Privacy-Aware Generative Augmentation for Imbalanced Intrusion Detection Using Differential Privacy

Trung Ha[1], Tran Khanh Dang[2]*, Dinh Thi Hong Loan[3]
Faculty of Information Systems, University of Information Technology, Ho Chi Minh City, Vietnam[1]
Vietnam National University Ho Chi Minh City, Ho Chi Minh City, Vietnam[1]
Institute for Experiential Technology, Van Lang University, Ho Chi Minh City, Vietnam [2]
Faculty of Mechanical, Electrical, and Computer Engineering-Van Lang School of Technology,
Van Lang University, Ho Chi Minh City, Vietnam[2]
Faculty of Information Technology, Ho Chi Minh University of Natural Resources and Environment,
Ho Chi Minh City, Vietnam[3]

***Abstract*—Machine learning-based intrusion detection systems are often constrained by severe class imbalance, while generative augmentation may memorize distinctive network-flow records and expose membership information. This study evaluates privacy-aware augmentation for CICIDS2017 by integrating differentially private stochastic gradient descent (DP-SGD) into a conditional generative adversarial network and a variational autoencoder, yielding DP-CGAN and DP-VAE. The formal record-level privacy guarantee applies to the generative component and is tracked with an RDP accountant implemented through Opacus. Experiments use five target Benign-to-Attack augmentation ratios from one-to-one to one-to-five and five target privacy budgets of 0.5, 1, 2, 4, and 8. Random forest (RF) and XGBoost are evaluated using accuracy, Macro-F1, mean minority-class recall, and per-class recall for Bot, BruteForce, and Infiltration. Synthetic-data quality is examined through t-SNE, Wasserstein distance, mean squared error, and mean absolute error, while empirical privacy is assessed through a black-box membership-inference protocol using ROC-AUC, attack accuracy, and an accuracy-derived attack advantage. DP-CGAN preserves high downstream utility across both classifiers. DP-VAE remains competitive with random forest but degrades substantially with XGBoost, particularly for BruteForce and Infiltration. Membership-inference ROC-AUC remains close to 0.5 for both private and non-private generators under the implemented attack, indicating that no reliable membership signal was detected rather than proving the absence of leakage. These results show that private generative augmentation is feasible for imbalanced IDS learning, but its utility depends strongly on the interaction between generator architecture and downstream classifier.**

***Keywords*—Intrusion detection system; class imbalance; synthetic data generation; differential privacy; DP-SGD**

I. Introduction

Machine learning-based intrusion detection systems (IDSs) are increasingly used to protect cloud platforms, Internet of Things infrastructures, enterprise networks, and software-defined environments. Their effectiveness, however, is frequently constrained by highly skewed traffic distributions, in which benign flows dominate while security-critical attacks remain sparsely represented. Data-level augmentation based on conditional generative adversarial networks (CGANs), variational autoencoders (VAEs), and related generative architectures has therefore become an important strategy for improving minority-class detection [1–7].

The usefulness of synthetic data cannot be established from downstream accuracy alone. Recent studies emphasize that synthetic-data generation should be assessed jointly in terms of quality, distributional similarity, utility, and privacy [8–10]. Broader analyses of generative models also show that replacing real records with synthetic samples does not, by itself, eliminate the risk of exposing sensitive information [11].

A learned generator may overfit distinctive training records and preserve information that enables an adversary to infer whether a particular sample contributed to model training. Membership-inference studies against generative models have demonstrated leakage under black-box and white-box access, reconstruction-based evidence, and density-based overfitting detection [12–16]. This risk is especially relevant to rare network-flow records because their scarcity and distinctiveness make them important for detection while potentially increasing their susceptibility to disclosure.

Differential privacy (DP) provides a formal mechanism for bounding the influence of an individual training record. Private synthetic-data approaches include Bayesian-network synthesis, teacher-ensemble methods, differentially private adversarial training, and privacy-preserving distributed generation [17–21]. In deep generative models, DP is commonly implemented through DP-SGD [22], with cumulative privacy loss tracked through Rényi differential privacy accounting [23] and software frameworks such as Opacus [24]. Formal privacy accounting nevertheless addresses a different question from empirical disclosure testing; broader attack-based frameworks therefore recommend reporting both forms of evidence [25].

Despite these advances, three gaps remain relevant to intrusion detection. First, most IDS augmentation studies prioritize predictive utility without formal privacy accounting. Second, privacy-preserving synthetic-data methods are

*Corresponding author.

commonly evaluated on general tabular datasets rather than highly imbalanced, multi-class network-flow data. Third, limited evidence is available on the joint effects of generator architecture, privacy budget, augmentation level, rare-class utility, and downstream classifier. These limitations make it difficult to identify a defensible privacy-utility operating point for IDS augmentation.

This study addresses these gaps through a domain-specific empirical evaluation on CICIDS2017 [26]. Rather than introducing a new generic private generator, established DP-SGD-based CGAN and VAE architectures are adapted for class-conditional network-flow augmentation. The evaluation spans five target Benign-to-Attack ratios and five privacy budgets and jointly considers downstream utility, rare-attack detection, optimization behavior, synthetic-data fidelity, and empirical membership leakage.

The main contributions of this study are summarized as follows:

1) Privacy-aware adaptation with explicit accounting. DP-SGD-based CGAN and VAE training is adapted for class-conditional augmentation on CICIDS2017. DP-CGAN privatizes the discriminator, whereas DP-VAE privatizes the complete encoder-decoder model. For each target budget, the achieved (ε, δ)-DP guarantee, noise multiplier, clipping norm, sample rate, and number of private optimization steps are reported.

2) Systematic evaluation of privacy-aware augmentation. A comparative study of conventional, non-private generative, and differentially private augmentation methods across various class-balance settings and privacy budgets. Their impact is evaluated using multiple downstream classifiers, with particular attention to overall detection performance and the recognition of minority attack classes.

3) Comprehensive analysis of the privacy-utility trade-off. Optimization behavior, synthetic-data fidelity, and empirical privacy leakage are examined within a unified evaluation framework. This analysis clarifies how generator architecture, privacy strength, augmentation level, and downstream classifier jointly influence the effectiveness of privacy-preserving intrusion detection.

The remainder of this study is organized as follows. Section II reviews related work on imbalanced IDS augmentation, private synthetic data, and privacy attacks against generators. Section III describes the threat model, privacy objectives, and DP-SGD implementation. Section IV presents the experimental protocol, exact tabular results, visual analyses, and a discussion of the privacy-utility-fidelity trade-off. Section V concludes the study and outlines future research directions.

II. Related Work

Class imbalance is a persistent limitation of machine learning-based IDSs because abundant benign traffic can dominate model optimization while rare attacks contribute only a small number of training examples. Conventional remedies include random oversampling, undersampling, cost-sensitive learning, and SMOTE. Although SMOTE is straightforward to integrate with standard classifiers, its samples are restricted to local interpolation and may not capture heterogeneous or multimodal attack distributions.

Deep generative models provide a more flexible alternative. Zhao et al. used CGAN-based class-conditional synthesis for imbalanced intrusion detection [1], whereas Lopez-Martin et al. developed a variational generative model for network-traffic augmentation [2]. An earlier study compared SMOTE, CGAN, and VAE across multiple augmentation ratios, demonstrating that their utility depends on both the generator and the downstream classifier [3]. More recent work has extended this direction through fully synthetic IDS training, feature-regularized GANs, and conditional encoder-decoder GAN architectures [4–6]. Surveys confirm the increasing use of GAN-based synthesis in IDS, while also highlighting training instability, data fidelity, and generalization as continuing challenges [7].

Synthetic-data evaluation increasingly adopts a multidimensional perspective. Dang and Hoang examined GAN-based synthesis through the complementary criteria of quality, similarity, and privacy [8]. Adams et al. likewise showed that fidelity, utility, and privacy may move in different directions across synthetic-data generators [9]. A recent systematic review further identified substantial variation in how privacy is implemented and measured for tabular synthetic data [10], while broader generative-AI research continues to document memorization and disclosure concerns [11].

Formal privacy can be incorporated into synthetic-data generation through several mechanisms. PrivBayes learns privatized Bayesian dependencies [17], PATE-GAN transfers information through noisy teacher aggregation [18], and DP-CGAN applies differential privacy to conditional adversarial training [19]. Related studies have examined model-inversion resistance [20] and distributed private generation [21]. These methods establish the feasibility of private synthesis, but their behavior on high-dimensional and highly imbalanced network-flow data remains insufficiently characterized. Moreover, DP-SGD, RDP accounting, and practical implementations such as Opacus introduce design choices whose effect on low-frequency attack classes requires domain-specific evaluation [22–24].

Synthetic records should not be assumed to be private merely because they are not exact copies of the source data. LOGAN demonstrated membership inference against GANs [12], while Monte Carlo and reconstruction-based attacks used generated samples or reconstruction quality to distinguish members from non-members [13]. GAN-Leaks subsequently organized such threats according to adversarial access and auxiliary knowledge [14]. DOMIAS showed that density-based overfitting detection can be particularly effective against uncommon records [15], a finding that is directly relevant to rare attack flows. Recent VAE-GAN research has also evaluated synthetic-data generation as a defense against membership inference while preserving downstream utility [16].

Membership inference captures only one dimension of disclosure risk. Giomi et al. proposed a broader attack-based framework covering singling out, linkability, and inference [25]. These studies motivate the combined use of formal privacy guarantees and empirical attacks: the privacy budget provides a

worst-case protection statement, whereas attack results characterize the evidence available to a specified adversary under an implemented protocol.

Table I summarizes the separation among IDS-oriented augmentation, general synthetic-data evaluation, formal private synthesis, and empirical privacy attacks. Recent IDS studies offer increasingly sophisticated augmentation mechanisms, but they generally do not provide formal privacy accounting. Conversely, private synthesis and privacy-risk studies are commonly evaluated on general tabular datasets without examining rare network-attack classes or the interaction with downstream IDS classifiers.

The present study connects these research directions in a common experimental pipeline. Its novelty is empirical and domain-specific: established DP-SGD-based CGAN and VAE architectures are adapted to imbalanced network-flow augmentation and assessed using formal privacy budgets, downstream utility, rare-class recall, synthetic-data fidelity, and black-box membership inference. The resulting analysis focuses on when private augmentation remains useful and how that outcome depends on the generator, privacy strength, augmentation ratio, and classifier.

TABLE I. COMPARISON OF REPRESENTATIVE WORK ON IMBALANCED IDS AUGMENTATION, PRIVATE SYNTHESIS, AND GENERATIVE-MODEL PRIVACY

Study	Scope and Method	IDS Imbalance	Formal DP	Privacy Evaluation
Zhao et al. [1]	CGAN-based IDS augmentation	Yes	No	Utility-focused; no formal privacy or attack evaluation.
Lopez-Martin et al. [2]	VAE-based generation for IDS	Yes	No	Evaluates detection utility; privacy leakage is not studied.
Nhan et al. [3]	SMOTE, CGAN, and VAE on CICIDS2017	Yes	No	Compares imbalance settings but does not include DP or privacy attacks.
Rahman et al. [4]; Li et al. [5]; Yang et al. [6]	Recent GAN-based augmentation for NIDS	Yes	No	Improves minority-class learning, but remains primarily utility-focused.
Dang and Hoang [8]; Adams et al. [9]; Hyrup et al. [10]	Quality, fidelity, utility, and privacy evaluation	No	Mixed	Supports multidimensional evaluation; not tailored to rare classes.
PrivBayes [17]	Bayesian-network tabular synthesis	No	Yes	Formal privacy for general tabular data; not tailored to rare IDS classes.
PATE-GAN [18]; DP-CGAN [19]	Differentially private adversarial synthesis	No	Yes	Private generation is established, but evidence for highly imbalanced multi-class IDS remains limited.
LOGAN, Monte Carlo, GAN-Leaks, DOMIAS [12-15]	Membership inference against generators	No	No	Provides attack methodologies, including risks to uncommon records.
Yan et al. [16]	VAE-GAN synthesis evaluated against MIA	No	No	Joint utility and empirical privacy evaluation outside the IDS domain.
Giomi et al. [25]	Unified synthetic-data privacy-risk framework	No	Method-agnostic	Covers multiple disclosure risks but does not evaluate rare-class IDS utility.

III. METHODOLOGY

This study develops a privacy-aware generative augmentation methodology for addressing class imbalance in machine-learning-based intrusion detection systems. Previous studies have demonstrated that SMOTE, conditional generative adversarial networks (CGANs), and variational autoencoders (VAEs) can improve the representation and detection of minority attack classes, although their effectiveness depends on the generator, imbalance severity, and downstream classifier [1–7]. Building on these findings, the present study incorporates DP into conditional generative training and evaluates whether the resulting synthetic data remain useful for intrusion detection while limiting the influence of individual training records.

The overall research workflow is illustrated in Fig. 1. The framework begins with an imbalanced network-flow dataset, which is cleaned and transformed using a fixed, data-independent preprocessing procedure. The preprocessed data are then divided into stratified training, validation, and test partitions before any augmentation is performed. This ordering ensures that no information from the validation or test partitions is used to generate synthetic records or balance the training data.

The training partition is processed under six data configurations: the original unaugmented setting, SMOTE, CGAN, VAE, DP-CGAN, and DP-VAE. SMOTE provides a conventional interpolation-based oversampling baseline, whereas CGAN and VAE represent non-private generative baselines. Their privacy-preserving counterparts incorporate DP-SGD during model training. In DP-CGAN, private optimization is applied to the discriminator, which directly accesses real training records. In DP-VAE, the complete encoder-decoder model is trained using DP-SGD. Per-sample gradient clipping and Gaussian noise injection are implemented through Opacus, while an RDP accountant tracks the cumulative privacy loss and reports the achieved (ε, δ)-differential privacy guarantee for each private model.

The original and augmented training datasets are subsequently used to train random forest and XGBoost classifiers. All configurations are evaluated on the same untouched real test set. The assessment covers aggregate utility (accuracy and Macro-F1), minority-sensitive utility (mean and per-class recall), synthetic-data fidelity (Wasserstein distance, mean squared error, mean absolute error, and t-SNE), and privacy protection (formal (ε, δ)-DP accounting and empirical membership inference).

The selected feature space is treated as fixed before private model training. A data-independent transformation is applied to prepare numerical network-flow features for the generative models without estimating normalization statistics from the private training records. After augmentation, random forest and XGBoost classifiers are trained on the resulting datasets and evaluated using classification utility, minority-class detection, synthetic-data quality, and empirical privacy-risk measures. The methodology therefore considers privacy and utility as interconnected evaluation dimensions rather than independent objectives.

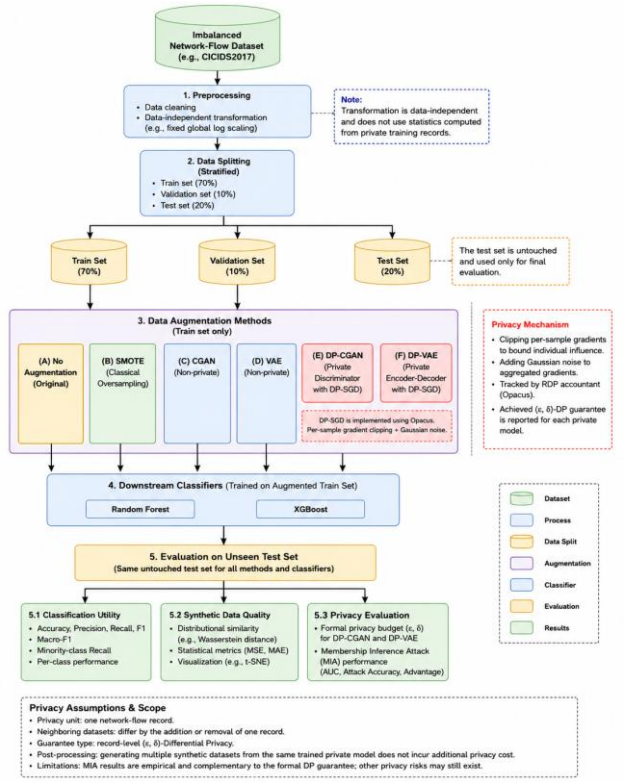

Fig. 1. Overall workflow of the proposed privacy-aware generative augmentation framework for imbalanced intrusion detection.

A. Attacker Knowledge and Access

Membership inference attacks against generative models aim to determine whether a candidate record contributed to the training data of a generator or its released synthetic dataset. Prior studies have demonstrated that membership information may be inferred using discriminator outputs, reconstruction behavior, generated samples, or distance-based evidence under different black-box and white-box access conditions [12–16].

The present study considers a black-box synthetic-data release setting. The adversary can access the synthetic records generated by the trained model and is assumed to know the feature representation, class-label space, preprocessing procedure, general generator architecture, and declared privacy parameters. This follows a conservative security assumption in which the protection of the source data does not depend on concealing the model architecture or training procedure.

The attacker does not have direct access to the original private training partition, record-level gradients, optimizer states, model-update history, or other internal training information. Candidate attack-flow records may be compared with the released synthetic data to estimate whether they were members of the generator's training partition.

The protected unit in this study is one network-flow record. Accordingly, the formal guarantee is interpreted as record-level differential privacy under an add-or-remove-one neighboring relation. It does not automatically provide user-level, host-level, session-level, or organization-level protection because a single entity may contribute multiple correlated network flows.

Membership inference is used as an empirical complement to formal privacy accounting. An attack result close to random guessing indicates that the implemented adversary was unable to reliably distinguish members from non-members under the evaluated protocol. However, such a result does not prove the absence of leakage under stronger, adaptive, or alternative privacy attacks. Synthetic-data privacy may also involve singling-out, linkability, reconstruction, or attribute-inference risks [25]. The formal privacy budget therefore remains the primary protection statement, whereas the attack results provide additional empirical evidence.

B. Privacy and Utility Objectives

The methodology is designed around two competing objectives: protecting individual training records and preserving the usefulness of the generated data for intrusion detection.

The privacy objective is to limit the effect that the inclusion or exclusion of one attack-flow record can have on a trained private generator. Differential privacy provides a formal mechanism for expressing this protection through the parameters ϵ and δ. A smaller value of ϵ represents a stronger privacy constraint, although stronger constraints generally require greater perturbation during model training. Previous private synthetic-data approaches [17–21], including PrivBayes, PATE-GAN, and DP-CGAN, demonstrate that explicit privacy guarantees can be incorporated into data generation, but they also reveal a trade-off between privacy strength and data utility.

The utility objective is to preserve the information required by downstream IDS classifiers. Because aggregate accuracy can conceal failures on rare attacks, the evaluation emphasizes Macro-F1, mean minority-class recall, and class-specific recall for Bot, BruteForce, and Infiltration, together with accuracy across augmentation ratios.

Synthetic-data quality is evaluated through complementary qualitative and quantitative evidence. t-SNE is used only as a visualization of local class structure, while Wasserstein distance, mean squared error, and mean absolute error quantify global discrepancies between real and synthetic samples. These metrics are interpreted together with downstream utility because low global distance does not necessarily preserve class-conditional decision boundaries.

The methodology therefore does not assume a single monotonic privacy-utility relationship. Instead, it evaluates how generator architecture, target privacy budget, Benign-to-Attack augmentation ratio, and downstream classifier jointly determine the operating point.

C. Differentially Private Stochastic Gradient Descent

Differentially private stochastic gradient descent is the main privacy-preserving optimization mechanism used in this study. DP-SGD extends conventional stochastic gradient descent by bounding the contribution of each training record and injecting calibrated Gaussian noise into the model-update process [22].

For each private update, gradients are computed separately for individual training examples. Their norms are clipped to a predefined maximum value so that no single flow record can exert an arbitrarily large influence on the update. Gaussian noise is then added to the aggregated clipped gradients before the optimizer modifies the model parameters. Repeating this procedure across the training process limits the distinguishability between models trained on neighboring datasets.

The revised implementation uses Opacus to perform per-sample gradient computation, gradient clipping, noise injection, private data sampling, and privacy accounting [24]. An RDP accountant tracks the cumulative privacy loss and converts it to an achieved (ε, δ)-DP guarantee [23]. This is important because clipping an already aggregated batch gradient does not provide the same record-level sensitivity control as standard DP-SGD. DP-SGD is applied only to DP-CGAN and DP-VAE. The conventional CGAN and VAE are trained without privacy perturbation and serve as non-private generative baselines. This separation enables the effect of differential privacy to be distinguished from differences caused by generator architecture alone.

IV. Experimental Evaluation

A. Experimental Setup

The CICIDS2017 dataset [26] was divided into training, validation, and test subsets, with test and validation proportions of 0.2 and 0.1, respectively. Privacy-aware generators were evaluated under $\varepsilon \in \{0.5, 1, 2, 4, 8\}$ with $\delta = 10^{-5}$ [3]. CGAN and VAE used a latent dimension of 16 and were trained for 40 epochs with a nominal batch size of 256. For DP-CGAN, DP-SGD was applied to the discriminator; for DP-VAE, it was applied to the complete encoder-decoder model. The clipping norm was fixed at 1.0, δ was set to 10^{-5}, and privacy loss was tracked with the Opacus RDP accountant.

Table II shows that the achieved privacy budgets closely match the corresponding targets for both private architectures. As ε increases from 0.5 to 8, the required noise multiplier decreases from 10.0000 to 1.0968, while δ, the clipping norm, sample rate, number of epochs, and number of private optimization steps remain fixed. The table therefore provides explicit, reproducible privacy accounting and confirms that the formal record-level guarantee is attached to the component that directly accesses private training records.

Fig. 2 complements the formal settings in Table II by showing how the two architectures respond to the calibrated clipping and noise. The non-private CGAN exhibits the oscillatory behavior expected from adversarial optimization, whereas the non-private VAE converges more smoothly. DP-CGAN remains trainable at all evaluated budgets, although its generator and discriminator losses continue to fluctuate. DP-VAE is more sensitive: at the strongest privacy settings, its reconstruction and KL terms change only marginally, while more visible latent-space learning emerges at $\varepsilon = 4$ and $\varepsilon = 8$. These trajectories demonstrate that the same formal privacy budget can affect adversarial and variational training differently; however, they are interpreted only as optimization evidence and not as a substitute for downstream utility.

TABLE II. Differential-Privacy Configurations used for DP-CGAN and DP-VAE

Model	Target ε	Achieved ε	Noise multiplier	Private component
DP-CGAN	0.5	0.4991	10.0000	Discriminator
DP-CGAN	1	0.9957	5.3516	Discriminator
DP-CGAN	2	1.9922	2.9297	Discriminator
DP-CGAN	4	3.9983	1.6992	Discriminator
DP-CGAN	8	7.9950	1.0968	Discriminator
DP-VAE	0.5	0.4991	10.0000	Encoder-decoder
DP-VAE	1	0.9957	5.3516	Encoder-decoder
DP-VAE	2	1.9922	2.9297	Encoder-decoder
DP-VAE	4	3.9983	1.6992	Encoder-decoder
DP-VAE	8	7.9950	1.0968	Encoder-decoder

Target and achieved ε values are reported separately. Common settings were $\delta = 10^{-5}$, clipping norm = 1.0, accountant sample rate = 0.0417, 40 epochs, and 960 private optimization steps. The formal guarantee applies to the private generative component under the add-or-remove-one record relation.

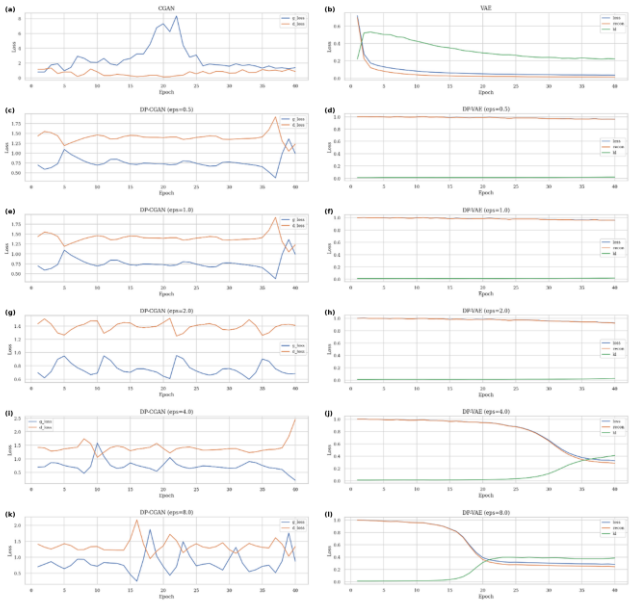

Fig. 2. Training-loss curves of CGAN, VAE, DP-CGAN, and DP-VAE. Panels (a)–(b) show the non-private models, whereas the remaining panels show target $\varepsilon \in \{0.5, 1, 2, 4, 8\}$.

B. Optimization Behavior and Synthetic-Data Structure

Fig. 3 provides a qualitative view of the local class structure produced by the augmentation methods. SMOTE most closely

follows the local neighborhoods of the original samples, CGAN produces denser class-conditioned regions, and VAE introduces a smoother representation. The private variants preserve recognizable local patterns but display greater class mixing, with DP-VAE showing the strongest smoothing. Because t-SNE is a local, low-dimensional projection, these plots are not treated as proof of global distributional fidelity.

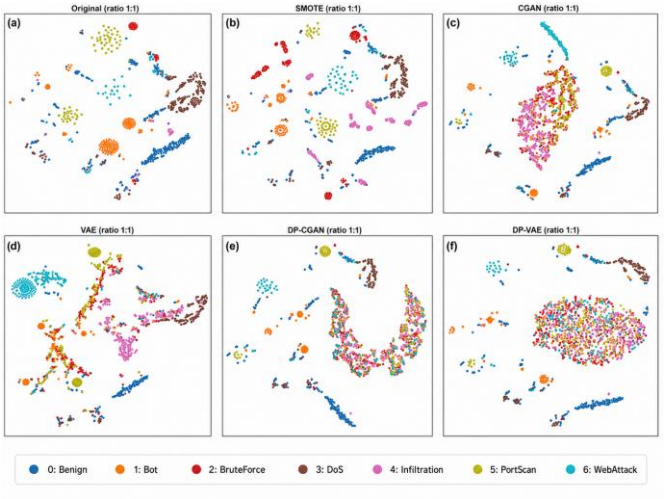

Fig. 3. t-SNE visualization of the original, SMOTE, CGAN, VAE, DP-CGAN, and DP-VAE class distributions at the Benign: Attack ratio of 1:1.

Fig. 4 shows that augmentation ratio and privacy budget have only a small effect on random forest accuracy [Fig. 4(a)] but a clearer effect on XGBoost [Fig. 4(b)]. With random forest, both private generators remain within the narrow range 0.9922–0.9927 across all ratios and budgets. With XGBoost, DP-CGAN ranges from 0.9847 to 0.9920, whereas DP-VAE ranges from 0.9756 to 0.9897 and generally weakens as the Benign-to-Attack ratio moves toward 1:5. The contrast between the two heatmaps demonstrates that accuracy alone is highly saturated and may obscure architecture-specific degradation; consequently, Tables III and IV and Fig. 5 and Fig. 6 provide the more informative minority-sensitive evidence.

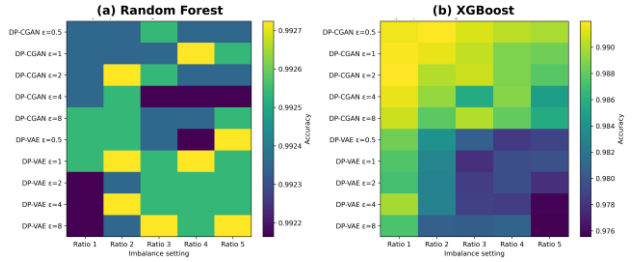

Fig. 4. Accuracy of DP-CGAN and DP-VAE across target Benign-to-Attack augmentation ratios from 1:1 to 1:5: (a) Random forest; (b) XGBoost.

Tables III and IV provide the exact Macro-F1 and mean minority-recall values, while Fig. 5 and Fig. 6 visualize the same trends across achieved privacy budgets. DP-CGAN is stable across both classifiers: Macro-F1 ranges from 0.9736 to 0.9809 with random forest and from 0.9694 to 0.9768 with XGBoost. DP-VAE remains competitive with random forest (0.9736–0.9793) but degrades markedly with XGBoost (0.7822–0.8066). The classifier interaction is even stronger for mean minority recall: DP-CGAN and DP-VAE with random forest, and DP-CGAN with XGBoost, remain near 0.949, whereas DP-VAE with XGBoost ranges only from 0.4787 to 0.5218. The tables and figures therefore support the second contribution by demonstrating a systematic comparison across privacy budgets and downstream classifiers, while also showing that the effect of ε is not monotonic and is less influential than the generator–classifier pairing within the evaluated range.

Tables V–VII and Fig. 7 refine the aggregate results by reporting recall separately for Bot, BruteForce, and Infiltration under the strongest evaluated augmentation setting, Benign: Attack = 1:5. Table V establishes the non-private reference for SMOTE, CGAN, and VAE, while Tables VI and VII isolate the effect of privacy budget for DP-CGAN and DP-VAE. The paired heatmaps in Fig. 7 make the classifier-dependent class failures visually explicit.

TABLE III. MACRO-F1 OF DP-CGAN AND DP-VAE ACROSS PRIVACY BUDGETS, AVERAGED OVER THE FIVE AUGMENTATION RATIOS

Target ε	**DP-CGAN**		**DP-VAE**	
	Random forest	***XGBoost***	***Random forest***	***XGBoost***
0.5	**0.9809**	0.9757	0.9774	0.8042
1	**0.9792**	0.9768	0.9757	0.7873
2	**0.9792**	0.9747	0.9736	0.8066
4	0.9753	0.9694	**0.9774**	0.8004
8	0.9736	0.9752	**0.9793**	0.7822

TABLE IV. MEAN MINORITY-CLASS RECALL OF DP-CGAN AND DP-VAE ACROSS PRIVACY BUDGETS, AVERAGED OVER THE FIVE AUGMENTATION RATIOS

Target ε	**DP-CGAN**		**DP-VAE**	
	Random forest	***XGBoost***	***Random forest***	***XGBoost***
0.5	0.9492	0.9495	**0.9497**	0.5086
1	0.9493	0.9490	**0.9498**	0.4799
2	0.9492	**0.9498**	**0.9498**	0.5218
4	0.9490	0.9488	**0.9495**	0.5096
8	0.9493	0.9486	**0.9498**	0.4787

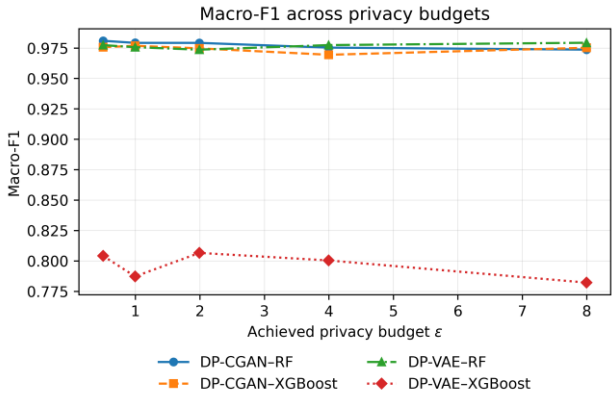

Fig. 5. Macro-F1 across achieved privacy budgets, averaged over the five augmentation ratios.

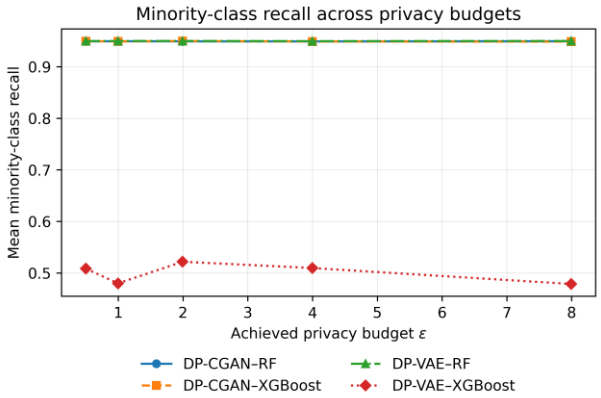

Fig. 6. Mean minority-class recall across achieved privacy budgets, averaged over the five augmentation ratios.

The exact values in Tables V–VII are consistent with the visual patterns in Fig. 7. Random forest is highly stable across all private configurations: Bot recall is approximately 0.99, BruteForce recall is 1.00, and Infiltration recall is 0.8571. DP-CGAN also preserves these classes with XGBoost, including BruteForce recall of 1.00 and Infiltration recall of 0.8571 at every evaluated budget. DP-VAE behaves differently. With random forest, it retains the same strong recalls, but with XGBoost, its Bot recall falls to 0.9517–0.9669, BruteForce recall remains at 0.1579, and Infiltration recall is 0.0. This class-level evidence explains the low DP-VAE–XGBoost Macro-F1 and minority recall, confirms that aggregate accuracy can conceal complete failure on a security-critical class, and strengthens the second contribution by demonstrating the importance of class-specific evaluation.

The privacy and fidelity evidence is reported separately because the two dimensions answer different questions. Table VIII and Fig. 8 evaluate whether the implemented black-box adversary can distinguish training members from non-members, whereas Table IX evaluates global agreement between real and synthetic data.

TABLE V. RARE-ATTACK RECALL OF NON-PRIVATE AUGMENTATION METHODS AT BENIGN: ATTACK = 1:5

Attack class	SMOTE		CGAN		VAE	
	RF	*XGB*	*RF*	*XGB*	*RF*	*XGB*
Bot	0.9924	**0.9949**	0.9898	0.9847	0.9924	0.9720
BruteForce	**1.0000**	**1.0000**	**1.0000**	0.8421	**1.0000**	0.7895
Infiltration	**0.8571**	**0.8571**	**0.8571**	**0.8571**	**0.8571**	0.7143

TABLE VI. RARE-ATTACK RECALL OF DP-CGAN ACROSS PRIVACY BUDGETS AT BENIGN: ATTACK = 1:5

Target ε	Bot		BruteForce		Infiltration	
	RF	*XGB*	*RF*	*XGB*	*RF*	*XGB*
0.5	**0.9898**	0.9847	**1.0000**	**1.0000**	**0.8571**	**0.8571**
1	**0.9924**	0.9847	**1.0000**	**1.0000**	**0.8571**	**0.8571**
2	0.9898	**0.9949**	**1.0000**	**1.0000**	**0.8571**	**0.8571**
4	**0.9898**	0.9847	**1.0000**	**1.0000**	**0.8571**	**0.8571**
8	**0.9924**	0.9873	**1.0000**	**1.0000**	**0.8571**	**0.8571**

TABLE VII. RARE-ATTACK RECALL OF DP-VAE ACROSS PRIVACY BUDGETS AT BENIGN: ATTACK = 1:5

Target ε	Bot		BruteForce		Infiltration	
	RF	*XGB*	*RF*	*XGB*	*RF*	*XGB*
0.5	**0.9924**	0.9669	**1.0000**	0.1579	**0.8571**	0.0000
1	**0.9924**	0.9618	**1.0000**	0.1579	**0.8571**	0.0000
2	**0.9924**	0.9517	**1.0000**	0.1579	**0.8571**	0.0000
4	**0.9924**	0.9593	**1.0000**	0.1579	**0.8571**	0.0000
8	**0.9924**	0.9542	**1.0000**	0.1579	**0.8571**	0.0000

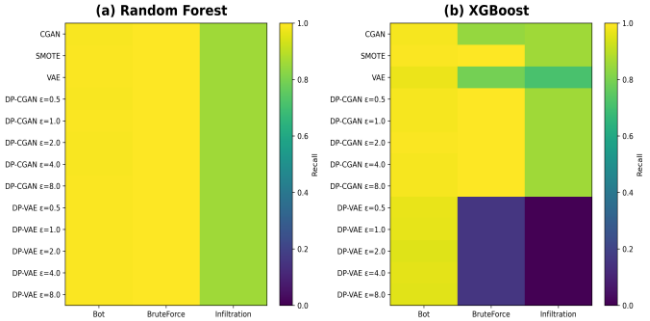

Fig. 7. Rare-attack recall at Benign: Attack = 1:5: (a) Random forest; (b) XGBoost.

TABLE VIII. EMPIRICAL MEMBERSHIP-INFERENCE RESULTS ACROSS PRIVACY BUDGETS, AVERAGED OVER THE FIVE AUGMENTATION RATIOS

Target ε	DP-CGAN			DP-VAE		
	ROC-AUC mean ± SD	*Accuracy*	*A_acc*	*ROC-AUC mean ± SD*	*Accuracy*	*A_acc*
0.5	0.5137 ± 0.0012	0.5085	0.0169	**0.5021 ± 0.0075**	**0.5033**	**0.0140**
1	0.5141 ± 0.0026	0.5109	0.0217	**0.4975 ± 0.0069**	**0.5011**	**0.0097**
2	0.5018 ± 0.0008	0.5009	**0.0017**	**0.4986 ± 0.0052**	**0.5001**	0.0088
4	0.4866 ± 0.0005	0.4891	0.0219	**0.5034 ± 0.0072**	**0.5009**	**0.0067**
8	0.4934 ± 0.0009	**0.4993**	**0.0024**	**0.4993 ± 0.0090**	0.4989	0.0117

A_acc = |2 × attack accuracy − 1|. Values close to ROC-AUC = 0.5, accuracy = 0.5, and A_acc = 0 indicate chance-level discrimination. SD is computed across augmentation ratios.

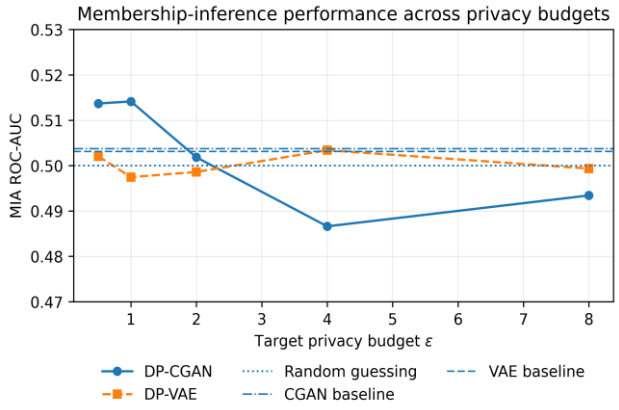

Fig. 8. Membership-inference ROC-AUC across target privacy budgets. The dotted line denotes random guessing; horizontal references denote the non-private CGAN and VAE baselines.

TABLE IX. SYNTHETIC-DATA FIDELITY ACROSS PRIVACY BUDGETS, AVERAGED OVER THE FIVE AUGMENTATION RATIOS

Target ε	DP-CGAN			DP-VAE		
	Wasserstein	*MSE*	*MAE*	*Wasserstein*	*MSE*	*MAE*
0.5	0.0572	0.0075	0.0586	**0.0439**	**0.0037**	**0.0475**
1	**0.0282**	0.0038	**0.0301**	0.0377	**0.0028**	0.0414
2	0.1098	0.0195	0.1104	**0.0322**	**0.0021**	**0.0361**
4	0.0575	0.0072	0.0583	**0.0275**	**0.0016**	**0.0314**
8	0.0318	0.0020	0.0330	**0.0236**	**0.0012**	**0.0276**

Table VIII provides the exact ROC-AUC, attack accuracy, and accuracy-derived advantage values, and Fig. 8 shows their behavior relative to random guessing and the non-private CGAN and VAE references. MIA ROC-AUC remains close to 0.5: DP-CGAN ranges from 0.4866 to 0.5141 and DP-VAE from 0.4975 to 0.5034. Attack accuracy is also near 0.5, with small advantages. However, the non-private references are themselves close to chance level. The defensible conclusion is therefore that the implemented attack did not detect a reliable membership signal in either private or non-private generators. These results provide empirical evidence under the stated threat model, but they do not establish that DP caused the low attack performance.

Table IX shows that synthetic-data fidelity does not follow the same pattern as downstream utility. DP-VAE generally obtains lower global distance and error as ε increases, reaching Wasserstein = 0.0236, MSE = 0.0012, and MAE = 0.0276 at ε = 8. DP-CGAN is non-monotonic, with particularly high discrepancy at ε = 2. Nevertheless, DP-VAE–XGBoost remains weak even when the global fidelity metrics are favorable. The mismatch between Table IX and the class-level results in Tables VI and VII indicates that Wasserstein distance, MSE, and MAE do not fully represent class-conditional separability or the decision boundaries required by the downstream classifier.

V. CONCLUSION AND FUTURE WORK

This study evaluated whether differentially private generative augmentation can address class imbalance in intrusion detection while providing a formal record-level privacy guarantee for the generative component. DP-SGD was integrated into CGAN and VAE training and evaluated on CICIDS2017 across five target Benign-to-Attack augmentation ratios, five privacy budgets, and two downstream classifiers. The revised evaluation combines exact utility tables, class-specific rare-attack recall, optimization curves, t-SNE visualization, quantitative fidelity metrics, and empirical membership-inference results.

The results show that DP-CGAN is the most robust private augmentation configuration across classifier choices. It preserves high Macro-F1 and stable recall for Bot, BruteForce, and Infiltration with both random forest and XGBoost. DP-VAE remains competitive with random forest but degrades substantially with XGBoost, including near-complete failure on BruteForce and Infiltration. This finding demonstrates that the downstream classifier can either tolerate or expose weaknesses in private synthetic data. Global fidelity metrics provide useful complementary evidence but do not fully predict class-conditional utility. Membership-inference ROC-AUC remains near random guessing for both private and non-private generators under the implemented protocol; therefore, the attack results indicate an absence of detected membership signal, not proof that differential privacy alone caused the empirical resistance.

Several limitations remain. The study uses one benchmark dataset, one random-split protocol, a single black-box membership attack, and utility summaries that do not yet include repeated generator-training confidence intervals. The evaluation also lacks a complete real-only, synthetic-only, and real-plus-synthetic ablation, which is needed to quantify the marginal contribution of the generated records and to explain why random forest remains strong when DP-VAE optimization is weak. Future work should therefore include session- or time-disjoint validation, additional IDS datasets, stronger and calibrated privacy attacks with a positive control, per-class fidelity and nearest-neighbor disclosure measures, repeated generator seeds, and ablations over the amount of synthetic data. These extensions would provide stronger evidence for deployment-oriented privacy-utility operating points.

ACKNOWLEDGMENT

This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number C2024-26-15.

REFERENCES

[1] G. Zhao, P. Liu, K. Sun, Y. Yang, T. Lan, and H. Yang, "Research on data imbalance in intrusion detection using CGAN," PLoS ONE, vol. 18, no. 10, Art. no. e0291750, 2023, doi: 10.1371/journal.pone.0291750.

[2] M. Lopez-Martin, B. Carro, and A. Sanchez-Esguevillas, "Variational data generative model for intrusion detection," Knowledge and Information Systems, vol. 60, no. 1, pp. 569–590, 2019, doi: 10.1007/s10115-018-1306-7.

[3] V. T. Nhan, H. Y. Nhi, T. Ha, D. T. H. Loan, and T. K. Dang, "Evaluating sampling strategies and generative models for addressing class imbalance in intrusion detection systems," in Proc. International Conference on Future Data and Security Engineering, 2025, pp. 110–125.

[4] M. S. Rahman, S. Pal, S. Mittal, T. Chawla, and C. Karmakar, "SYN-GAN: A robust intrusion detection system using GAN-based synthetic data for IoT security," Internet of Things, vol. 26, Art. no. 101212, 2024, doi: 10.1016/j.iot.2024.101212.

[5] J. Li, W. Zong, Y.-W. Chow, and W. Susilo, "Mitigating class imbalance in network intrusion detection with feature-regularized GANs," Future Internet, vol. 17, no. 5, Art. no. 216, 2025, doi: 10.3390/fi17050216.

[6] Y. Yang, X. Liu, D. Wang, Q. Sui, C. Yang, H. Li, Y. Li, and T. Luan, "A CE-GAN-based approach to address data imbalance in network intrusion detection systems," Scientific Reports, vol. 15, no. 1, Art. no. 7916, 2025, doi: 10.1038/s41598-025-90815-5.

[7] M. Alauthman, N. Aslam, A. Al-Qerem, A. Aldweesh, and P. Sureephong, "Generative adversarial networks for intrusion detection systems: A comprehensive survey of applications, challenges, and research directions," Arabian Journal for Science and Engineering, 2026, doi: 10.1007/s13369-026-11103-6.

[8] T. K. Dang and D. Q. Hoang, "Synthesizing realistic data with GANs: Balancing quality, similarity, and privacy," in Proc. 18th International Conference on Advanced Computing and Analytics (ACOMPA), 2024, pp. 123–131, doi: 10.1109/ACOMPA64883.2024.00025.

[9] T. Adams, C. Birkenbihl, K. Otte, H. G. Ng, J. A. Rieling, A. F. Näher, U. Sax, F. Prasser, and H. Fröhlich, "On the fidelity versus privacy and utility trade-off of synthetic patient data," iScience, vol. 28, no. 5, Art. no. 112382, 2025, doi: 10.1016/j.isci.2025.112382.

[10] T. Hyrup, A. D. Lautrup, A. Zimek, and P. Schneider-Kamp, "A systematic review of privacy-preserving techniques for synthetic tabular health data," Discover Data, vol. 3, no. 1, Art. no. 5, 2025, doi: 10.1007/s44248-025-00022-w.

[11] A. Golda, K. Mekonen, A. Pandey, A. Singh, V. Hassija, V. Chamola, and B. Sikdar, "Privacy and security concerns in generative AI: A comprehensive survey," IEEE Access, vol. 12, pp. 48126–48144, 2024.

[12] J. Hayes, L. Melis, G. Danezis, and E. De Cristofaro, "LOGAN: Membership inference attacks against generative models," Proceedings on Privacy Enhancing Technologies, vol. 2019, no. 1, pp. 133–152, 2019.

[13] B. Hilprecht, M. Härterich, and D. Bernau, "Monte Carlo and reconstruction membership inference attacks against generative models," Proceedings on Privacy Enhancing Technologies, vol. 2019, no. 4, pp. 232–249, 2019.

[14] D. Chen, N. Yu, Y. Zhang, and M. Fritz, "GAN-Leaks: A taxonomy of membership inference attacks against generative models," in Proc. 2020 ACM SIGSAC Conference on Computer and Communications Security, 2020, pp. 343–362.

[15] B. van Breugel, H. Sun, Z. Qian, and M. van der Schaar, "Membership inference attacks against synthetic data through overfitting detection," in Proc. International Conference on Artificial Intelligence and Statistics, vol. 206, 2023, pp. 3493–3514.

[16] J. Yan, H. Huang, K. Yang, H. Xu, and Y. Li, "Synthetic data for enhanced privacy: A VAE-GAN approach against membership inference attacks," Knowledge-Based Systems, vol. 309, Art. no. 112899, 2025, doi: 10.1016/j.knosys.2024.112899.

[17] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao, "PrivBayes: Private data release via Bayesian networks," ACM Transactions on Database Systems, vol. 42, no. 4, pp. 1–41, 2017.

[18] J. Jordon, J. Yoon, and M. van der Schaar, "PATE-GAN: Generating synthetic data with differential privacy guarantees," in Proc. International Conference on Learning Representations, 2019.

[19] R. Torkzadehmahani, P. Kairouz, and B. Paten, "DP-CGAN: Differentially private synthetic data and label generation," in Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019.

[20] D. Chen, S. C. S. Cheung, C.-N. Chuah, and S. Ozonoff, "Differentially private generative adversarial networks with model inversion," in Proc. IEEE International Workshop on Information Forensics and Security, 2021, pp. 1–6.

[21] A. Triastcyn and B. Faltings, "Federated generative privacy," IEEE Intelligent Systems, vol. 35, no. 4, pp. 50–57, 2020.

[22] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in Proc. 2016 ACM SIGSAC Conference on Computer and Communications Security, 2016, pp. 308–318, doi: 10.1145/2976749.2978318.

[23] I. Mironov, "Rényi differential privacy," in Proc. IEEE 30th Computer Security Foundations Symposium, 2017, pp. 263–275.

[24] A. Yousefpour, I. Shilov, A. Sablayrolles, D. Testuggine, K. Prasad, M. Malek, J. Nguyen, S. Ghosh, A. Bharadwaj, J. Zhao, G. Cormode, and I. Mironov, "Opacus: User-friendly differential privacy library in PyTorch," in NeurIPS 2021 Workshop on Privacy in Machine Learning, 2021.

[25] M. Giomi, F. Boenisch, C. Wehmeyer, and B. Tasnádi, "A unified framework for quantifying privacy risk in synthetic data," Proceedings on Privacy Enhancing Technologies, vol. 2023, no. 2, pp. 312–328, 2023, doi: 10.56553/popets-2023-0055.

[26] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in Proc. 4th International Conference on Information Systems Security and Privacy, 2018, pp. 108–116, doi: 10.5220/0006639801080116.