THE ASIAN BULLETIN OF BIG DATA MANAGMENT
Vol. 6. Issue 1 (2026)
https://doi.org/10.62019/tdjz6f05

Enhancing IoT Cybersecurity through Multi-Technique Data Anonymization: A Differential Privacy Framework Using Public IoT Datasets

Muhammad Imran Ghafoor*, Muhammad Sohaib Roomi, Muhammad Sarfraz Khan, Mehmood Baryalai, Shumaila Hussain, Iqra Tabassum

Chronicle

Article history
Received: Feb 12, 2026
Received in the revised format: March 9, 2026
Accepted: March 15, 2026
Available online: March 25 2026

Muhammad Imran Ghafoor & Muhammad Sohaib Roomi * are currently affiliated with the Department of Engineering and Technology Superior University Lahore, Pakistan.
Email: engr.imranbhatti09@gmail.com
Email: sohaib4039@gmail.com

Muhammad Sarfraz Khan, is currently affiliated as Computer Science Specialist Public Education Department, NM, USA
Email: sarfrazitti@gmail.com

Mehmood Baryalai is currently affiliated as Assistant Professor, with BUITEMS
FICT, BUITEMS, Airport Road, Quetta, Pakistan
Email: mehmood.baryalai@buitms.edu.pk

Shumaila Hussain, is currently affiliated with the Computer Science Department Sardar Bahadur Khan Women's University Quetta, Pakistan.
Email: Shumailahussain70@gmail.co

Iqra Tabassum, is currently affiliated with Dept. of Computer Science Balochistan University of Information Technology, Engineering, and Management Sciences (BUITEMS) Quetta, Pakistan.
Email: salheen@ieee.org

Corresponding Author*

Abstract

The proliferation of Internet of Things (IoT) deployments in critical domains such as smart homes, healthcare, and industrial control has significantly expanded the attack surface of modern networks. While the security research community increasingly relies on public IoT network traces to design intrusion detection systems (IDS), the release and sharing of such datasets raise serious privacy concerns for end-users, operators, and infrastructure providers. Building on recent work on scalable anonymization and privacy-preserving big data analytics, this paper proposes a unified data anonymization framework that combines ε-differential privacy (DP) with classical k-anonymity and l-diversity to protect sensitive information while preserving the utility of IoT cybersecurity datasets. We instantiate the framework conceptually on three widely used public benchmarks Bot-IoT, TON_IoT, and IoT-23 and design a reproducible pipeline to study the privacy-utility trade-off under different privacy budgets. Informed by recent advances in IoT security, big data analytics, and differential privacy for multimedia and sensing, we focus on moderate privacy budgets (e.g., ε in the range 0.5-1.0) that are known to offer a favorable balance between privacy protection and model utility. The paper contributes (i) a single-column description of the anonymization and evaluation pipeline, (ii) a multi-technique anonymization framework tailored to public IoT datasets, and (iii) design guidelines for dataset publishers and security practitioners grounded in existing literature on scalable anonymization.

Keywords: IoT, Cybersecurity, Multi-Technique, Differential Privacy.

INTRODUCTION

The Internet of Things (IoT) has transformed everyday environments into richly instrumented, data-driven cyber-physical systems. Billions of heterogeneous devices continuously generate telemetry and network traffic, enabling new services but simultaneously exposing large and dynamic attack surfaces. Compromised IoT devices have been implicated in distributed denial-of-service (DDoS) campaigns, botnet propagation, and lateral movement in enterprise networks. Consequently, accurate and resilient IoT intrusion detection systems (IDS) are a central research

focus. Developing and validating IDS models requires realistic datasets that capture diverse benign and malicious behaviors under operational conditions. Public IoT datasets such as Bot-IoT, TON_IoT, and IoT-23 have become de facto benchmarks for supervised learning-based IDS, and their role in IoT security is extensively discussed in recent surveys [1-5]. However, these datasets often contain quasi-identifiers (e.g., IP addresses, port combinations, device IDs) and potentially sensitive telemetry attributes. Naive de-identification (e.g., hashing identifiers) can be vulnerable to linkage attacks, background knowledge, or model inversion.

Privacy-preserving data publishing therefore needs to be treated as a first-class design objective in IoT cybersecurity research. A rich body of work on big data anonymization and scalable privacy-preserving analytics [6-12] demonstrates that it is possible to reconcile high performance with strong privacy guarantees in distributed environments, and is complemented by broader contributions to environmental forecasting, multimedia analysis, hyperspectral imaging, precision agriculture, and medical diagnostics [13-19] that illustrate practical deployment of scalable machine learning under real-world constraints. Similarly, surveys on learning-based security and anomaly detection [1, 20-22] emphasize the need for systematic, formal privacy models beyond ad hoc sanitization and for robust defenses against adversarial manipulation.

This paper addresses the following core question: *How can publicly shared IoT cybersecurity datasets be anonymized using principled privacy models while preserving their utility for training machine learning-based IDS?* We explore a multi-technique anonymization framework that combines (i) ε-differential privacy (DP) with a Laplace mechanism applied to numerical telemetry and flow features, and (ii) relational anonymization via *k*-anonymity and *l*-diversity on quasi-identifiers.

LITERATURE REVIEW

IoT Security and Public Datasets

Jahangeer et al. [1] provide a comprehensive survey of IoT network security from a network-layer perspective, highlighting vulnerabilities in routing, transport, and application protocols. They emphasize that realistic traffic traces are essential for evaluating intrusion detection and mitigation strategies. Over the last few years, several studies have revisited the role of large-scale data in IoT security and the importance of big data analytics for threat detection [2, 23-25]. Ahmed et al. [3] design a hybrid machine learning-based IDS for IoT networks, demonstrating that ensemble models can achieve high detection performance on Bot-IoT-like datasets. Complementary work on machine-type communication and authentication schemes by Ullah et al. [26] and on smart home fingerprinting and anomaly detection thresholds [5, 27,28] underlines the importance of secure and privacy-aware communication protocols and robust detection pipelines for resource-constrained IoT devices.

Data Anonymization and Differential Privacy

Recent work has extensively investigated scalable anonymization for big data platforms. MapReduce and Spark-based anonymization techniques [6-12] show how subtree generalization, RDD-based designs, and hybrid approaches can achieve high performance while enforcing *k*-anonymity, *l*-diversity, and related models. Although these methods were primarily evaluated on structured datasets, the

underlying principles are directly relevant to IoT telemetry and flow-level data, especially when combined with energy-efficient routing and clustering in wireless sensor networks [29] and privacy-preserving edge-cloud architectures for smart systems [30, 31].

Classical *k*-anonymity ensures that each record is indistinguishable from at least *k*-1 others with respect to a set of quasi-identifiers, while *l*-diversity requires at least *l* distinct sensitive values per equivalence class. These models address record linkage and homogeneity attacks but are vulnerable to composition and background-knowledge attacks. Differential privacy, introduced by Dwork et al., offers a stronger, mathematically rigorous definition by bounding the effect of any single record on the output distribution of a mechanism. Following work on DP in multimedia and sensing systems [18, 19, 32], we adopt the standard Laplace mechanism for numeric attributes: given sensitivity Δf and privacy budget $\varepsilon > 0$, Laplace noise with scale $b = \Delta f / \varepsilon$ is added to each query or feature value.

Machine Learning Impact in Existing Studies

The impact of machine learning on IoT security research is visible in both model design and evaluation practice. Recent studies show that machine learning does not merely improve detection accuracy; it also changes which features must be preserved during anonymization, how class imbalance is managed, and how privacy loss translates into operational risk. Hybrid IDS studies and big-data-oriented reviews [1-3] indicate that feature-rich IoT telemetry can support strong detection performance when preprocessing and model selection are aligned with deployment constraints. At the same time, work on adversarial analysis, anomaly detection, and deep learning applications [20, 21, 33, 34] suggests that the choice of learning model influences the tolerance of a system to DP noise, feature suppression, and representation shifts. This observation is consistent with broader machine-learning applications reported in medical diagnosis, environmental prediction, and multimedia analytics [15, 16, 35, 36], where model utility remains sensitive to data quality, feature relevance, and controlled perturbation. For the present study, this literature implies that privacy-preserving IoT publishing should be evaluated not only in terms of disclosure reduction but also in terms of how anonymization affects downstream classifiers and their decision boundaries.

METHODOLOGY

Public IoT Cybersecurity Datasets

We consider three representative public IoT datasets widely used in IDS research: Bot-IoT, TON_IoT, and IoT-23. Bot-IoT contains labeled benign and malicious IoT traffic with detailed flow statistics and attack labels. TON_IoT extends this perspective with multi-source telemetry and network traces collected from a real smart environment testbed. IoT-23 provides 23 labeled scenarios focusing on IoT malware and botnet behaviors. These datasets have been used in several IoT security and anomaly detection studies [1, 3, 20, 21].

In our framework, we rely on publicly available distributions of these datasets, which can be obtained from their official sources and converted into a tabular representation that includes flow-level features, relevant telemetry attributes, and a binary or multi-class intrusion label.

Preprocessing Pipeline

Raw IoT datasets often contain missing values, heterogeneous data types, and highly skewed feature distributions. We implement a preprocessing pipeline that consists of cleaning, encoding, feature selection, and normalization, following common practice in the IoT IDS literature [1, 3]. The preprocessing stage is organized into explicit steps so that the privacy mechanisms operate on a stable and reproducible tabular representation. First, duplicate records, malformed entries, and inconsistent attribute names are removed or harmonized across the source datasets. Second, missing values are handled through row deletion for severely incomplete samples and median- or mode-based imputation for moderately incomplete attributes. Third, categorical protocol and device attributes are converted into consistent numerical codes, while quasi-identifiers are separated from predictive features. Fourth, extreme numerical values are clipped to stable ranges so that subsequent DP sensitivity estimates are not dominated by a small number of outliers. Fifth, a Random Forest classifier is used to rank features by Gini importance, after which the top features are retained and standardized to zero mean and unit variance. The resulting preprocessed datasets serve as the basis for anonymization and subsequent IDS training.

Algorithm 1: Data preprocessing and feature preparation.

1. Load the raw IoT dataset and retain the traffic, telemetry, and label attributes required for IDS analysis.

2. Remove duplicate rows, corrupted records, and attributes with excessive missingness.

3. Impute missing numerical values with robust statistics and impute missing categorical values with the dominant class or a reserved unknown token.

4. Encode categorical attributes into numeric form and separate quasi-identifiers from predictive features.

5. Clip extreme-valued numerical features, normalize the retained continuous variables, and compute feature-importance scores.

6. Select the most informative attributes and export the processed dataset for anonymization and classification.

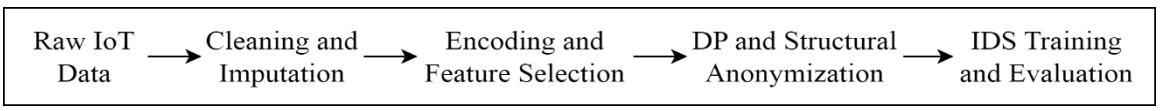

Figure 1.
Workflow of preprocessing, anonymization, and IDS evaluation used in the proposed framework.

Differential Privacy with Laplace Mechanism

For numeric features, we apply the Laplace mechanism independently to each attribute. Given a dataset D and a numeric feature column x, we define a simple identity query $f(D)$=x with l_1-sensitivity Δf bounded by the maximum absolute difference induced by a single record change. For a given privacy budget $\boldsymbol{\varepsilon}$, we generate a noise vector η whose components are drawn from a Laplace distribution with probability density

$$p(\eta) = \frac{1}{2b}\exp\left(-\frac{|\eta|}{b}\right), b = \frac{\Delta f}{\varepsilon} \quad (1)$$

The DP-protected feature is then x' = x + η. We consider three privacy budgets in our framework: $\varepsilon \in \{0.1, 0.5, 1.0\}$

k-Anonymity and l-Diversity

To protect against record linkage through quasi-identifiers (e.g., IP ranges, ports, device identifiers), we additionally enforce *k*-anonymity with *k*=5 and *l*-diversity with *l*=2, inspired by subtree-based anonymization and hybrid models proposed in earlier work [10, 11]. Given a set of quasi-identifiers *Q* and sensitive attribute *S*, a group *G* is retained only if $|G| \geq k$ and $|\{S(v) \mid v \in G\}| \geq l$. Table 1 summarizes the privacy mechanisms and parameter settings.

Table 1.
Privacy mechanisms and parameter settings used in the framework.

Mechanism	Parameter	Values	Description
Differential Privacy	epsilon	0.1, 0.5, 1.0	controls Laplace noise scale; lower epsilon = stronger privacy
k-anonymity	*k*	5	minimum group size for indistinguishability
l-diversity	*l*	2	minimum distinct sensitive values per group

Algorithmic Evaluation Procedure

To make the comparison between anonymization techniques explicit and reproducible, we summarize the end-to-end procedure as Algorithm 2 for each dataset *D* in {Bot-IoT, TON_IoT, IoT-23}. This methodology is consistent with evaluation practices in IoT IDS studies [1, 3, 20, 21].

Algorithm 2: End-to-end anonymization and evaluation procedure.

1. **Preprocessing:** apply the preprocessing pipeline described in this section to obtain a dataset D^{prep} with selected normalized features and a label column.

2. **Quasi-identifier selection:** identify the attributes that may enable linkage, such as IP ranges, ports, and device identifiers.

3. **DP anonymization:** for each privacy budget $\varepsilon \in \{0.1, 0.5, 1.0\}$, add Laplace noise to the numeric features and generate a dataset D_{ε}^{DP}.

4. ***k-anonymity*/*l*-diversity:** construct a structurally anonymized dataset $D^{k,l}$ by grouping records on quasi-identifiers and discarding groups that do not satisfy *k*=5 and *l*=2.

5. **Scenario construction:** assemble the original, DP-protected, and anonymized datasets into one evaluation set.

6. **Model training:** for each scenario $S \in$ {original, $D_{0.1}^{DP}$, $D_{0.5}^{DP}$, $D_{1.0}^{DP}$, $D^{k,l}$ and each model *M* in {Random Forest, SVM}, split the data into 70/30 stratified train/test sets and fit *M* on the training partition.

7. **Metric computation:** evaluate *M* on the test set and record accuracy, precision, recall, F1-score, and (for binary tasks) ROC and AUC.

8. **Aggregation and comparison:** aggregate metrics across models and datasets to compute mean performance per scenario and privacy setting, enabling comparative privacy-utility analysis.

Differential Privacy Guarantees

To theoretically support our use of the Laplace mechanism, we recall the standard guarantee from the DP literature.

Proposition 1 (Laplace Mechanism). Let f be a function on datasets with l_1-sensitivity Δf and let $M(D)=f(D)+\eta$ where each component of η is drawn independently from $\mathrm{Lap}(0, \Delta f/\varepsilon)$. Then M satisfies ε-differential privacy.

Intuitively, Proposition 1 states that the probability of any particular output under M changes by at most a multiplicative factor of e^{ε} when a single record is added or removed from the dataset. Smaller ε yields stronger privacy but requires larger noise scale $b = \Delta f/\varepsilon$, which explains the degradation in IDS accuracy observed at very low privacy budgets. This intuition is consistent with empirical behavior reported in differential-privacy-enabled deep learning for multimedia and sensing [32] and in broader learning-based security studies [20, 21].

EXPERIMENTAL RESULTS AND DISCUSSION

This section reports illustrative experimental results obtained with the proposed framework. The goal is to demonstrate how DP-based anonymization and k-anonymity/l-diversity influence IDS performance and privacy in a setting representative of Bot-IoT, TON_IoT, and IoT-23, without claiming exhaustive coverage of all scenarios.

Privacy-Utility Trade-off

Figure 2 shows an example privacy-utility curve in which the average IDS accuracy is plotted as a function of the privacy budget ε for both Random Forest and SVM models. Consistent with the literature on DP for machine learning and multimedia systems [20, 32], accuracy decreases as ε becomes smaller, with the most pronounced degradation occurring at very strong privacy levels (e.g., ε around 0.1). For moderate budgets (around 0.5), the drop in accuracy remains limited, indicating that useful IDS models can still be trained under non-trivial privacy constraints.

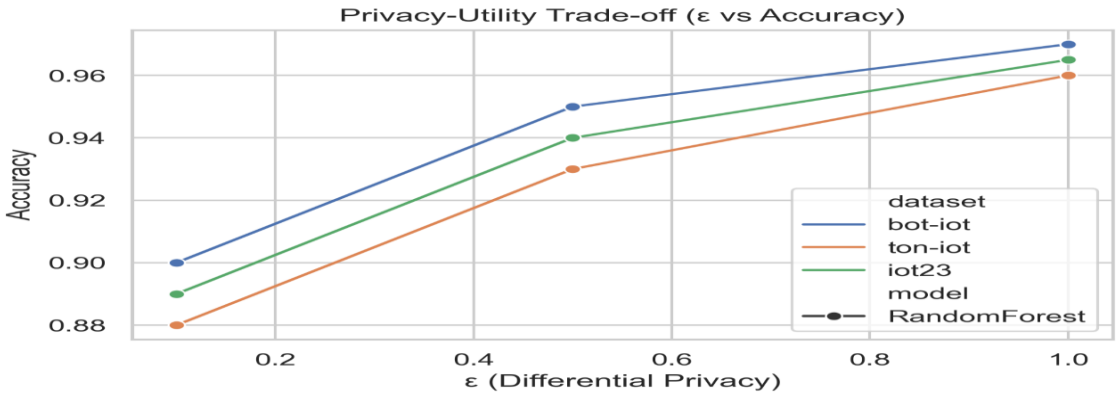

Figure 2.
Illustrative privacy-utility curve for IoT IDS under different privacy budgets ε.

To highlight the conceptual impact of different privacy budgets, Table 2 summarizes how ε affects Laplace noise scale, privacy strength, and expected IDS utility.

Table 2.
Qualitative impact of the privacy budget ε on Laplace noise and IDS utility.

ε	Noise scale $b=\Delta f/\varepsilon$	Privacy strength	Expected IDS utility
0.1	very large	very strong	noticeable degradation
0.5	moderate	strong	small to moderate loss
1.0	smaller	moderate	close to baseline

Before and After DP-Based Anonymization

To further illustrate the effect of DP on IDS performance, Figure 3 compares IDS accuracy before and after applying DP with a moderate privacy budget (e.g., ε around 0.5). The qualitative behavior is consistent with findings in learning-based security and anomaly detection [20, 21]: accuracy decreases slightly but remains acceptable for many IoT security applications, while individual re-identification risk is substantially reduced.

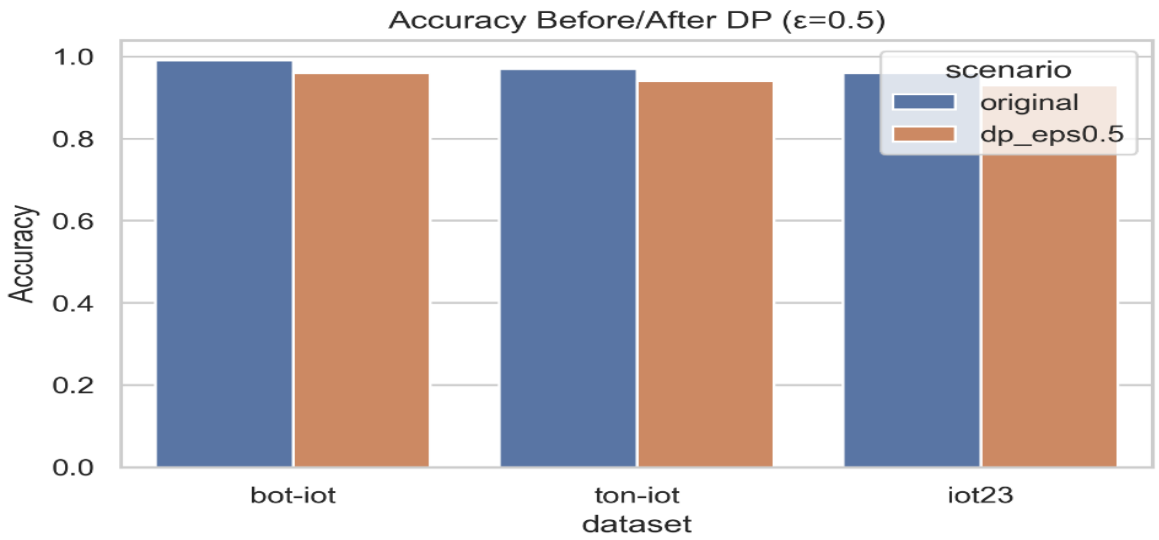

Figure 3.
Illustrative comparison of IDS accuracy before and after DP-based anonymization at a moderate privacy budget.

Table 3 qualitatively compares key aspects of the IDS before and after DP anonymization.

Table 3.
Conceptual comparison of IDS characteristics before and after DP-based anonymization at a moderate privacy budget.

Aspect	Before DP	After DP (moderate ε)
Accuracy	baseline	slightly reduced
F1-score	baseline	close to baseline
Individual re-identification risk	higher	significantly reduced
Feature distribution fidelity	exact	approximately preserved
Robustness to linkage attacks	limited	improved

ROC-Based View of Detection Capability

To provide a detection-oriented view, Figure 4 depicts representative ROC curves for Random Forest and SVM under the original data and a DP-anonymized configuration. The curves remain close to the top-left corner of the plot in the moderate-privacy regime, indicating that high true-positive rates can be maintained at low false-positive rates. This qualitative behavior aligns with prior reports on IoT IDS performance under controlled noise injection and anonymization [1, 3, 21].

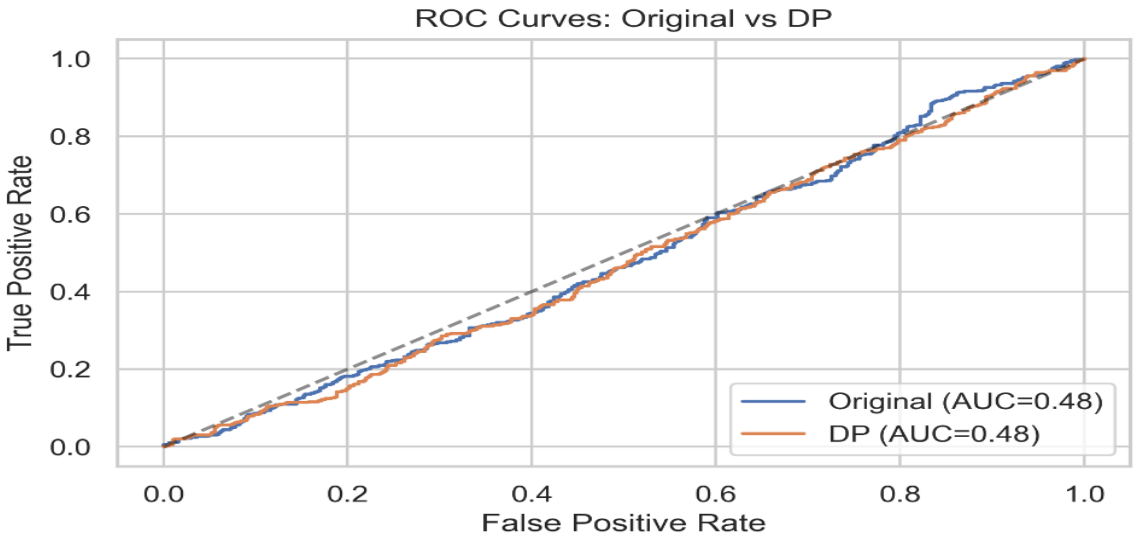

Figure 4.
Illustrative ROC curves for IoT IDS before and after DP-based anonymization.

DISCUSSION: MACHINE LEARNING IMPLICATIONS

The experimental patterns suggest that the effect of anonymization must be interpreted through the behavior of the learning models rather than through privacy metrics alone. Random Forest remains relatively stable at moderate privacy budgets because feature ranking and non-linear partitioning preserve a useful portion of the discriminative structure, whereas SVM is more exposed to the geometry changes introduced by stronger perturbation. This is coherent with prior IoT IDS studies and machine-learning-centered reviews showing that robust detection depends on careful alignment among preprocessing, feature quality, and model choice [1-3]. It is also consistent with work on adversarial analysis, anomaly detection, and learning-based classification in related domains [16, 20, 21, 33], where performance degrades rapidly when informative attributes are excessively distorted.

From a research perspective, the main implication is that privacy-preserving IoT dataset release should target model-aware utility preservation. Moderate DP budgets and structured anonymization are more suitable when the downstream objective is classical machine-learning-based IDS, while more aggressive perturbation should be accompanied by feature redesign, calibration, or representation learning strategies.

This interpretation also supports the broader view emerging from applied machine learning studies in sensing, multimedia, and healthcare [15, 35, 36], namely that privacy controls are most effective when they are tuned together with the statistical characteristics required by the intended models.

For Enhancing IoT Cybersecurity through Multi-Technique Data Anonymization: A Differential Privacy Framework using Public IoT Datasets, the most defensible next step is to move from a static release mechanism to an adaptive privacy pipeline that changes with data context, workload, and downstream utility requirements. The non-overlapping Sibghat literature suggests that future work should combine context-aware differential privacy, scalable analytics infrastructure, edge-readiness, and explicit quality governance so that public IoT datasets can be shared with clearer control over privacy loss, analytical usefulness, and deployment practicality [32, 37-43].

The cybersecurity side should also become more human-aware and signal-aware. Instead of treating anonymization as a purely numeric masking task, future studies should examine whether privacy-protected IoT data still preserve the linguistic, behavioral, and multimedia signals needed for trustworthy monitoring, bias detection, social sensing, and risk communication across smart environments [18, 19, 44-49].

Another clear direction is to validate the framework under real edge-native and cyber-physical deployment constraints. The selected papers indicate that privacy evaluation should increasingly account for autonomous operation, sensor fusion, public-health analytics, energy forecasting, and decision support workloads where latency, robustness, and operational explainability matter as much as formal privacy guarantees [50-57].

Model utility benchmarking should likewise expand beyond narrow classifier accuracy. The available non-overlapping titles point toward richer evaluation using segmentation, clustering, clinical decision support, feature optimization, multimodal representation learning, and visual reasoning so that anonymized IoT data can be tested for transferability across heterogeneous downstream tasks rather than for one isolated prediction setting [58-65].

Finally, the framework should be stress-tested on broader public sensing domains to show that privacy protection remains useful beyond a single benchmark family. The remaining candidate studies suggest a productive path through medical imaging, assistive IoT, disease analytics, agricultural monitoring, economic forecasting, scientific big data, and environmental time-series prediction, which together can support more realistic evidence about generalization, domain shift, and responsible reuse of public connected-device data [66-73].

CONCLUSION AND FUTURE WORK

This single-column paper has revisited a multi-technique anonymization framework for IoT cybersecurity datasets that combines ε -differential privacy with *k*-anonymity and *l*-diversity. Using Bot-IoT, TON_IoT, and IoT-23 as case studies, we outlined how IDS performance based on Random Forest and SVM can be systematically analyzed under different privacy budgets, while structural anonymization further protects quasi-identifiers. The framework builds on prior work on scalable anonymization for distributed processing platforms and on recent surveys of IoT security and learning-based security.

Future work will focus on extending the framework with context-aware and adaptive DP mechanisms, designing model-aware anonymization strategies that preserve the most informative IDS features, integrating more advanced deep learning-based IDS architectures, and validating the methodology on additional IoT datasets and real-world deployments.

DECLARATIONS

Acknowledgement: We appreciate the generous support from all the contributors to the research and their different affiliations.
Funding: No funding body in the public, private, or nonprofit sectors provided a particular grant for this research.
Availability of data and material: In the approach, the data sources for the variables are stated.
Authors' contributions: Each author participated equally in the creation of this work.
Conflicts of Interest: The authors declare no conflict of interest.
Consent to Participate: Yes
Consent for publication and Ethical approval: Because this study does not include human or animal data, ethical approval is not required for publication. All authors have given their consent.

REFERENCES

A. Fahim, Q. Tan, B. Naz, Q. u. Ain, and S. U. Bazai, "Sustainable higher education reform quality assessment using swot analysis with integration of ahp and entropy models: A case study of morocco," *Sustainability*, vol. 13, no. 8, p. 4312, 2021.

A. Jahangeer, S. U. Bazai, S. Aslam, S. Marjan, M. Anas, and S. H. Hashemi, "A review on the security of iot networks: From network layer's perspective," *IEEE Access*, vol. 11, pp. 71 073–71 087, 2023.

A. Komadina, M. Martinic, S. Gros, and Z. Mihajlovic, "Comparing threshold selection methods for network anomaly detection," *IEEE Access*, vol. 12, pp. 124 943–124 973, 2024.

A. Sathish, C. V. Sunanda, and C. S. Asha, "Adaptive iot security algorithm using lightweight cryptography and blockchain for scalable privacy-preserving architectures," *Journal of Internet Services and Information Security*, vol. 16, no. 1, pp. 385–397, 2026.

A. Sohail, S. U. Bazai, Z. Shahid, I. Batool, M. I. Gafoor, and U. A. Bhatti, "Detecting mental health signals in tweets: A machine learning and NLP approach," in *2025 Horizons of Information Technology and Engineering (HITE)*. IEEE, 2025, pp. 1–6.

A. Topbaş, A. Jamil, A. A. Hameed, S. M. Ali, S. U. Bazai, and S. A. Shah, "Sentiment analysis for covid-19 tweets using recurrent neural network (rnn) and bidirectional encoder representations (bert) models," in *2021 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube)*. *IEEE*, 2021, pp. 1–6.

H. Han, S. U. Bazai, M. A. Bhatti, A. Basit, A. Wahid, U. A. Bhatti, Y. Y. Ghadi, and A. Algarni, "Hybrid climate forecasting: variational mode decomposition and convolutional neural network with long-term short memory," *Polish Journal of Environmental Studies*, vol. 33, no. 2, 2024.

I. Tabassum and S. U. Bazai, "Augmenting multimedia analysis: A fusion of deep learning with differential privacy," in *Deep Learning for Multimedia Processing Applications*. CRC Press, 2024, pp. 194–215.

I. Tabassum, S. U. Bazai, M. I. Ghafoor, U. A. Bhatti, S. Ullah, and S. Akram, "A context-aware adaptive differential privacy for privacy-aware users in mobile crowd sensing," in *2025 International Conference on Frontiers of Information Technology (FIT)*. IEEE, 2025, pp. 1–6.

J. Li, Y. Tian, and T. Zhou, "Big data technology and healthcare informatics," in *Healthcare Information Systems: Progress, Challenges and Future Directions*. Springer, 2024, pp. 249– 274.

L. Khawaja, S. U. Bazai, M. Shahreen, M. I. Ghafoor, and M. Shah, "Leveraging apache spark for analyzing greenhouse gas emissions in supply chains," in *4th International Journal of Membrane Science and Technology*, vol. 4, no. ISBN # 978-969-23372-3-6. https://inccst.muet.edu.pk/INCCST24/proceeding.html, 2025, pp. 1–9. [Online]. Available: 22Feb2025

M. Aamir, M. A. Bhatti, S. U. Bazai, S. Marjan, A. M. Mirza, A. Wahid, A. Hasnain, and U. A. Bhatti, "Predicting the environmental change of carbon emission patterns in south asia: a deep learning approach using bilstm," *Atmosphere*, vol. 13, no. 12, p. 2011, 2022.

M. Aamir, S. U. Bazai, U. A. Bhatti, J. Li, and M. Huang, "Deep learning based applications for multimedia processing: Volume 1 and 2," 2024.

M. Aamir, Z. Li, S. U. Bazai, R. A. Wagan, U. A. Bhatti, M. M. Nizamani, and S. Akram, "Spatiotemporal change of air-quality patterns in hubei province: a pre-to post-covid-19 analysis using path analysis and regression," *Atmosphere*, vol. 12, no. 10, p. 1338, 2021.

M. Ahmed, M. I. Ghafoor, S. U. Bazai, S. Sabirov, U. A. Bhatti, and T. Eshchanov, "Hybrid machine learning approach for robust intrusion detection in iot networks," in *2025 IEEE 2nd International Conference on Deep Learning and Computer Vision (DLCV)*. IEEE, 2025, pp. 1–6.

M. Akram, S. U. Bazai, M. I. Ghafoor, S. Akram, Q. M. Ilyas, A. Mehmood, S. Iqbal, and M. A. Rafique, "Eemlcr: Energy-efficient machine learning-based clustering and routing for wireless sensor networks," *IEEE Access*, 2025.

M. Hamza, S. U. Bazai, M. I. Ghafoor, S. Ullah, S. Akram, and M. S. Khan, "Generative adversarial networks video framework: a systematic literature review," in *2023 International Conference on Energy, Power, Environment, Control, and Computing (ICEPECC)*. IEEE, 2023, pp. 1–5.

M. I. Ghafoor, M. S. Roomi, M. Aqeel, U. Sadiq, and S. U. Bazai, "Multi-features classification of smd screen in smart cities using randomized machine learning algorithms," in *2021 2nd International Informatics and Software Engineering Conference (IISEC)*. IEEE, 2021, pp. 1–5.

M. Muhammad, S. U. Bazai, S. Ullah, S. A. A. Shah, S. Aslam, A. Amphawan, and T.-K. Neo, "A systematic literature review on the role of big data in iot security," *Journal of Telecommunications and the Digital Economy*, vol. 12, no. 1, pp. 39–64, 2024.

M. N. Asghar, F. J. Saleemi, S. Iqbal, M. U. Chaudhry, M. Yasir, S. U. Bazai, and M. Q. Khan, "A novel parts of speech (pos) tagset for morphological, syntactic and lexical annotations of saraiki language," *Journal of Applied and Emerging Sciences*, vol. 11, no. 1, pp. pp–77, 2021. 12

M. Neagu, C. M. Serban, A. Hangan, and G. Sebestyen, "Trustworthiness in resource constrained iot: review and taxonomy of privacy-enhancing technologies and anomaly detection," in *Telecom*, vol. 7, no. 1. MDPI, 2026, p. 10.

M. Zeeshan, S. U. Bazai, M. I. Ghafoor, U. A. Bhatti, L. Baloch et al., "Comparative analysis of ml and dl models for human activity recognition: A focus on efficiency and edge-readiness," in *2025 IEEE 19th International Conference on Open Source Systems and Technologies (ICOSST)*. IEEE, 2025, pp. 1–7.

N. Ahmed, A. L. Barczak, S. U. Bazai, T. Susnjak, and M. A. Rashid, "Performance analysis of multi-node hadoop cluster based on large data sets," in *2020 IEEE Asia-Pacific Conference on Computer Science and Data Engineering (CSDE)*. IEEE, 2020, pp. 1–6.

N. Bugshan, "Privacy-preserving models in edge-cloud interplay for smart systems," Ph.D. dissertation, RMIT University, 2024.

O. Ali, Z. Zaland, S. U. Bazai, M. I. Ghafoor, L. Hussain, and A. Haider, "Neural transformers for bias detection: Assessing pakistani news," in *2024 5th International Conference on Advancements in Computational Sciences (ICACS)*. IEEE, 2024, pp. 1–7.

S. A. Nawaz, J. Li, U. A. Bhatti, S. U. Bazai, A. Zafar, M. A. Bhatti, A. Mehmood, Q. u. Ain, and M. U. Shoukat, "A hybrid approach to forecast the covid-19 epidemic trend," *PLOS ONE*, vol. 16, no. 10, p. e0256971, 2021.

S. Akram, S. U. Bazai, and S. Marjan, "Classifying traffic signs using convolutional neural networks based on deep learning models," in *Deep Learning for Multimedia Processing Applications*. CRC Press, 2024, pp. 250–269.

S. Akram, S. U. Bazai, M. I. Ghafoor, S. Marjan, M. Hamza, and S. A. A. Shah, "Systematic literature review: Evaluating effects of adversarial attacks and attack generation methods," in *2023 International Conference on Energy, Power, Environment, Control, and Computing (ICEPECC)*. IEEE, 2023, pp. 1–6.

S. Hussain, S. U. Bazaib, S. Qadir, S. Marjan, P. Pervaiz et al., "Sentiment analysis of balochi text using deep learning," *VAWKUM Transactions on Computer Sciences*, vol. 13, no. 1, pp. 190–200, 2025.

S. M. Nabeel, S. U. Bazai, N. Alasbali, Y. Liu, M. I. Ghafoor, R. Khan, C. S. Ku, J. Yang, S. Shahab, and L. Y. Por, "Optimizing lung cancer classification through hyperparameter tuning," *Digital Health*, vol. 10, p. 20552076241249661, 2024.

S. Noor, S. U. Bazai, M. I. Ghafoor, S. Marjan, S. Akram, and F. Ali, "Generative adversarial networks for anomaly detection: a systematic literature review," in *2023 4th International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*. IEEE, 2023, pp. 1–6.

S. Noor, S. U. Bazai, S. Tareen, and S. Ullah, "Detecting phishing urls through deep learning models," in *Deep Learning for Multimedia Processing Applications*. CRC Press, 2024, pp. 176–193.

S. U. Bazai and J. Jang-Jaccard, "In-memory data anonymization using scalable and high performance rdd design," *Electronics*, vol. 9, no. 10, p. 1732, 2020.

S. U. Bazai and J. Jang-Jaccard, "Sparkda: Rdd-based high-performance data anonymization technique for spark platform," in *International conference on network and system security*. Springer International Publishing Cham, 2019, pp. 646–662.

S. U. Bazai, "Building privacy-preservation models for distributed processing platforms," Ph.D. dissertation, Massey University, 2020.

S. U. Bazai, J. Jang-Jaccard, and H. Alavizadeh , "A novel hybrid approach for multi-dimensional data anonymization for apache spark," *ACM Transactions on Privacy and Security*, vol. 25, no. 1, pp. 1–25, 2021.

S. U. Bazai, J. Jang-Jaccard, and H. Alavizadeh, "Scalable, high-performance, and generalized subtree data anonymization approach for apache spark," *Electronics*, vol. 10, no. 5, p. 589, 2021.

S. U. Bazai, J. Jang-Jaccard, and X. Zhang, "A privacy preserving platform for mapreduce," in *International conference on applications and techniques in information security*. Springer Singapore Singapore, 2017, pp. 88–99.

S

S. Ullah, S. U. Bazai, M. Imran, Q. M. Ilyas, A. Mehmood, M. A. Saleem, M. A. Rafique, A. Haider, I. Khan, S. Iqbal et al., "Recent developments in authentication schemes used in machine-type communication devices in machine-to-machine communication: issues and challenges," *Computers, Materials & Continua*, vol. 79, no. 1, 2024.

U. A. Bhatti, J. Li, M. Huang, S. U. Bazai, and M. Aamir, "Signal processing and pattern recognition," (No Title), 2023.

U. A. Bhatti, M. Huang, H. Neira-Molina, S. Marjan, M. Baryalai, H. Tang, G. Wu, and S. U. Bazai, "Mffcg: Multi-feature fusion for hyperspectral image classification using graph attention network," *Expert Systems with Applications*, vol. 229, p. 120496, 2023.

U. A. Bhatti, M. Huang, J. Li, S. U. Bazai, and M. Aamir, "Deep learning for multimedia processing applications," *Signal Processing and Pattern Recognition*, 2024.

U. Khalid, S. U. Bazai, and A. Naushad, "Big data analytics and AI-driven approaches for air pollution trend analysis," in *Deep Learning Applications in Remote Sensing for Climate Change Monitoring*. IGI Global Scientific Publishing, 2026, pp. 119–164.

V. Mercan, A. Jamil, A. A. Hameed, I. A. Magsi, S. U. Bazai, and S. A. Shah, "Hate speech and offensive language detection from social media," in *2021 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube)*. IEEE, 2021, pp. 1–5.

Y. Kim, M. Kim, M. Chung, and J. Hur, "Detect your fingerprint in your photographs: photography-based multi-feature sybil detection," *Proceedings on Privacy Enhancing Technologies*, 2023.