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

Haruto Ishii[1], Kunio Akashi[2], Yuji Sekiya[1]

[1]*Graduate School of Information Science and Technology, The University of Tokyo (Japan)*

[2]*Information Technology Center, The University of Tokyo (Japan)*

Contact Email: **harutowoody@g.ecc.u-tokyo.ac.jp**

UTokyo

IARIA

Haruto Ishii

- Education & Current Affiliation
  - Ph. D Student at the Graduate School of Information Science and Technology, The University of Tokyo
  - Received my M. S. in Information Science and Technology from The University of Tokyo in March 2026
- Research Focus
  - Specializing in Cybersecurity
- Research Interests
  - Network Intrusion Detection System (NIDS)
  - Deep Learning (also Machine Learning)
  - Tabular Data Synthesis and Augmentation

Network Intrusion Detection System (NIDS)

- Monitors network traffic to detect suspicious activities

Detection Target

Packet-based

Analyzes individual packet payloads

Flow-based

Utilizes flow statistics for detection
Flow: Defined by a 5-tuple (Src/Dst IP, Port, Protocol)

Detection Method

Signature-based

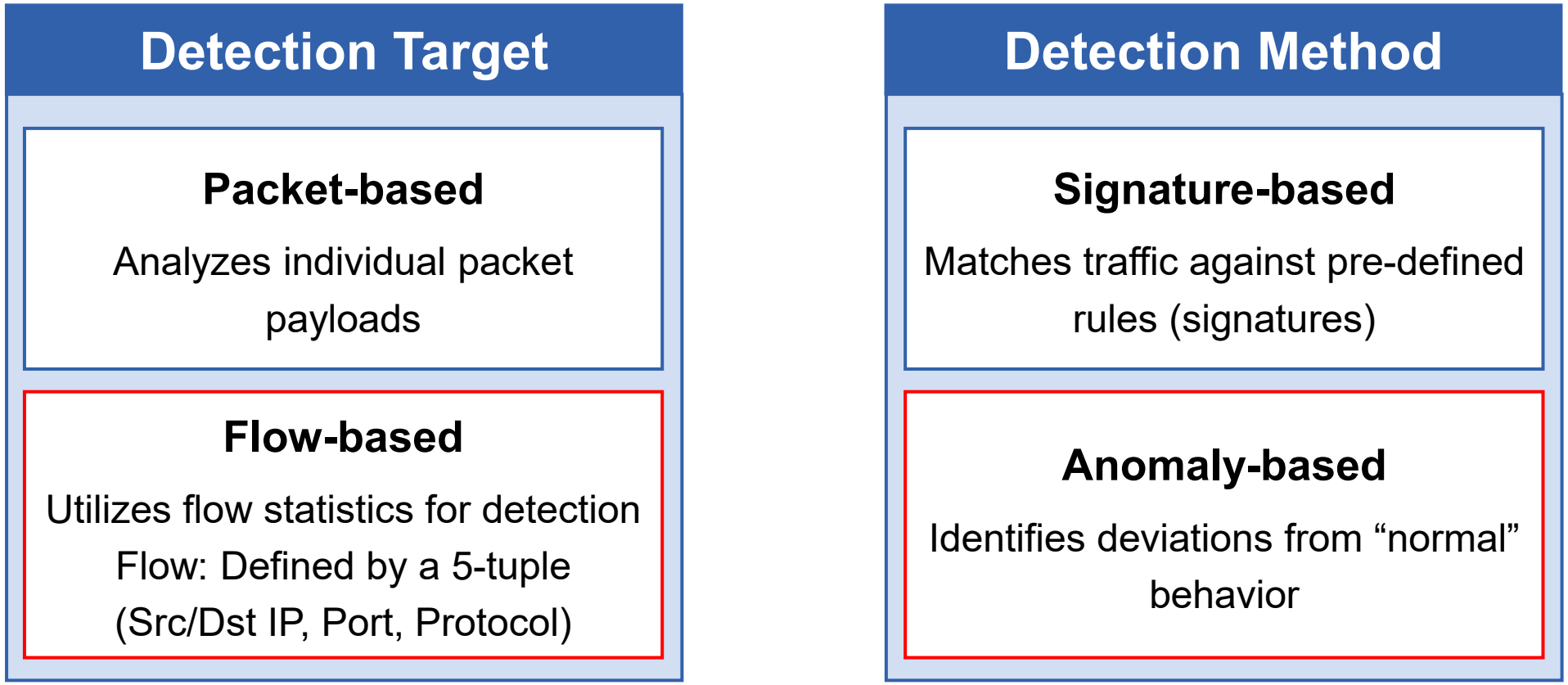
Matches traffic against pre-defined rules (signatures)

Anomaly-based

Identifies deviations from “normal” behavior

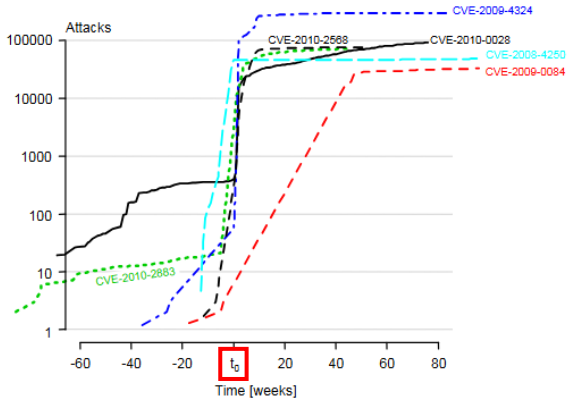
Network Intrusion Detection System (NIDS)

- Monitors network traffic to detect suspicious activities

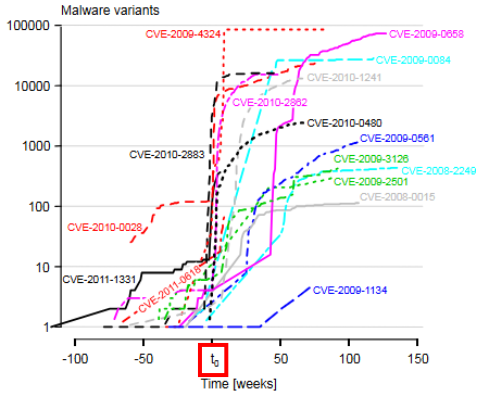


Zero-Day Attacks

- Exploits before patches are released
- Rapid surge in attacks after disclosure[1]
  - Up to 100,000x increase in attack volume
  - Only 65% of patches available at the time of disclosure
  - Up to 85,000 variants emerge
  - Zero-day window: Avg. 312 days (Max. 30 months)



Trends in attack volume after disclosure[1]



Trends in number of variants[1]

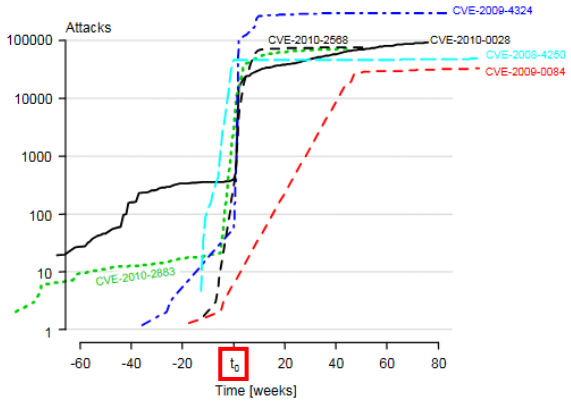
[1] Bilge, Leyla, and Tudor Dumitraş. "Before we knew it: an empirical study of zero-day attacks in the real world." *Proceedings of the 2012 ACM conference on Computer and communications security*. 2012.

Zero-Day Attacks

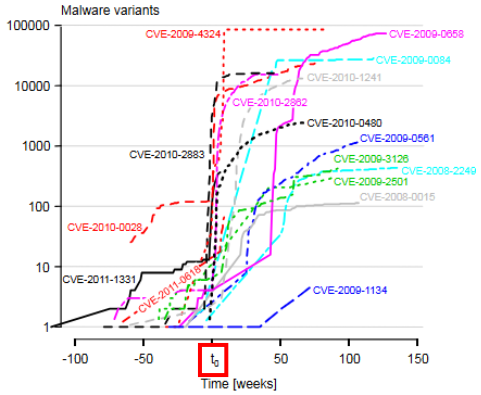
- Exploits before patches are released
- Rapid surge in attacks after disclosure[1]
  - Up to 100,000x increase in attack volume
  - Only 65% of patches available at the time of disclosure
  - Up to 85,000 variants emerge
  - Zero-day window: Avg. 312 days (Max. 30 months)

Enabling ML detection with limited data is essential for:

- Detection of emerging threats
- Forensics for incident analysis



Trends in attack volume after disclosure[1]



Trends in number of variants[1]

[1] Bilge, Leyla, and Tudor Dumitraş. "Before we knew it: an empirical study of zero-day attacks in the real world." *Proceedings of the 2012 ACM conference on Computer and communications security*. 2012.

Approaches to Zero-Day Attack Detection

- Detection is difficult when data is extremely limited
  - Zero-day attacks typically involve severely scarce data
- There are two main approaches:

Algorithm-level	Data-level
Modifies model architectures or loss functions	Augments minority class data to balance the class distribution
Stable detection performance	**Highly generalizable**
Low generalizability	**Accuracy heavily depends on the quality of generated data**

Approaches to Zero-Day Attack Detection

- Detection is difficult when data is extremely limited
  - Zero-day attacks typically involve severely scarce data
- There are two main approaches:

Algorithm-level	Data-level
Modifies model architectures or loss functions	Augments minority class data to balance the class distribution
Stable detection performance	**Highly generalizable**
Low generalizability	**Accuracy heavily depends on the quality of generated data**

Comparison of Data Augmentation Methods

Comparison of existing methods

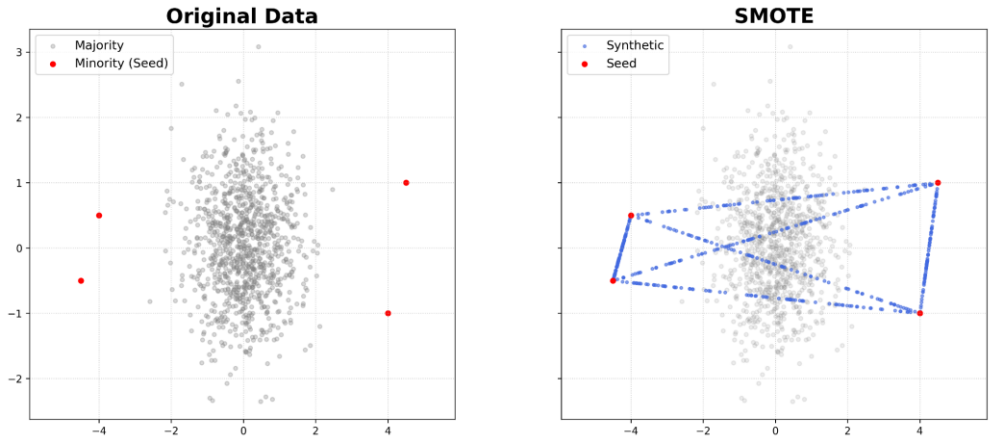
Comparison criteria	Non-ML methods	ML methods
Execution Time	+	-
Data Diversity	~	+
Few-shot Adaptability	~	-
Prior Applications	+	~
Controllability	-	-

- Examples of Non-ML methods:
  - Random Oversampling (RO)
  - SMOTE [2]
  - ADASYN [3]
- Examples of ML methods:
  - TVAE [4]
  - CTGAN [4]
  - Diffusion Model [5]
- ML methods are not effective when data is extremely limited

[2] Chawla, Nitesh V., et al. "SMOTE: synthetic minority over-sampling technique." Journal of artificial intelligence research 16 (2002): 321-357.
[3] He, Haibo, et al. "ADASYN: Adaptive synthetic sampling approach for imbalanced learning." 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence). Ieee, 2008.
[4] Xu, Lei, et al. "Modeling tabular data using conditional gan." Advances in neural information processing systems 32 (2019).
[5] Zhang, Hengrui, et al. "Mixed-type tabular data synthesis with score-based diffusion in latent space." arXiv preprint arXiv:2310.09656 (2023).

Representative Non-ML Methods: SMOTE

- Proposed by Chawla et al. in 2002
  - The most widely used method
- SMOTE works as follows：
  1. Find nearest neighbors within the minority class
  2. Generate synthetic data by randomly selecting a point on the line segment between two points
- Fast and effective, but lacks controllability and diversity, risking decision boundary violation



Example of augmentation using SMOTE
Left: Original data
Right: Data after augmentation
(Red points: Minority data,
Blue points: Augmented data)

Objective of this Study

Establishing a data augmentation method for rapid zero-day response with high controllability and diversity

Objective of this Study

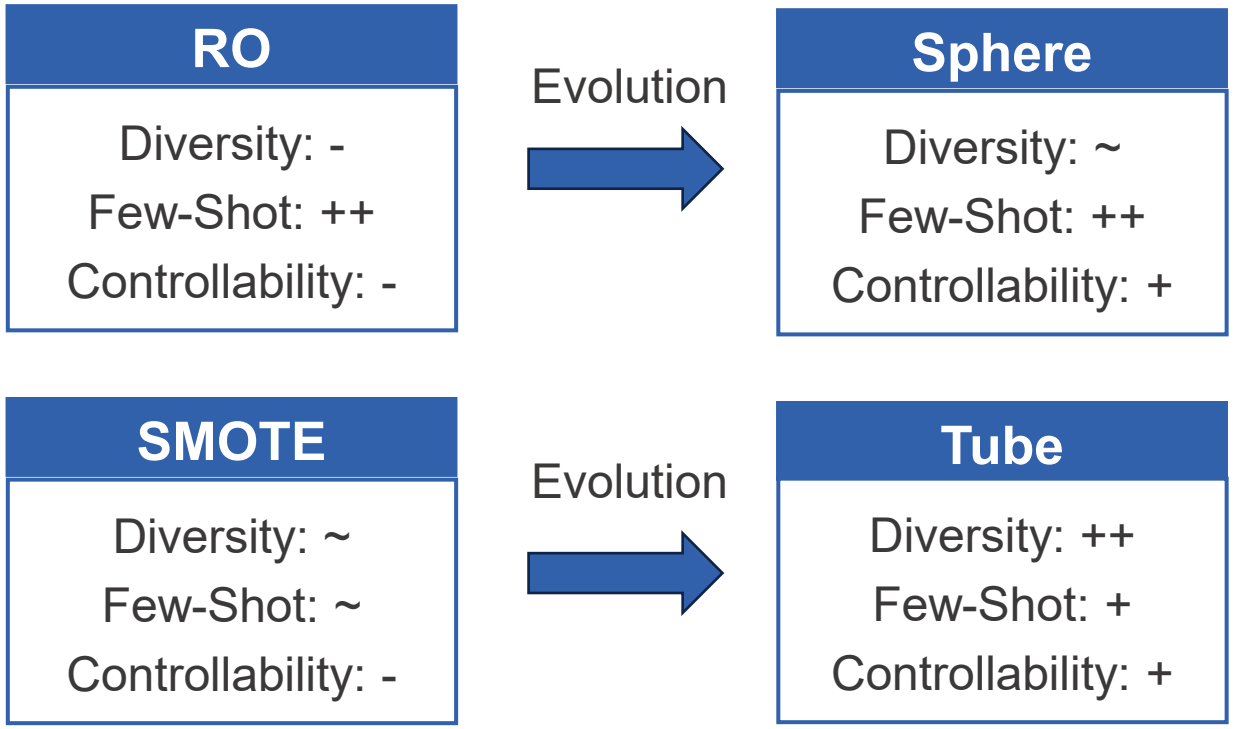
Establishing a data augmentation method for rapid zero-day response with high controllability and diversity

RO
Diversity: -
Few-Shot: ++
Controllability: -

SMOTE
Diversity: ~
Few-Shot: ~
Controllability: -

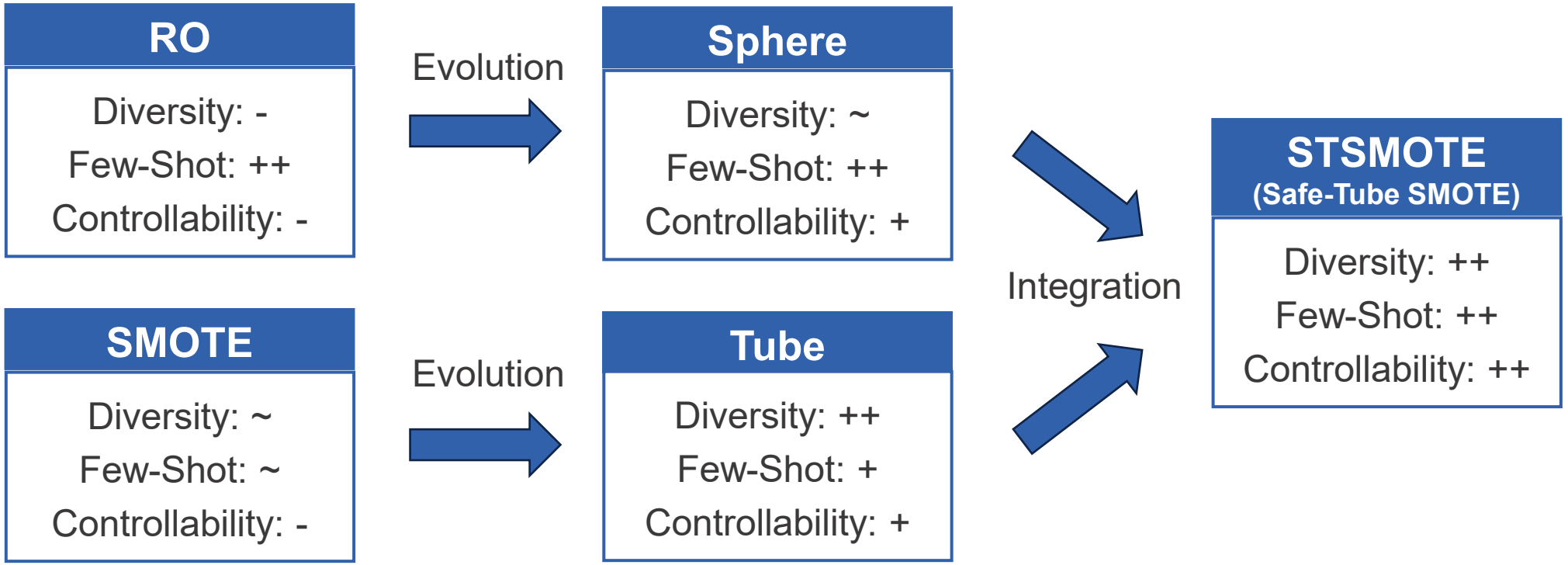
Objective of this Study

Establishing a data augmentation method for rapid zero-day response with high controllability and diversity



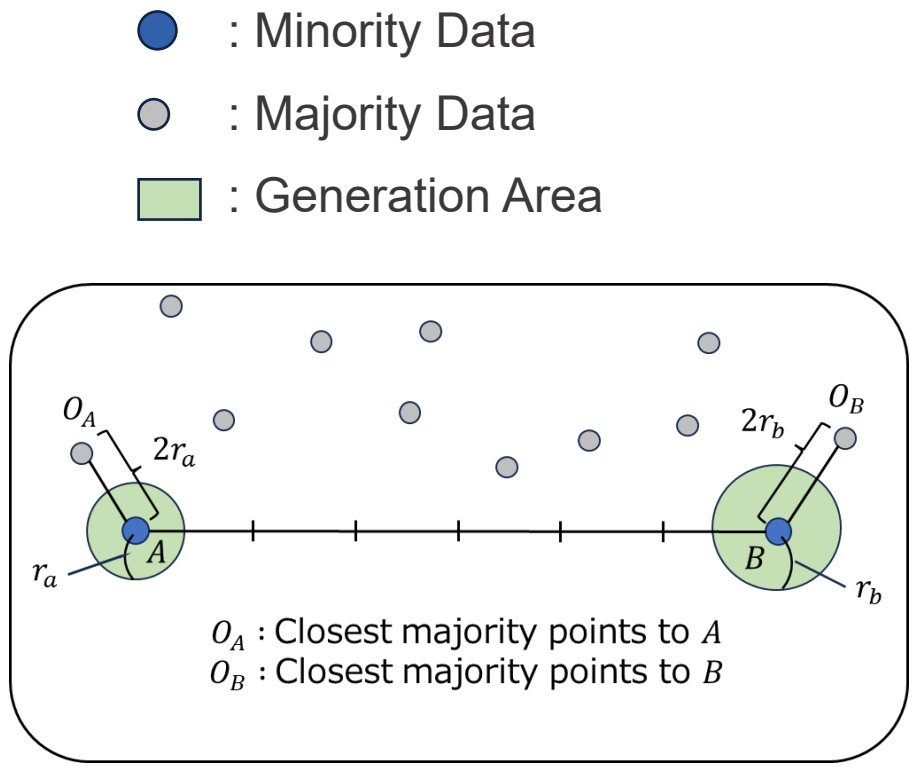
Objective of this Study

Establishing a data augmentation method for rapid zero-day response with high controllability and diversity



Data Generation by Sphere

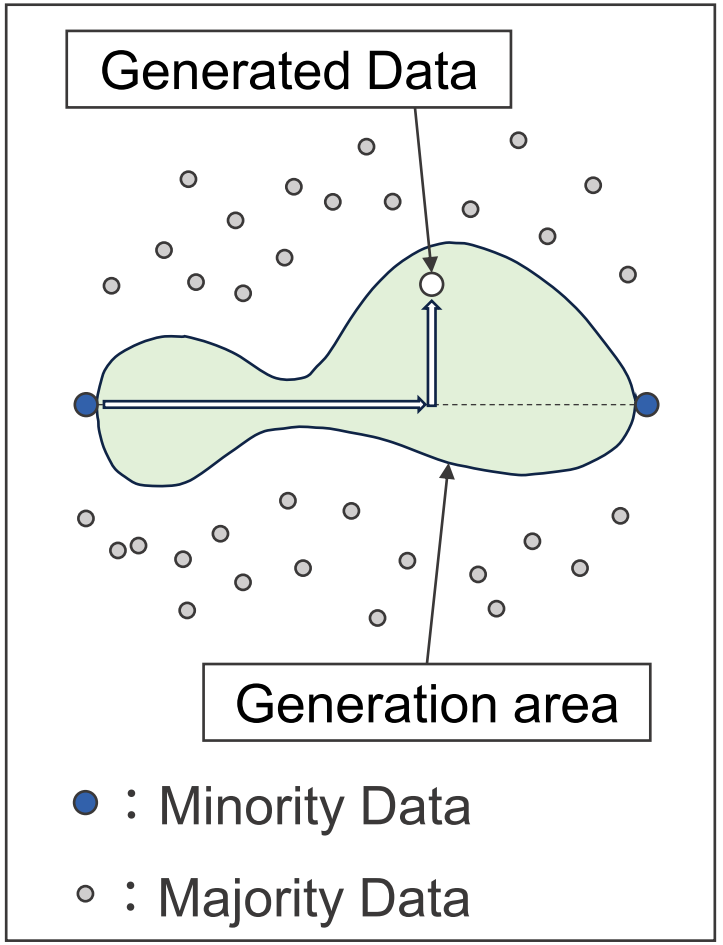
- Generate data in the local neighborhood rather than simple duplication
  - Ensures diversity and avoids overfitting
- Sphere works as follows:
  1. Apply Quantile Transformation to numerical features (excluding categorical features)
  2. Find nearest majority class neighbor for each minority instance
  3. Generate data within a hypersphere centered at the minority point (radius: half the distance to the neighbor)



Data generation by Sphere

Data Generation by Tube: Overview

- Expands the generation area compared to SMOTE
  - Interpolation + Orthogonal perturbation
- Controls generation area via surrounding majority data
  - Limits orthogonal perturbation size
  - Skips generation if majority data is too close (perturbation size ≤ 0)
- Calculates area using 3 constraints:
  - Cone Constraint, Local Obstacles, Monotonicity



Overview of Tube

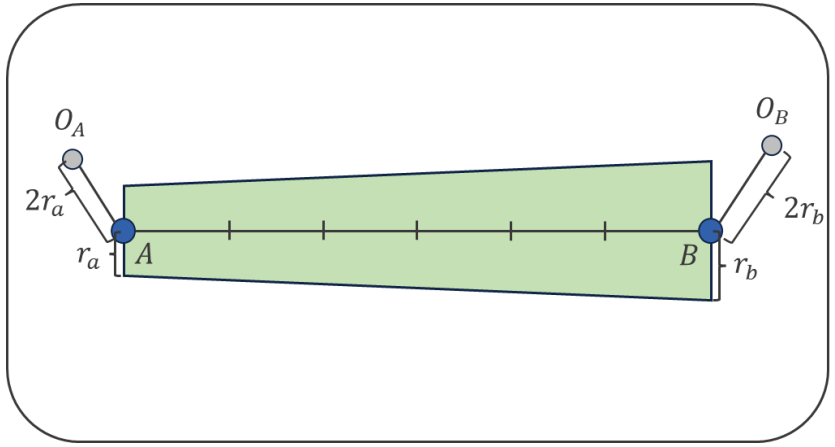
Tube：Cone Constraint

- Initial condition for maximum orthogonal perturbation size
- Sets generation area as a truncated cone using Sphere radii
  - Approximated by segmented cylinders between two points (6 segments shown in the figure)
- Let h_i be the cylinder radius of each segment
  - h_1 to h_6 are defined in the figure
  - h_i is dynamically adjusted by subsequent constraints

: Minority Data

: Majority Data

: Generation Area



Initial generation area defined by Cone Constraint

Tube：Local Obstacle

Adjust h_i based on majority class distribution

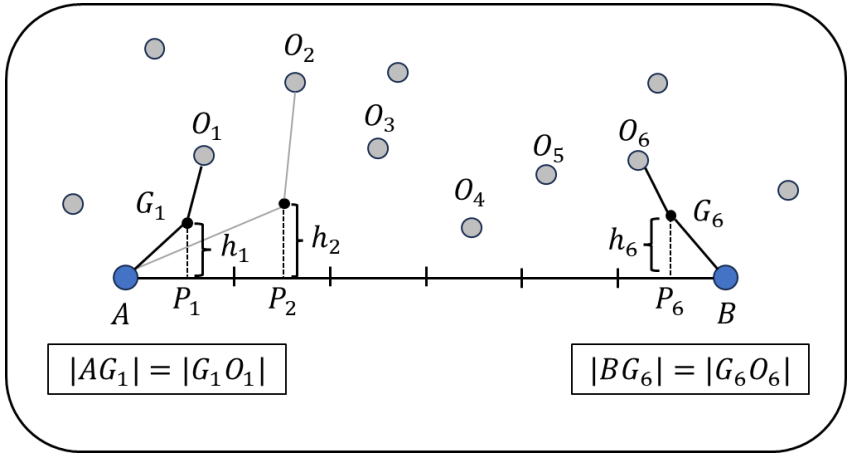
Find which segment contains the orthogonal projection from majority data
(Let O_i be the closest point)

Calculate h_i so that G_i (midpoint P_i + orthogonal perturbation) is closer to the endpoint than O_i

Adopt the smaller h_i compared to the Cone
(Skip generation in this segment if $h_i < 0$)

: Minority Data

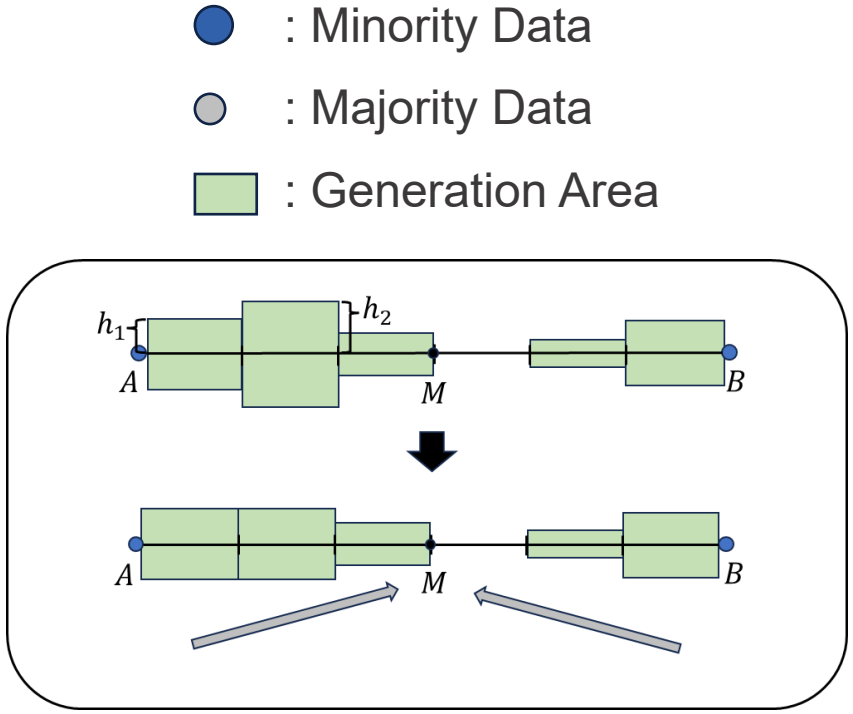
: Majority Data



Adjustment of the generation area by Local Obstacle

Tube：Monotonicity

- Some segments remain unadjusted by Local Obstacle
  - When no orthogonal projection from majority data
  - Causes unnatural shapes in the generation area
- Uncertainty increases further from endpoints
  - Generation area should shrink toward the midpoint
- Enforces h_i to decrease monotonically toward the midpoint

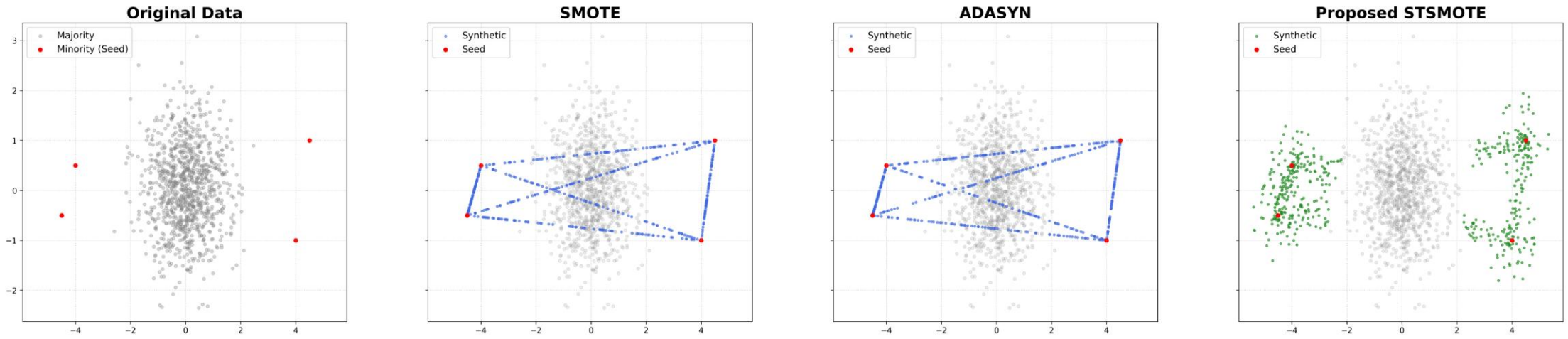


Refining the generation area via the Monotonicity constraint

Data Generation by STSMOTE

- Generate data after finalizing generation area for Sphere and Tube
  - Sphere to Tube generation ratio is 1:1
  - Generate data randomly within each generation area
- Adopt a simple copy strategy for categorical features
  - Sphere: Copy from the center instance
  - Tube: Copy from the closer endpoint
- Apply Sphere to all minority data, and Tube to all minority data pairs
  - Computational complexity is $\mathcal{O}(N^2)$
  - Set $K = 1{,}000$ segments per Tube

Illustrative Example of Generated Data



- Existing methods violate decision boundaries, while the proposed method avoids them
  - Local Obstacle effectively restricts generation area
- Generates data more broadly compared to existing methods

Experimental Setup

- Simulate zero-day attack scenarios with limited data
  - Dataset: CIC-IDS2018 [6] (Majority training data: 5,000 samples × 11 classes)
  - Minority training data: 4 classes, undersampled to $N = 2, 5$
  - Evaluated over 25 independent runs with different random seeds
  - Test data: 1,250 samples × 11 majority classes + all remaining minority data
- Compared against 8 baseline methods
  - No-ML：No Augmentation (No Aug.), RO, SMOTE, ADASYN, Borderline-SMOTE [7]
  - ML：TVAE, CTGAN, TabSyn (Diffusion Model)
- Evaluated using four classifiers (CatBoost, XGBoost, LightGBM, TabM [8])
  - Measured Average F1-Score of the minority class in a multi class setting

[6] Sharafaldin, Iman, Arash Habibi Lashkari, and Ali A. Ghorbani. "Toward generating a new intrusion detection dataset and intrusion traffic characterization." ICISSp 1.2018 (2018): 108-116.
[7] Han, Hui, Wen-Yuan Wang, and Bing-Huan Mao. "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning." International conference on intelligent computing. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005.
[8] Gorishniy, Yury, Akim Kotelnikov, and Artem Babenko. "Tabm: Advancing tabular deep learning with parameter-efficient ensembling." arXiv preprint arXiv:2410.24210 (2024).

Experimental Results (N=5)

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.)	***	***	**	*	***	***	-	*	-	**	-	-	**	***	***	-

- **Proposed method ranked 1st in 14/16 cases** (4 classes × 4 classifiers)
  - Statistically outperformed all baselines in 11 cases (Holm-Corrected Wilcoxon test)
  - Achieved $p < 0.001$ in 6 of these cases
  - Recorded an average rank of **2.01** (Best baseline: ADASYN at 4.00)

Experimental Results (N=2)

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.)	*	***	-	*	*	***	-	-	-	**	-	-	**	***	***	**

- **Proposed method ranked 1st in 14/16 cases**
  - 10 of these cases showed statistical significance
  - Recorded an average rank of **2.19** (Best baseline: SMOTE at 4.25)
- Our method performs effectively even when the data is extremely limited

Operational Efficiency & Safety

- **Execution time: ~1 sec (N=5)**, comparable to SMOTE
  - TabSyn (Diffusion model) takes ~55 mins
- **False Positive Rate (FPR): < 0.01%** (1.2625 / 13,750)
  - ~0.04% without augmentation (4.845 / 13,750)
  - Extremely low FPR is crucial to minimize SOC burden
  - Validates the effectiveness of safety strategies in Sphere & Tube
- **Geometrically defined generation areas**
  - Provides high controllability (lacking in existing methods)
  - Allows dynamic expansion/shrinkage based on real-world accuracy

Conclusion

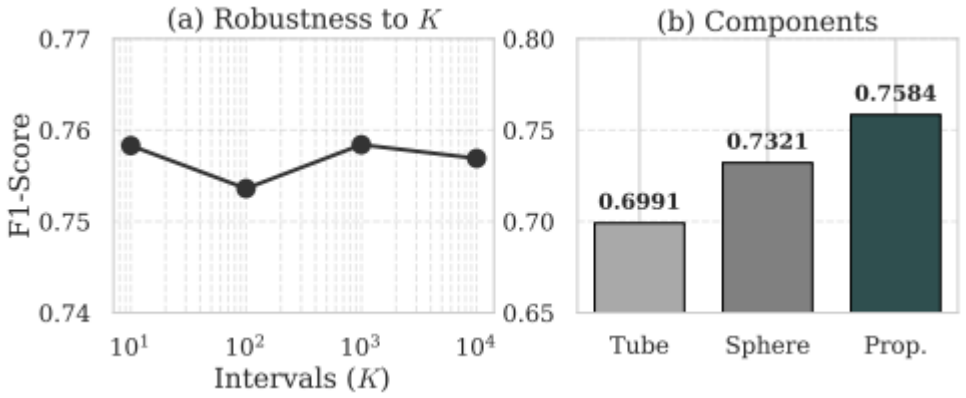
- Detection is challenging under extremely limited data (e.g. zero-day attacks)
- Existing augmentation methods lack diversity and controllability
- Proposed STSMOTE: Integrates “Sphere” (extended RO) and “Tube” (extended SMOTE)
- Significantly outperformed baselines in F1-Score, while maintaining high speed and a low FPR
- Provides controllability by defining the generation area with geometric parameters

Appendix

Ablation Study

- Evaluated the contribution of each component focusing on:
  - Robustness to the number of segments ($K \in \{10, 100, 1000, 10000\}$)
  - Individual strategies (Sphere only, Tube only, and Combine)
- Experimental Setup:
  - Target: Brute Force Web ($N = 5$)
  - Classifier: CatBoost

Results of Ablation Study



Results of Ablation Study
(Left) Robustness to the number of segments K
(Right) Comparison of Tube, Sphere, STSMOTE

- Robustness to K
  - No significant performance differences across$K$
  - Highly robust to K in Tube
  - A sufficiently large value is recommended for better controllability
- Effectiveness of Components:
  - Combining Sphere and Tube achieved the best results
  - Confirmed the effectiveness of each component (Tube > SMOTE, Sphere > RO)

Additional Dataset Results (F1-Score)

TABLE VIII. COMPARISON OF F1-SCORES ON ADDITIONAL DATASETS. 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	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.)	-	***	***	**	-	***	***	***	***	***	***	***	**	-	-	-

Additional Dataset Results (Average Rank)

TABLE IX. COMPARISON OF AVERAGE RANK ON ADDITIONAL DATASETS. 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	Shellcode ($N = 5$)				Shellcode ($N = 2$)				Bot ($N = 5$)				Bot ($N = 2$)				Avg.
	Cat	XGB	LGB	TabM	Cat	XGB	LGB	TabM	Cat	XGB	LGB	TabM	Cat	XGB	LGB	TabM	
No Aug.	8.20	8.18	8.38	7.70	8.20	8.30	8.22	7.72	8.38	8.46	8.46	7.98	8.18	8.38	8.30	8.22	8.20
RO	5.04	5.52	4.60	5.36	4.18	4.32	4.34	5.70	3.02	5.04	5.98	4.88	2.88	3.00	3.46	5.08	4.53
SMOTE	2.70	2.80	3.92	2.88	2.44	3.08	3.90	2.34	5.22	4.06	4.90	4.48	4.24	4.04	3.98	3.30	3.64
BSMOTE	6.98	7.02	7.22	6.66	8.20	8.30	8.22	7.72	5.10	4.62	5.40	4.22	7.64	7.92	7.94	7.92	6.94
ADASYN	2.18	2.56	3.30	2.28	2.40	2.90	4.28	2.54	3.24	3.66	4.48	3.36	3.84	4.54	4.50	3.44	3.34
CTGAN	7.90	7.80	6.36	7.24	7.60	7.24	6.36	6.66	8.20	7.48	5.40	7.32	7.56	6.16	5.08	5.52	6.87
TVAE	6.20	5.56	5.00	7.08	5.32	5.02	4.02	5.70	7.16	6.26	3.68	6.50	5.82	5.06	**2.96**	4.68	5.38
TabSyn	3.96	4.44	4.78	4.16	4.74	4.52	4.22	5.38	3.38	4.34	5.26	4.74	3.30	3.48	4.54	4.36	4.35
STSMOTE	**1.84**	**1.12**	**1.44**	**1.64**	**1.92**	**1.32**	**1.44**	**1.24**	**1.30**	**1.08**	**1.44**	**1.52**	**1.54**	**2.42**	4.24	**2.48**	**1.75**
(Sig.)	-	***	***	**	-	***	***	***	***	***	***	***	**	-	-	-	

Experimental Results (N=5, Average Rank)

TABLE IV. COMPARISON OF AVERAGE RANK 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				Avg.
	Cat	XGB	LGB	TabM	Cat	XGB	LGB	TabM	Cat	XGB	LGB	TabM	Cat	XGB	LGB	TabM	
No Aug.	7.96	7.76	8.92	7.50	7.90	8.12	8.18	7.32	8.64	8.40	8.84	5.44	8.74	8.62	8.92	8.06	8.08
RO	2.98	6.04	5.30	4.16	4.28	6.38	5.50	5.52	5.68	6.98	6.76	5.48	3.42	5.64	5.90	5.30	5.33
SMOTE	5.14	5.52	5.16	5.10	5.06	5.52	5.96	5.00	4.14	3.84	4.52	4.10	4.72	4.44	4.18	4.26	4.79
BSMOTE	4.74	5.28	5.80	3.96	4.18	4.80	4.84	3.52	4.48	3.28	4.30	3.84	4.66	4.58	5.26	4.06	4.47
ADASYN	3.30	2.88	4.80	3.46	3.16	3.82	6.04	3.14	5.00	4.96	5.78	**3.52**	3.42	3.74	4.26	2.64	4.00
CTGAN	8.06	6.22	4.26	7.24	7.86	5.66	3.30	7.40	5.24	5.00	3.72	7.50	7.86	6.66	4.80	7.40	6.14
TVAE	7.58	6.20	4.70	7.44	7.40	4.26	3.82	6.90	2.72	3.28	**2.24**	7.32	6.88	5.88	4.96	7.72	5.58
TabSyn	4.02	4.02	4.26	4.34	3.20	5.24	4.78	4.20	6.54	7.16	5.60	4.08	3.42	4.24	4.78	3.68	4.60
STSMOTE	**1.22**	**1.08**	**1.80**	**1.80**	**1.96**	**1.20**	**2.58**	**2.00**	**2.56**	**2.10**	3.24	3.72	**1.88**	**1.20**	**1.94**	**1.88**	**2.01**
(Sig.)	***	***	**	*	***	***	-	*	-	**	-	-	**	***	***	-	

Experimental Results (N=2, Average Rank)

TABLE VI. COMPARISON OF AVERAGE RANK ON $N = 2$. 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 $^{***}p < 0.001$. † STATISTICAL SIGNIFICANCE IS CALCULATED BASED ON PAIRWISE F1-SCORE DISTRIBUTIONS. IN THIS SPECIFIC CASE, STSMOTE SIGNIFICANTLY OUTPERFORMS OTHER METHODS IN F1-SCORE DESPITE A LOWER AVERAGE RANK CAUSED BY RELATIVE PERFORMANCE FLUCTUATIONS AMONG BASELINES.

Method	Web				XSS				DDoS				SQL				Avg.
	Cat	XGB	LGB	TabM	Cat	XGB	LGB	TabM	Cat	XGB	LGB	TabM	Cat	XGB	LGB	TabM	
No Aug.	8.12	7.74	8.38	7.88	7.04	6.26	8.26	7.26	8.40	8.00	8.92	6.54	8.34	7.98	8.28	8.00	7.84
RO	2.58	3.94	5.10	4.50	**3.18**	6.16	4.68	4.98	4.52	5.18	5.60	4.14	2.94	4.08	5.58	4.70	4.49
SMOTE	4.58	4.74	4.14	3.22	3.96	4.98	4.32	3.48	4.62	5.54	4.82	4.38	3.72	4.70	3.74	3.04	4.25
BSMOTE	7.52	7.94	7.74	7.62	6.80	7.32	6.90	7.12	4.60	5.10	4.90	4.10	8.10	7.62	7.88	7.56	6.80
ADASYN	4.26	3.96	5.26	3.54	4.28	3.84	6.04	3.58	5.10	4.90	5.36	4.90	4.28	4.78	4.32	3.04	4.47
CTGAN	7.16	5.72	3.74	6.50	6.46	5.64	3.88	6.02	5.44	5.46	3.96	7.12	7.14	6.04	4.20	6.32	5.68
TVAE	5.74	4.76	3.34	5.98	5.78	4.62	3.48	5.16	**2.16**	3.88	**2.48**	6.50	5.64	4.26	3.74	6.68	4.64
TabSyn	3.16	4.40	4.82	4.14	4.26	5.08	5.02	4.72	6.72	5.30	5.32	3.96	3.36	4.36	5.68	4.22	4.66
STSMOTE	**1.88**	**1.80**	**2.48**	**1.62**	3.24†	**1.10**	**2.42**	**2.68**	3.44	**1.64**	3.64	**3.36**	**1.48**	**1.18**	**1.58**	**1.44**	**2.19**
(Sig.)	*	***	-	*	*	***	-	-	-	**	-	-	**	***	***	**	