STSMOTE: A Controllable Data Augmentation Method for Few-Shot Zero-Day Attack Detection in Network Intrusion Detection Systems

Haruto Ishii*, Kunio Akashi†, and Yuji Sekiya*
*Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan
†Information Technology Center, The University of Tokyo, Tokyo, Japan
e-mail: harutowoody@g.ecc.u-tokyo.ac.jp, {k-akashi, sekiya}@nc.u-tokyo.ac.jp

***Abstract*—Deep learning-based Network Intrusion Detection Systems are effective in detecting cyberattacks; however, they suffer from significant performance degradation due to class imbalance. This issue is particularly critical in scenarios with minimal data, such as the initial response to zero-day attacks. This necessitates effective data augmentation methods tailored for such few-shot situations. In this paper, we propose Safe-Tube Synthetic Minority Over-sampling Technique (STSMOTE), a controllable data augmentation method designed to function effectively even under these extreme conditions. STSMOTE generates safe and diverse synthetic data by integrating two geometric strategies: Sphere, which augments the local neighborhood of minority samples, and Tube, which captures the structural relationships between samples. Experimental results using the CIC-IDS2018 dataset and 8 baseline methods demonstrated that STSMOTE achieved the best performance in 14 out of 16 cases, using only 5 training samples per attack class. Notably, statistical significance ($p < 0.05$) was confirmed in 11 of those cases, with 6 cases achieving $p < 0.001$. Furthermore, even under the extreme condition of only 2 training samples, it achieved the best performance in 14 out of 16 cases, demonstrating exceptional robustness. Crucially, unlike black-box generative models, our method allows for quantitative control of the generation range via geometric parameters, representing a significant contribution to the security domain, where reliability is paramount.**

Keywords-network intrusion detection system; zero-day attacks; data augmentation; controllability; security; IoT.

I. Introduction

While the convenience of the Internet has improved dramatically in recent years, this progress has been accompanied by an escalation in cyber threats. Network Intrusion Detection Systems (NIDS) serve as a critical defense mechanism by continuously monitoring traffic to alert administrators of malicious behavior. Recently, research utilizing Machine Learning (ML) and Deep Learning (DL) has gained significant traction to address the intrinsic limitations of conventional signature-based methods against unknown or zero-day attacks [1][2]. These technologies hold the potential to detect unknown attack patterns; however, their performance is heavily dependent on the quality and quantity of the training data. Specifically, NIDS datasets often suffer from severe "class imbalance," where benign traffic constitutes the vast majority, while malicious traffic is extremely rare. In particular, samples of zero-day attacks or sophisticated targeted attacks in the real world are exceedingly scarce. Constructing high-precision detection models in such "few-shot" scenarios—where only a handful of attack samples are available—remains a pressing challenge.

To address this imbalance, oversampling techniques, such as Synthetic Minority Over-sampling Technique (SMOTE) [3] and Adaptive Synthetic Sampling (ADASYN) [4], are widely adopted as standard approaches. These methods synthetically augment data by performing linear interpolation between samples of the minority class. However, in environments characterized by complex distributions or extreme data scarcity (as is common in NIDS datasets), these methods exhibit critical flaws. The first is the "encroachment of decision boundaries." As shown in Figure 1, when minority class samples are sparsely distributed, existing methods often generate data across the distribution region of the majority class, creating noise that hinders classifier learning. The second is the "lack of generative diversity." Since simple linear interpolation generates data strictly along the line segments connecting original samples, it lacks the spatial expansion necessary to adequately represent the variations of unknown attacks.

The primary objective of this study is to propose a novel data augmentation method capable of generating diverse data without compromising the classifier's decision boundary, even in few-shot scenarios where available attack samples are extremely limited. The proposed method, Safe-Tube SMOTE (STSMOTE), balances safety and diversity by geometrically defining a "Safe Tube" around source data points and applying probabilistic perturbations within this constrained region. In this paper, we evaluate and discuss the effectiveness and robustness of the proposed method, particularly under severe conditions with minimal sample sizes (e.g., $N = 2$).

The main contributions of this study are as follows:

- Proposal of a novel data augmentation method effective in scenarios with extremely scarce attack samples.
- Comprehensive performance comparison against 8 existing methods (including state-of-the-art techniques) across diverse experimental settings with varying sample sizes, attack types, and classifiers.
- Release of the source code to facilitate reproducibility and further research within the community [5].

The remainder of the paper is organized as follows: Section II reviews related work. Section III details the proposed STSMOTE. Section IV presents the experimental evaluation, and Section V concludes the paper.

II. Related Work

In NIDS, class imbalance significantly degrades detection accuracy for specific attacks. The most common approach

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

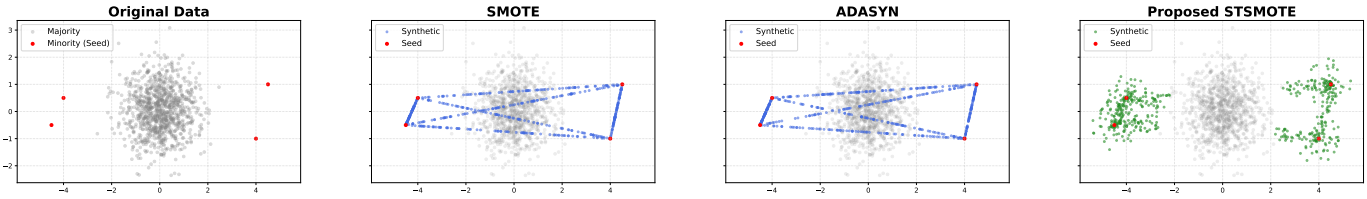

Figure 1. Illustrative example showing the behavior of oversampling methods under extremely limited minority samples. SMOTE and ADASYN generate samples across the majority region, whereas the proposed STSMOTE produces diverse samples while avoiding decision boundary violation.

to mitigate this issue is data augmentation (oversampling), which artificially increases the minority class. Among these techniques, SMOTE [3], which generates data by performing linear interpolation between neighboring minority class samples, is a representative method used in numerous studies [1][2]. Subsequently, SMOTE derivatives (e.g., ADASYN [4][6], Borderline-SMOTE (BSMOTE) [7][8]) have been applied to improve NIDS performance. While these methods have proven effective when a sufficient number of minority class samples are available, their effectiveness has not been sufficiently verified in "few-shot" scenarios where training data is extremely scarce. Furthermore, it has been pointed out that these methods tend to generate data lacking in diversity [9].

In recent years, DL-based methods like Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) have been widely studied for resolving NIDS class imbalance [9][10]. Moreover, research utilizing diffusion model-based methods, which demonstrate superior performance over GAN and VAE-based approaches [11], has emerged and shown promising results [12]. However, these DL-based methods generally require large datasets, and their stability in extremely limited, zero-day attack scenarios remains insufficiently validated.

As countermeasures for class imbalance, approaches other than data augmentation have been proposed, such as redesigning loss functions (represented by Focal Loss) [13] and improving model architectures [14]. While these are promising approaches that address imbalance at the algorithm level, they have the drawback of requiring designs dependent on specific model structures or loss functions. In contrast, data augmentation offers the advantage of being model-agnostic, allowing it to be easily integrated into any existing classifier or pipeline. Therefore, this study focuses on data augmentation techniques specifically designed for few-shot scenarios ($N = 2 - 5$), which represent the initial response phase to zero-day attacks where immediate model deployment is required with minimal available samples.

III. Proposed Method: STSMOTE

While SMOTE risks violating decision boundaries in sparse data, and Random Oversampling (RO) often induces overfitting due to simple duplication, we propose STSMOTE to integrate their advantages. STSMOTE consists of two generation strategies: Sphere, which spatially extends the concept of RO, and Tube, which safely applies the concept of SMOTE. Crucially, in few-shot scenarios ($N \leq 5$), statistical distribution estimation (e.g., via Mahalanobis distance) becomes highly unstable due to data scarcity; therefore, we employ deterministic geometric constraints that function robustly without relying on density estimation. This design enables the generation of diverse data while avoiding collisions with the majority class, even under such extreme conditions. Safety is defined conservatively with respect to observed majority samples, as guaranteeing separation from unseen majority distributions is fundamentally infeasible.

A. Sphere Generation Strategy

First, we describe the Sphere strategy, which is an evolution of RO. While conventional RO merely duplicates data, Sphere introduces spatial expansion by adding perturbations to data points, thereby preventing model overfitting. However, since indiscriminate noise injection invites intrusion into the majority class region, this method defines a geometrically safe generation region using the following procedure (illustrated in Figure 2(a)):

1. **Preprocessing:** We apply the Quantile Transform to numerical features. This technique is widely adopted in synthetic data generation to effectively mitigate scale discrepancies and handle non-Gaussian distributions. Both Tube and Sphere strategies operate exclusively on these transformed numerical features.
2. **Determination of Safety Radius:** For each minority class data point m_i, we search for the nearest majority class data point O_i and calculate the Euclidean distance $d_i = \|m_i - O_i\|$. We then define the safety radius r_i for generated data as $r_i = d_i/2$.
3. **Data Generation:** Random data points are generated within a hypersphere of radius r_i centered at m_i.

The radius setting of $d_i/2$ geometrically guarantees that any generated data point remains closer to its source minority sample than to any observed majority sample. This process is applied to all minority samples to achieve safe and diverse data augmentation.

B. Tube Generation Strategy

The Tube strategy spatially extends the concept of linear interpolation used in SMOTE. It defines a safe cylindrical (or conical) generation region around the line segment connecting two minority class data points without encroaching on the distribution of the majority class. In SMOTE, data is generated only on the line segment, which tends to limit diversity. In contrast, Tube aims to generate data that maintains safety while enhancing diversity. This is achieved by allowing perturbations in the direction orthogonal to the line segment.

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

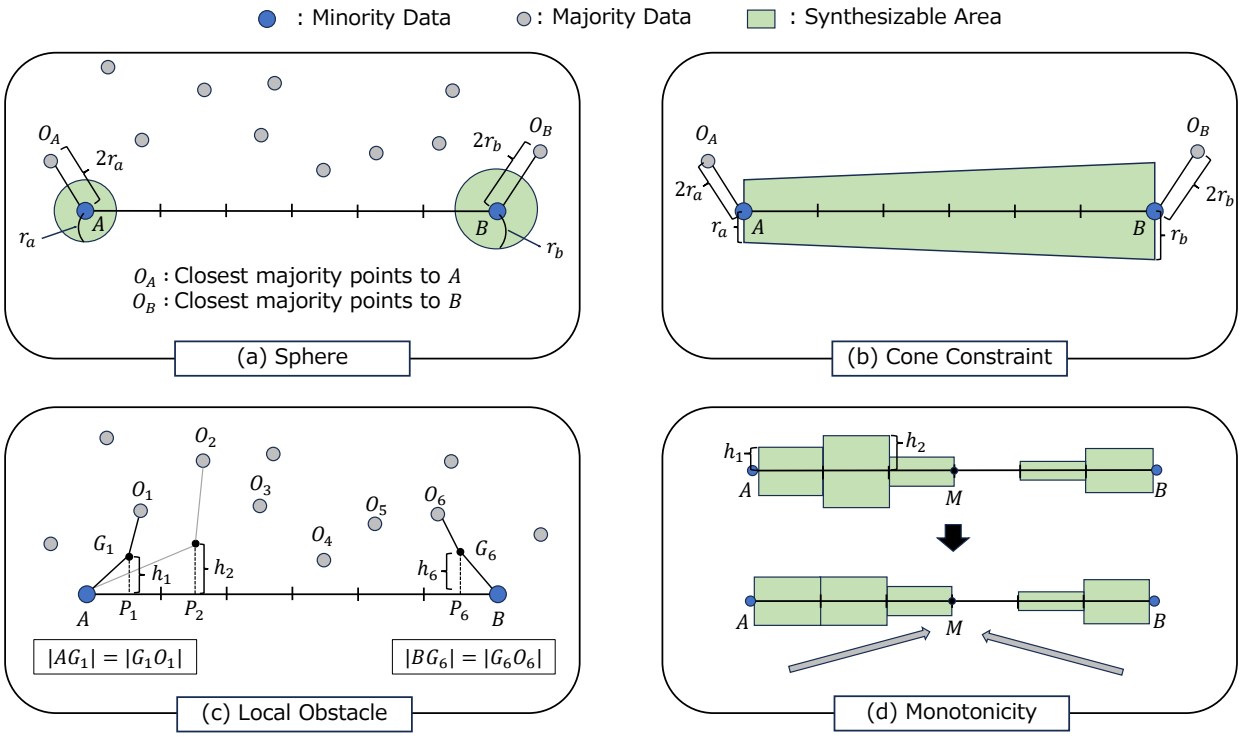

Figure 2. Geometric illustration of data generation strategies and safety constraints. (a) Sphere strategy generates data within a hypersphere of radius r_a around each minority point, where r_a is half the distance to the nearest majority point O_A. (b) Cone constraint limits the perturbation magnitude to form a truncated cone connecting Sphere radii at endpoints A and B. (c) Local obstacle constraint ensures generated points (e.g., G_1, G_6) remain closer to the minority segment than to nearby majority points (O_1, O_6). (d) Monotonicity constraint enforces decreasing perturbation magnitude from endpoints toward midpoint M. Constraints are applied in the order (b) → (c) → (d).

1) Calculation of Perturbation Magnitude: First, two arbitrary points A and B are selected from the minority class dataset to form a line segment AB. The segment is divided into K intervals, where a sufficiently large K approximates complex majority boundaries while managing computational costs. We then determine the safe perturbation magnitude h_i for each interval i through the following three-step process.

Step 1: Cone Constraint (Initialization). To ensure a smooth connection between Tube and Sphere, we initialize the perturbation magnitude using the Sphere radii r_a and r_b at endpoints A and B. Assume interval i is located at a position dividing the segment AB in the ratio $t : (1-t)$. The initial magnitude h_i is set to linearly interpolate the radii:

$$h_i = (1-t)r_a + tr_b \tag{1}$$

This constrains the Tube shape to resemble a truncated cone connecting the two Spheres, as illustrated in Figure 2(b).

Step 2: Local Obstacle Constraint. While the Cone constraint defines the global shape, it does not account for specific majority samples encroaching into this region. Therefore, we tighten the constraint based on nearby majority class data ("local obstacles"), as shown in Figure 2(c). We calculate the orthogonal projection of all majority points onto the segment AB. For interval i, if a majority point O projects onto it, we calculate the maximum safe height h_{safe}. Let P_i, h_{dist}, r_{dist}, and r_i be the midpoint of interval i, the perpendicular distance from obstacle O, the distance along the segment from O's projection to the nearest endpoint A, and the distance from P_i to A, respectively.

The value h_{safe} is derived from the condition that the generated point G (at height h_{safe}) must be closer to the nearest minority endpoint A than to the obstacle O (i.e., $\|G-A\| \leq \|G-O\|$). We express the squared Euclidean distances as $\|G-A\|^2 = r_i^2 + h_{safe}^2$ and $\|G-O\|^2 = (r_i - r_{dist})^2 + (h_{safe} - h_{dist})^2$. By solving the inequality derived from these distances for h_{safe}, we obtain:

$$h_{safe} = \frac{r_{dist}^2 + h_{dist}^2 - 2r_i r_{dist}}{2h_{dist}} \tag{2}$$

We update $h_i \leftarrow \min(h_i, h_{safe})$. If multiple obstacles exist in the same interval, the minimum value among them is adopted.

Step 3: Monotonicity Constraint. Finally, we apply a smoothing constraint. The midpoint M of the segment is the furthest point from the known minority samples and has high uncertainty. Therefore, we enforce that h_i decreases monotonically from the endpoints toward the midpoint. Specifically, comparing an interval i closer to the endpoint and its inner adjacent interval $i+1$, if $h_i < h_{i+1}$, we update $h_{i+1} \leftarrow h_i$. This prevents the tube from expanding unnaturally near the midpoint, as depicted in Figure 2(d).

2) Data Generation: Through the calculations above, the safe generation region is determined. During data generation, we randomly select an interval where the allowable perturbation magnitude $h > 0$. For a point on the segment within that

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

interval, we generate and add random noise in the range [0, h] in the direction perpendicular to the segment to create a new data point. Note that if $h \leq 0$ in all intervals, or if points A and B are identical, we perform Sphere generation for the same number of samples instead of Tube generation. This serves as a fallback strategy to prioritize safety while ensuring data generation does not stop.

C. Generation Ratio and Handling of Categorical Data

In the proposed method, the ratio of data generation between Sphere and Tube is set to $1:1$. Although various allocation strategies could be considered based on the number of minority samples N or the theoretical volume of the generation space, we adopt a simple equal allocation strategy. This approach ensures a balanced contribution from both local neighborhood augmentation (Sphere) and structural connectivity augmentation (Tube), avoiding bias towards either strategy while minimizing the complexity of hyperparameter tuning.

Furthermore, the geometric perturbations in this method are applied only to numerical features. Note that these operations are performed on features scaled by the Quantile Transform. After the generation process, the inverse Quantile Transform is applied to restore the synthetic numerical data to its original scale. For categorical features, we adopt a strategy of copying values from the nearest minority class data point without exploring new combinations. Specifically, we adopt the values of the center point m_i for Sphere, and the values of the endpoint closer to the generated point for Tube.

In the context of network traffic, categorical features (e.g., protocols, services, and flags) often possess strict semantic dependencies. Generating arbitrary combinations of these features poses a risk of creating unrealistic packet headers that do not exist in real-world environments (e.g., setting Transmission Control Protocol (TCP) flags on a User Datagram Protocol (UDP) packet). Such semantic noise can lead to an increase in False Positives (FPs) in the detection model. In actual security operations, excessive FPs cause alert fatigue for Security Operations Center (SOC) personnel, significantly degrading operational efficiency. Therefore, we prioritized operational reliability by adopting a conservative approach that strictly preserves observed categorical combinations, thereby ensuring that the augmented data remains semantically valid and safe.

IV. Experiments

This section presents the experimental evaluation of the proposed method. We detail the dataset, baseline methods, and classifiers used, followed by a discussion of the results.

A. Dataset and Preprocessing

In this experiment, we utilized the CIC-IDS2018 dataset [15], which is widely used in the field of NIDS. First, the entire dataset was split into training and testing sets with a 4:1 ratio via stratified sampling to maintain class distribution. Among the classes, four types with fewer than 5,000 original training samples (Brute Force Web (Web), Brute Force Cross-Site Scripting (XSS), Distributed Denial of Service attack Low Orbit Ion Cannon UDP (DDoS), and Structured Query Language Injection (SQL)) were defined as minority classes, while the remaining 11 classes, all of which exceeded this threshold, were defined as majority classes.

While real-world NIDS environments exhibit even more extreme imbalance, to unify experimental conditions and isolate the performance of augmentation methods, the majority classes were undersampled to 5,000 samples each for training and 1,250 samples each for testing, and were fixed as a common dataset throughout all experiments. This design choice serves two purposes: (1) it eliminates bias from pre-existing class imbalances among majority classes, which could confound the evaluation of augmentation effectiveness; (2) it establishes a clear target for minority class augmentation (5,000 samples), enabling fair comparison across methods. On the other hand, to simulate a few-shot scenario, one of the four minority classes was selected as the target. We conducted separate experiments for two distinct conditions where the training samples were randomly subsampled to $N = 5$ and $N = 2$, respectively (55,000 majority vs. N minority samples). All remaining target class data were added to the test set to ensure sufficient samples for statistically reliable evaluation. Subsequently, the subsampled minority data were oversampled using the augmentation methods described below until they reached 5,000 samples to match the majority class count. This evaluation process was conducted individually for each of the four minority classes as the target.

B. Baseline Methods and Implementation

To verify the effectiveness of the proposed STSMOTE, we compared it with a total of 8 baseline methods:

- **Classical Methods:** No Augmentation (No Aug.), RO, SMOTE [3], ADASYN [4], BSMOTE [8].
- **Deep Learning Methods:** Tabular VAE (TVAE) [16], Conditional Tabular GAN (CTGAN) [16], and TabSyn [11].

Our selection comprehensively covers the spectrum of augmentation techniques: from widely adopted density-based and boundary-aware SMOTE variants in the NIDS domain (e.g., ADASYN, Borderline-SMOTE) to state-of-the-art deep generative models, including the diffusion-based TabSyn. Regarding the handling of categorical features, since the standard implementation of SMOTE does not support categorical data, we utilized SMOTE-Nominal Continuous (SMOTENC) as an alternative. For other classical methods that inherently lack support for categorical features, we applied the augmentation only to numerical features, while categorical values were imputed by copying those from the nearest neighbor in the minority class. For the proposed STSMOTE, the number of Tube intervals was set to $K = 1000$. This value was empirically chosen as a sufficiently large number to approximate complex majority-class boundaries while keeping the computational cost manageable. Regarding the hyperparameters of the baseline methods, we primarily adopted the recommended values or default settings of each method. However, for methods that do not function with extremely small datasets ($N < 5$)—such as

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

TABLE I. COMPARISON OF F1-SCORES ON $N = 5$. BOLD INDICATES THE BEST PERFORMANCE. ASTERISKS IN THE ROW BELOW STSMOTE DENOTE STATISTICAL SIGNIFICANCE LEVELS COMPARED TO ALL OTHER METHODS (HOLM-CORRECTED WILCOXON SIGNED-RANK TEST): $^{*}p < 0.05$, $^{**}p < 0.01$, AND $^{***}p < 0.001$.

Method	Web				XSS				DDoS				SQL			
	Cat	**XGB**	**LGB**	**TabM**	**Cat**	**XGB**	**LGB**	**TabM**	**Cat**	**XGB**	**LGB**	**TabM**	**Cat**	**XGB**	**LGB**	**TabM**
No Aug.	0.4422	0.4820	0.0153	0.4550	0.6905	0.6669	0.2245	0.7270	0.7551	0.6577	0.0000	0.6138	0.4485	0.3876	0.0001	0.5815
RO	0.7296	0.5522	0.5389	0.6552	0.8288	0.7142	0.7323	0.8014	0.8682	0.7165	0.6991	0.8452	0.8165	0.5399	0.5063	0.7233
SMOTE	0.6214	0.5711	0.5565	0.5973	0.8138	0.7469	0.7489	0.8056	0.9065	0.8292	0.7832	0.8851	0.7570	0.5846	0.5589	0.7547
BSMOTE	0.6334	0.5714	0.4374	0.6284	0.8476	0.7487	0.7723	0.8408	0.9033	0.8305	0.7849	**0.8876**	0.7394	0.5762	0.4810	0.7557
ADASYN	0.6907	0.6527	0.5797	0.6564	0.8621	0.7598	0.6948	0.8478	0.8458	0.7587	0.5664	0.8554	0.8056	0.6020	0.5576	0.8027
CTGAN	0.4234	0.5301	0.5793	0.4740	0.6993	0.7207	0.8156	0.7008	0.8847	0.7754	0.8376	0.5033	0.5179	0.4946	0.5466	0.5928
TVAE	0.4233	0.5287	0.5523	0.4756	0.7128	0.7424	0.7969	0.7388	0.9333	0.8224	**0.8752**	0.6771	0.5977	0.5196	0.5393	0.5759
TabSyn	0.6770	0.6218	0.5921	0.6410	0.8674	0.7376	0.7827	0.8305	0.8610	0.7081	0.7435	0.8779	0.8100	0.5921	0.5397	0.7783
STSMOTE	**0.7584**	**0.7215**	**0.6686**	**0.7203**	**0.9004**	**0.8412**	**0.8371**	**0.8663**	**0.9441**	**0.8742**	0.8741	0.8817	**0.8441**	**0.7653**	**0.7168**	**0.8212**
(Sig.)	***	***	**	*	***	***	-	*	-	**	-	-	**	***	***	-

those requiring a minimum number of neighbors for k-Nearest neighbors (k-NN)—minimal adjustments were made to ensure operation. For full details of these hyperparameters, please refer to the publicly available implementation code [5]. All experiments were conducted on a computing node equipped with an AMD EPYC 7532 32-Core Processor and an NVIDIA A100 40GB Graphics Processing Unit (GPU).

C. Classifiers and Evaluation Metrics

To evaluate the quality of the data generated by augmentation, we employed three gradient boosting decision tree models widely used in NIDS (CatBoost, XGBoost, LightGBM) and TabM [17], a state-of-the-art Multi-Layer Perceptron (MLP)-based deep learning model for tabular data. These classifiers were trained on a multi-class classification task involving 12 classes (11 majority classes and 1 target minority class). To fairly assess the impact of the augmentation methods, the hyperparameters of the classifiers were set to their defaults.

We adopted the F1-Score as the primary metric to evaluate performance on imbalanced data. To focus on the detection capability for the specific target class within the multi-class framework, we calculated the metric in a one-vs-rest setting, treating the target minority class as Positive and all other classes (majority classes) as Negative.

To mitigate the impact of variability due to the randomness of undersampling, we repeated the experiment 25 times for each target class, each time using a different random seed for minority class extractions. Performance was evaluated based on the average F1-Score across these trials. Given the high sensitivity to data selection, we employed Average Rank across trials as a robust stability metric instead of standard deviation. Furthermore, to verify statistical significance, we performed the Wilcoxon signed-rank test with Holm's correction (significance level $p < 0.05$, $p < 0.01$, $p < 0.001$).

D. Results and Discussion

First, Table I presents the F1-Scores when the number of training samples was set to $N = 5$. Across the 16 experimental cases combining four types of attacks and four types of models, STSMOTE achieved the best performance in 14 cases. Notably, in 11 of these cases, a statistically significant difference ($p < 0.05$) was confirmed against all other methods, with 6 cases meeting the stricter level of $p < 0.001$. Even in the two cases where STSMOTE did not rank first, the margin was negligible (< 0.01); conversely, in successful cases, it outperformed the runner-up by over 0.16 in some instances. Furthermore, STSMOTE recorded an overall average rank of 2.01, marking a substantial improvement over the best baseline method, ADASYN (rank 4.00), and demonstrating high stability.

Regarding trends by attack class, while STSMOTE ranked first in all classes other than DDoS, for DDoS, statistical significance was confirmed only with XGBoost, and performance was slightly inferior to other methods in two cases. This is likely attributable to the characteristics of the DDoS data. As shown in Table I, the baseline F1-Scores for DDoS are high—reaching 0.9441 with CatBoost—and the performance gaps among methods are narrow. This suggests that the DDoS samples retain sufficient representativeness to predict the overall distribution even with limited data; thus, any augmentation method contributes to accuracy improvement provided it does not generate data that deviates significantly from the original distribution. However, even under such conditions, STSMOTE maintained the best average rank across the four models, demonstrating stable performance.

Next, Table II presents the results under the even more severe condition where the number of training samples was set to $N = 2$. Even in this extreme scenario, STSMOTE achieved the best performance in 14 out of 16 cases, with statistical significance ($p < 0.05$) confirmed in 10 cases. In terms of average rank, it recorded 2.19, significantly outperforming the best baseline method, SMOTE, which scored 4.25. This suggests that the proposed method functions stably without performance collapse even when the data is extremely scarce.

Furthermore, we evaluated the risk of over-generation encroaching on the majority class. STSMOTE recorded an average of only 1.2625 FPs (0.01% of the 13,750 majority test samples), which is even lower than the 4.845 average of the No Augmentation. This result quantitatively proves that the geometric constraints of our Sphere and Tube strategies successfully achieved "safe" augmentation without violating

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

TABLE II. COMPARISON OF F1-SCORES ON $N = 2$. STATISTICAL SIGNIFICANCE NOTATIONS AND FORMATTING FOLLOW TABLE I.

Method	Web				XSS				DDoS				SQL			
	Cat	**XGB**	**LGB**	**TabM**	**Cat**	**XGB**	**LGB**	**TabM**	**Cat**	**XGB**	**LGB**	**TabM**	**Cat**	**XGB**	**LGB**	**TabM**
No Aug.	0.2435	0.2076	0.0272	0.1853	0.5646	0.4540	0.0215	0.3826	0.3244	0.3212	0.0000	0.2879	0.1433	0.0640	0.0097	0.1884
RO	0.4763	0.3143	0.2597	0.3632	0.6940	0.4966	0.5713	0.6399	0.7150	0.4769	0.4199	0.6385	0.5842	0.2246	0.1609	0.4428
SMOTE	0.4081	0.2973	0.2822	0.4218	0.6900	0.5280	0.5996	0.6923	0.6982	0.4788	0.4789	0.5778	0.5355	0.2011	0.2821	0.5589
BSMOTE	0.2634	0.2089	0.0621	0.1990	0.5748	0.4243	0.2234	0.3891	0.6997	0.4868	0.4786	0.6621	0.1628	0.0700	0.0186	0.2012
ADASYN	0.4105	0.2995	0.2363	0.4203	0.6795	0.5588	0.3820	0.6816	0.6415	0.4830	0.4116	0.5705	0.5180	0.2132	0.2711	0.5519
CTGAN	0.2729	0.2561	0.3446	0.2709	0.6064	0.5338	0.6398	0.5849	0.6479	0.4711	0.5504	0.3020	0.2552	0.1280	0.2702	0.3261
TVAE	0.2978	0.2706	0.3493	0.2983	0.6239	0.5522	0.6432	0.6347	**0.8096**	0.5763	**0.6322**	0.3347	0.3799	0.2096	0.2583	0.3117
TabSyn	0.4588	0.3042	0.2778	0.3852	0.6950	0.5552	0.5612	0.6554	0.6677	0.4820	0.4397	0.6642	0.5608	0.1905	0.1684	0.4694
STSMOTE	**0.4831**	**0.4178**	**0.3662**	**0.4575**	**0.7265**	**0.6968**	**0.6888**	**0.7031**	0.7765	**0.7499**	0.5588	**0.7250**	**0.6230**	**0.4888**	**0.4603**	**0.6234**
(Sig.)	*	***	-	*	*	***	-	-	-	**	-	-	**	***	***	**

TABLE III. COMPARISON OF F1-SCORES FOR REPRESENTATIVE METHODS ON ADDITIONAL DATASETS. MAJORITY SAMPLES WERE ADJUSTED TO 5,000 (UNSW-NB15) AND 4,000 (CIC-IDS2017). STATISTICAL SIGNIFICANCE NOTATIONS AND FORMATTING FOLLOW TABLE I.

Method	**Shellcode** ($N = 5$)				**Shellcode** ($N = 2$)				**Bot** ($N = 5$)				**Bot** ($N = 2$)			
	Cat	**XGB**	**LGB**	**TabM**	**Cat**	**XGB**	**LGB**	**TabM**	**Cat**	**XGB**	**LGB**	**TabM**	**Cat**	**XGB**	**LGB**	**TabM**
No Aug.	0.0252	0.1090	0.0307	0.1600	0.0012	0.0060	0.0053	0.0292	0.5880	0.6423	0.0725	0.5935	0.2939	0.3069	0.0361	0.2369
RO	0.4323	0.3296	0.3038	0.3364	0.1660	0.0980	0.0615	0.0846	0.8943	0.7712	0.7200	0.8120	0.6064	0.5332	0.5279	0.4930
SMOTE	0.5533	0.4282	0.3260	0.5401	0.2320	0.1410	0.0762	0.2193	0.7977	0.7881	0.7355	0.8037	0.5335	0.5228	0.5152	0.5285
BSMOTE	0.1319	0.1818	0.1030	0.2196	0.0012	0.0060	0.0053	0.0292	0.8091	0.7743	0.6461	0.7664	0.3142	0.3354	0.0719	0.2526
ADASYN	0.5618	0.4501	0.3441	0.5647	0.2342	0.1412	0.0735	0.2232	0.8746	0.7916	0.7444	0.8340	0.5568	0.5107	0.4982	0.5164
CTGAN	0.0428	0.1275	0.1972	0.1870	0.0056	0.0188	0.0284	0.0621	0.6137	0.6959	0.7179	0.6726	0.3594	0.3979	0.4786	0.4392
TVAE	0.2655	0.2666	0.2484	0.1905	0.1217	0.0774	0.0628	0.0787	0.6596	0.7326	0.7755	0.7381	0.4211	0.4606	**0.5394**	0.4659
TabSyn	0.4985	0.3857	0.2959	0.4302	0.1556	0.0980	0.0702	0.0830	0.8727	0.7863	0.7288	0.8049	0.6009	0.5339	0.5130	0.4971
STSMOTE	**0.5821**	**0.5817**	**0.4853**	**0.6081**	**0.2591**	**0.2293**	**0.1757**	**0.2984**	**0.9222**	**0.8649**	**0.8488**	**0.8715**	**0.6403**	**0.5778**	0.5142	**0.5478**
(Sig.)	-	***	***	**	-	***	***	***	***	***	***	***	**	-	-	-

the majority class boundaries.

Finally, we address the generalizability of the proposed method. While this study evaluated performance using four distinct attack classes within CIC-IDS2018, we conducted additional experiments on the UNSW-NB15 [18] and CIC-IDS2017 [15] datasets, targeting the Shellcode class and the Bot class as minority classes, respectively. These classes were selected because they possess different characteristics from the attack classes used in the main experiments. For these experiments, we adhered to the same evaluation protocol and applied all 8 baseline methods used in the main analysis. Table III presents these comprehensive comparison results. Across 16 cases ($N = 5, 2$), STSMOTE achieved the best F1-Score in 15 cases with an average rank of 1.75 (surpassing the best baseline ADASYN at 3.34). Furthermore, statistical significance was confirmed in 11 cases, 9 of which achieved the strictest level of $p < 0.001$. Notably, for the Bot class at $N = 5$, statistical significance ($p < 0.001$) was confirmed across all classifiers. Moreover, the overall average rank of 1.75 in these additional experiments even surpassed the average ranks obtained in the main experiments. These results robustly verify that the effectiveness of the proposed method is not dataset-dependent. Detailed experimental data, such as raw p-values and the source code, are available in our repository [5].

In our experimental environment, STSMOTE completed in approximately one second for $N = 5$, comparable to SMOTE and significantly faster than the diffusion-based TabSyn (55 minutes). Although the Tube strategy involves $O(N^2)$ complexity, this cost is negligible in few-shot scenarios and can be mitigated by limiting Tube combinations to k-NN.

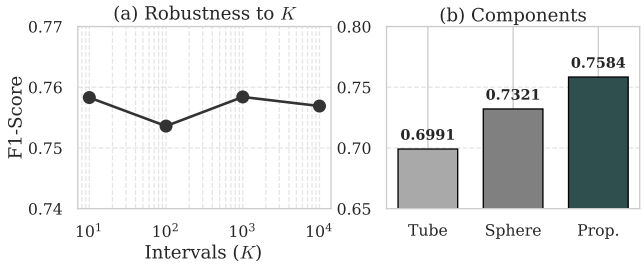

Figure 3. Ablation study on Web ($N = 5$, CatBoost). (a) Robustness to tube interval parameter K. (b) Contribution of generation strategies.

Furthermore, even with $N = 100$, the process completed within approximately 4 minutes, demonstrating practical scalability for scenarios where attack samples accumulate over time.

E. Ablation Study

This section analyzes the sensitivity of the hyperparameter K and verifies the effectiveness of the Sphere and Tube strategies. We selected CatBoost and the Brute Force Web class for this evaluation, as they exhibited the most significant performance improvement. First, we compared performance across $K \in \{10, 100, 1000, 10000\}$, and then compared the proposed method against individual strategies. The results are shown in Figure 3.

As shown in Figure 3(a), F1-Score fluctuations were negligible even when varying K widely, confirming the method's robustness to this parameter. Since the increase in computational time was marginal even with larger K, setting

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

a sufficiently large K to approximate decision boundaries is feasible, provided it remains within acceptable computational costs.

Regarding the impact of components (Figure 3(b)), the proposed method (0.7584) outperformed both Tube-only (0.6991) and Sphere-only (0.7321), confirming that combining strategies maximizes performance. Notably, Tube-only recorded a higher F1-Score than SMOTE, and Sphere-only outperformed RO. These results demonstrate that the Sphere and Tube components are powerful extensions of their respective predecessors.

Ultimately, the superior performance of the combined method demonstrates that Sphere and Tube are mutually complementary. We deliberately refrained from conducting a granular ablation study on the generation ratio, as the optimal balance between Sphere and Tube is inherently data-dependent; thus, a "best" ratio derived specifically for this target class would likely lack generalizability. However, the fact that the fixed 1:1 ratio achieved a high score of 0.7584 validates the effectiveness of this simple allocation strategy.

F. Controllability and Iterative Strategy

Experiments demonstrate that STSMOTE achieves high performance and robustness across diverse datasets. Notably, this performance was achieved under a conservative design philosophy: we employed a fixed 1:1 allocation ratio between Sphere and Tube, and adopted a strict copying strategy for categorical features to maintain semantic validity. The fact that such a "sub-optimal" configuration outperformed state-of-the-art methods while maintaining an extremely low FP rate (0.01%) indicates that STSMOTE functions as a highly reliable baseline for initial zero-day responses.

However, this inherent conservatism suggests a trade-off where the potential diversity of generated samples may be limited to ensure safety. Addressing this trade-off highlights the key advantage of our method over black-box generative models: the explicit controllability of the generation process. Unlike latent-space methods, STSMOTE allows operators to quantitatively control three key components: (1) the magnitude of geometric noise, (2) the exploration range of categorical features, and (3) the allocation ratio between Sphere and Tube.

This controllability enables an iterative optimization strategy for dynamic risk management. For instance, if a deployed model exhibits excessive FPs, operators can tighten the safety radius r_i or revert categorical generation to the strict copying strategy. Conversely, if the FP rate is well below the operational tolerance threshold (as observed in our experiments), the system can utilize this "operational margin" to proactively expand the generation range or explore novel combinations of categorical features to enhance diversity. Simultaneously, the Sphere-Tube ratio can be incorporated into this optimization loop, allowing the algorithm to automatically discover the optimal ratio that maximizes detection performance for the specific attack topology.

In conclusion, STSMOTE is not merely a static augmentation algorithm but a controllable framework that adapts to operational needs. Future work will focus on developing automated algorithms to iteratively optimize these parameters based on model feedback, further reducing the burden on security analysts.

V. Conclusion and Future Work

In this paper, we propose STSMOTE, a controllable data augmentation method designed for constructing high-precision detection models from extremely limited data, specifically targeting the initial response to zero-day attacks. By integrating the Sphere and Tube strategies, the proposed method enables the generation of safe and diverse data while preserving the geometric structure of the minority samples.

In our experiments, STSMOTE demonstrated superior performance over 8 baseline methods and achieved a practical execution speed suitable for real-time operations. Crucially, the ability to explicitly describe and control the generation process via geometric parameters is a distinct advantage over existing black-box methods. This controllability is particularly effective in the security domain, where rigorous risk management is required.

In future work, we first plan to evaluate the proposed method under more extreme real-world imbalance scenarios (e.g., $1:10^6$) and conduct comprehensive ablation studies across a wider range of datasets and attack types. Subsequently, we aim to develop the iterative algorithms described in Section IV-F to automatically optimize the generation regions and allocation ratios of Sphere and Tube. Within this iterative framework, we will also explore methods for generating novel combinations of categorical features while strictly controlling the FP rate. Finally, we aim to demonstrate the effectiveness of this advanced approach in other security domains characterized by extreme data scarcity.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP23K28051.

References

[1] F. Ullah, S. Ullah, G. Srivastava, and J. C.-W. Lin, “IDS-INT: Intrusion detection system using transformer-based transfer learning for imbalanced network traffic”, *Digital Communications and Networks*, vol. 10, no. 1, pp. 190–204, Feb. 2024, ISSN: 2352-8648. DOI: 10.1016/j.dcan.2023.03.008

[2] H. Zhang, L. Huang, C. Q. Wu, and Z. Li, “An effective convolutional neural network based on SMOTE and Gaussian mixture model for intrusion detection in imbalanced dataset”, *Computer Networks*, vol. 177, p. 107 315, Aug. 2020, ISSN: 1389-1286. DOI: 10.1016/j.comnet.2020.107315

[3] 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, Jun. 1, 2002, ISSN: 1076-9757. DOI: 10.1613/jair.953

[4] H. He, Y. Bai, E. A. Garcia, and S. Li, “ADASYN: Adaptive synthetic sampling approach for imbalanced learning”, in *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, Jun. 2008, pp. 1322–1328. DOI: 10.1109/IJCNN.2008.4633969

[5] H. Ishii, “STSMOTE”, Jan. 2026, Accessed: Mar. 13, 2026. [Online]. Available: https://github.com/Haruto-Ishii/STSMOTE

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

[6] J. Liu, Y. Gao, and F. Hu, "A fast network intrusion detection system using adaptive synthetic oversampling and LightGBM", *Computers & Security*, vol. 106, p. 102 289, Jul. 2021, ISSN: 0167-4048. DOI: 10.1016/j.cose.2021.102289

[7] Y. Sun et al., "Borderline SMOTE Algorithm and Feature Selection-Based Network Anomalies Detection Strategy", *Energies*, vol. 15, no. 13, p. 4751, Jan. 2022, ISSN: 1996-1073. DOI: 10.3390/en15134751

[8] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning", in *Advances in Intelligent Computing*, vol. 3644, 2005, pp. 878–887, ISBN: 978-3-540-28226-6 978-3-540-31902-3. DOI: 10.1007/11538059_91

[9] O. H. Abdulganiyu, T. A. Tchakoucht, Y. K. Saheed, and H. A. Ahmed, "XIDINTFL-VAE: XGBoost-based intrusion detection of imbalance network traffic via class-wise focal loss variational autoencoder", *The Journal of Supercomputing*, vol. 81, no. 1, p. 16, Oct. 2024, ISSN: 1573-0484. DOI: 10.1007/s11227-024-06552-5

[10] J. Lee and K. Park, "GAN-based imbalanced data intrusion detection system", *Personal and Ubiquitous Computing*, vol. 25, no. 1, pp. 121–128, Feb. 2021, ISSN: 1617-4917. DOI: 10.1007/s00779-019-01332-y

[11] H. Zhang et al., "Mixed-Type Tabular Data Synthesis with Score-based Diffusion in Latent Space", arXiv, May 2024. DOI: 10.48550/arXiv.2310.09656

[12] C. Wang, J. Wu, and L. Wang, "LRT-DDPM: A Diffusion Model-Based Approach for Network Traffic Data Generation in Intrusion Detection", *IEEE Access*, vol. 13, pp. 149 054–149 070, 2025, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2025.3598777

[13] A. S. Dina, A. B. Siddique, and D. Manivannan, "A deep learning approach for intrusion detection in Internet of Things using focal loss function", *Internet of Things*, vol. 22, p. 100 699, Jul. 2023, ISSN: 2542-6605. DOI: 10.1016/j.iot.2023.100699

[14] P. Bedi, N. Gupta, and V. Jindal, "I-SiamIDS: An improved Siam-IDS for handling class imbalance in network-based intrusion detection systems", *Applied Intelligence*, vol. 51, no. 2, pp. 1133–1151, Feb. 2021, ISSN: 1573-7497. DOI: 10.1007/s10489-020-01886-y

[15] I. Sharafaldin, A. Habibi Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization", in *Proceedings of the 4th International Conference on Information Systems Security and Privacy*, 2018, pp. 108–116, ISBN: 978-989-758-282-0. DOI: 10.5220/0006639801080116

[16] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling Tabular data using Conditional GAN", in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[17] Y. Gorishniy, A. Kotelnikov, and A. Babenko, "TabM: Advancing Tabular Deep Learning with Parameter-Efficient Ensembling", arXiv, Feb. 2025. DOI: 10.48550/arXiv.2410.24210

[18] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)", in *2015 Military Communications and Information Systems Conference (MilCIS)*, Jan. 2015, pp. 1–6. DOI: 10.1109/MilCIS.2015.7348942

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org