BEST
Journal of Applied Electrical & Science Technology – University of PGRI Adi Buana Surabaya
p-ISSN 2715-2871
e-ISSN 2714-5247

Handling Imbalanced Fraudulent Transaction Data Using SMOTE-Tomek and Random Forest: A Classification Approach

Mohamad Ilham
Electrical Engineering
University of PGRI Adi Buana
Surabaya, Indonesia
mohamadilham@unipasby.ac.id

Adi Winarno
Electrical Engineering
University of PGRI Adi Buana
Surabaya, Indonesia
adiwinarno@unipasby.ac.id

Moch. Lutfi
Informatics Engineering
Yudharta University
Pasuruan, Indonesia
moch.lutfi@yudharta.ac.id

Artanti Indrasetianingsih
Statistics
University of PGRI Adi Buana
Surabaya, Indonesia
artanti.indra@unipasby.ac.id

***Abstract*— This research aims to address the class imbalance problem in fraud detection using hybrid resampling techniques, specifically SMOTE-Tomek, combined with Random Forest classifiers. Imbalanced data in fraud detection tasks can severely hinder model performance, resulting in poor detection of minority (fraud) cases. By employing SMOTE to oversample minority class instances and Tomek links to clean the borderline majority class samples, this study evaluates the effectiveness of this hybrid method in improving classification metrics. Using a benchmark credit card fraud dataset, we compare the performance of Random Forest models with and without the hybrid sampling approach. The experimental results show that SMOTE-Tomek significantly enhances recall and F1-score without sacrificing accuracy. This finding underscores the importance of using appropriate resampling strategies for improving model robustness in fraud detection.**

***Keywords*—**Fraud Detection, Imbalanced Data, SMOTE-Tomek, Random Forest, Clasisification

I. INTRODUCTION

In the modern digital era, cashless transactions—particularly credit card usage—have become an integral part of daily financial activities. However, the convenience of digital payments is accompanied by a growing threat: credit card fraud. Fraudulent activities not only result in significant financial losses but also compromise user trust and the integrity of financial systems. As a result, detecting fraudulent transactions early has become a critical priority for banks, financial institutions, and digital payment providers.

One of the core challenges in building an effective fraud detection system is the class imbalance problem, where the number of legitimate transactions overwhelmingly outnumbers fraudulent ones. This imbalance causes classification models to become biased toward the majority class, often failing to recognize patterns associated with the minority (fraud) class [1], [3], [7]. Consequently, the rate of false negatives increases, meaning more fraudulent transactions are misclassified as legitimate, which is especially dangerous in financial contexts.

To mitigate this problem, various resampling techniques have been introduced, primarily oversampling and undersampling strategies. Among these, SMOTE (Synthetic Minority Oversampling Technique) is a widely used method that generates synthetic samples of the minority class based on the feature space of nearest neighbors [2], [4], [9]. Studies have shown that SMOTE significantly improves recall and F1-score, making classifiers more sensitive to fraud instances [6], [8].

Nevertheless, SMOTE on its own can introduce overfitting, particularly when dealing with noisy or borderline data. To address this, hybrid sampling techniques such as SMOTE-Tomek have been proposed. In this approach, SMOTE is used to oversample the minority class, while Tomek Links are used to remove ambiguous samples from the majority class, enhancing data quality and reducing overlap between classes [10].

In terms of classification algorithms, Random Forest has gained popularity in the fraud detection domain due to its robustness, ability to handle high-dimensional data, and strong generalization capabilities [1], [5]. Several studies confirm that Random Forest, when combined with SMOTE or hybrid sampling methods, yields superior performance compared to traditional algorithms [2], [7], [8].

Based on this foundation, the present study proposes the use of SMOTE-Tomek hybrid sampling in combination with the Random Forest classifier to improve fraud detection performance. The model is evaluated on a widely used benchmark credit card fraud dataset using standard performance metrics such as accuracy, precision, recall, and F1-score. The goal is to examine the extent to which hybrid sampling techniques enhance the model's ability to detect minority class instances without compromising overall classification performance.

Several studies have addressed class imbalance in fraud detection using resampling methods and ensemble classifiers.

Sundaravadivel et al. [1] investigated the effectiveness of Random Forest and SMOTE for detecting fraudulent transactions, achieving high accuracy (99.5%) and recall.

p-ISSN 2715-2871
e-ISSN 2714-5247

Their findings highlighted Random Forest's robustness and SMOTE's usefulness in balancing minority classes.

Oluwatoyin and Akinola [5] designed a real-time fraud detection system combining SMOTE and Random Forest, which achieved 97% accuracy and integrated real-time notifications using Twilio. Their approach validated the practicality of hybrid resampling in operational systems.

Marimuthu et al. [6] developed a hybrid ensemble model combining Random Forest and XGBoost, augmented with SMOTE, to minimize false negatives. Their results demonstrated improved recall and F1-scores, showing SMOTE's effectiveness in maintaining sensitivity in fraud detection.

Samant et al. [2] emphasized Random Forest's superior performance over Decision Tree and Logistic Regression in fraud detection scenarios, especially when enhanced with SMOTE. They noted that Random Forest with SMOTE improved precision and recall to 91% and 81% respectively, supporting its applicability in imbalanced environments.

These studies collectively support the hypothesis that combining ensemble models like Random Forest with hybrid resampling techniques such as SMOTE-Tomek can significantly improve the detection of fraudulent transactions.

II. METHODS

This study uses the Random Forest classifier integrated with SMOTE-Tomek to mitigate class imbalance. The key steps include:

1. Data Preprocessing:
   - Use a credit card fraud dataset with significant imbalance (fraudulent cases < 0.2%).
   - Apply normalization to continuous attributes.

2. Resampling Strategy:
   - Apply SMOTE to synthetically generate minority class (fraud) samples.
   - Apply Tomek links to remove borderline majority class (non-fraud) instances, reducing overlapping.

3. Model Training:
   - Split dataset (70% training, 30% testing).
   - Train Random Forest classifier on:
   - Original (imbalanced) data
   - Resampled (SMOTE-Tomek) data

4. Evaluation Metrics:
   - Accuracy
   - Precision
   - Recall
   - F1-Score
   - Confusion Matrix

The process flow is summarized in Figure 1, showing the transition from preprocessing to evaluation.

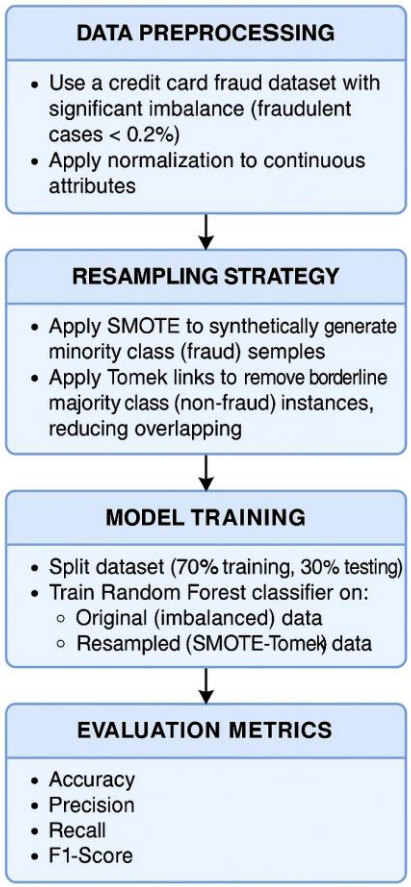

Fig. 1. Method Proposed Chart

III. RESULT AND DISCUSSION

4.1 Dataset Description

The dataset used in this study is the well-known Credit Card Fraud Detection Dataset from Kaggle. It contains 284,807 transaction records, of which only 492 are fraudulent, resulting in a highly imbalanced distribution with fraud cases representing only 0.173% of the data. This severe imbalance poses significant challenges for classification models, as they tend to favor the majority (non-fraud) class. A summary of the dataset is provided in Table below.

TABEL 1 Dataset Description

Feature	Value
Total Transactions	284.807
Non-Fraud Transactions	284.315
Fraud Transactions	492
Fraud Percentage	0,173%

4.2 Data Preprocessing

Before applying classification, the dataset underwent several preprocessing steps. Since the majority of the dataset features were anonymized using PCA, only two features—Amount and Time—required standard preprocessing.

The Amount feature was normalized using standard scaling to bring all values into a comparable range and

p-ISSN 2715-2871
e-ISSN 2714-5247

reduce the impact of extreme outliers. The result of normalization for a sample of transactions is shown in Table 2.

TABEL 2 Dataset Preprocessing

Transaction ID	Original Amount	Normalized Amount
1	149,62	0,52
2	2,69	0,01
3	378,66	1.00
4	123,50	0,41
5	69,99	0,81

4.3 Resampling with SMOTE-Tomek

To address the issue of class imbalance, the SMOTE-Tomek hybrid sampling method was applied to the training set. SMOTE was first used to generate synthetic samples for the fraud class, increasing its representation to match the non-fraud class. Following this, Tomek Links were applied to remove ambiguous borderline samples from the majority class, thereby reducing overlap and improving class separability.

The impact of SMOTE is illustrated in Table 3, showing that the minority class was increased from 394 to 227,451 instances, effectively balancing the dataset.

TABEL 3 Resampling with SMOTE-Tomek

Class	Before SMOTE	After SMOTE
Non-Fraud (0)	227.451	227.452
Fraud (1)	394	227.451

The performance comparison between the original and SMOTE-Tomek sampled datasets is shown in the following:

TABEL 4 Performance Comparison

Model	Accuracy	Trecision	Recall	F1Score
RF (original)	0,9952	0,91	0,62	0,73
RF+SMOTE	0,9938	0,88	0,89	0,88
RF+SMOTE-Tomek	0,9941	0,90	0,91	0,905

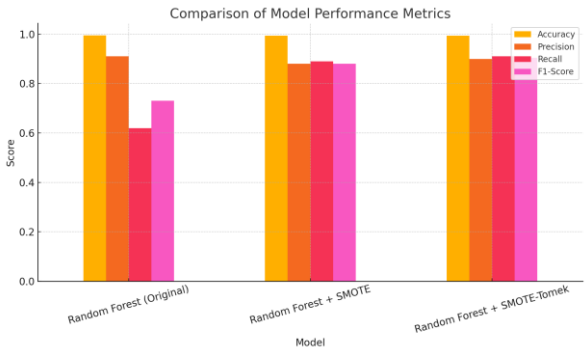

Figure 2. Comparison of model performance metrics

From the chart, it is evident that:

- The original model, despite its high accuracy, has the lowest recall, indicating poor fraud detection capability.
- The SMOTE-only model improves recall significantly but shows a slight trade-off in precision
- The SMOTE-Tomek model achieves the most balanced performance across all metrics, confirming that the removal of borderline noise enhances generalization.

The Random Forest with SMOTE-Tomek consistently outperformed other configurations in recall and F1-score, which are critical in fraud detection. Although accuracy slightly decreased, it remains within acceptable limits, and the increased recall indicates better detection of rare fraud cases. The confusion matrix also shows a reduction in false negatives, addressing the core issue of imbalance.

IV. CONCLUSION

This study demonstrates that the integration of SMOTE-Tomek with Random Forest significantly improves the performance of fraud detection models, especially in recognizing rare fraudulent transactions. Compared to baseline models, SMOTE-Tomek increases recall and F1-score, which are essential for minimizing financial risks due to undetected fraud. Future work may focus on integrating other hybrid resampling techniques such as ADASYN or combining with deep learning models for comparative analysis.

V. ACKNOWLEDGEMENT

Conceptualization of paper topics, M. Ilham; research methodology, M. Ilham and Moch Lutfi; validation of research results, M. Ilham; the formal analysis and the research investigation, M. Ilham and Artanti Indrasetianingsih ; the resources, M. Ilham and Moch Lutfi; writing—original draft preparation, M. Ilham; writing—review and editing, M. Ilham and Wildan Surya Wijaya; visualization data and the research results, M. Ilham; attribute data collector, Wildan Surya Wijaya

REFERENCES

[1] A. Sundaravadivel, A. Adithya, S. P. Soundararajan, and A. Gopal, “Optimizing Credit Card Fraud Detection with Random Forests and SMOTE,” International Journal of Engineering and Advanced Technology (IJEAT), vol. 8, no. 6, pp. 2174–2180, 2019.

[2] [2] S. Samant, P. Joshi, S. Bankar, S. Jain, and S. Ahuja, “SMOTE Based Credit Card Fraud Detection for Imbalanced Data: Performance Analysis,” 2024 International Technology Conference (OTCON) on Smart Computing for Industry 4.0, pp. 1–6, 2024. doi: 10.1109/OTCON60325.2024.10688312.

[3] [3] Q. Zou, J. Liu, Z. Shao, and Y. Wang, “A Credit Card Fraud Detection Method Based on Mahalanobis Distance Hybrid Sampling and Random Forest Algorithm,” Sensors, vol. 22, no. 7, pp. 1–18, 2022. doi: 10.3390/s22072627.

[4] [4] N. S. S. Pranavi et al., “Transaction Fraud Detection Using SMOTE Oversampling,” 2022 3rd International Conference for

p-ISSN 2715-2871
e-ISSN 2714-5247

Emerging Technology (INCET), pp. 1–6, 2022. doi: 10.1109/INCET54531.2022.9824146.

[5] [5] O. A. Ogunleye and T. O. Akinola, "Real-Time Credit Card Fraud Detection and Reporting System Using Machine Learning," 2022 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS), pp. 46–51, 2022. doi: 10.1109/AIMS56470.2022.00017.

[6] [6] S. M. Marimuthu et al., "A Comparative Study of Sampling Techniques for Imbalanced Credit Card Fraud Detection," 2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA), pp. 121–127, 2023. doi: 10.1109/ICIDCA57996.2023.10198379.

[7] [7] P. Thakur and N. Joshi, "Enhancing the Random Forest Model via Synthetic Minority Oversampling Technique for Credit-Card Fraud Detection," 2023 4th International Conference on Intelligent Engineering and Management (ICIEM), pp. 1–5, 2023. doi: 10.1109/ICIEM58232.2023.10169375.

[8] [8] M. Patel, B. Shah, and N. Dholakia, "Credit Card Fraud Detection Using Enhanced Random Forest Classifier for Imbalanced Data," International Journal of Engineering Trends and Technology (IJETT), vol. 69, no. 5, pp. 46–52, 2021.

[9] [9] R. Latha and C. V. Priya, "A Machine Learning Approach for Credit Card Fraud Detection in Massive Datasets Using SMOTE and Random Sampling," 2023 International Conference on Intelligent Sustainable Systems (ICISS), pp. 15–20, 2023. doi: 10.1109/ICISS57995.2023.10194925.

[10] [10] H. Zhou, C. Su, Y. Xie, and W. Zhang, "An Improved Unbalanced Data Classification Method Based on Hybrid Sampling Approach," Mathematics, vol. 10, no. 10, pp. 1–16, 2022. doi: 10.3390/math10101757.

[11] [11] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 20–29, 2004.

[12] [12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, 2002.

[13] [13] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. doi: 10.1023/A:1010933404324.

[14] [14] B. Pozzolo, O. Caelen, Y.-A. Le Borgne, S. Waterschoot, and G. Bontempi, "Learned lessons in credit card fraud detection from a practitioner perspective," Expert Systems with Applications, vol. 41, no. 10, pp. 4915–4928, 2014.

[15] [15] X. Li, Y. He, and J. Zhu, "Class imbalance problem in credit card fraud detection: Based on SMOTE-Tomek Link method," Journal of Physics: Conference Series, vol. 1213, no. 4, pp. 1–7, 2019.

[16] [16] I. Tomek, "Two modifications of CNN," IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-6, no. 11, pp. 769–772, 1976.

[17] [17] K. Raghuwanshi and A. Dutta, "Fraud detection in credit card transactions using machine learning algorithms," Procedia Computer Science, vol. 167, pp. 2261–2270, 2020. doi: 10.1016/j.procs.2020.03.276.

[18] [18] N. Ahmed, A. Mahmood, and M. R. Islam, "A survey of machine learning techniques for credit card fraud detection," Financial Innovation, vol. 9, no. 1, pp. 1–21, 2023.

[19] [19] M. Pathak and P. Srivastava, "Hybrid sampling and feature selection-based credit card fraud detection model," International Journal of Information Technology, vol. 15, pp. 1691–1698, 2023.

[20] [20] J. Brownlee, "SMOTE for Imbalanced Classification with Python," Machine Learning Mastery, 2020. [Online]. Available: https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/