Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.082332

ARTICLE

Differential Privacy for Security Telemetry: An Empirical Study of Utility Loss in Intrusion Detection Systems

Sajad Homayoun*

Cyber Security Group, Department of Electronic Systems, Aalborg University, Copenhagen, Denmark

*Corresponding Author: Sajad Homayoun. Email: sajadh@es.aau.dk

Received: 16 March 2026; Accepted: 09 May 2026; Published: 15 June 2026

ABSTRACT: Intrusion detection systems depend on detailed security telemetry, yet such telemetry is often too sensitive to share or reuse outside controlled environments. Differential Privacy (DP) offers formal protection by injecting randomness, but its practical impact on detection utility is not well understood, especially under class imbalance and for rare attacks. This paper presents a controlled empirical study of feature-level DP applied to security telemetry for intrusion detection. Using a fixed model and a fixed train–test split, we vary only the privacy budget and quantify how performance changes across standard metrics, including macro-averaged scores and per-class recall. While aggregate metrics such as accuracy and Micro-F1 remain comparatively high, class-balanced metrics degrade substantially under stronger privacy constraints. In particular, the detection of rare and low-volume attacks is severely affected, with some classes becoming undetectable under feature-level DP perturbation. These results indicate that privacy–utility trade-offs in intrusion detection are highly class-dependent and that aggregate performance measures may hide operationally relevant degradation.

KEYWORDS: Differential privacy; security telemetry; privacy–utility trade-off; rare attack detection; class imbalance

1 Introduction

Intrusion detection systems (IDS) rely on detailed security telemetry to identify malicious activity in networked environments. Such telemetry typically includes network flows, packet-level statistics, logs, and alerts that capture how hosts communicate and how services are used over time. Access to fine-grained telemetry is often essential for distinguishing normal behavior from attacks, particularly in complex and dynamic networks.

At the same time, security telemetry is inherently sensitive. Network traces and logs may expose internal addressing schemes, communication patterns, user behavior, and operational characteristics of an organization's infrastructure. This sensitivity limits how telemetry can be shared, reused, or analyzed outside tightly controlled environments, even when broader access could improve intrusion detection methods or support collaborative security efforts.

Privacy-preserving mechanisms inevitably affect the information available to IDSs. Approaches such as aggregation, feature reduction, or the introduction of uncertainty can reduce the risk of sensitive information disclosure, but they also modify the characteristics of the data on which detection depends. Since many attacks manifest as subtle deviations within otherwise normal traffic, even modest distortions of telemetry may have a disproportionate impact on detection performance, particularly for rare or low-volume attacks [1].

Differential Privacy (DP) has emerged as a principal framework for protecting sensitive data while enabling statistical analysis. By providing formal guarantees that limit what can be inferred about individual records, DP has become an attractive option for handling security telemetry and for training machine learning models on sensitive network data. Consequently, DP has been increasingly considered in intrusion detection and related security analytics [2–5].

However, the guarantees offered by DP are achieved through the introduction of controlled randomness [6]. In intrusion detection settings, where models often rely on fine-grained patterns and small variations in telemetry, DP-induced noise may alter the signals required for effective detection. While existing work demonstrates that DP can be integrated into IDS pipelines, evaluations rely on aggregate performance metrics and limited privacy configurations [7–9]. Only a subset of studies explicitly examines privacy–utility trade-offs by analyzing detection performance across varying privacy budgets [3,10–12].

As a result, the practical impact of DP on intrusion detection utility remains insufficiently understood. In particular, there is limited empirical guidance on how DP-induced noise affects the detection of rare or low-volume attacks, which are often of greatest concern in operational security environments such as Security Operation Centers (SOC). This lack of systematic understanding introduces uncertainty for both researchers and practitioners seeking to deploy privacy-preserving IDS.

In this paper, we present an empirical study of the impact of DP on intrusion detection when applied directly to security telemetry features. Rather than proposing a new detection model or privacy mechanism, we focus on quantifying how varying levels of DP-induced noise influence detection performance. By systematically evaluating intrusion detection under different privacy settings, we aim to characterize the resulting utility loss and its implications for practical deployment.

This paper makes the following contributions:

- We design a controlled experimental framework that isolates the impact of feature-level DP on intrusion detection by holding the model, data split, and evaluation procedure fixed.
- We provide empirical evidence that commonly reported aggregate metrics (e.g., accuracy and Micro-F1) can mask substantial utility degradation under DP, particularly in class-imbalanced intrusion detection tasks.
- We demonstrate that feature-level DP can severely impair the detection of rare and low-volume attacks, with some attack classes becoming effectively undetectable across a range of privacy budgets.

Rather than proposing another detection model, we focus on understanding how DP noise affects intrusion detection behavior under realistic class imbalance. By deliberately avoiding privacy-aware model tuning or adaptive mitigation strategies, the analysis is designed to expose how sensitive common intrusion detection pipelines are to feature-level DP under realistic class imbalance. This perspective complements existing work focused on achieving acceptable aggregate performance under DP, by highlighting failure modes that may remain hidden when evaluation emphasizes overall accuracy or Micro-F1 alone.

The remainder of this paper is organized as follows. Section 2 reviews related work and Section 3 presents the methodology. Section 4 shows the experimental setup and the results, while Section 5 discusses the implications and limitations. Finally, Section 6 concludes the paper.

2 Related Work

This section follows a PRISMA-guided systematic mapping approach, adhering to the PRISMA 2020 reporting guidelines [13], to identify and categorize existing research on the use of DP in the context of security telemetry and intrusion detection. Rather than conducting a full systematic literature review with quantitative synthesis, the goal of this mapping is to provide a structured overview of how DP has been

applied in intrusion detection and related security analytics, and to identify gaps that motivate the empirical analysis presented in this paper.

The mapping focuses on three research questions (RQs), defined as follows:

- RQ1: In which stages of intrusion detection and security telemetry pipelines has DP been applied?
- RQ2: What types of security telemetry and utility metrics are evaluated in DP-based intrusion detection studies?
- RQ3: To what extent do existing studies empirically analyze the impact of DP on detection utility, particularly for rare attacks?

Search Strategy and Queries

A structured keyword-based search was used across these databases: *IEEE Xplore, ACM Digital Library, Web of Science*. The search terms were designed to balance precision and recall, focusing on DP, intrusion detection, and security telemetry, while avoiding overly broad terms such as "SOC" or "SIEM" that tend to produce a high number of irrelevant results.

The literature search was restricted to studies published between 2010 and 31 January 2026. We used the following query to retrieve all related papers:

```
TITLE("differential privacy" OR "differentially private")
AND TITLE-ABS("intrusion detection" OR IDS OR NIDS)
AND TITLE-ABS(telemetry OR traffic OR flows OR logs)
AND PUBYEAR > 2009
```

The query syntax was adapted to the requirements of each database while preserving the logical structure and keywords. Table 1 shows the number of papers found in each database.

Table 1: Number of records retrieved from each bibliographic database using the defined search query (prior to screening and deduplication).

Database	Number of Records
Scopus	11
IEEE Xplore	7
ACM Digital Library	4
Web of Science	24
Total	**46**

Fig. 1 shows the paper selection process using the PRISMA flow diagram. To ensure a consistent and unbiased selection process, a set of predefined exclusion criteria was applied:

- **Reason 1 (R1): DP not applied to intrusion detection telemetry or features.** Studies in which DP was present but not applied to data, features, or learning processes directly related to intrusion detection were excluded.
- **Reason 2 (R2): No quantitative evaluation of intrusion detection utility.** Studies that proposed DP mechanisms but did not report quantitative intrusion detection performance metrics, or did not enable comparison between DP and non-DP settings, were excluded.
- **Reason 3 (R3): Detection task not evaluated on security-relevant telemetry.** Studies whose evaluation relied on datasets or tasks that could not reasonably be interpreted as security telemetry for intrusion detection were excluded.

- **Reason 4 (R4): Insufficient experimental or methodological detail.** Studies lacking sufficient detail regarding datasets, DP parameters, or evaluation methodology to support interpretation or comparison were excluded.

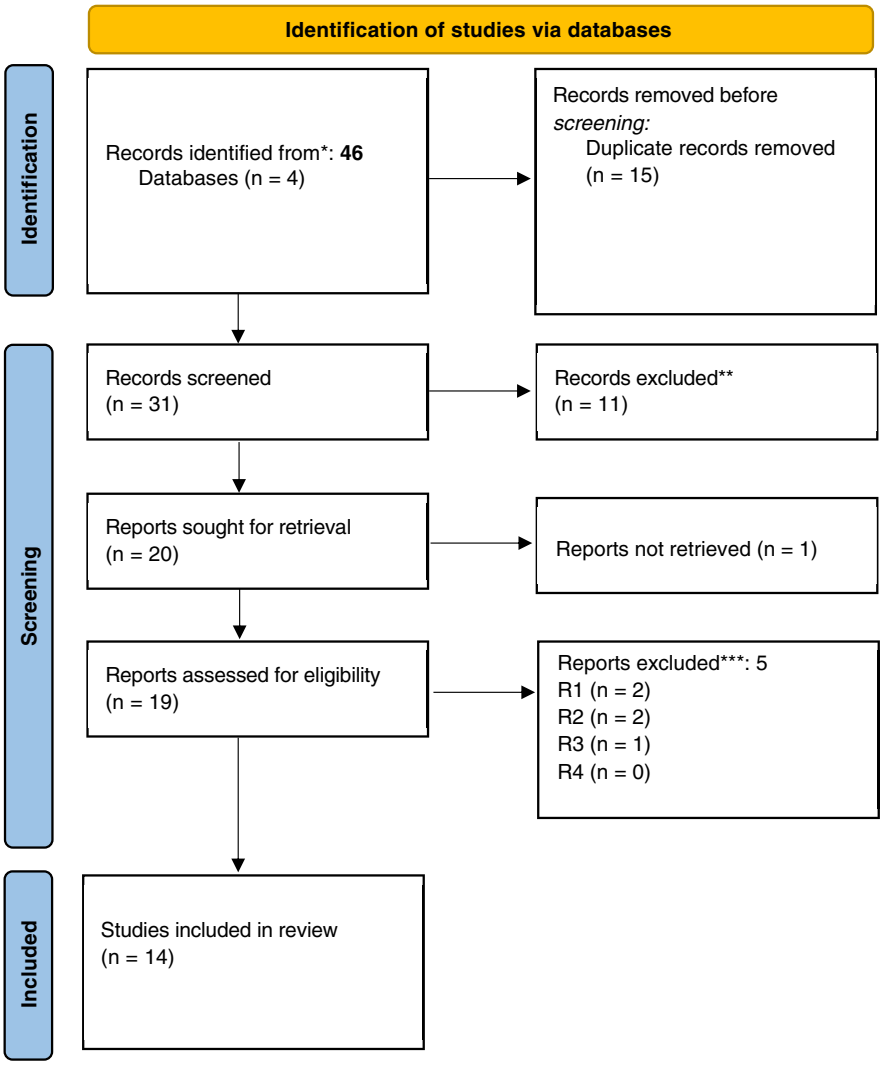


Figure 1: PRISMA flow diagram illustrating the study selection process for the systematic mapping of DP in intrusion detection and security telemetry. *Records identified from electronic database searches. **Records excluded at title/abstract screening stage. ***Reasons (R1-R4) are explained in the text.

Table 2 summarizes the reviewed studies across dimensions directly relevant to this work. Domain indicates the application context of the intrusion detection system (e.g., Internet of Things (IoT), Industrial Internet of Things (IIoT), vehicular ad-hoc network (VANET)). Telemetry describes the primary data representation used for detection. DP Placement specifies where Differential Privacy is integrated within the detection pipeline. Utility Evaluation reflects whether the study systematically analyzes privacy–utility degradation under varying privacy configurations. Metrics Type indicates whether evaluation relies primarily on aggregate performance measures or includes class-balanced metrics.

Table 2 reveals several consistent patterns. DP is most frequently integrated at the learning stage, either during centralized training or within federated learning updates. This pattern is particularly common in IoT,

industrial, and vehicular intrusion detection systems [2,3,11,12,14,15]. In contrast, comparatively fewer studies apply DP directly to telemetry representations prior to model training, such as through feature perturbation or graph-level privacy mechanisms [4,5,16,17].

Table 2: Systematic mapping of included studies on DP in intrusion detection and security telemetry.

Study	Domain	Telemetry	DP Placement	Utility Evaluation	Metrics Type
[7]	VANET	Traffic features	Telemetry/training	Partial	Aggregate
[4]	NIDS	Graph-based	Telemetry	Partial	Aggregate
[3]	IoT	Traffic features	FL updates/training	Explicit	Aggregate
[2]	IIoT	Traffic features	Training/FL updates	Partial	Aggregate
[16]	VANET	Graph (flow relations)	Telemetry	Partial	Aggregate
[5]	VANET	Logs (CAN bus)	Telemetry	Partial	Aggregate
[11]	IoT	Logs/traffic features	Training/FL updates	Explicit	Aggregate
[8]	NIDS	Traffic features	Training	Explicit	Aggregate + class metrics
[9]	IoT	Traffic features	Training	Explicit	Aggregate + class metrics
[10]	NIDS	Synthetic features	Synthetic data generation	Explicit	Aggregate
[12]	IoT	Traffic features	FL updates/training	Explicit	Aggregate + class metrics
[14]	VANET	Sensor dataset	FL updates/training	Explicit	Aggregate
[17]	VANET	Simulated traffic	Telemetry	Partial	Aggregate + class metrics
[15]	NIDS	Traffic features	FL updates/training	Partial	Aggregate

Moreover, the majority of studies evaluate detection performance primarily using aggregate metrics, most commonly overall accuracy or F1-score. Although some works report additional measures, systematic class-balanced analysis remains less common. Only a subset of studies explicitly examines privacy–utility trade-offs by evaluating performance across multiple privacy configurations [3,10–12], while others provide limited or partial evaluation.

In general, the mapping suggests that while privacy-preserving intrusion detection has been actively explored across multiple domains, a structured and comparable evaluation of privacy-induced utility degradation remains limited. In particular, the dominance of aggregate metrics and the relatively infrequent use of class-sensitive analysis make it difficult to assess how DP affects minority classes and rare attack detection in operational settings.

3 Methodology

The purpose of this section is to empirically assess how the application of DP to security telemetry affects the utility of intrusion detection systems. Rather than proposing a new detection model or privacy mechanism, we design a controlled experimental setup in which DP is applied directly to telemetry features prior to model training, and its impact on detection performance is measured systematically.

The study follows a comparative evaluation approach. A baseline intrusion detection model is trained on original, non-perturbed telemetry features and evaluated using standard detection metrics. DP is then applied to the same telemetry features under varying privacy budgets, and the model is retrained and re-evaluated under identical conditions. By keeping the detection model and experimental parameters constant and varying only the privacy configuration, the resulting changes in performance can be attributed to the effects of DP.

Particular attention is paid to the detection of rare and low-volume attacks, which are often critical in operational security environments, but may be especially sensitive to noise introduced for privacy protection. Detection performance is therefore analyzed both in aggregate and at the level of individual attack classes, enabling a more fine-grained assessment of utility loss under DP.

Fig. 2 shows the controlled experimental pipeline used to isolate the effect of feature-level DP. The baseline and privacy-preserving paths differ only in the application of DP to training features, while the model architecture, hyperparameters, and test data remain unchanged. This design allows observed performance differences to be attributed primarily to privacy-induced perturbation. To support reproducibility, the implementation of the experimental pipeline, including data preprocessing, differential privacy perturbation, and model evaluation scripts, is available in an online repository[1].

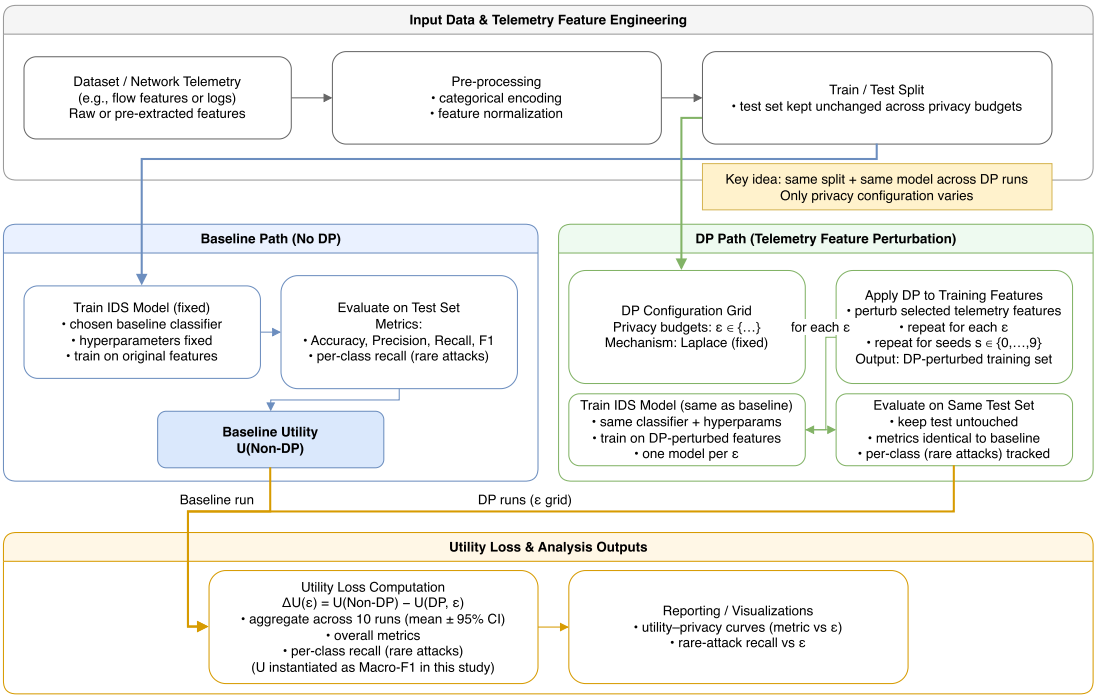


Figure 2: Experimental pipeline for assessing intrusion detection utility loss under DP applied to security telemetry features.

[1] https://github.com/sajadhomayoun/dp-ids-utility

In this paper we employ the Laplace mechanism [6] to perturb telemetry features. The Laplace mechanism provides ε-differential privacy under well-defined sensitivity assumptions and represents a standard approach for feature-level data perturbation. The focus of the analysis is on the impact of controlled noise injection rather than on comparing alternative DP mechanisms.

For a feature value x, the privatized value $\tilde{x}$ is obtained as

$$\tilde{x} = x + \mathrm{Lap}\left(0, \frac{\Delta f}{\varepsilon}\right), \tag{1}$$

where ε denotes the privacy budget and Δf represents the global sensitivity of the feature as the maximum possible change in the feature value due to the inclusion or exclusion of a single record. Since features are normalized to the range $[0,1]$ prior to perturbation, the sensitivity is bounded by $\Delta f = 1$ for each feature.

DP is applied after all preprocessing steps. The perturbation is applied independently per feature dimension, with categorical features encoded prior to normalization and treated as numerical inputs. After noise injection, feature values may fall outside the original $[0,1]$ range. To maintain consistency with the model input space, all perturbed values are clipped back to this range. This post-processing step does not violate differential privacy guarantees, but may introduce additional distortion near the boundaries.

Consequently, noise is scaled inversely with ε, so that smaller privacy budgets introduce larger perturbations.

4 Experimental Design and Results

4.1 Experimental Setup

4.1.1 Dataset Description

The experiments in this study are conducted using two intrusion detection datasets, UNSW-NB15 [1,18] and CICIDS2017 [19]. UNSW-NB15 is a widely used benchmark for evaluating intrusion detection systems. The dataset was created to address the limitations of previous intrusion detection datasets by incorporating modern attack scenarios, realistic background traffic, and a diverse set of network flow features. We used CICIDS2017 for cross-dataset validation to show generality of our results (explained in Section 4.2.4).

UNSW-NB15 consists of network traffic records represented by flow-level telemetry features extracted from packet captures. These features describe statistical and behavioral characteristics of network communications, such as packet counts, byte volumes, and timing information. Such representations closely resemble flow-level telemetry commonly used in network intrusion detection systems [1], making the primary dataset well suited for studying the effects of privacy mechanisms applied at the feature level.

The dataset includes both benign traffic and multiple attack categories, with a distribution that is notably imbalanced across classes. Several attack types occur infrequently relative to normal traffic, reflecting conditions commonly encountered in operational environments. This class imbalance is particularly important for our study, as rare and low-volume attacks are often of high security relevance and may be especially sensitive to perturbations introduced for privacy protection.

The UNSW-NB15 dataset serves as the primary dataset for evaluating the impact of DP on intrusion detection utility. Its combination of realistic telemetry features, modern attack scenarios, and imbalanced class distributions enables a meaningful analysis of how privacy-induced noise affects detection performance, both overall and for individual attack classes. Compared to alternative benchmarks, UNSW-NB15 offers a combination of modern attack scenarios, flow-level telemetry, and pronounced class imbalance, making it particularly suitable for examining privacy effects on rare-event detection.

Prior to model training, categorical features were encoded using standard encoding techniques, and numerical features were normalized. No additional feature selection was performed, and the same preprocessing configuration was applied across all privacy budgets.

Table 3 shows the class distribution of the training data used in the experiments, illustrating the pronounced class imbalance, particularly for minority attack categories such as Shellcode and Worms.

Table 3: Class distribution of the UNSW-NB15 training set used in the experiments.

Attack Category	Samples	Percentage (%)
Normal	93,000	36.09
Generic	58,871	22.84
Exploits	44,525	17.28
Fuzzers	24,246	9.41
DoS	16,353	6.35
Reconnaissance	13,987	5.43
Analysis	2,677	1.04
Backdoor	2,329	0.90
Shellcode	1,511	0.59
Worms	174	0.07
Total	**257,673**	100

4.1.2 Experimental Configuration

To ensure reproducibility and isolate the effect of differential privacy, we employ a fixed Random Forest classifier with 100 estimators, no maximum depth constraint, and parallel execution enabled. All experiments are conducted using a fixed train–test split, where the scaler is fitted on the training data and applied consistently to both training and test sets. DP is applied exclusively to the training data, while the test set remains unperturbed to ensure fair evaluation. The privacy budget is evaluated over the range $\varepsilon \in \{0.1, 0.5, 1, 2, 5, 10\}$. To account for stochastic variability, each configuration is repeated across 10 independent runs with different random seeds.

4.2 Experimental Results

This section reports the empirical impact of applying DP to security telemetry features used for intrusion detection. We keep the train–test split and the model configuration fixed across runs; only the privacy budget ε varies. For each privacy budget, experiments are repeated over 10 independent random seeds, with DP noise re-sampled in each run. Reported curves show the mean across runs, and shaded regions indicate 95% confidence intervals. The non-private model serves as the baseline for comparison. The selected privacy budget range ($\varepsilon \in \{0.1, 0.5, 1, 2, 5, 10\}$) is intended to cover a broad spectrum of privacy-utility trade-offs rather than to reflect a single deployment setting. While smaller values of ε (e.g., $\varepsilon < 0.1$) are often associated with strong privacy guarantees in high-sensitivity applications, they can lead to severe degradation in model utility, as observed in our results. Conversely, larger values such as $\varepsilon = 10$ correspond to weaker privacy but provide a useful reference point for understanding the transition between private and non-private performance regimes. Therefore, the chosen range enables a systematic exploration of how privacy strength impacts detection performance, rather than prescribing specific operational settings.

Experiments are conducted on a workstation with an Apple M2 Max processor and 32 GB RAM. The implementation is in Python and scikit-learn.

4.2.1 Overall Detection Performance

Fig. 3 summarizes the main performance trends using F1-based metrics. Macro-F1 drops sharply under strong privacy (small ε) and recovers only gradually as ε increases, but remains below the non-private baseline across all evaluated privacy budgets. At the same time, Micro-F1 remains consistently higher than Macro-F1 across privacy budgets, indicating that aggregate performance can remain relatively strong while class-balanced performance degrades substantially.

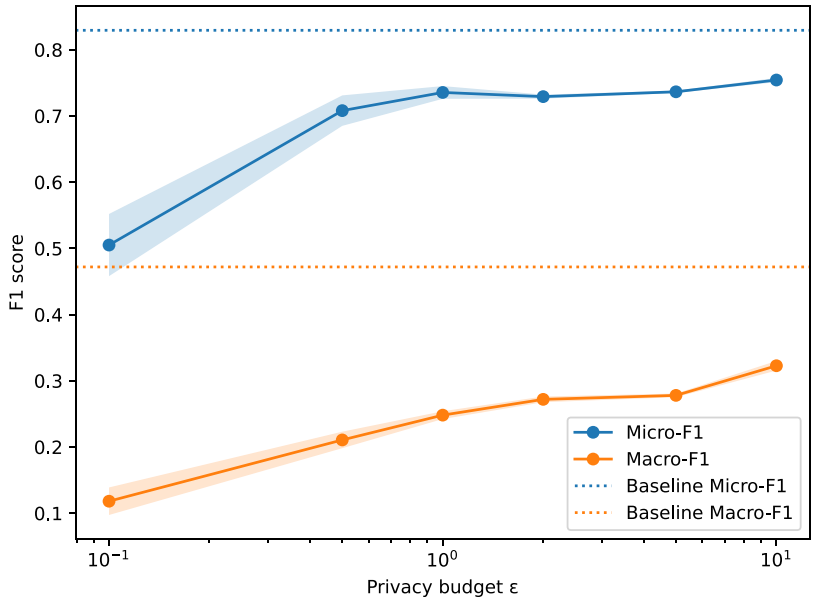
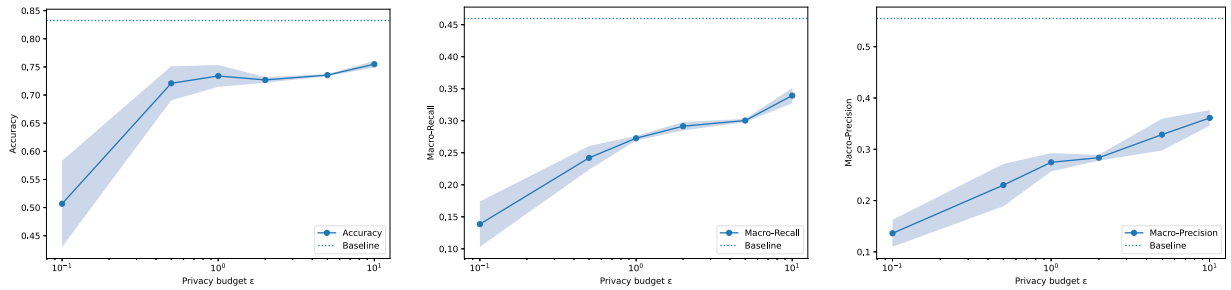


Figure 3: Macro-F1 and Micro-F1 under differential privacy across privacy budgets ε (log-scale). Dotted lines indicate the non-private baselines. Lines show mean performance across 10 runs; shaded regions indicate 95% confidence intervals.

The variability across runs is also privacy-dependent. The confidence intervals are widest at small privacy budgets ($\varepsilon \leq 0.5$), indicating increased instability when strong noise is injected into the training features. As ε increases, the confidence bands narrow, suggesting that model behavior becomes more stable as the magnitude of perturbation decreases. This pattern indicates that privacy strength affects not only mean performance but also the reliability of the learned decision boundaries.

To complement the F1 view, Fig. 4 reports accuracy, macro-recall, and macro-precision. Accuracy improves monotonically with larger privacy budgets, but it remains less sensitive to failures on minority attack classes (Fig. 4a). Macro-recall and macro-precision increase with ε as well (Fig. 4b,c), but both remain below the baseline, indicating that DP affects both missed detections and false alarms.



a Accuracy vs. ε. **b** Macro-recall vs. ε. **c** Macro-precision vs. ε.

Figure 4: Overall metrics under DP across privacy budgets ε (log-scale). (**a**) Accuracy increases with larger ε, (**b**) macro-recall improves as privacy constraints are relaxed, (**c**) macro-precision increases with higher ε. Dotted lines indicate the non-private baselines. Lines show mean performance across 10 runs; shaded regions indicate 95% confidence intervals.

4.2.2 Privacy–Utility Trade-off

In this study, utility $U(\cdot)$ is instantiated as the Macro-F1 score. Privacy-induced utility loss is computed as $\Delta U(\varepsilon) = U(\text{Non-DP}) - U(\text{DP}, \varepsilon)$, where $U(\text{Non-DP})$ denotes the baseline performance without privacy perturbation and $U(\text{DP}, \varepsilon)$ denotes the performance obtained under differential privacy with privacy budget ε.

Fig. 5 quantifies the privacy–utility trade-off using Macro-F1 loss relative to the non-private baseline. Utility loss decreases as ε increases, but remains non-negligible even at larger privacy budgets (showing that class-balanced detection performance remains costly under DP).

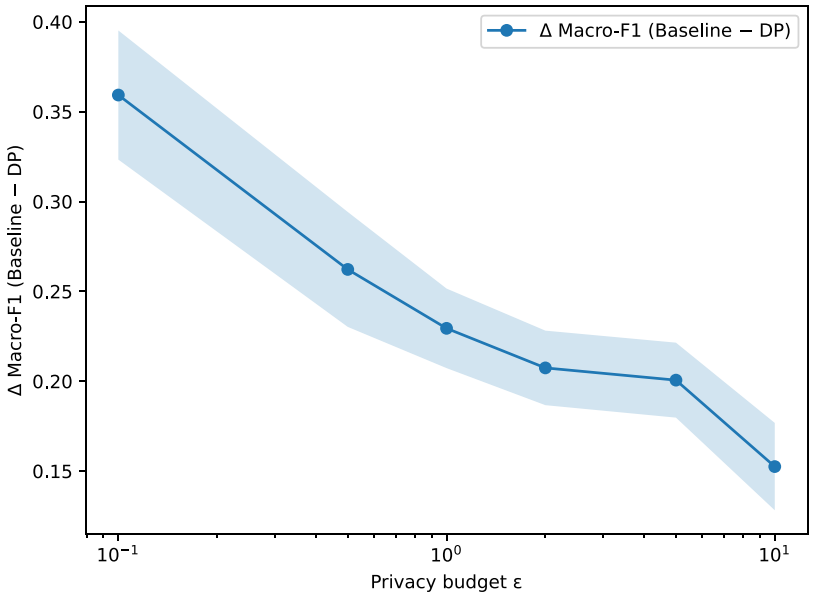


Figure 5: Utility loss measured as ΔMacro-F1 (baseline minus DP) across privacy budgets ε (log-scale). Lines show mean performance across 10 runs; shaded regions indicate 95% confidence intervals.

4.2.3 Impact on Rare and Minority Attacks

The effect of DP is most pronounced for rare attack classes. Fig. 6 shows recall for the least frequent classes in the dataset. Across the evaluated privacy budgets, recall for the least frequent attack classes collapses to zero under feature-level DP, and does not recover as the privacy budget increases.

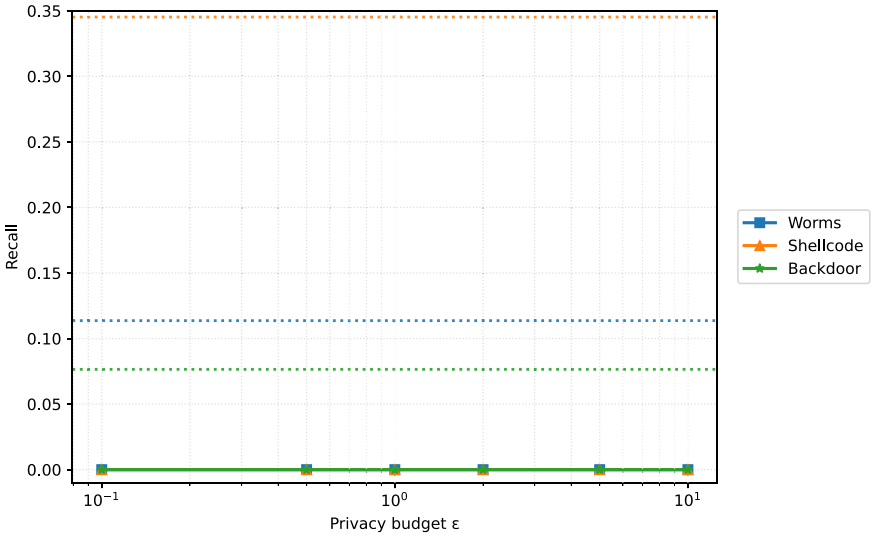


Figure 6: Recall for rare attack classes across privacy budgets ε (log-scale). Dotted horizontal lines indicate their non-private recall values (shown in matching colors).

Fig. 7 further shows class-dependent behavior across a broader set of classes. High-frequency classes partially recover as ε increases, while several less frequent or harder-to-separate classes remain sensitive to perturbation. DP introduces strongly class-dependent utility loss. High-frequency classes partially recover as ε increases, whereas several minority classes remain consistently undetected. This matters in practice because SOC value is often tied to performance on the minority and high-impact events.

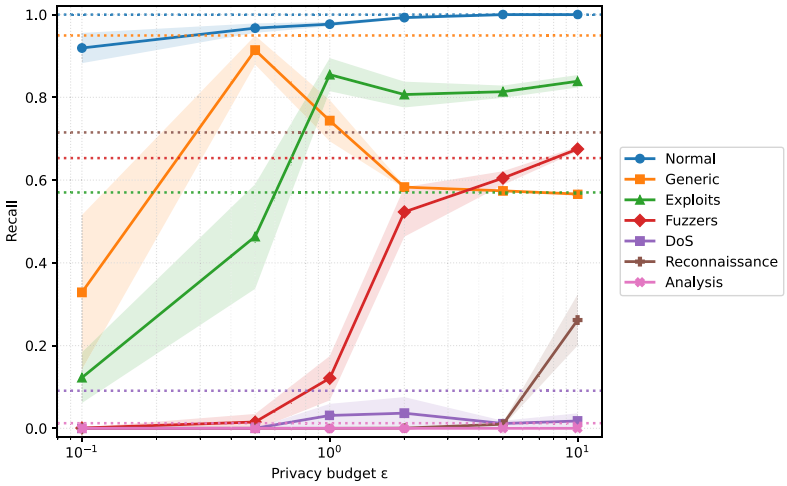


Figure 7: Per-class recall across privacy budgets ε (log-scale). Solid lines represent recall under differential privacy, while dotted horizontal lines indicate the corresponding non-private baseline recall for each class shown in matching colors. Lines show mean performance across 10 runs; shaded regions indicate 95% confidence intervals.

4.2.4 Cross-Dataset Validation on CICIDS2017

To evaluate the generalizability of our findings, we conduct additional experiments on the CICIDS2017 dataset [19], which differs from UNSW-NB15 in terms of traffic characteristics, attack diversity, and class distribution. We apply the same data pipeline, model configuration, and differential privacy mechanisms to ensure consistency across datasets. As part of the preprocessing pipeline, we first aggregate fine-grained attack labels into broader categories by grouping semantically related attack types (e.g., multiple Denial of Service (DoS) variants under *DoS*, Web-based attacks under *WebAttack*, and File Transfer Protocol (FTP)/Secure Shell (SSH) attacks under *BruteForce*). This aggregation reduces label fragmentation and ensures sufficient class support for minority classes. Subsequently, we apply light undersampling to the majority class (Normal traffic) while preserving all minority attack classes, in order to mitigate class imbalance and maintain consistency with the experimental setup used for UNSW-NB15. The resulting dataset exhibits the following class distribution: Normal (300,000, 35.0%), DoS (251,712, 29.4%), PortScan (158,804, 18.5%), Distributed Denial-of-Service (DDoS) (128,025, 14.9%), BruteForce (13,832, 1.6%), WebAttack (2180, 0.25%), and Bot (1956, 0.23%).

Fig. 8 shows the impact of differential privacy on Micro-F1 and Macro-F1. Consistent with the results on UNSW-NB15, Micro-F1 remains relatively robust across privacy budgets, while Macro-F1 degrades substantially at low values of ε. This divergence indicates that aggregate performance metrics can obscure significant degradation in minority class detection under strong privacy constraints.

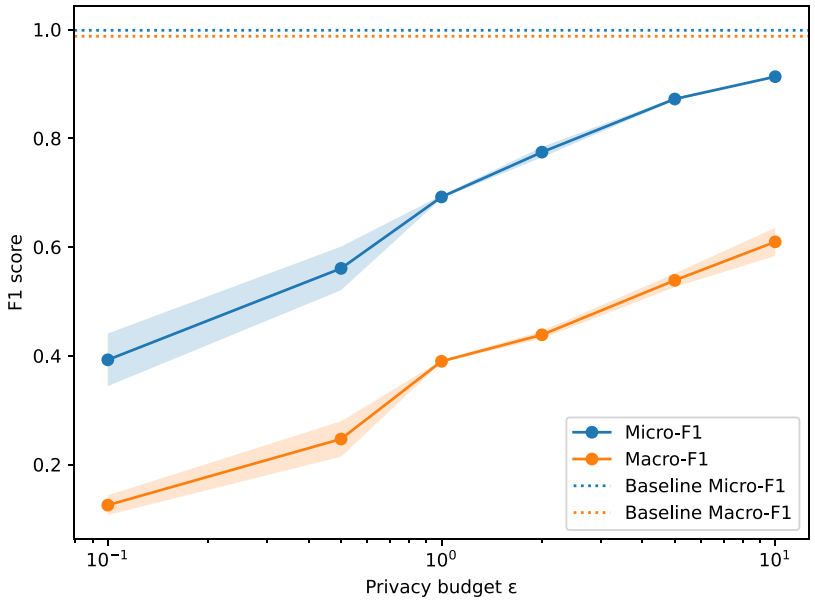


Figure 8: Impact of differential privacy on Micro-F1 and Macro-F1 for CICIDS2017. While Micro-F1 remains relatively stable, Macro-F1 degrades significantly at low privacy budgets, indicating reduced performance on minority classes.

Fig. 9 shows class-dependent recall across representative attack categories in CICIDS2017. While majority classes such as Normal, DoS, and PortScan maintain relatively high recall even under stronger privacy constraints, minority classes including Bot, WebAttack, and BruteForce exhibit significantly degraded performance, in some cases approaching near-zero recall at low privacy budgets. This contrast highlights the non-uniform impact of differential privacy across classes with different support levels.

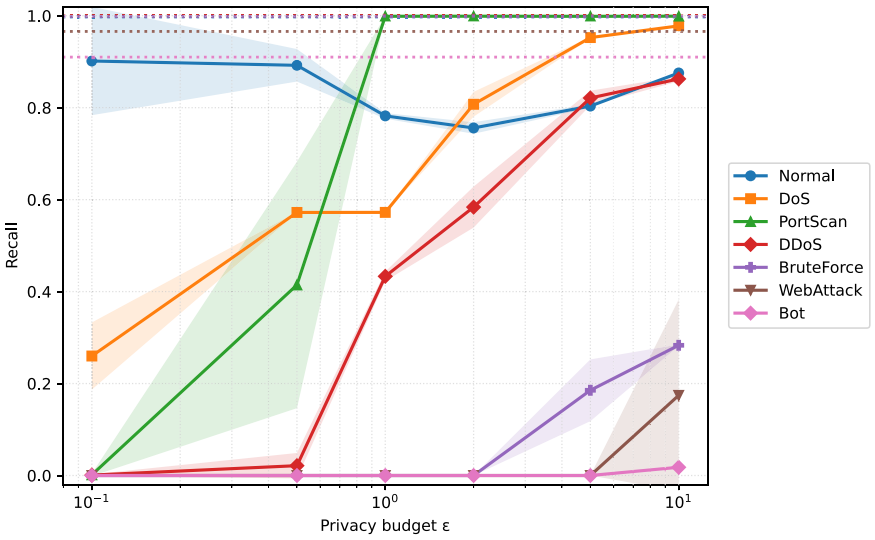


Figure 9: Class-wise recall vs. privacy budget ε for CICIDS2017, illustrating the stronger impact of differential privacy on minority classes compared to majority classes.

This behavior mirrors the collapse of rare attack detection observed in UNSW-NB15, indicating that the vulnerability of minority classes under differential privacy is consistent across datasets, despite differences in traffic characteristics and attack composition.

5 Discussion

The results provide a consistent picture of how Differential Privacy affects intrusion detection when applied directly to security telemetry features. Increasing the privacy budget improves performance across all evaluated metrics, but the recovery is uneven and strongly dependent on class distribution. In particular, class-balanced metrics and rare attack detection remain substantially degraded even at relatively large values of ε.

A clear divergence emerges between accuracy and Macro-F1 under privacy constraints. Accuracy remains comparatively high across privacy budgets because it is dominated by majority classes, especially normal traffic. In contrast, Macro-F1, macro-recall, and macro-precision reveal severe degradation at low and moderate privacy budgets. This indicates that DP-induced noise disproportionately affects minority classes, which contribute equally to macro-averaged metrics but have limited representation in the data. Accuracy alone therefore does not provide a reliable picture of intrusion detection utility in privacy-preserving settings.

The impact of DP is most pronounced for rare and low-volume attacks. Across 10 independent runs, recall for the least frequent classes remains at zero under feature-level DP across all evaluated privacy budgets, indicating a robust and consistent failure rather than a single-run artifact. Even when overall performance improves with increasing ε, the detection signal for these attacks does not recover. In this setting, the injected noise appears sufficient to overwhelm subtle statistical patterns associated with rare attacks, effectively suppressing their separability. From a security perspective, this behavior is notable, as rare attacks may correspond to stealthy or high-impact activity.

The observed collapse in recall for rare attack classes can be attributed to multiple interacting factors. First, differential privacy noise may overwhelm subtle statistical deviations that characterize low-frequency attacks, making them indistinguishable from background traffic. Second, the perturbation can disrupt correlations between features that are critical for separating minority classes from normal behavior. Since

rare attacks are represented by a limited number of samples, their decision boundaries are inherently fragile and highly sensitive to noise. As a result, the model may consistently misclassify these instances, leading to near-zero recall. We further note that this behavior is consistent with the observed degradation in macro-level metrics, including macro-precision and macro-F1, which reflect performance across all classes. This suggests that the degradation is not limited to missed detections, but reflects a broader difficulty in distinguishing minority classes under strong privacy constraints.

The class-dependent behavior observed across privacy budgets further shows that DP does not introduce uniform degradation. This uneven recovery indicates that privacy noise interacts with class imbalance and feature distributions in non-trivial ways. Selecting a privacy budget based solely on aggregate performance metrics may therefore introduce unintended blind spots in detection coverage.

Taken together, these findings characterize the privacy–utility trade-off associated with feature-level Differential Privacy in intrusion detection. While DP provides formal privacy guarantees, its operational impact depends on the distributional properties of the data and the detection objectives of the system. In environments where rare or low-volume attacks are operationally critical, feature-level DP may require additional mitigation strategies or alternative integration approaches. More broadly, the results highlight the importance of evaluation practices that go beyond accuracy and explicitly account for class imbalance and rare-event detection when assessing privacy-preserving IDS.

Limitations and Scope

The analysis is conducted on two publicly available intrusion detection datasets (UNSW-NB15 and CICIDS2017). While this extends the evaluation beyond a single benchmark, the observed effects of differential privacy, particularly the degradation in rare attack detection, may still be influenced by dataset-specific feature distributions, traffic characteristics, and class imbalance patterns. As such, the findings may not fully generalize to other datasets or real-world environments with different properties.

Further validation across a broader range of datasets, including IoT- and OT-oriented benchmarks, remains part of future work to assess the robustness and general applicability of the observed privacy–utility trade-offs.

DP is applied directly to preprocessed telemetry features. This design isolates the effect of feature-level perturbation, but it does not address alternative integration strategies such as privacy applied during model training, at data collection time, or within federated learning frameworks. Different integration points may lead to different privacy–utility trade-offs.

The evaluation relies on a fixed classifier and hyperparameter configuration across privacy budgets. This controlled setup isolates the impact of DP but does not explore privacy-aware model tuning or architectural adaptations that could partially compensate for utility loss. The reported degradation reflects behavior under fixed-model assumptions.

While multiple metrics are reported, including macro-averaged scores and per-class recall, the analysis does not incorporate cost-sensitive evaluation or downstream alert-handling considerations. Operational impact may vary depending on how missed detections and false positives are prioritized in practice.

Within these boundaries, the experimental design provides a controlled setting for examining how feature-level Differential Privacy interacts with class imbalance and rare attack detection in intrusion detection systems.

6 Conclusion

This paper presented an empirical study of the impact of DP on intrusion detection when applied directly to security telemetry features. Rather than proposing a new detection model, the focus was on quantifying how privacy-induced perturbation affects detection utility across different privacy budgets.

The results show that while aggregate performance metrics such as accuracy and Micro-F1 may remain relatively stable under DP, class-balanced metrics and per-class recall reveal substantial degradation. In particular, the detection of rare and low-volume attacks is severely affected, with some attack classes becoming entirely undetectable even at relatively large privacy budgets. These findings highlight that utility loss under DP is not uniform and is strongly influenced by class imbalance and feature sensitivity.

From a practical perspective, the study suggests that the application of DP in IDS requires careful evaluation beyond aggregate metrics. In settings where the detection of rare or subtle attacks is critical, feature-level DP may introduce blind spots that are not immediately visible through standard performance measures.

We further validated these findings on the CICIDS2017 dataset, observing consistent behavior and confirming that differential privacy disproportionately impacts minority attack classes across datasets.

This work contributes empirical evidence on the privacy–utility trade-offs associated with DP in intrusion detection and underscores the importance of evaluation practices that explicitly account for class imbalance and rare-event detection. Future work may extend this analysis to additional datasets, alternative privacy integration strategies, and operational settings to further refine the understanding of privacy-preserving intrusion detection.

Acknowledgement: Not applicable.

Funding Statement: The author received no specific funding for this study.

Availability of Data and Materials: The data used in this study are publicly available. The UNSW-NB15 dataset can be accessed at https://research.unsw.edu.au/projects/unsw-nb15-dataset, and the CICIDS2017 dataset is available at https://www.unb.ca/cic/datasets/ids-2017.html. All reported results were obtained through analysis of this dataset as described in the paper.

Ethics Approval: Not applicable.

Conflicts of Interest: The author declares no conflicts of interest.

References

1. Moustafa N, Slay J. 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). Piscataway, NJ, USA: IEEE; 2015. p. 1–16. doi:10.1109/MilCIS.2015.7348942.
2. Ruzafa-Alczar P, Fernndez-Saura P, Mrmol-Campos E, Gonzlez-Vidal A, Hernndez-Ramos JL, Bernal-Bernabe J, et al. Intrusion detection based on privacy-preserving federated learning for the industrial IoT. IEEE Trans Ind Inform. 2023;19(2):1145. doi:10.1109/TII.2021.3126728.
3. Gutti C, Thumula K, Balbudhe P. Federated learning for distributed IoT security: a privacy-preserving approach to intrusion detection. IEEE Access. 2025;13:135863–75.
4. Pei X, Deng X, Tian S, Jiang P, Zhao Y, Xue K. A privacy-preserving graph neural network for network intrusion detection. IEEE Trans Dependable Secure Comput. 2025;22(1):740–56. doi:10.1109/tdsc.2024.3417853.
5. Franke P, Kreutzer M, Simo H. Privacy-preserving IDS for in-vehicle networks with local differential privacy. In: IFIP advances in information and communication technology. Vol. 619. Berlin/Heidelberg, Germany: Springer; 2021. p. 58–77.

6. Dwork C, McSherry F, Nissim K, Smith A. Calibrating noise to sensitivity in private data analysis. In: Halevi S, Rabin T, editors. Theory of cryptography. Berlin/Heidelberg, Germany: Springer; 2006. p. 265–84.
7. Cui J, Xiao J, Zhong H, Zhang J, Wei L, Bolodurina I, et al. LH-IDS: lightweight hybrid intrusion detection system based on differential privacy in VANETs. IEEE Trans Mobile Comput. 2024;23(12):12195–210.
8. Siachos I, Kaltakis K, Papachristopoulou K, Giannoulakis I, Kafetzakis E. Comparison of machine learning algorithms trained under differential privacy for intrusion detection systems. In: Proceedings of the 2023 IEEE International Conference on Cyber Security and Resilience, CSR 2023. Piscataway, NJ, USA: IEEE; 2023. p. 654–8.
9. Machooka D, Yuan X, Roy K, Chen G. Differential privacy with DP-SGD and PATE for intrusion detection: a comparative study. In: 2025 IEEE 4th International Conference on AI in Cybersecurity, ICAIC 2025. Piscataway, NJ, USA: IEEE; 2025. p. 1–7.
10. Amin MARA, Shetty S, Formicola V, Otto M. Assessing the quality of differentially private synthetic data for intrusion detection. In: Security and Privacy in Communication Networks (SecureComm 2022). Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering, LNICST. Vol. 462. Berlin/Heidelberg, Germany: Springer; 2023. p. 473–90.
11. Anastasakis Z, Psychogyios K, Velivassaki T, Bourou S, Voulkidis A, Skias D, et al. Enhancing cyber security in IoT systems using FL-based IDS with differential privacy. In: 2022 Global Information Infrastructure and Networking Symposium, GIIS 2022. Piscataway, NJ, USA: IEEE; 2022. p. 30–4.
12. Mosaiyebzadeh F, Pouriyeh S, Han M, Liu L, Xie Y, Zhao L, et al. Privacy-preserving federated learning-based intrusion detection system for IoHT devices. Electronics. 2025;14(1):67. doi:10.1109/infocomwkshps57453.2023.10225932.
13. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. 2021;372:n71.
14. Anyanwu GO, Karimipour H. CLDP=FATD: secure federated averaging threat detection framework for intelligent vehicle sensor networks based on client-level differential privacy. IEEE Internet Things J. 2025;12(7):7693–707.
15. Soumya TR, Thamba MW, Jayasree P, Poonkuzhali P, Sathya V, Lathif SIA. PrivGuardNet: a federated transformer-based intrusion detection system with differential privacy for scalable cybersecurity defense. In: Proceedings of 8th International Conference on Computing Methodologies and Communication, ICCMC 2025. Piscataway, NJ, USA: IEEE; 2025. p. 1920–7.
16. Wen Q, Li Z. DDoS attack detection based on differential privacy graph data protection in IoV. In: ADMIT 2024—Conference Proceedings: 2024 3rd International Conference on Algorithms, Data Mining, and Information Technology. New York, NY, USA: ACM; 2025. p. 281–7.
17. Mahalakshmi G, Sumathi G, Prakash M, Solainayagi P. Enhancing VANET security through ANFIS-based intrusion detection and element perturbation. Int J Commun Syst. 2026;39(1):e70305. doi:10.1002/dac.70305.
18. Sarhan M, Layeghy S, Moustafa N, Portmann M. NetFlow datasets for machine learning-based network intrusion detection systems. In: Deze Z, Huang H, Hou R, Rho S, Chilamkurti N, editors. Big Data Technologies and Applications (BDTA 2020, WiCON 2020). Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Vol. 371. Berlin/Heidelberg, Germany: Springer; 2021. p. 117–35. doi:10.1007/978-3-030-72802-1_9.
19. Sharafaldin I, Habibi Lashkari A, Ghorbani AA. Toward generating a new intrusion detection dataset and intrusion traffic characterization. In: Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP). Setbal, Portugal: SciTePress; 2018. p. 108–16.