https://doi.org/10.31436/iiumej.v26i1.3547

IMPROVING MODEL PERFORMANCE FOR PREDICTING EXFILTRATION ATTACKS THROUGH RESAMPLING STRATEGIES

ARIF RAHMAN HAKIM, KALAMULLAH RAMLI*, MUHAMMAD SALMAN, ESTI RAHMAWATI AGUSTINA

Department of Electrical Engineering, Faculty of Engineering, Universitas Indonesia, Depok, Indonesia

**Corresponding author: kalamullah.ramli@ui.ac.id*

(Received: 17 November 2024; Accepted: 1 January 2025; Published online: 10 January 2025)

ABSTRACT: Addressing class imbalance is critical in cybersecurity applications, particularly in scenarios like exfiltration detection, where skewed datasets lead to biased predictions and poor generalization for minority classes. This study investigates five Synthetic Minority Oversampling Technique (SMOTE) variants, including BorderlineSMOTE, KMeansSMOTE, SMOTEENC, SMOTEENN, and SMOTETomek, to mitigate severe imbalance in our customized tactic-labeled dataset with dominant majority class influence and weak class separability class imbalance. We use seven imbalance metrics to assess each SMOTE variant's impact on class distribution stability and separability. Furthermore, we evaluate model performance across five classifiers: Logistic Regression, Naïve Bayes, Support Vector Machine, Random Forest, and XGBoost. Findings reveal that SMOTEENN consistently enhances performance metrics (accuracy, precision, recall, F1-score, and geometric mean) on an average of 99% across most classifiers, establishing itself as the most adaptable variant for handling imbalance. This study provides a comprehensive framework for selecting resampling strategies to enhance classification efficacy in cybersecurity tasks with imbalanced data.

ABSTRAK: Menangani ketidakseimbangan kelas adalah penting dalam aplikasi keselamatan siber, terutama dalam senario seperti pengesanan eksfiltrasi, di mana set data yang condong membawa kepada ramalan yang berat sebelah dan generalisasi yang lemah untuk kelas minoriti. Kajian ini menyiasat lima varian Teknik Sintetik Pencontohan Lebihan Minoriti (SMOTE), termasuk BorderlineSMOTE, KMeansSMOTE, SMOTEENC, SMOTEENN, dan SMOTETomek, untuk mengurangkan ketidakseimbangan teruk dalam set data berlabel taktik yang disesuaikan dengan pengaruh kelas majoriti dominan dan ketidakcerahan kelas yang lemah. Kami menggunakan tujuh metrik ketidakseimbangan untuk menilai kesan setiap varian SMOTE terhadap kestabilan dan ketidakcerahan taburan kelas. Selain itu, kami menilai prestasi model merentasi lima pengelas: Regresi Logistik, Naïf Bayes, Mesin Vektor Sokongan, Hutan Rawak, dan XGBoost. Penemuan menunjukkan bahawa SMOTEENN secara konsisten meningkatkan metrik prestasi (ketepatan, ketepatan, pengingatan, skor F1, dan purata geometri) sebanyak 99% secara purata merentasi kebanyakan pengelas, menegaskan dirinya sebagai varian yang paling boleh disesuaikan untuk menangani ketidakseimbangan. Kajian ini menyediakan rangka kerja komprehensif untuk memilih strategi pencontohan semula bagi meningkatkan keberkesanan klasifikasi dalam tugas keselamatan siber dengan data yang tidak seimbang.

KEYWORDS: *Machine Learning, Imbalance Data, SMOTE, and Exfiltration*

https://doi.org/10.31436/iiumej.v26i1.3547

1. INTRODUCTION

Cybersecurity has become a critical issue across various sectors, affecting businesses, government entities, and individuals. The rapid advancement of technology has created new opportunities but also facilitated increasingly sophisticated cyber threats. The frequency of cyber-attacks has surged, with a notable rise in global incidents. Projections estimate that cybercrime will cost USD 9.5 trillion by 2024, growing at an annual rate of about 15%, reaching USD 10.5 trillion by 2025 [1].

The evolution of cyber threats requires organizations to assess their cybersecurity posture and adapt strategies accordingly and proactively. Common Vulnerabilities and Exposures (CVEs) are vital for identifying and responding to potential cyber incidents. CVEs provide a structured framework for categorizing vulnerabilities, aiding effective threat management and incident response. Integrating CVEs with frameworks like MITRE ATT&CK helps organizations understand adversary tactics and prioritize vulnerabilities based on their impact. Automated tools for vulnerability detection and remediation can reduce the time and resources required for effective cybersecurity management. Machine Learning (ML) also plays a crucial role in combating sophisticated cyber threats. ML algorithms improve accuracy by learning from new data and adapting to evolving threats. Additionally, ML is essential for incident response, automating threat mitigation to reduce the attack window and limit damage.

However, implementing ML in cybersecurity faces challenges from imbalanced datasets, which impact classification model performance. Imbalanced datasets occur when one class (usually the majority) outnumbers the other (minority) class. This imbalance can lead to biased predictions, poor generalization, and higher error rates for the minority class, which is often critical in cybersecurity, such as fraud or intrusion detection systems [2]. Several resampling methods have been developed to address class imbalance and improve classifier performance. These methods include undersampling, oversampling, and hybrid techniques. Our study focuses on addressing the class imbalance in a custom dataset designed to classify whether a sequence of tactics indicates exfiltration. We used a tactic-labeled dataset from CVE descriptions, which shows severe imbalance, with the minority class comprising less than 1%.

This paper examines the dataset's imbalance characteristics in detail and investigates oversampling-based solutions to address this issue. Oversampling was chosen for its ease of implementation and wide applicability. The most popular oversampling technique, SMOTE [3], is commonly used to balance class distribution and has proven effective in many studies and real-world applications. We investigate how five SMOTE variants affect imbalance characteristics: BorderlineSMOTE [4], KMeansSMOTE [5], SMOTEENC [6], SMOTEENN [7], and SMOTETomek [8]. We evaluate each dataset's imbalance using seven metrics: class distribution, imbalance ratio, minority class percentage, Coefficient of Variation (CV), Gini index, entropy, and Fischer's ratio. These metrics are applied to both the original and resampled datasets to assess the impact of each SMOTE variant.

The datasets from the five SMOTE variants are also used to assess model performance across five classifiers: Logistic Regression (LR) [9], Naïve Bayes (NB) [10], Support Vector Machine (SVM) [11], Random Forest (RF) [12], and Extreme Gradient Boosting (XGBoost) [13]. Performance is measured using six metrics: accuracy, precision, recall, specificity, F1-score, and geometric mean. These metrics facilitate a comparative analysis across models. The evaluation of imbalance characteristics is enhanced by including both seven imbalance metrics and six classifier performance metrics.

https://doi.org/10.31436/iiumej.v26i1.3547

Our study provides the following contributions:

- *A customized tactic-labeled dataset for exfiltration classification and measuring its imbalance characteristics*: Developed a new dataset specifically for classifying the Exfiltration tactic from ten other tactics. This dataset was derived from the originally used for mapping CVE description into MITRE Tactics and Techniques. Additionally, a focused analysis is conducted to identify and understand imbalance characteristics within a dataset using seven metrics (class distribution, imbalance ratio, minority class percentage, Coefficient of Variation, Gini index, entropy, and Fischer's ratio).
- *Investigation of SMOTE Variants*: Five SMOTE variants (BorderlineSMOTE, KMeansSMOTE, SMOTEENC, SMOTEENN, and SMOTETomek) are applied, and their impact on dataset characteristics is systematically assessed.
- *Multi-Metric Evaluation of Imbalance*: The imbalance of the original and resampled datasets is evaluated using the seven imbalanced metrics, providing an in-depth view of resampling effects.
- *Cross-Classifier Performance Comparison*: The impact of SMOTE resampling on model performance is assessed across five classifiers (LR, NB, SVM, RF, and XGBoost) using six classifier performance metrics (accuracy, precision, recall, specificity, F1-score, geometric mean) to establish a benchmark on how resampling strategies influence model efficacy.

This paper is divided into the following five sections. Section 2 reviews existing studies, focusing on existing studies that utilized the CVE dataset. Section 3 outlines the creation of our dataset and analysis of its imbalance characteristics. Section 4 investigates SMOTE variants by providing an in-depth view of resampling effects on imbalanced characteristics. In addition, we provide cross-classifier performance comparisons to establish a benchmark on how resampling strategies influence model efficacy. Lastly, conclusions are presented in Section 5.

2. RELATED WORKS

This section provides a review of existing methods or models that leverage the mapping of tactics from CVEs. Our review focuses on the aims, method used, how CVEs play a role in the method, and key findings, as provided in Table 1.

Table 1. Tactics distribution in the created dataset

Study	Aims & Methods	Role of CVEs	Key Findings
[14]	Systematic mapping of CVEs to ATT&CK techniques using neural networks and unsupervised labeling.	CVEs are systematically connected to techniques for efficient threat mitigation.	Enhanced prioritization and threat management.
[15]	ML and DL approach for multi-label classification of CVEs to ATT&CK techniques.	CVEs form the basis for linking vulnerabilities to adversarial approaches.	Improved accuracy and reliability in mapping.
[16]	Transformer-based models (e.g., SecRoBERTa) for mapping CVEs to ATT&CK techniques.	CVEs are utilized for improving cybersecurity measures.	SecRoBERTa achieves 78.88% weighted F1; improved cybersecurity understanding.
[17]	BERT-based models with TextAttack for data augmentation to map CVEs to ATT&CK techniques.	CVEs are used to create a tagged corpus for ATT&CK Enterprise Matrix strategies.	F1-score of 47.84%; improved training set balance with augmentation.

A study in [14] developed a systematic approach for mapping CVEs to MITRE ATT&CK techniques to enhance the understanding and management of cybersecurity threats. CVEs are crucial to the paper's investigation of how vulnerabilities might be systematically connected to attack strategies, improving defenders' ability to prioritize and mitigate threats efficiently. The research method combines advanced neural network modeling, unsupervised labeling, and enriched data representation to effectively map CVEs to ATT&CK techniques, enhancing cybersecurity threat management. Here are some research limitations: reproducibility and inconsistencies, concept drift, limited coverage of techniques, lack of granular categories, and dependence on unsupervised labeling.

Authors in [15] enhanced the process of linking CVEs to MITRE ATT&CK techniques using ML and deep learning (DL) approaches. They highlight the role of CVE as a foundational piece for linking vulnerabilities to adversarial approaches, hence improving organizations' ability to defend against cyber-attacks effectively. They employed multi-label classifications, Multi-Layer Perceptron (MLP), data augmentation, cross-validating techniques, hyperparameter tuning, reproduction, and comparison to enhance accuracy and reliability. The research limitations include dataset limitations, reproducibility issues, framework updates, limited technique coverage, and model comparison challenges.

Transformer-based models are used to automatically map CVEs to MITRE ATT&CK techniques to better cybersecurity understanding and protection measures [16]. An extended dataset of 9985 entries, security auditing tools, and MITRE ATT&CK methods are included in the study. The top models were SecRoBERTa, SecBERT, CyBERT, and TARS, with SecRoBERTa scoring 78.88% in weighted F1. Their study's main shortcoming is that conceptual linkages between tactics make it difficult to deduce a technique sequence from textual descriptions for specific vulnerabilities. Likewise, BERT-based language models are used to match 1813 MITRE ATT&CK-annotated CVEs to approaches in the study [17]. TextAttack-based data augmentation algorithms correct the training set imbalance, achieving an F1-score of 47.84%. The methodology creates a tagged corpus by manually mapping CVEs to MITRE ATT&CK Enterprise Matrix strategies and procedures, while the F1-score measures model performance.

Table 1 highlights that prior research has predominantly leveraged CVEs as a critical source for mapping MITRE ATT&CK tactics or techniques. However, these studies have largely overlooked the potential of analyzing sequences of tactics derived from CVEs to predict attacker objectives. To address this gap, our work focuses on processing CVE data into actionable sequences that predict specific attack goals, such as exfiltration. Moreover, existing literature emphasizes the lack of consensus on the optimal approach to handle data imbalance [18]. This variability underscores the pressing need for standardized evaluation frameworks to ensure consistency and comparability across datasets.

3. DATASET CREATION FOR EXFILTRATION CLASSIFICATION AND IMBALANCE CHARACTERISTICS

This section explains the dataset processing steps we employed and details the imbalance characteristics of the dataset. Additionally, we introduce the seven metrics used to measure the imbalance characteristics of the initial dataset in this section and the various datasets generated by the resampling methods discussed in the subsequent section. Furthermore, we present the measurement results of the imbalanced characteristics of the initial dataset using these seven metrics. These results serve as a baseline for analyzing the impact of resampling methods on imbalance characteristics by examining changes in the measured metrics.

https://doi.org/10.31436/iiumej.v26i1.3547

3.1. Modifying Tactics Mapping Dataset Based on CVE Description

The dataset source used in this study is created by the study [19] aimed to automatically map CVEs to the 14 MITRE ATT&CK tactics using transformer-based models and is publicly available in [20]. The data cleaning phase involved filtering out rows with missing tactic labels. In this research, two tactics, Reconnaissance, and Resources, were excluded as they predominantly occur on the attacker's side. We retained the tactic Initial Access to represent how attackers infiltrate the target environment. Consequently, ten tactics were chosen as features for this study: (1) Initial Access, (2) Execution, (3) Persistence, (4) Privilege Escalation, (5) Defense Evasion, (6) Credential Access, (7) Discovery, (8) Lateral Movement, (9) Collection, and (10) Command and Control. We also excluded the impact since we focused on exfiltration as the adversarial objective and designated exfiltration as the target variable. This custom-built dataset was selected to capture patterns within these ten tactics and assess whether their presence leads to exfiltration tactics. Fig. 1 shows 14 tactics derived from MITRE ATT&CK, where in this study, ten tactics were used as features, one tactic as a target feature, and three tactics were excluded.

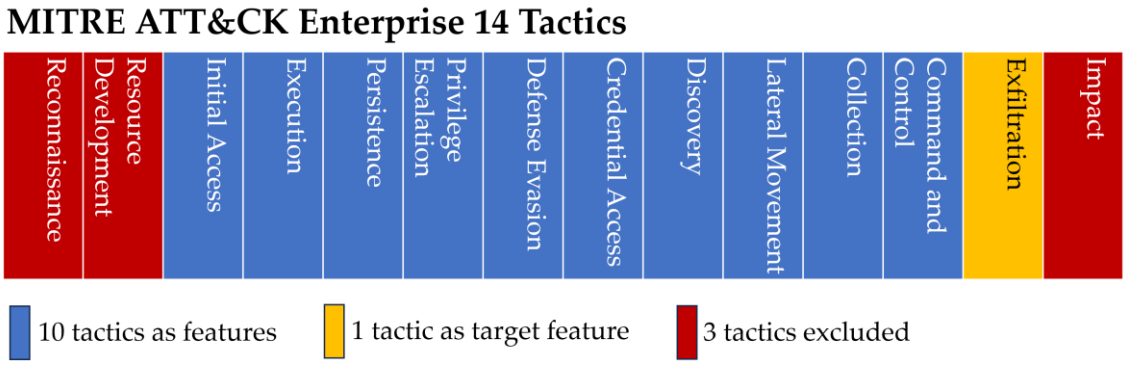

Figure 1. Adversarial tactics and features mapping in our modified dataset.

Table 2. Tactics distribution in the created dataset

Features	Tactics	Number of 1
Feature_1	Initial Access	722
Feature_2	Execution	264
Feature_3	Persistence	3016
Feature_4	Privilege Escalation	3218
Feature_5	Defense Evasion	7552
Feature_6	Credential Access	614
Feature_7	Discovery	2369
Feature_8	Lateral Movement	1932
Feature_9	Collection	663
Feature_10	Command and Control	427
Target	**Exfiltration**	**51**

On the other hand, Table 2 shows the distribution of the ten tactics used as features, with the Exfiltration Tactic serving as the target feature in our 9602-row dataset. The table shows that "Defense Evasion" is the most frequently occurring tactic among the features, while "Command and Control" is the least frequent tactic in the dataset. On the other hand, "Exfiltration," as the target feature, appears only 51 times. The created dataset is referred to as the initial dataset, particularly in comparative analyses with datasets produced by various resampling methods to provide a sense of the initial state versus the post-resampling conditions.

https://doi.org/10.31436/iiumej.v26i1.3547

3.2. Imbalance Characteristics of the Initial Dataset

This section explains the characteristics of the initial dataset as measured by seven commonly used metrics to illustrate dataset imbalance. Each metric explains its purpose, accompanied by its calculation formula. Additionally, we provide an interpretation of the calculation results for each metric as applied to the initial dataset.

3.2.1. The Imbalance Metrics

The class distribution indicates the number of instances present in each dataset class. In a balanced dataset, the instances across classes are generally equal, while an imbalanced dataset exhibits a noticeable difference in instance counts. Recognizing the class distribution is critical for choosing the appropriate modeling techniques and evaluation metrics, as standard metrics like accuracy may not provide a reliable assessment in imbalanced situations. Meanwhile, the imbalance ratio is a quantitative measure to represent the imbalance between the majority and minority classes. This ratio is determined by dividing the count of instances in the majority class by that in the minority class. A higher imbalance ratio suggests a more pronounced disparity between classes, which can create additional challenges for classifiers. This metric is essential for assessing the imbalance's severity and informing the selection of techniques to address it effectively. In addition, the minority class percentage is a significant metric representing the proportion of instances in the minority class relative to the total dataset size. When the minority class percentage is low, models tend to favor the majority class, potentially causing reduced recall for the minority class.

The CV quantifies the ratio of the standard deviation to the mean, offering insight into the variability within class distributions in imbalanced datasets. A high CV suggests significant variability relative to the mean, indicating an imbalance and inconsistent distribution within classes. The Gini index, commonly used to measure inequality, assesses class distribution disparity in imbalanced datasets. With values ranging from 0 (complete equality) to 1 (maximum inequality), a higher Gini index reveals a more significant disparity between majority and minority classes. This metric is beneficial for evaluating sampling strategies or model adjustments aimed at reducing class imbalance.

Entropy, measuring uncertainty or disorder, assesses the impurity of a dataset's class distribution. High entropy indicates a more balanced class distribution, while low entropy suggests dominance by a single class. When an excessive number of majority samples fall within the overlapping region, the probability for the majority class nears 1. In contrast, the minority class approaches 0, driving entropy toward its minimum value of 0. Conversely, as the count of majority samples approaches that of minority samples, entropy tends towards its maximum value of 1. Fischer's ratio, also known as Fischer's discriminant ratio [21], assesses class separability by comparing between-class variance to within-class variance. Higher Fischer's ratios indicate clearer class separability, an advantage in classification tasks. However, achieving high Fischer's ratios is more difficult with imbalanced datasets, as minority classes may be underrepresented

3.2.2. The Initial Dataset

The calculation results for the seven imbalance metrics of the initial dataset are presented in Table 3. Overall, the measurements indicate that the initial dataset exhibits severe imbalance characteristics. The class distribution is highly skewed, with 9,551 instances in the majority class and only 51 in the minority class. An imbalance ratio of 187 means that the majority class is 187 times larger than the minority class, accompanied by a minority class percentage of merely 0.53%. Additionally, a relatively high CV of 2.44 reflects a significant disparity

between the majority and minority classes. Likewise, the entropy value of only 0.05 shows a strong dominance by the majority class. Furthermore, a very low Fisher's ratio of 0.05 indicates poor class separability, which could result in an even more significant underrepresentation of the minority class.

Table 3. Imbalance characteristics of the initial dataset

Metrics	Result
Class Distribution	955:51
Imbalance Ratio	187.27
Minority Class Percentage	0.53
Coefficient of Variation	2.44
Gini Index	0.01
Entropy	0.05
Fischer's Ratio	0.05

4. INVESTIGATION OF SMOTE VARIANTS

This study employed five SMOTE variants: BorderlineSMOTE, KMeansSMOTE, SMOTEENC, SMOTEENN, and SMOTETomek. This section provides a detailed description of each SMOTE variant utilized, along with the post-resampling dataset size and the feature distribution within each resulting dataset. We then present the imbalance characteristics of each dataset, which are evaluated using seven metrics consistent with those applied to the initial dataset. Based on the values derived from these metrics, we conduct a comparative analysis between the initial and resampled datasets. Furthermore, we compare these metric values across the five SMOTE variants to determine the most suitable variant for our dataset.

Additionally, we present a comprehensive evaluation of model performance across each of the five SMOTE variants using five commonly adopted classifiers: LR, NB, SVM, RF, and XGBoost. Model performance is assessed using six metrics: accuracy, precision, recall, specificity, F1-score, and geometric mean. By comparing model performance across the five SMOTE variants, we aim to identify the most effective resampling strategy for enhancing the model's efficacy with SMOTE techniques. Fig. 2 illustrates the block diagram of our study's investigation of SMOTE variants.

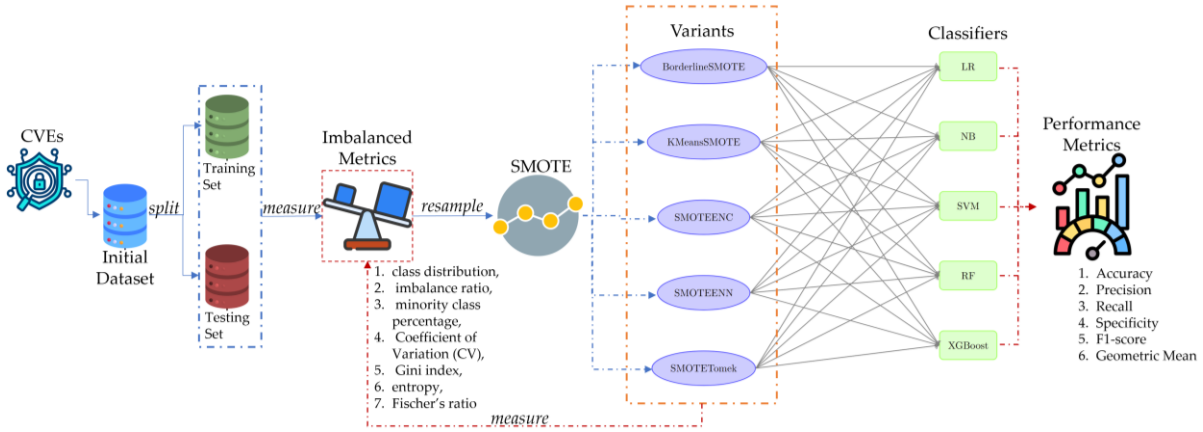

Figure 2. Block diagram of our investigation on SMOTE variants.

4.1. Resampled Datasets using SMOTE Variants

The cleaned dataset from the previous stage is clearly imbalanced, with the majority class outnumbering the minority class by 93%:7%. Table 4 compares the initial dataset

characteristics without resampling and the five datasets generated from each SMOTE Variants. The table displays the number of rows for each dataset and the ratio of majority to minority class in percentages.

Table 4. Comparison of dataset size and class distribution ratios after resampling

Resampling	# Rows	*Ratio (%)
Initial (without resampling)	9602	99.47:0.53
BorderlineSMOTE	19102	50:50
KMeansSMOTE	19104	50.1:49.9
SMOTEENC	19102	50:50
SMOTEENN	9520	74.61:25.39
SMOTETomek	19102	50:50

*Ratio between majority class and minority class

Notably, BorderlineSMOTE, SMOTEENC, and SMOTETomek datasets have the same number of rows and a precisely balanced ratio. Meanwhile, the KMeansSMOTE dataset has the most significant number of rows and has an almost balanced ratio. On the other hand, the SMOTEENN dataset has the lowest number of rows, with the majority class 3 times more rows than the minority class. However, this does not guarantee optimal model performance.

Additionally, to assess the number of features present, we provide each dataset's distribution of features (tactics), including the initial dataset and the five resampled datasets generated using SMOTE variants in Table 5. The table illustrates a significant variation in the effectiveness of different SMOTE techniques in balancing class distributions across features (tactics), with "Exfiltration" as the target feature. For several tactics, such as "Initial Access," "Execution," and "Defense Evasion," most SMOTE variants, especially SMOTENC, KMeansSMOTE, and BorderlineSMOTE, demonstrate substantial increases in sample counts. This increase indicates the strategy of these techniques in addressing class imbalance across non-target tactics.

Table 5. Comparison of features (tactics) distribution across the initial dataset and SMOTE variants

Tactic	Initial	SMOTE Variants (Oversampling)				
		B-SMOTE[1]	K-SMOTE[1]	SMOTEENC	SMOTENN	SMOTE-T[1]
Initial Access	722	2367	3831	5451	1241	1453
Execution	2642	8096	8721	7385	1366	4970
Persistence	3016	5445	3016	4209	2206	4192
Privilege Escalation	3218	6103	3218	4411	2404	4556
Defense Evasion	7552	10085	7553	14238	7008	13925
Credential Access	614	614	614	614	609	614
Discovery	2369	3312	3150	6890	4284	6669
Lateral Movement	1932	3737	1932	2717	840	2717
Collection	663	3824	3578	2232	1163	2030
Command and Control	427	627	427	553	498	505
Exfiltration[2]	**51**	**9551**	**9553**	**9551**	**7103**	**9551**

[1] B-SMOTE (BordelineSMOTE), K-SMOTE (KMeansSMOTE), SMOTE-T (SMOTETomek)
[2] Target feature

Notably, SMOTENC achieves the highest augmentation for "Defense Evasion," a tactic with high initial counts, while "Execution" also sees a marked rise under KMeansSMOTE and BorderlineSMOTE, reflecting these methods' responsiveness to the initial dataset's imbalance patterns. Interestingly, "Exfiltration," the target feature and the rarest class in the initial dataset (with only 51 samples), is consistently upsampled across all SMOTE variants. The substantial upsampling of "Exfiltration" by each SMOTE variant highlights the importance of this target

https://doi.org/10.31436/iiumej.v26i1.3547

class in the resampling process. Fig. 3 shows the tactics count comparison across SMOTE variants datasets and the initial dataset.

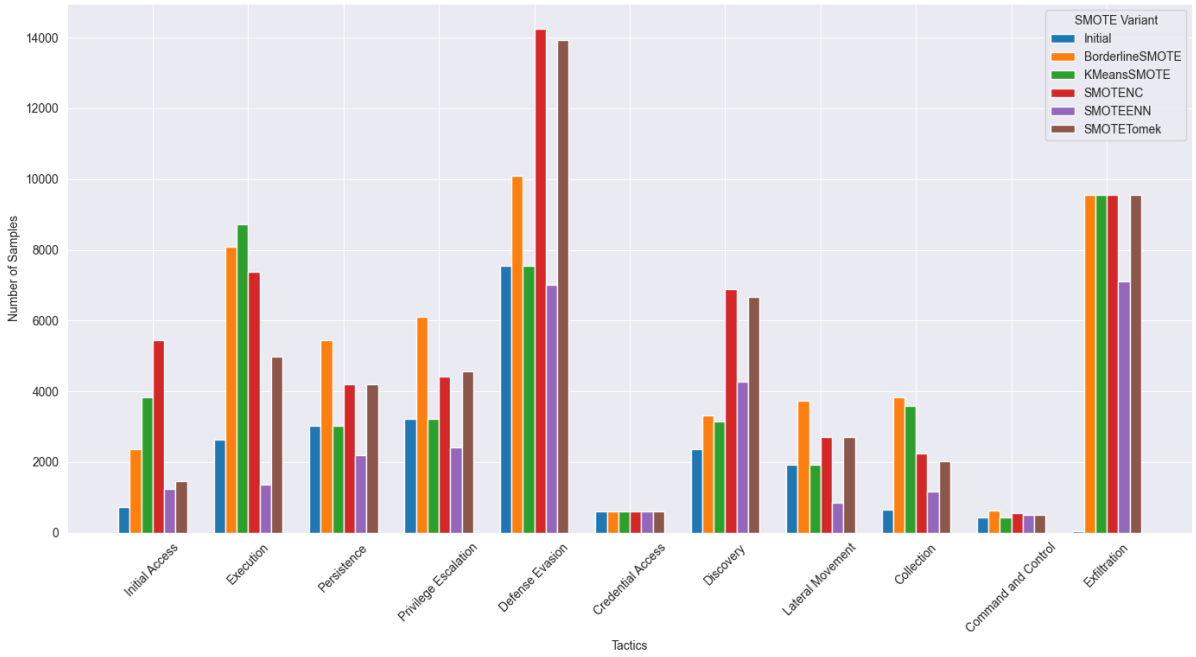

Figure 3. Tactics count comparison across SMOTE variants datasets and the initial dataset.

4.2. Imbalanced Metrics Value Across All Datasets

In this section, we employed a multi-metric evaluation of dataset imbalance characteristics. The imbalance of the original and resampled datasets is evaluated using the seven imbalanced metrics, including class distribution, imbalance ratio, minority class percentage, Coefficient of Variation, Gini index, entropy, and Fischer's ratio. This investigation provides an in-depth view of the resampling effects.

4.2.1. Comparison of Imbalance Metric Values

We present the imbalance metric measurements for the initial dataset and the five SMOTE variant datasets in Table 6. The table shows that the initial dataset exhibits a severe imbalance, with a class distribution ratio of 9551:51, resulting in a high imbalance ratio of 99.47:0.53. This leads to a minority class percentage of only 0.53%, indicating the rarity of the "Exfiltration" feature. Consequently, the metrics of the Gini Index and Entropy are shallow at 0.01 and 0.05, respectively, reflecting the minimal uncertainty and high predictability due to the overwhelming presence of the majority class. The low Fischer's Ratio (0.05) suggests limited separability between the majority and minority classes in the initial dataset.

Substantial changes are observed in the class distribution and associated metrics after applying different SMOTE variants. BorderlineSMOTE, SMOTENC, and SMOTETomek successfully achieve a balanced class distribution with a 1:1 imbalance ratio and a 50% minority class percentage. These methods also show uniform Gini Index and Entropy values (0.5 and 1, respectively), representing increased class uncertainty and balance. However, variations in the CV indicate that SMOTETomek has a higher variability in class distribution at 2.77, compared to SMOTENC (2.49) and BorderlineSMOTE (2.49), suggesting differing impacts on the distribution's stability across these techniques. Fischer's ratio is slightly higher for BorderlineSMOTE (0.14), indicating better class separability than SMOTETomek (0.05).

https://doi.org/10.31436/iiumej.v26i1.3547

Table 6. Imbalances metric values

Dataset	CD*	IR*	MCP*	CV*	GI*	E*	FR*
Initial (without resampling)	9551:51	99.47 : 0.53	0.53	2.44	0.01	0.05	0.05
BorderlineSMOTE	9551:9551	50 : 50	50	2.49	0.50	1	0.14
KMeansSMOTE	9551:9553	50.1 : 49.9	49.99	2.82	0.50	1	0.57
SMOTENC	2417:7103	50 : 50	25.39	2.42	0.38	0.82	0.32
SMOTEENN	9551:9551	74.61 : 25.39	50	2.49	0.50	1	0.12
SMOTETomek	9551:9551	50 : 50	50	2.77	0.50	1	0.05

*CD (class distribution), IR (imbalance ratio), MCP (minority class percentage), CV (Coefficient of Variation), GI (Gini index), E (entropy), FR (Fischer's ratio)

In contrast, KMeansSMOTE and SMOTEENN result in distinct class distributions. KMeansSMOTE achieves a near-perfect 1:1 class distribution with a slight minority deviation, yielding a high Fischer's Ratio of 0.57, suggesting strong class separability. Meanwhile, SMOTEENN produces an imbalance ratio of 2.94 with a minority class percentage of 25.39%, positioning it as a less balanced but potentially more stable approach for specific models. The Entropy (0.82) and Gini Index (0.38) values for SMOTEENN indicate a moderate increase in class uncertainty and separability, highlighting its unique impact on dataset structure compared to other SMOTE techniques.

4.2.2. Best Strategy Based on Imblalance Metric Values Across SMOTE Variants

This section focuses on identifying the most effective SMOTE variant based on the results of the imbalance metrics measurements. Based on the comparison metrics presented in Table 5, KMeansSMOTE demonstrates the most effective performance in balancing the target class, "Exfiltration." This method nearly achieves a 1:1 distribution with a class ratio of 9551:9553, yielding an Imbalance Ratio close to 1. Additionally, the Minority Class Percentage reaches 49.99%, closely approximating 50%, indicating an almost perfect balance between the majority and minority classes. This balance is essential for enhancing the quality of training data, enabling the model to better recognize patterns within the minority "Exfiltration" class.

Moreover, KMeansSMOTE achieves the highest CV at 2.82, reflecting consistent distribution across classes after rebalancing. Its Gini Index and Entropy values are near optimal at 0.5 and 1, respectively, indicating a balanced level of uncertainty between classes. With the highest Fischer's Ratio (0.57), this method also exhibits superior separability between majority and minority classes, which is critical for improving the model's ability to distinguish between them. Based on these metrics, KMeansSMOTE provides an optimal balance of distribution consistency, separability, and class uncertainty, making it the preferred choice for addressing the imbalance in the target class, "Exfiltration." We further examine cross-classifier performance comparison using model performance metrics to investigate resampling strategies and their impact on model efficacy.

4.3. Cross-Classifier Performance Comparison

In this section, we investigate the impact of SMOTE resampling on model performance assessed across five classifiers (LR, NB, SVM, RF, and XGBoost) using six classifier performance metrics (accuracy, precision, recall, specificity, F1-score, geometric mean) to establish a benchmark on how resampling strategies influence model efficacy. We present each classifier's model performance individually, then compare SMOTE variant performance based on six performance metrics within each classifier. This approach focuses on comparing the performance of the five SMOTE variants rather than identifying the best classifier among the five used. This way, we can further determine which SMOTE variant performs most effectively for these models.

https://doi.org/10.31436/iiumej.v26i1.3547

4.3.1. Logistic Regression Classifier

Based on the performance metrics across different resampling methods using LR in Fig. 4, several insights emerge regarding the efficacy of each SMOTE variant. The initial dataset without resampling exhibits high accuracy (0.995) and specificity (1) but has zero values for precision, recall, F1 score, and geometric mean, indicating that it fails to detect minority class instances. This outcome underscores the limitations of unbalanced data in handling minority classes effectively. Among the resampling methods, KMeansSMOTE and SMOTEENN stand out, as they yield superior metrics across the board, with KMeansSMOTE achieving a balanced performance across accuracy (0.962), precision (0.933), recall (0.996), and specificity (0.929). Meanwhile, SMOTEENN produces the highest F1 score (0.979) and an impressive recall (0.989), suggesting a robust ability to detect minority instances effectively while maintaining good overall accuracy.

Figure 4. Model performance of SMOTE variants using LR classifier.

When examining the overall effectiveness of each SMOTE variant, SMOTEENN emerges as the most effective method for this classifier model. It balances detection performance across metrics with high precision, recall, and F1 score values while achieving a notable geometric mean of 0.946, indicating balanced performance across both classes. KMeansSMOTE, with a slightly lower F1 score and geometric mean, still performs admirably and may be a suitable alternative depending on specific goals. In contrast, BorderlineSMOTE, SMOTENC, and SMOTETomek exhibit lower values in several metrics, particularly in recall and F1 score, suggesting they are less effective for minority class detection in this context. Consequently, SMOTEENN and KMeansSMOTE are recommended for optimizing this model's performance across balanced metrics.

4.3.2. Naïve Bayes Classifier

Fig. 5 shows that the NB classifier yields varying results across resampling methods, with KMeansSMOTE and SMOTEENN demonstrating substantial improvements compared to others. Without resampling, the classifier achieves high specificity (0.945) but performs poorly in terms of recall (0.067), precision (0.006), and F1 score (0.012), highlighting its ineffectiveness in capturing the minority class. Among the resampling techniques, KMeansSMOTE shows the most balanced performance, with high scores across accuracy (0.950), precision (0.912), recall (0.996), and F1 score (0.952). This balanced metric profile

https://doi.org/10.31436/iiumej.v26i1.3547

reflects its ability to handle class imbalance effectively in NB. SMOTEENN also performs well, particularly in recall (0.989) and F1 score (0.907), but its lower specificity (0.436) suggests some trade-offs in its balance between classes.

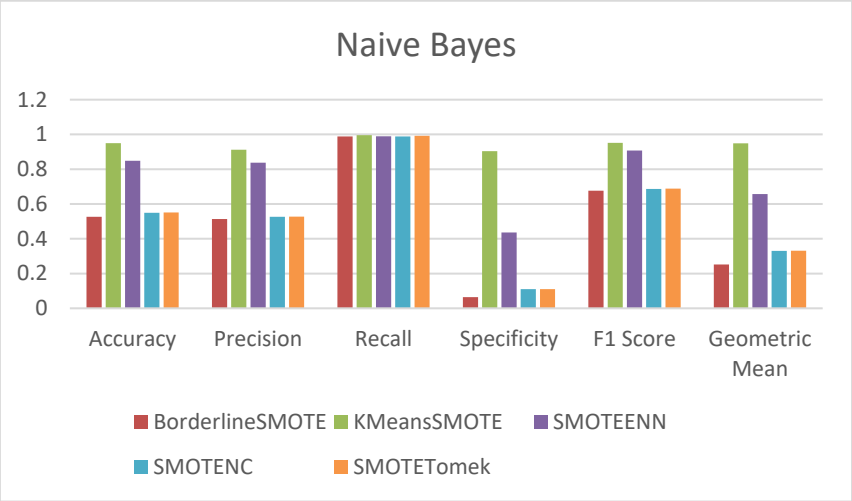

Figure 5. Model performance of SMOTE variants using NB classifier.

Considering the overall metrics, KMeansSMOTE is the best-performing variant for the NB classifier. It maintains high precision, recall, F1-score, and geometric mean (0.949), indicating stable and reliable performance across different metrics. While SMOTEENN also shows strong recall and F1-score, its reduced specificity could impact performance in specific applications. BorderlineSMOTE, SMOTENC, and SMOTETomek display lower accuracy and specificity values, making them less effective for NB in this context. KMeansSMOTE is recommended for optimal model efficacy with NB, as it provides balanced detection across both minority and majority classes.

4.3.3. Random Forest Classifier

The RF classifier shows substantial improvement across various metrics when combined with resampling methods, particularly in recall and F1-score. Without resampling, RF achieves perfect specificity (1.0) but fails to identify any instances of the minority class, resulting in zero values for precision, recall, F1 score, and geometric mean. This demonstrates that the model is highly imbalanced and fails to generalize to the minority class. When resampling methods are applied, metrics like precision, recall, and F1 score improve significantly, with SMOTEENN and KMeansSMOTE showing the most robust results, as illustrated in Fig. 6.

Among the resampling methods, SMOTEENN provides the best performance across all metrics, achieving near-perfect values for accuracy (0.997), precision (1.0), recall (0.996), specificity (1.0), F1 score (0.999), and geometric mean (0.999). This indicates that SMOTEENN allows RF to balance minority and majority classes exceptionally well. KMeansSMOTE also performs well, with high scores in precision (0.944), recall (0.995), F1 score (0.969), and geometric mean (0.968), showing a solid balance across metrics, though slightly lower than SMOTEENN. Other methods, like BorderlineSMOTE and SMOTENC, show good but comparatively lower performance, particularly in recall and F1 score, while SMOTETomek lags with a lower overall balance. Therefore, SMOTEENN is the most effective resampling technique for the RF classifier, followed closely by KMeansSMOTE for robust, balanced performance across metrics.

https://doi.org/10.31436/iiumej.v26i1.3547

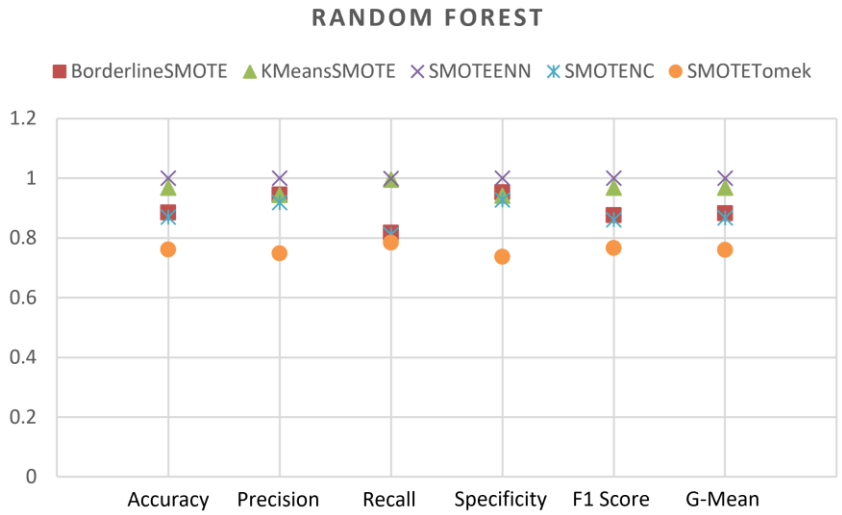

Figure 6. Model performance of SMOTE variants using RF classifier.

4.3.4. Support Vector Machine Classifier

Using the SVM classifier, we can see in Fig. 7 that the performance metrics improve considerably when resampling methods are applied, particularly regarding recall and F1 score. Without resampling, the classifier achieves a high specificity (1.0) but ultimately fails to detect the minority class, resulting in zero values for precision, recall, F1 score, and geometric mean. This highlights that the model is severely biased toward the majority class. However, applying resampling techniques will significantly enhance the classifier's ability to capture the minority class, reflected by notable improvements in recall and F1 score across most methods.

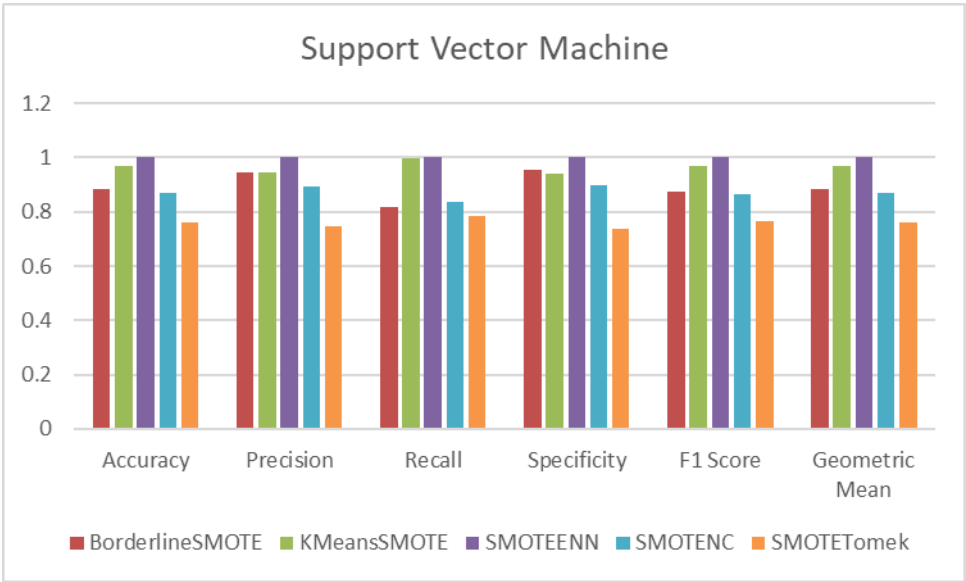

Figure 7. Model performance of SMOTE variants using SVM classifier.

Among the resampling techniques, SMOTEENN yields the best results, with near-perfect scores in accuracy (0.999), precision (1.0), recall (0.999), specificity (1.0), F1 score (0.999), and geometric mean (0.999). This indicates that SMOTEENN provides the SVM classifier an exceptional balance between both classes. KMeansSMOTE also performs well, showing high precision (0.944), recall (0.995), F1 score (0.969), and geometric mean (0.967), though slightly lower than SMOTEENN across these metrics. BorderlineSMOTE and SMOTENC offer

https://doi.org/10.31436/iiumej.v26i1.3547

moderate performance, with BorderlineSMOTE achieving a good balance (e.g., precision of 0.945 and F1 score of 0.876) but still falling short of SMOTEENN and KMeansSMOTE. SMOTETomek has the lowest performance across metrics, indicating a relatively weaker balance in addressing class imbalance. Overall, SMOTEENN stands out as the optimal resampling technique for SVM, followed closely by KMeansSMOTE for strong and balanced results across metrics.

4.3.5. XGBoost Classifier

Using the XGBoost classifier and applying resampling techniques will significantly improve capturing the minority class compared to the baseline model (without resampling). Without resampling, XGBoost performs similarly to the SVM classifier, achieving perfect specificity (1.0) but with zero values for precision, recall, F1 score, and geometric mean, indicating that the model is heavily biased toward the majority class. This result highlights the necessity of resampling methods to enable XGBoost to recognize and classify instances from the minority class effectively.

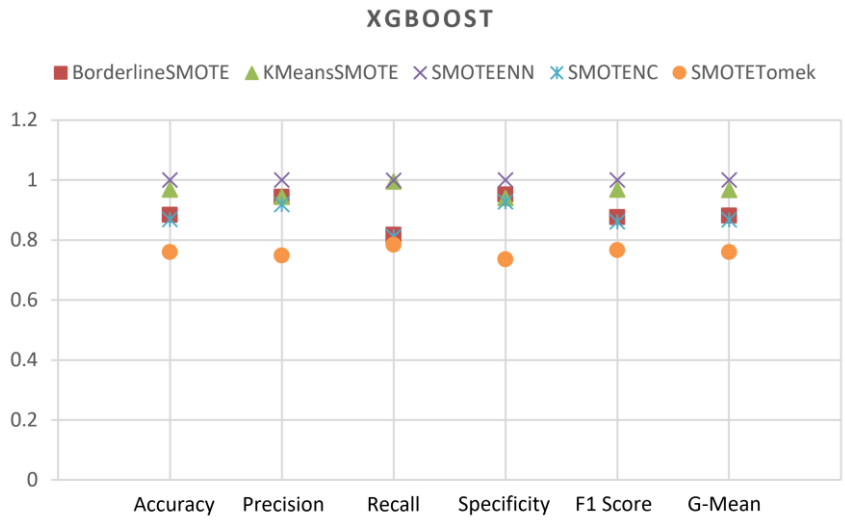

Figure 8. Model performance of SMOTE variants using XGBoost classifier.

Based on Fig. 8, we can see that among the resampling techniques, SMOTEENN again emerges as the best-performing method, achieving near-perfect metrics across the board with an accuracy of 0.9996, precision of 1.0, recall of 0.9995, specificity of 1.0, F1 score of 0.9998, and geometric mean of 0.9998. This suggests that SMOTEENN provides XGBoost with an optimal balance, allowing it to perform exceptionally well in both classes. KMeansSMOTE also performs strongly, with high values across precision (0.9434), recall (0.9951), F1 score (0.9686), and geometric mean (0.9673), making it a close alternative to SMOTEENN. BorderlineSMOTE and SMOTENC yield moderate performance, with BorderlineSMOTE achieving a good balance between metrics (e.g., F1 score of 0.8761), though lower than SMOTEENN and KMeansSMOTE. SMOTETomek shows the lowest performance, with lower scores across all metrics, indicating it may be less effective for balancing the dataset in this context. Overall, SMOTEENN is the optimal resampling technique for XGBoost, followed by KMeansSMOTE, as these two methods provide the most balanced and robust performance across all metrics.

https://doi.org/10.31436/iiumej.v26i1.3547

4.3.6. Results Across Different Classifiers

Based on the analysis results across different classifiers in Section 4.4.1. to 4.4.5, SMOTEENN consistently outperforms other SMOTE variants across four classifiers, demonstrating its strong effectiveness in handling class imbalance across different model types. SMOTEENN achieved near-perfect accuracy, precision, recall, F1 score, and geometric mean for most classifiers, demonstrating its ability to create a balanced representation of majority and minority classes. KMeansSMOTE, although not consistently dominant, excels in one classifier and provides a reliable alternative with high performance in most metrics, especially in recall and F1 scores. The comparison in Table 7 highlights SMOTEENN as the most versatile and reliable variant of SMOTE across a wide range of classifiers, making it the preferred choice for unbalanced data sets in various machine-learning models.

Table 7. The six model performance metrics of five SMOTE variants across five classifiers

Classifiers	SMOTE variants	Accuracy	Precision	Recall	Specificity	F1-score	G-Mean
LR	Initial	0.995	0.000	0.000	1.000	0.000	0.000
	BorderlineSMOTE	0.823	0.895	0.732	0.914	0.806	0.818
	KMeansSMOTE	0.962	0.933	0.995	0.928	0.963	0.961
	SMOTEENN	0.968	0.968	0.989	0.905	0.979	0.946
	SMOTENC	0.814	0.897	0.710	0.919	0.793	0.808
	SMOTETomek	0.720	0.720	0.720	0.720	0.720	0.720
NB	Initial	0.941	0.006	0.067	0.945	0.012	0.251
	BorderlineSMOTE	0.526	0.514	0.989	0.064	0.676	0.252
	KMeansSMOTE	0.950	0.912	0.995	0.904	0.952	0.949
	SMOTEENN	0.849	0.838	0.989	0.436	0.907	0.657
	SMOTENC	0.549	0.526	0.988	0.111	0.687	0.331
	SMOTETomek	0.551	0.527	0.992	0.111	0.688	0.331
RF	Initial	0.995	0.000	0.000	1.000	0.000	0.000
	BorderlineSMOTE	0.885	0.945	0.817	0.953	0.877	0.882
	KMeansSMOTE	0.968	0.944	0.995	0.941	0.969	0.968
	SMOTEENN	0.997	1.000	0.996	1.000	0.999	0.999
	SMOTENC	0.869	0.918	0.811	0.928	0.861	0.868
	SMOTETomek	0.761	0.749	0.785	0.737	0.766	0.760
SVM	Initial	0.995	0.000	0.000	1.000	0.000	0.000
	BorderlineSMOTE	0.885	0.945	0.817	0.953	0.876	0.882
	KMeansSMOTE	0.968	0.944	0.995	0.941	0.969	0.968
	SMOTEENN	0.999	1.000	0.999	1.000	0.999	0.999
	SMOTENC	0.869	0.893	0.838	0.900	0.865	0.868
	SMOTETomek	0.760	0.748	0.785	0.736	0.766	0.760
XGBoost	Initial	0.995	0.000	0.000	1.000	0.000	0.000
	BorderlineSMOTE	0.884	0.944	0.817	0.952	0.876	0.882
	KMeansSMOTE	0.968	0.943	0.995	0.940	0.969	0.967
	SMOTEENN	1.000	1.000	1.000	1.000	1.000	1.000
	SMOTENC	0.869	0.918	0.811	0.927	0.861	0.867
	SMOTETomek	0.760	0.748	0.785	0.736	0.766	0.760

In terms of F1-scores the very low initial results are primarily due to severe class imbalance in the dataset, where the minority class is vastly underrepresented. This imbalance causes classifiers to prioritize the majority class, resulting in high specificity (1.000) but near-zero recall and precision, leading to F1-scores of 0.000. The classifiers struggle to learn patterns for the minority class without sufficient examples, highlighting their limitations when applied to imbalanced datasets without preprocessing. The marked improvement in F1-scores after applying resampling strategies, such as SMOTEENN and KMeansSMOTE, underscores the critical role of addressing class imbalance to enhance model performance.

5. CONCLUSION

This study addresses the class imbalance problem in datasets labeled with tactics derived from CVE descriptions for exfiltration classification. Our analysis highlights the severe imbalance in the initial dataset, marked by skewed class distribution, dominant majority classes, and poor separability between classes. By applying five SMOTE variants and evaluating their impact on imbalance characteristics and model performance, we identify SMOTEENN as the most effective method for achieving consistent, near-perfect performance across multiple classifiers. SMOTEENN's ability to maintain balance while enhancing accuracy, precision, recall, and F1 scores demonstrates its adaptability to diverse model types.

Among alternative approaches, KMeansSMOTE also shows potential, particularly for scenarios requiring high recall and class separability. These findings underline the importance of selecting appropriate resampling techniques based on specific use cases. Moreover, our study contributes to addressing the lack of consensus on the best imbalance handling method by providing a multi-metric evaluation framework that reveals the strengths and weaknesses of each technique across different classifiers. This framework establishes a foundation for standardized evaluation metrics, promoting comparability and reproducibility across datasets.

While resampling strategies offer a promising solution to address data imbalance, several challenges remain. These include computational overhead and training time, the sequence and proportion of oversampling and undersampling steps that can significantly influence classification outcomes, and the scalability and generalizability of methods across varying datasets and attack types. Additionally, the handling of sensitive data and the development of privacy-preserving approaches for imbalanced datasets are critical challenges that warrant further research. Future work will focus on integrating these techniques into real-world enterprise environments and advancing privacy-preserving SMOTE methods for sensitive data, ensuring both accuracy and privacy. By addressing these challenges, this research contributes to the development of more robust machine learning models capable of handling imbalanced datasets effectively.

ACKNOWLEDGEMENT

The work of Arif Rahman Hakim was supported by the Lembaga Pengelola Dana Pendidikan (LPDP), Ministry of Finance of the Republic of Indonesia. Universitas Indonesia supported this work through the Hibah Publikasi Terindeks Internasional (PUTI) Kolaborasi Internasional (Q2) 2023 Scheme under Contract NKB-820/UN2.RST/HKP.05.00/2023.

REFERENCES

[1] Cybersecurity ventures, "Cybercrime to cost the world $9,5 Trillion USD Annually in 2024," 2024. [Online]. Available: https://www.esentire.com/resources/library/2023-official-cybercrime-report

[2] P. Verma, J. G. Breslin, D. O'Shea, N. Mehta, N. Bharot, and A. Vidyarthi, "Leveraging Gametic Heredity in Oversampling Techniques to Handle Class Imbalance for Efficient Cyberthreat Detection in IIoT," *IEEE Trans. Consum. Electron.*, vol. 70, no. 1, pp. 1940–1951, 2024, doi: 10.1109/TCE.2023.3319439.

[3] W. P. Chawla, Nitesh V and Bowyer, Kevin W and Hall, Lawrence O and Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[4] H. Han, W. Wang, and B. Mao, "Borderline-SMOTE : A New Over-Sampling Method in Imbalanced Data Sets Learning," pp. 878–879, 2005.

[5] G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE," *Inf. Sci. (Ny).*, vol. 465, pp. 1–20, 2018, doi: 10.1016/j.ins.2018.06.056.

[6] M. Mukherjee and M. Khushi, "Smote-enc: A novel smote-based method to generate synthetic data for nominal and continuous features," *Appl. Syst. Innov.*, vol. 4, no. 1, 2021, doi: 10.3390/asi4010018.
[7] 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 Explor. Newsl.*, vol. 6, no. 1, pp. 20–29, 2004, doi: 10.1145/1007730.1007735.
[8] G. E. a P. a Batista, a L. C. Bazzan, and M. C. Monard, "Balancing Training Data for Automated Annotation of Keywords: a Case Study," *Rev. Tecnol. da Informação*, vol. 3, no. 2, pp. 15–20, 2004.
[9] N. H. N. B. M. Shahri, S. B. S. Lai, M. B. Mohamad, H. A. B. A. Rahman, and A. Bin Rambli, "Comparing the performance of adaboost, xgboost, and logistic regression for imbalanced data," *Math. Stat.*, vol. 9, no. 3, pp. 379–385, 2021, doi: 10.13189/ms.2021.090320.
[10] C. K. Aridas, S. Karlos, V. G. Kanas, N. Fazakis, and S. B. Kotsiantis, "Uncertainty Based Under-Sampling for Learning Naive Bayes Classifiers under Imbalanced Data Sets," *IEEE Access*, vol. 8, pp. 2122–2133, 2020, doi: 10.1109/ACCESS.2019.2961784.
[11] V. Ganganwar, "An overview of classification algorithms for imbalanced datasets," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 2, no. 4, pp. 42–47, 2012, [Online]. Available: http://www.ijetae.com/files/Volume2Issue4/IJETAE_0412_07.pdf
[12] B. Pes, "Learning from high-dimensional and class-imbalanced datasets using random forests," *Information*, vol. 12, no. 8, p. 286, 2021.
[13] S. He, B. Li, H. Peng, J. Xin, and E. Zhang, "An Effective Cost-Sensitive XGBoost Method for Malicious URLs Detection in Imbalanced Dataset," *IEEE Access*, vol. 9, pp. 93089–93096, 2021, doi: 10.1109/ACCESS.2021.3093094.
[14] A. Kuppa, L. Aouad, and N. A. Le-Khac, "Linking CVE's to MITRE ATT and CK Techniques," *ACM Int. Conf. Proceeding Ser.*, 2021, doi: 10.1145/3465481.3465758.
[15] S. EL JAOUHARI, N. TAMANI, and R. I. JACOB, "Improving ML-based Solutions for Linking of CVE to MITRE ATT &CK Techniques," *2024 IEEE 48th Annu. Comput. Software, Appl. Conf.*, pp. 2442–2447, 2024, doi: 10.1109/compsac61105.2024.00392.
[16] I. Branescu, O. Grigorescu, and M. Dascalu, "Automated Mapping of Common Vulnerabilities and Exposures to MITRE ATT&CK Tactics," *Inf.*, vol. 15, no. 4, pp. 1–20, 2024, doi: 10.3390/info15040214.
[17] M. D. and R. R. Octavian Grigorescu , Andreea Nica, "CVE2ATT & CK : BERT-Based Mapping of CVEs to MITRE," *Algorithms*, pp. 1–22, 2022.
[18] T. Al-Shehari and R. A. Alsowail, "Random resampling algorithms for addressing the imbalanced dataset classes in insider threat detection," *Int. J. Inf. Secur.*, vol. 22, no. 3, pp. 611–629, 2023, doi: 10.1007/s10207-022-00651-1.
[19] I. Branescu, O. Grigorescu, and M. Dascalu, "Automated Mapping of Common Vulnerabilities and Exposures to MITRE ATT&CK Tactics," *Inf.*, vol. 15, no. 4, pp. 1–19, 2024, doi: 10.3390/info15040214.
[20] I. Branescu, "Open source the dataset and the code used for Automated Mapping CVE to ATT&CK," Github. Accessed: Oct. 16, 2024. [Online]. Available: https://github.com/readerbench/CVE2ATT-CK-tactics
[21] T. K. Ho and M. Basu, "Complexity measures of supervised classification problems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 289–300, 2002, doi: 10.1109/34.990132.