LSTM[8] and CNN[10], were widely used and performed well in financial fraud classification[5] [15] [23].

2.2.1 | Long-Short Term Memory (LSTM)

Recurrent Neural Networks (RNN), a type of deep neural network mainly utilized for time series data, are expanded into Long Short-Term Memory (LSTM). Unlike the RNN, which uses the current input from the previous hidden state to build a new hidden state, each neuron in the LSTM is a cell with memory that can store information and maintain its state. The input, output, and forget gates in cells increase the capacity of the LSTM memory[8]. Applying LSTM in financial fraud can assess the sequence of transactions as a whole/one entity rather than as individuals[19].

2.2.2 | Convolutional Neural Network (CNN)

Another form of Deep Learning architecture is the Convolutional Neural Network (CNN), which is described as a multi-layer network divided into input, pooling, fully connected, and output layers. Overfitting is reduced by the pooling layer, which makes input easier to handle and lowers the number of calculations and parameters in the network. The fully connected and output layer, which uses a multi-layer perceptron or densely connected feedforward neural network as its input, uses convolutional and pooling functions to categorize data[19]. Concise configuration, which executes a 1D convolution process, allows for the advantage of real-time and low-cost implementation[24].

2.2.3 | Optimization Algorithm

Deep Learning algorithms are more sensitive in selecting hyperparameters which can affect model performance. Therefore, optimization techniques are necessary, which are applied to choose the optimum hyperparameters automatically. Genetic Algorithm (GA) is one of the optimization techniques that can be used to select appropriate hyperparameters that are modeled based on natural evolution, which selects the best option from a group of candidates represented as binary strings called chromosomes[1]. Chromosome refers to a point in the hyperparameter as a potential solution. Iterative development refers to evolution toward the optimum solution. Generation in the evolution of all chromosomes includes the selection of parents, crossover, and mutation. Competition is used to determine the selection of parents[11]. Each solution is represented by a chromosome made from various strings, with a unique fitness value. The fitness value defines how effective and satisfactory the solution's performance outcomes are. The probability of each string in the current generation population is calculated by considering the fitness value to choose parent chromosomes from the population. Crossover creates a new generation by crossing the parent solution and the chromosomes at some time to create children. Changes in the value of particular bits in the chromosomal string caused by mutation frequently introduce variety into the population. This optimization algorithm aims to maximize the fitness value based on the most suitable criteria[12]. The output of the optimization algorithm is to display the most suitable hyperparameters in the population for better performance[25].

2.3 | Performance Evaluation

The challenge of fraud detection is widely considered of highly imbalanced data classification, the confusion matrix and Kappa statistics may be used to assess a classification model's success[26]. The confusion matrix explains how a dataset's ground truth and a shorter model prediction diverge, which can give more detailed information regarding classifier performance. Positive prediction Accuracy is measured by Precision. Recall, often referred to as sensitivity, is the percentage of positive instances that the classifier correctly recognizes. The number of overall accurate predictions made by the model is known as Accuracy. The F1-Score, which is the average of Accuracy and Recall, is a better metric to aim for[20].

2.3.1 | Confusion Matrix

Confusion Matrix are four categories of model outcomes, including Accuracy, Precision, Recall, and F1-Score are performance assessment measures displayed in classification report[7] [27] [28]. The following list includes some of the performance calculation equations applied in the model. The F1-Score metric is the average harmonic value of each prediction class by Equation 1 2 3 4.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

2.3.2 | Kappa Statistics

Kappa Statistics is used to evaluate the suitability of the classification results. The results of the classification are considered very agreeable if the Kappa score is 0.85, 0.6-0.85 indicates agree, and 0.45-0.6 indicates moderate, <0.45 is considered a bad agreement. The Kappa value is used to determine the performance of the classifier[29]. Kappa is based on the agreement coefficient K corrected for opportunities with the following formula by equation 5.

$$K = \frac{P_o - P_e}{1 - P_e} \tag{5}$$

Where Pe is the percentage of observed agreements that are predicted by chance and Po is the overall frequency of observed agreements. The degree that occurs when the observed frequency of agreement exceeds the frequency of agreement Pe, as would be predicted based on random classification, is indicated by the letter K. With reference to the qualitative description of agreement strength, the relative strength of agreement associated to the Kappa statistic was evaluated.

3 | MATERIAL AND METHOD

The proposed method consists of two phases, data preparation, and financial fraud classification, as shown in Figure 1 . Credit card and bitcoin fraud are used as the case study, containing imbalanced data. In the data preparation phase, imbalanced data is handled using the Generative Oversampling GAN algorithm combined with the traditional SMOTE or ADASYN. The combination is proposed to generate new data on minority classes (fraud). After the data is balanced for fraud and non-fraud classes, the classification phase is conducted using CNN and LSTM.

3.1 | Data Collection

The first dataset contains 284,807 transaction records from credit card activities. The data have 31 features in the form of numeric, which are the results of the PCA transformation. Due to security and confidentiality issues, original data with native raw features and more background knowledge is not provided[1]. The second dataset is the collection of anonymous transactions from the Bitcoin blockchain. The data provided were also transformed to maintain security and intellectual property. Thus, the description of 165 attributes is not available. Several features represent local information about transactions, including the time steps, input/output amounts, transaction fees, output volumes, and aggregate figures such as the average bitcoin received by input/output and the average number of incoming transactions related with a total of 203,769 transactions[3].

3.2 | Data Preparation: Oversampling

The first challenge to prepare the gathered data is balancing the class instances. Three oversampling methods are applied, i.e., GAN, SMOTE and ADASYN. GAN is a generative oversampling technique that involves generating entirely new instances using Deep Learning as the approach. This technique has the advantage of producing more realistic and diverse synthetic data than traditional oversampling, which could improve generalization performance. However generative can also be computationally expensive. The generative model's input is random noise z, which is processed by generator G and discriminator D. Generator is used to generate new samples. The discriminator learns to better distinguish between the real and generated examples by minimizing its prediction errors as shown in equation 6[18] [19].

$$Kx'_{\text{GAN}} = \min_{\theta_G} \max_{\theta_D} \left(\mathbb{E}_{xP_d}[\log D(x)] + \mathbb{E}_{zP_z}[\log(1 - D(G(z)))] \right) \tag{6}$$

where, θG is parameter of the generator, θD is parameter of the discriminator,Pd is the data distribution, and $P\zeta$ is the prior data distribution. SMOTE algorithm generates synthetic data based on a k-Nearest Neighbor to increase the number of minority class samples to be equal to the majority class at random observations. The synthetic sample quality is calculated using Eq. 7 by considering the distance between the two observation vectors, weight (w) and distance (μ),.

$$\beta(x, y) = w_x w_y \sum_{i=1}^{N} \mu(x_i, y_i)^{\rho} \tag{7}$$

$\beta(x, y)$ represents the observation distance between vectors X and Y, $w_x w_y$ is the weight of distance, N is the number of predictors, and ϕ indicates which distance algorithm used for measuring the proximity between each instance. According to Eq. 8[26], the value of $\mu(x_i, y_i)$ represents the distance between X and Y vector observations for each explanatory variable. Manhattan Distance is used for categorical data, indicated by $\phi = 1$, while Euclidean Distance for numerical data $\phi = 2$.

$$x'_{\text{SMOTE}} = \mu(x_i, y_i) = \sum_{i=1}^{n} \left| \frac{j_{1i}}{j_1} - \frac{j_{2i}}{j_2} \right|^{\theta} \tag{8}$$

n represents the number of class categories in the first variable, j_1 is the number of values for the first category in each x_i occurrence, $j_1 i$ is the number for the first category in each x_i that belongs to the i-th class, j_2 is the number of values for the second category in each y_i occurrence, $j_2 i$ is the number for the second category in each y_i that belongs to the i-th class, and variable θ is a constant value that is typically set to 1. ADASYN algorithm synthesizes samples by determining the weight size for each minority class sample xi as shown in Eq. 9.

$$x'_{\text{ADASYN}} = x_i + (x_{xi} - x_i) \times \theta \tag{9}$$

xi represents the minority class examples for each sample of data, xxi represents the minority data chosen from the k-Nearest Neighbor on data xi. xxi-xi indicates the difference between the raw and synthetic data, and the $theta$ variable represents a random value of $\theta \epsilon 0, 1$.

The proposed ensemble oversampling combines two oversampling techniques. This process is written through Eq. 10. This technique, which combines generative and traditional approaches, can produce a greater variety of data results. The combined data will create a larger pool of information and perhaps increase the diversity of the generated data. It can introduce more variation while preserving the integrity and characteristics of both datasets.

$$x'_{\text{Ens}} = \left\{ x \,\middle|\, x \in \left(X_{\text{GAN}} \cup X_{\text{trad}} \right),\ X_{\text{trad}} \in \left\{ X_{\text{SMOTE}}, X_{\text{ADASYN}} \right\} \right\} \tag{10}$$

Apart from dealing with imbalanced data, financial fraud with balanced data is also investigated in this study. The aim is to compare whether classification using synthetic instances (from imbalanced data) can give as good results as the original data. On the other hand, on the original dataset, getting balanced data is very rare and difficult. Bitcoin dataset is used by taking 10,000 samples as starting imbalanced data from 203,769 rows. The scenario is started by splitting data D into D_Train and D_Test. For example, given the ratio 70%:30% and 10,000 imbalanced data, |D_Train|is 7,000 and |D_Test| is 3,000. D_Test is not oversampled as test data, while D_Train is added from the pool of 193,769 samples until the ratio of |D_Train^0 |:|D_Train^1 | is 1:1. Hyperparameter optimization in classification phase is carried out using the Genetic Algorithm (GA). This phase aims to enhance the model's performance on a validation set while preventing overfitting on the training set. In this study, selection of the best parameters in the search process is based on the number of populations and generations as many as 5 generations and 50 populations, that produce chromosomes in the form of parameter candidates with the highest fitness values. The parameter with the highest fitness value indicates that the solution given is good according to the performance results.

4 | RESULT AND DISCUSSION

This section discusses the experimental setup and the evaluation results carried out, described in Sections 4 as follows.

4.1 | Experiment Setup

Performance measurements were done by applying Accuracy, Precision, Recall, F1-Score, and Kappa Statistics. Train and test data were divided using the hold-out method, which contains three splits of train: test ratio, i.e., 55:45, 70:30, and 90:10. In each split, oversampling is carried out on the train data, which is then evaluated using the test data. Accuracy metrics are used as the basis for evaluating model performance in general. The Kappa metric is employed to anticipate biased results from accuracy. The mean and standard deviation of three splits were used as the performance score.

1. Twelve combination scenarios were proposed for credit card fraud according to oversampling and classification algorithms. While in bitcoin fraud, fourteen scenarios were conducted. Details of the scenarios are described as follows:
2. CNN and LSTM without oversampling were used as two baseline models.
3. Single oversampling (SMOTE, ADASYN, GAN) combined with Deep Learning algorithms (CNN, LSTM) constructed six models.
4. Ensemble Oversampling (GAN+SMOTE, GAN+ADASYN) combined with CNN and LSTM created four combinations.
5. Two more models in bitcoin fraud were generated from balanced train data combined with CNN and LSTM.

4.2 | Evaluation Results

Three methods of evaluation metrics are applied in these sections based on metrics referred to: Eq. 1, Eq. 4, and Eq. 5, i.e., F1-Score, Accuracy, and Kappa Statistics shown in Table 1 . In the case of imbalanced data, the low F1-Score is can be caused by the high Precision value and the low Recall value (the ability of the model to correctly predict minority classes/fraud), because this metric uses the average of Precision and Recall harmonics. The good accuracy score might also reflect the biased outputs. Therefore, the Kappa metric is essential to minimize the biased result from accuracy by considering fraud and non-fraud predictions equally. Kappa Statistics is employed initially to measure the confidence level of prediction. Each metric evaluation result represents the average results/mean (high score means good performance) and Standard Deviation (StD) values (low score means good performance) of the distribution data of three split train test results.

In the case of credit card fraud data, GAN oversampling exhibits superior performance compared to traditional approaches, as shown by the higher scores in GAN-CNN, Ensemble methods, and GAN-LSTM than SMOTE or ADASYN models. It demonstrates that GAN is capable of generating simulated instance data that closely resembles the test data, surpassing other methods. The Ensemble Oversampling combined with CNN models yielded the highest scores, with an average F1-Score of 0.995, average Accuracy of 0.996, and average Kappa of 0.990. Meanwhile, the Ensemble Oversampling in combination with LSTM models achieved the highest F1-Score of 0.972 and highest Kappa of 0.945 on average. These findings indicate that the proposed Ensemble Oversampling technique is effective in generating a diverse and realistic set of training instances Table 1 .

TABLE 1 Evaluation Results of Data and Method Algorithm

No	lgorithm	F1-Score	F1-Score	Accuracy	Accuracy	Kappa	Kappa
		Mean	StD	Mean	StD	Mean	StD
	Data Credit Card Fraud						
1	Baseline 1D CNN	0.906	0.007	0.992	0.001	0.901	0.008
2	SMOTE Oversampling CNN	0.865	0.015	0.988	0.002	0.858	0.016
3	ADASYN Oversampling CNN 0.874	0.030	0.989	0.001	0.868	0.031	
4	GAN Oversampling CNN	0.985	0.008	0.991	0.005	0.980	0.007
5	Ensemble Oversampling GAN + SMOTE CNN	0.995	0.001	0.995	0.001	0.990	0.002
6	Ensemble Oversampling GAN + ADASYN CNN	0.994	0.001	0.994	0.001	0.989	0.001
7	Baseline LSTM	0.880	0.011	0.990	0.002	0.875	0.012
8	SMOTE Oversampling LSTM	0.872	0.022	0.989	0.001	0.867	0.022
9	ADASYN Oversampling LSTM	0.870	0.017	0.988	0.001	0.864	0.017
10	GAN Oversampling LSTM	0.951	0.008	0.945	0.014	0.933	0.008
	Ensemble Oversampling						
11	Ensemble Oversampling GAM + SMOTE + LSTM	0.971	0.005	0.972	0.004	0.943	0.009
12	Ensemble Oversampling GAN + ADASYN + LSTM	0.972	0.004	0.973	0.004	0.945	0.008
	Data Bitcoin Fraud						
1	Baseline CNN	0.840	0.015	0.970	0.003	0.823	0.017
2	SMOTE Oversampling CNN	0.803	0.018	0.959	0.006	0.780	0.022
3	ADASYN Oversampling CNN	0.765	0.032	0.948	0.007	0.736	0.035
4	Balanced Train Data CNN	0.853	0.024	0.974	0.004	0.839	0.026
5	GAN Oversampling CNN	0.937	0.012	0.952	0.017	0.900	0.021
6	Ensemble Oversampling GAN +SMOTE CNN	0.954	0.021	0.957	0.019	0.913	0.037
7	Ensemble Oversampling GAN + ADASYN CNN	0.948	0.020	0.947	0.021	0.894	0.042
8	Baseline LSTM	0.806	0.018	0.965	0.003	0.787	0.019
9	SMOTE Oversampling LSTM	0.709	0.038	0.931	0.018	0.671	0.047
10	ADASYN Oversampling LSTM	0.673	0.034	0.917	0.018	0.629	0.042
11	Balanced Train Data LSTM	0.767	0.014	0.960	0.003	0.745	0.015
12	GAN Oversampling LSTM	0.909	0.001	0.965	0.019	0.874	0.004
13	Ensemble Oversampling GAN + SMOTE LSTM	0.943	0.005	0.942	0.005	0.885	0.011
14	Ensemble Oversampling GAN + ADASYN LSTM	0.944	0.007	0.943	0.005	0.886	0.010

Similarly, in the case of bitcoin fraud data, GAN oversampling also exhibits superior performance compared to traditional approaches. The Ensemble Oversampling in combination with CNN models yielded the highest F1-Score and Kappa metrics, with an average F1-Score of 0.954 and average Kappa of 0.913. On the other hand, the highest Accuracy metric was achieved using the balanced original data approach, with an average of 0.974. This suggests that the original balanced data performs well in terms of accuracy. The Ensemble Oversampling combined with LSTM models yielded the highest F1-Score and Kappa metrics, with an average F1-Score of 0.944 and average Kappa of 0.886. Another interesting finding reveals that the baseline CNN (without oversampling) achieves better F1-Score, Accuracy, and Kappa scores compared to SMOTE-CNN, ADASYN-CNN, while the baseline LSTM also outperforms SMOTE-LSTM and ADASYN-LSTM in terms of performance. This suggests that oversampling does not guarantee improved results, particularly when employing Deep Learning classification methods.

In the context of classification, CNN utilizes maximum pooling to learn features, while LSTM processes data sequentially through gates. The CNN method offers the advantage of automatically extracting crucial features from each data segment, making it more efficient in terms of computational memory and complexity. On the other hand, LSTM benefits from its ability to process output data as input, allowing for iterative learning. Generally, the CNN algorithm outperforms LSTM, particularly when the data is transformed numerically and lacks a strong sequential relationship, as indicated in Financial Fraud dataset. Moreover, the evaluation results indicate that the quality of augmented or generated data can significantly impact the performance of Deep Learning models. When using the same classification algorithm but different simulation or generated data, the resulting performance will also vary.

5 | CONCLUSION

Using a combination of oversampling techniques and optimized Deep Learning algorithms has been found to be effective in addressing imbalanced data in Financial Fraud classification. In this study, various scenario methods were employed, specifically twelve combination scenarios for credit card fraud and fourteen scenarios for bitcoin fraud, considering different oversampling and classification algorithms. An approach called Ensemble Oversampling was introduced, which combines multiple oversampling methods. The results demonstrated that the proposed method outperforms relying solely on a single oversampling technique given several scenarios. In future work, the consideration of weighted instances and the certainty level of the generated oversampling data will be explored. Also, it is recommended to explore other imbalanced financial fraud datasets and consider multi-class classification scenarios.

ACKNOWLEDGMENT

CREDIT

Moch Deny Pratama: Conceptualization, methodology, formal analysis, investigation, writing—original draft and editing. Agus Budi Raharjo: Supervision, methodology, validation, formal analysis, writing—review and editing, funding acquisition. Diana Purwitasari: Supervision, methodology, formal analysis, writing—a review and editing.

References

1. Al-Hashedi KG, Magalingam P. Financial fraud detection applying data mining techniques: A comprehensive review from 2009 to 2019. Computer Science Review 2021;40. https://www.sciencedirect.com/science/article/pii/S1574013721000423?via%3Dihub.

2. Kim E, et al. Champion-challenger analysis for credit card fraud detection: Hybrid ensemble and Deep Learning. Expert Systems with Applications 2019;128:214–224. https://www.sciencedirect.com/science/article/pii/S0957417419302167?via%3Dihub.

3. Liu XF, Jiang XJ, Liu SH, Tse CK. Knowledge Discovery in Cryptocurrency Transactions: A Survey. IEEE Access 2021;9(2):37229–37254. https://www.sciencedirect.com/science/article/pii/S0957417419302167?via%3Dihub.

4. Weber M, et al. Anti-Money Laundering in Bitcoin: Experimenting with Graph Convolutional Networks for Financial Forensics. arXiv preprint arXiv:190802591 2019;1(10).

5. Nicholls J, Kuppa A, Le-Khac NA. Financial cybercrime: A comprehensive survey of Deep Learning approaches to tackle the evolving financial crime landscape. IEEE Access 2021;9:163965–163986. https://ieeexplore.ieee.org/document/9642993.

6. Makki S, Assaghir Z, Taher Y, Haque R, Hacid MS, Zeineddine H. An Experimental Study With Imbalanced Classification Approaches for Credit Card Fraud Detection. IEEE Access 2019;p. 93010–93022. https://ieeexplore.ieee.org/document/8756130.

7. Benchaji I, Douzi S, Ouahidi BE, Jaafari J. Enhanced credit card fraud detection based on attention mechanism and LSTM deep model. Journal of Big Data 2021;8(1). https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00541-8.

8. Zhang X, Han Y, Xu W, Wang Q. A novel sensitive staphylococcal enterotoxin C1 fluoroimmunoassay based on functionalized fluorescent core-shell nanoparticle labels. Food Chemistry 2007;105:1623–1629. https://www.sciencedirect.com/science/article/pii/S0308814607003135?via%3Dihub.

9. Alarab I, Prakoonwit S, Nacer MI. Competence of graph convolutional networks for anti-money laundering in bitcoin blockchain. In: Proceedings of the ACM International Conference Proceeding Series July; 2020. p. 23–27. https://dl.acm.org/doi/10.1145/3409073.3409080.

10. Xia P, Ni Z, Xiao H, Zhu X, Peng P. A Novel Spatiotemporal Prediction Approach Based on Graph Convolution Neural Networks and Long Short-Term Memory for Money Laundering Fraud. Arabian Journal for Science and Engineering 2022;47(2):1921–1937. https://link.springer.com/article/10.1007/s13369-021-06116-2.

11. Tani L, Rand D, Veelken C, Kadastik M. Evolutionary algorithms for hyperparameter optimization in Machine Learning for application in high energy physics. European Physical Journal C 2021;81(2):1–9. https://link.springer.com/article/10.1140/epjc/s10052-021-08950-y.

12. Hassan MR, Ismail WN, Chowdhury A, Hossain S, Huda S, Hassan MM. A framework of genetic algorithm-based CNN on multi-access edge computing for automated detection of COVID-19. Journal of Supercomputing 2022;78(7):10250–10274. https://link.springer.com/article/10.1007/s11227-021-04222-4.

13. Shamsudin H, Yusof UK, Jayalakshmi A, Khalid MNA. Combining oversampling and undersampling techniques for imbalanced classification: A comparative study using credit card fraudulent transaction dataset. In: Proceedings of the IEEE International Conference on Control and Automation (ICCA), vol. 2020-Octob; 2020. p. 803–808. https://ieeexplore.ieee.org/document/9264517.

14. Ishaq A, et al. Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques. IEEE Access 2021;9:39707–39716. https://ieeexplore.ieee.org/document/9370099.

15. Shen F, Zhao X, Kou G, Alsaadi FE. A new Deep Learning ensemble credit risk evaluation model with an improved synthetic minority oversampling technique. Applied Soft Computing 2021;98:106852. https://www.sciencedirect.com/science/article/pii/S1568494620307900?via%3Dihub.

16. Koziarski M. Radial-Based Undersampling for imbalanced data classification. Pattern Recognition 2020;102. https://www.sciencedirect.com/science/article/pii/S0031320320300674?via%3Dihub.

17. Huda S, et al. An Ensemble Oversampling Model for Class Imbalance Problem in Software Defect Prediction. IEEE Access 2018;6:24184–24195. https://www.sciencedirect.com/science/article/pii/S0031320320300674?via%3Dihub.

18. Hilal W, Gadsden SA, Yawney J. Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances. Expert Systems with Applications 2022;193:116429. https://www.sciencedirect.com/science/article/pii/S0957417421017164?via%3Dihub.

19. Engelmann J, Lessmann S. Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced learning. Expert Systems with Applications 2021;174(December 2020):114582. https://www.sciencedirect.com/science/article/pii/S0957417421000233?via%3Dihub.

20. Fajardo VA, et al. On oversampling imbalanced data with deep conditional Generative models. Expert Systems with Applications 2021;169:114463. https://www.sciencedirect.com/science/article/pii/S0957417420311155?via%3Dihub.

21. Tran TC, Dang TK. Machine Learning for Prediction of Imbalanced Data: Credit Fraud Detection. In: 2021 15th International Conference on Ubiquitous Information Management and Communication (IMCOM); 2021. p. 1–7. https://ieeexplore.ieee.org/document/9377352.

22. Mqadi N, Naicker N, Adeliyi T. A SMOTe based Oversampling Data-Point Approach to Solving the Credit Card Imbalanced data Problem in Financial fraud Detection. International Journal of Computing and Digital Systems 2021;1(1). https://journal.uob.edu.bh/items/15251cc8-0e71-4121-ad36-f858990a715b.

23. Dong S, Wang P, Abbas K. A survey on Deep Learning and its applications. Computer Science Review 2021;40:100379. https://www.sciencedirect.com/science/article/pii/S1574013721000198?via%3Dihub.

24. Kiranyaz S, Avci O, Abdeljaber O, Ince T, Gabbouj M, Inman DJ. 1D convolutional neural networks and applications: A survey. Mechanical Systems and Signal Processing 2021;151:107398. https://www.sciencedirect.com/science/article/pii/S0888327020307846?via%3Dihub.

25. Bakhashwain N, Sagheer A. Online Tuning of Hyperparameters in Deep LSTM for Time Series Applications. International Journal of Intelligent Engineering and Systems 2020;14(1):212–220. https://www.inass.org/2021/2021022821.pdf.

26. Vitianingsih AV, Othman Z, Baharin SSK, Suraji A, Maukar AL. Application of the Synthetic Over-Sampling Method to Increase the Sensitivity of Algorithm Classification for Class Imbalance in Small Spatial Datasets. International Journal of Intelligent Engineering and Systems 2022;15(5):676–690. https://inass.org/wp-content/uploads/2022/06/2022103158-2.pdf.

27. Pratama MD, Sarno R, Abdullah R. Sentiment Analysis User Regarding Hotel Reviews by Aspect Based Using Latent Dirichlet Allocation, Semantic Similarity, and Support Vector Machine Method. International Journal of Intelligent Engineering and Systems 2022;15(3):514–524. https://dx.doi.org/10.22266/ijies2022.0630.43.

28. Narayan V, Ganapathisamy S. Hybrid Sampling and Similarity Attention Layer in Bidirectional Long Short Term Memory in Credit Card Fraud Detection. International Journal of Intelligent Engineering and Systems 2022;15(6):35–44. https://inass.org/wp-content/uploads/2022/06/2022103158-2.pdf.

29. Wang L, et al. Multi-classifier-based identification of COVID-19 from chest computed tomography using generalizable and interpretable radiomics features. European Journal of Radiology 2021;136:109552. https://www.sciencedirect.com/science/article/pii/S0720048X21000322?via%3Dihub.

How to cite this article: Moch DP., Diana P., Agus BR., ENSEMBLE OVERSAMPLING FOR FINANCIAL FRAUD CLASSIFICATION OF IMBALANCED DATA , *IPTEK The Journal of Technology and Science*, .