



iJRASET

International Journal For Research in
Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** X **Month of publication:** October 2026

DOI: <https://doi.org/10.22214/ijraset.2026.85158>

www.ijraset.com

Call: ☎ 08813907089

E-mail ID: ijraset@gmail.com

Optimizing the Privacy Accuracy Trade-off in Data Stream Mining Using Random Cycle Cumulative Noise Addition

Anita A. Parmar¹, Chirag A. Patel²

¹Research Scholar, Gujarat Technological University, Ahmedabad, India,

²L. E. College, Morbi, India

Abstract: *Data stream mining has become increasingly important as a result of the continuous generation of data in real world applications including healthcare monitoring, sensor networks, transportation systems and online services. While stream classification techniques provide strong predictive performance, publishing or processing such streams raises serious privacy concerns. Perturbation based privacy preserving data stream mining (PPDSM) methods address this issue by adding noise to the data before it is analyzed. Cumulative noise addition accumulates noise across a fixed-size cycle of records give a favourable privacy–accuracy trade-off compared with independent noise addition. However, a fixed cycle size introduces a predictable and periodic structure in the perturbation process. This paper proposes Random-Cycle Cumulative Noise Addition (RC-CNA), a perturbation mechanism that adds noise to random cycle sizes instead of fix-cycle size. RC-CNA is evaluated against the fixed-cycle model. The proposed method raises classification accuracy, lowers privacy breach probability, and achieves the overall Privacy–Accuracy trade-off. Proposed method indicates that randomizing the cycle size, combined with the modulation, is a practical improvement that simultaneously strengthens privacy protection and improves classification utility.*

Keywords: *Privacy-Preserving Data Mining, Stream Classification, Cumulative Noise Addition, Random Cycle Size, Privacy–Accuracy Trade-off.*

I. INTRODUCTION

The rise of connected devices, online platforms, and sensor networks has led to an explosion of high-velocity, high-volume data streams that must be analyzed continuously rather than in static batches. Applications such as remote health monitoring, financial transaction processing, and location-based transportation services all generate data streams that are frequently mined in real time to support prediction and decision-making. Stream classification algorithms, most notably Hoeffding Trees [1] and Adaptive Random Forests (ARF) [2], have been developed specifically to learn incrementally from such streams while adapting to concept drift. However, the same streams often contain sensitive information about the individuals who generated them, and continuously releasing or processing this data for mining purposes creates a real risk that an adversary could recover private details about a specific person or record.

Privacy-Preserving Data Mining (PPDM) addresses this risk, and data perturbation is one of the most widely used PPDM strategies for streaming settings because it is computationally lightweight and can be applied online. Perturbation methods alter the original data values like adding random noise, so that individual records cannot easily be recovered, while still preserving the statistical properties needed for accurate classification. A well known limitation of perturbation, however, is that increasing the amount of noise to improve privacy tends to reduce classification accuracy, and vice versa. This creates an inherent privacy–accuracy trade-off that has motivated a substantial body of PPDM research.

Denham et al. [3] proposed random projection-based cumulative noise addition, in which noise is added cumulatively across successive records in the stream rather than independently to each record. This was shown to achieve a similar privacy benefit to independent noise addition while causing markedly less damage to classification accuracy, since the variance of the noise needed at each step is small. The drawback of unbounded cumulative addition is that, over a sufficiently long run, the accumulated noise can grow large enough to overwhelm the true data values, degrading accuracy as the classifier begins to learn the noise instead of the underlying signal.

Building directly on this issue, the base study for this work [4] proposed eight variations of cumulative noise addition that bound the maximum noise level using cycle-wise addition, noise resetting, absolute noise values, and a logistic control function. Their best-performing variant, Logistic Random Walk without Resetting (SRW), outperformed the original cumulative noise baseline (LRW) on both privacy and accuracy. Crucially, that study's own vulnerability analysis found that the start and end "flat" regions of each fixed-size noise cycle are disproportionately easy to attack, because the noise level or the noise addition rate is low precisely at those points in every cycle. The authors explicitly identified randomly varying the cycle size, instead of using a single fixed value, as a promising direction for future work to close this vulnerability.

We propose Random-Cycle Cumulative Noise Addition (RC-CNA), a chunk-wise perturbation mechanism built around a cycle length that is drawn independently and uniformly at random from a predefined range for every cycle, rather than fixed in advance. Because an attacker can no longer predict where one cycle ends and the next begins, the systematic vulnerability of the flat/boundary regions identified in prior work is disrupted. RC-CNA adds cumulative noise process that lets old noise fade geometrically rather than resetting sharply to zero at a cycle boundary and a chunk-variance-adaptive noise scale that adjusts the noise magnitude to the local variability of the data. We implement RC-CNA within the same experimental framework used in the base study, an Adaptive Random Forest classifier under prequential evaluation, attacked with a known input output attack and compare it against the fixed-cycle baseline (LRW).

The main contributions of this paper are as follows:

- A random cycle extension of cumulative noise addition (RC-CNA) that removes the fixed periodic structure exploited by known input–output attacks.
- An empirical evaluation on a real-world streaming data set showing that RC-CNA simultaneously improves classification accuracy and strengthens privacy protection relative to the fixed-cycle baseline.
- A quantitative analysis of the Privacy–Accuracy Magnitude (PAM) score, showing a substantial reduction relative to the fixed-cycle method.

The remainder of this paper is organized as follows. Section 2 reviews related work on privacy-preserving data stream mining and noise-based perturbation. Section 3 presents the system and threat model. Section 4 formalizes the noise addition model, the proposed random-cycle mechanism, and the privacy metrics used for evaluation. Section 5 describes the proposed RC-CNA algorithm in detail. Section 6 describes the stream classification methodology. Section 7 presents the experimental setup and Section 8 presents and discusses the results. Section 9 concludes the paper.

II. RELATED WORK

Privacy-preserving data stream mining has attracted considerable research attention because conventional PPDM techniques developed for static, batch-oriented data sets are frequently unsuitable for continuous, high-velocity streams [5]. Early approaches to privacy preservation relied on anonymization and aggregation, generalizing or suppressing identifying attributes so that individual records could not be distinguished. Such techniques typically require access to the whole data set at once, which conflicts with the single-pass, bounded-memory processing model required by streaming applications. Recent surveys and systematic reviews [6], presents this work and identify that achieving accuracy privacy trade-off as an open challenge.

Noise-based perturbation methods were introduced as a lighter-weight alternative that is naturally compatible with online, single-pass processing. Broadly, these methods fall into two families. Distance-preserving methods, such as random rotation and random projection [7-8], transform the data set with an orthogonal or near-orthogonal matrix so that the pairwise relationships between records, and hence classification accuracy are largely preserved. Additive methods, such as randomization and condensation, instead perturb individual attribute values directly with noise, which better protects record-level privacy but at a greater cost to accuracy. Across both families, a range of reconstruction attacks have been developed in the literature, including spectral filtering, principal component analysis, independent component analysis, maximum a posteriori estimation, and known input–output attacks [9], each exploiting a different structural weakness of the corresponding perturbation scheme.

Several perturbation methods have been designed for stream classification. Li and sun[10] preserved privacy on evolving streams by tracking correlation among streams. Chao et al. [11] introduced PCDS, two stage classification method , first perturbation and second classification, reported lower record base linkage and lower data error. Chamikara et al. [12] introduced p2RoCAI , a stream perturbation algorithm which offer higher accuracy and attack resilience. Fatlawi et al. [13] combined differential privacy with adaptive random forest to classify medical streams.

Random projection-based [3] cumulative noise addition combined random projection with a novel form of additive noise designed specifically for streaming environments rather than adding independent noise to every record, the same random noise values are carried forward and accumulated across a run of subsequent records. This was found to resist known input–output attacks more effectively than independent noise addition, because the attacker faces an increasing level of noise as the stream unfolds, while still keeping the per-record distortion and hence the accuracy penalty small. The key limitation identified in that work is that, left unbounded, cumulative noise inevitably grows large enough over a long stream to dominate the true signal, harming accuracy.

Hewage, Pears and Naeem [4], the base study for the present work addressed this limitation by bounding the maximum cumulative noise level through cycle-wise addition. They introduced eight variations combining a linear random-walk or a logistic-function-controlled noise schedule with optional noise resetting and optional use of absolute noise values, evaluated using Adaptive Random Forest classification under prequential evaluation and a known input–output attack model. Their results showed that Logistic Random Walk without Resetting (SRW) achieved the lowest overall Privacy–Accuracy Magnitude (PAM) score, outperforming the original fixed-cycle baseline (LRW) in both relative reconstruction error and breach probability across the cycle sizes tested (300, 600, 1000, and 2000 records). Their subsequent vulnerability analysis, however, showed that attacks targeting the starting and ending “flat” regions of each cycle where the rate of noise addition is lowest were noticeably more successful than randomly timed attacks, because the fixed cycle size makes the location of these regions predictable once an attacker infers the cycle boundary.

This vulnerability is a direct consequence of using a single, fixed cycle size throughout the stream: once an attacker has identified or guessed the cycle length, every subsequent cycle boundary in the stream can be predicted. To the best of our knowledge, no prior cumulative noise addition study has evaluated randomizing the cycle length itself as a countermeasure this randomization was proposed only as a direction for future work in the base study. Our work fills this gap by implementing and empirically evaluating a random-cycle variant of cumulative noise addition.

More recently, Hewage et al. (2025) [14] proposed an Accuracy Privacy Optimization Framework (APOF) which predicts the classification accuracy achievable for a user-specified privacy requirement. APOF extends cumulative noise addition with a regression-based accuracy predictor, but its underlying perturbation mechanism remains context-independent: it does not adapt the noise level to individual feature importance.

Stream classification [15] itself is supported by algorithms such as the Hoeffding Tree, which builds a decision tree incrementally using a small, bounded sample of each node's data, and the Adaptive Random Forest, an ensemble extension that additionally detects and adapts to concept drift. Both classifiers are designed for prequential (test-then-train) evaluation, in which each incoming instance is used to test the current model before being used to update it, closely mirroring real deployment conditions. Because of their strong incremental learning capability and drift resilience, both are well suited to evaluating the accuracy impact of privacy-preserving perturbation mechanisms in a genuinely streaming setting, and ARF in particular is adopted for the experiments in this paper for direct comparability with the base study.

III.SYSTEM AND THREAT MODEL

A. Data Stream Model

The data stream is modeled as a conceptually infinite, ordered sequence of instances arriving over time. Each instance consists of a vector of numerical attributes together with a target class label. Consistent with the streaming setting, instances are processed sequentially and incrementally using bounded memory; the entire stream is never materialized or stored. In the reference implementation, the stream is read and perturbed in fixed-size batches, or chunks, of b consecutive records at a time ($b = 10$ in the experiments reported here) rather than strictly one record at a time. This is a common practical accommodation for streaming I/O and does not alter the sequential, single-pass nature of the processing.

B. Noise Addition Model

Let X denote the original data stream and Y the corresponding perturbed stream released for mining. Noise is added to each non-target attribute of every record. Rather than resetting sharply to zero at a fixed cycle boundary, as in the base study's schemes, the perturbation used here carries noise forward from one chunk to the next using an exponentially-decayed running sum: a fraction γ of the previous chunk's accumulated noise persists into the current chunk, to which a freshly drawn Gaussian noise batch is added. This produces a smoothly bounded, non-resetting analogue of cumulative noise addition. The scale of the freshly generated noise at any point in the stream is itself modulated by two further factors: a linear modulation of the record's position within the current cycle, and the local variance of the chunk currently being processed.

C. Random Cycle Size

Prior cumulative noise addition schemes, including the base study's LRW/LRWR/LAR/LA and SRW/SRWR/SAR/SA variants, use a single, fixed cycle size c_s for the entire stream. In the mechanism proposed here, the cycle length is instead sampled independently for every new cycle from a discrete uniform distribution over a predefined range $[C_{\min}, C_{\max}]$. The cycle boundary is checked at every chunk (i.e., every b records); once the number of records processed since the last boundary reaches the current cycle length, a new cycle length is redrawn uniformly at random from $[C_{\min}, C_{\max}]$ and the linear modulation function (Section 4) restarts from its zero point. Because the length of each cycle is unknown in advance and varies across the stream, an attacker cannot reliably predict where one cycle ends and the next begins, which removes the periodic, predictable structure that fixed-cycle schemes expose.

D. Threat Model: Known Input–Output Attack

Consistent with the base study, the attacker is assumed to possess a small number of known original–perturbed record pairs, together with knowledge of the general perturbation mechanism (zero-mean Gaussian cumulative noise), but not the specific noise values realized for any individual record, the cycle length in effect at any point, or the boundaries of any cycle in the random-cycle setting. Using this partial knowledge, the attacker attempts to reconstruct unknown original records from their perturbed counterparts, and privacy is quantified by how often this reconstruction succeeds within a specified distance threshold.

IV. MATHEMATICAL FORMULATION AND PRIVACY METRICS

Table 1 summarizes the notation used throughout this section, consistent with the symbol conventions of the base study.

Symbol	Meaning
X	Original data stream
Y	Perturbed data stream
b	Chunk size (records processed per batch)
X_j, Y_j	Original / perturbed chunk j (a $b \times m$ matrix)
Ψ_j	Decayed cumulative noise matrix applied to chunk j
N_j	Freshly drawn Gaussian noise matrix for chunk j
γ	Decay factor for cumulative noise persistence ($0 < \gamma < 1$)
$\lambda(p), \lambda_{\max}$	Linear modulation factor at cycle position p , and its peak value
$\sigma_{\text{base}}, \sigma_j$	Base noise scale and chunk-adaptive noise scale
v_j	Mean feature-wise variance of chunk j
C_k	Length of the k -th (randomly drawn) cycle
$C_{\min}, C_{\max}$	Lower and upper bounds of the random cycle-size range
ϵ	Privacy-breach distance threshold
RE_t	Relative reconstruction error for record t
P_{breach}	ϵ -privacy breach probability over the stream
PAM	Privacy-Accuracy Magnitude score

Table 1: Symbol Table

E. Chunk-wise Cumulative Noise Addition

Let the original stream be partitioned into consecutive chunks of b records, $X = \{X_j, j = 0, 1, 2, \dots, m\}$, where m is the number of non-target (perturbable) attributes. Rather than resetting sharply to zero at the end of a fixed cycle, the cumulative noise matrix Ψ_j applied to chunk j is generated by an exponentially-decayed recursion:

$$\Psi_j = \gamma \cdot \Psi_{j-1} + N_j, \quad \Psi_{-1} = 0$$

where γ is a decay factor and N_j is a freshly drawn $(b \times m)$ matrix of independent zero-mean Gaussian noise values. The perturbed chunk is then:

$$Y_j = X_j + \Psi_j$$

Because $0 < \gamma < 1$, the influence of noise generated in earlier chunks decays geometrically rather than being discarded abruptly at a cycle boundary; this bounds the long-run growth of the cumulative noise without introducing the discontinuity that a hard reset creates.

F. Chunk- Linear Within-Cycle Modulation

The scale of the freshly generated noise N_j is modulated by a linear function of the record's position within the current cycle. Let C denote the current cycle length and p the number of records processed since the start of the current cycle ($0 \leq p < C$). The modulation factor is:

$$\lambda(p) = (2\lambda_{\max} / C) \cdot p, \quad \text{if } p < C/2$$

$$\lambda(p) = 2\lambda_{\max} \cdot (1 - p/C), \quad \text{otherwise}$$

where $\lambda_{\max}$ is the maximum modulation factor. $\lambda(p)$ rises linearly from 0 at the start of a cycle to $\lambda_{\max}$ at the cycle's midpoint, then falls linearly back to 0 at the cycle's end, giving a triangular noise-scale envelope in place of the flat or logistic envelopes used in the base study. This envelope resets to zero whenever a new cycle begins, but — unlike the base study's hard reset of the cumulative noise itself — it modulates only the scale of the newly generated noise; the underlying cumulative process Ψ_j continues to decay smoothly across the cycle boundary as described in Section 4.

G. Data-Adaptive Noise Scale

The standard deviation used to draw N_j additionally adapts to the local variability of the chunk currently being processed:

$$\sigma_j = \sigma_{\text{base}} \cdot (1 + 0.5 \cdot v_j)$$

where v_j is the mean, across features, of the within-chunk feature variance of X_j , and σ_{base} is a fixed base noise scale set as a small proportion of the average feature standard deviation over the whole stream ($\sigma_{\text{base}} = 0.01 \times \text{mean feature standard deviation}$ in the experiments reported here). Each entry of N_j is then drawn independently as $N(0, (\lambda(p_j) \cdot \sigma_j)^2)$, where p_j is the cycle position of chunk j . This makes the noise scale responsive to locally noisier or more variable regions of the stream, in addition to the triangular cycle-position modulation.

H. Proposed Random-Cycle Formulation

In fixed-cycle cumulative noise addition, the cycle length C is a constant and applied uniformly across the entire stream. In the proposed Random-Cycle Cumulative Noise Addition (RC-CNA) mechanism, the cycle length is instead an independent random variable, resampled every time a cycle boundary is reached:

$$C_k \sim \text{DiscreteUniform}(C_{\min}, C_{\max}), \quad k = 1, 2, 3, \dots$$

where C_k is the length of the k -th cycle in the stream, and $C_{\min}$ and $C_{\max}$ define the configurable range from which cycle lengths are drawn. Because C_k varies unpredictably from one cycle to the next, an attacker cannot anticipate the position of the chunk.

The attacker is assumed to know the general perturbation mechanism (Gaussian, zero mean) and a small number of original-perturbed record pairs, but not the realized noise values, the cycle boundaries, or the decay/modulation parameters in effect for any given record. Because Ψ_j is a linear combination of zero-mean Gaussian variables, it remains zero-mean regardless of the decay weighting or modulation applied to it, so the optimal unbiased point estimator of an unknown original record, given only its perturbed value, remains the perturbed value itself:

$$\hat{x}_t = y_t$$

I. Privacy-Accuracy Trade-off

The relative reconstruction error for a record is defined as the normalized distance between the true record and its estimate:

$$\text{RE}_t = |x_t - \hat{x}_t| / |x_t|$$

A record is said to suffer an ϵ -privacy breach if its relative reconstruction error falls below a specified threshold, $\text{RE}_t < \epsilon$. The ϵ -privacy breach probability over the evaluated stream, P_{breach} , is computed as the fraction of attacked records satisfying this condition. Following the base study, the overall trade-off between privacy and accuracy is summarized using the Privacy-Accuracy Magnitude score:

$$\text{PAM} = (\text{RE})^2 + (P_{\text{breach}})^2$$

where RE and P_{breach} denote the (normalized) relative error and breach probability aggregated over the evaluated stream. A lower PAM value indicates a more favorable joint privacy-accuracy outcome

V. PROPOSED METHODOLOGY: RANDOM-CYCLE CUMULATIVE NOISE ADDITION (RC-CNA)

RC-CNA processes the stream in fixed-size chunks and applies, to each chunk, a noise batch shaped by three coordinated mechanisms layered around the central proposal of this paper, a randomly drawn cycle length. First, the cycle length in effect at any point in the stream is redrawn independently and uniformly at random whenever a cycle boundary is reached, rather than fixed for the whole stream, which is what directly targets the fixed-cycle vulnerability identified in the base study's attack analysis. Second, the scale of the noise generated for a chunk is modulated by a linear function of that chunk's position within the current (randomly-sized) cycle, ramping up from zero, peaking at the cycle midpoint, and ramping back down to zero. Third, rather than resetting the cumulative noise to zero at a cycle boundary, the cumulative noise is carried forward with an exponential decay, so that old noise fades out smoothly instead of being discarded abruptly. The procedure is summarized in Algorithm 1.

Algorithm 1 Pseudo Code — Random-Cycle Cumulative Noise Addition (RC-CNA)

```

Input: Chunk size  $b$ , Cycle-size range  $[C_{\min}, C_{\max}]$ , Base noise scale  $\sigma_{\text{base}}$ ,
      Max modulation  $\lambda_{\text{max}}$ , Decay factor  $\gamma$ , Stream ( $X$ )
Initialize:  $C \sim \text{DiscreteUniform}(C_{\min}, C_{\max})$  // current cycle length
       $\Psi = 0$  ( $b \times m$  matrix) // cumulative noise state
      record_count = 0
Process:
  While End of  $X$ 
    Read next chunk  $X_j$  of  $b$  records
     $p \leftarrow \text{record\_count} \bmod C$  // position within current cycle
     $\lambda \leftarrow \text{LinearModulation}(p, C, \lambda_{\text{max}})$ 
     $v_j \leftarrow \text{mean feature-wise variance of } X_j$ 
     $\sigma_j \leftarrow \sigma_{\text{base}} \times (1 + 0.5 \times v_j)$  // chunk-adaptive noise scale
     $N_j \leftarrow \text{draw } (b \times m) \text{ matrix from } N(0, (\lambda \times \sigma_j)^2)$ 
     $\Psi \leftarrow \gamma \times \Psi + N_j$ 
     $Y_j \leftarrow X_j + \Psi$  // perturb the chunk
    record_count  $\leftarrow$  record_count +  $b$ 
    if record_count  $\bmod C == 0$ :
       $C \sim \text{DiscreteUniform}(C_{\min}, C_{\max})$  // redraw cycle length at the boundary
Output: Perturbed Stream ( $Y$ )

```

This is a low-cost modification relative to the base study's fixed-cycle schemes: it requires only one additional random draw per cycle (rather than per record) to determine the next cycle length, so the computational overhead is negligible. Unlike a minimal single-variable change to the base study's LRW scheme, RC-CNA is a composite mechanism, decay-based persistence, and adaptive scaling, since this is how the mechanism was implemented and evaluated.

VI. STREAM CLASSIFICATION METHODOLOGY

The stream classifier is evaluated using prequential evaluation [16-17]. For each incoming instance, a prediction is made first, the running accuracy metric is updated based on that prediction, and only then is the instance used to incrementally update the model. This mirrors real-world streaming deployment, in which the model must always be tested on data it has not yet been trained on, and it does not require a held-out test set.

Adaptive Random Forest (ARF), with Naive Bayes leaf prediction, is used as the classification algorithm, matching the base study to ensure the results reported here are directly comparable. ARF processes the perturbed stream generated by both the fixed-cycle baseline and the proposed RC-CNA mechanism under identical classifier settings, so that any difference in classification accuracy can be attributed to the perturbation mechanism rather than to the learning algorithm.

Classification performance is measured using cumulative (prequential) accuracy over the entire evaluated stream, incrementally updated after every instance and reported as the proportion of correctly classified instances.

VII. EXPERIMENTAL SETUP

The experiments were conducted on the New York City Taxi Trip Duration data set [18], the same real-world streaming benchmark used for the corresponding experiments in the base study. Records are treated as a stream ordered by the time they were produced; only the numerical features of the data set are used, and the data set contains potentially sensitive trip and location information, making it a realistic candidate for privacy-preserving processing.

To ensure a fair, like-for-like comparison, the classifier configuration, attack configuration, and noise variance were kept identical between the fixed-cycle baseline and the proposed method:

- Classification method — Adaptive Random Forest (ARF)
- Perturbation method— LRW (Linear Random Walk without Resetting),
- Proposed method — RC-CNA
- cycle length drawn from a discrete uniform distribution and redrawn at every cycle boundary
- Attacking method — Known Input/Output attack, with 4 known input–output pairs per attack and attacks performed on 5% of the total record count
- Privacy threshold — $\epsilon = 0.2$.

The RC-CNA parameters above match the reference implementation used to generate the perturbed stream for this study.

VIII. RESULTS AND ANALYSIS

A. Overall Performance

Table 2 reports the classification accuracy, relative error, breach probability, and PAM score obtained on the Taxi data stream for the fixed-cycle baseline (LRW) and the proposed random-cycle method (RC-CNA).

Method	Original Accuracy	Perturbed Accuracy	Relative Error	Breach Probability	PAM
Base paper (fixed cycle, LRW)	0.705	0.592	0.407	0.253	0.230
Proposed (random cycle, RC-CNA)	0.705	0.698	0.301	0.069	0.095

Table 2: Performance comparison on the Taxi data set

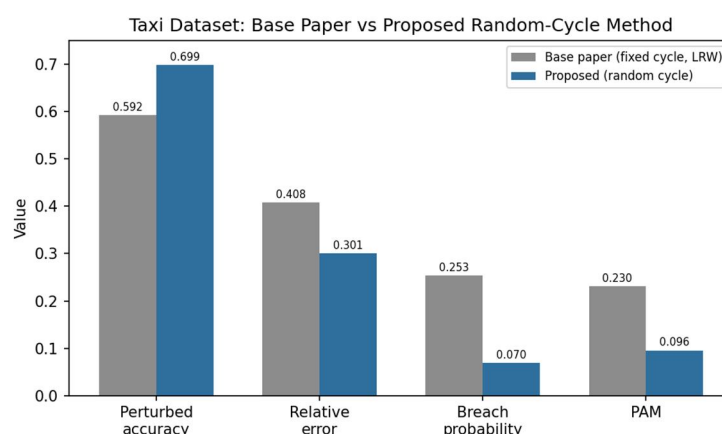


Figure 1: Base paper (LRW) vs. proposed random-cycle method on the Taxi data set

B. Privacy–Accuracy Trade-off

Under the fixed-cycle baseline, classification accuracy of the perturbed stream is decreased. Under the proposed RC-CNA mechanism, perturbed stream accuracy is higher than the fixed-cycle baseline. RC-CNA changes several aspects of the perturbation process together like randomized cycle length, the decay-based cumulative persistence, and the chunk-adaptive noise scale. So this accuracy gain reflects the combined effect of the proposed mechanism rather than cycle length randomization in isolation.

The known input–output attack model is used to evaluate privacy as described in Section 4. The proposed random-cycle method reduces relative error and also reduces breach probability relative to the fixed-cycle baseline. This confirms that randomizing the cycle boundary disrupts the attacker's ability to exploit the predictable, low-noise regions identified in the base study's vulnerability analysis: because the attacker can no longer anticipate where a cycle begins or ends, a much smaller proportion of attacked records fall within the ϵ -privacy breach threshold.

The PAM score reduced for the proposed method. As PAM combines relative error and breach probability, and both individual components improve under RC-CAN. Unlike many perturbation trade-offs, in which a gain in privacy is paid for with a loss in accuracy (or vice versa), the random-cycle mechanism improves both quantities simultaneously relative to the fixed-cycle baseline. This indicates that the fixed cycle boundary in the original scheme was a genuine, avoidable weakness rather than a necessary cost of achieving low classification error.

C. Discussion

These results are consistent with the hypothesis raised as future work in the base study, that a fixed cycle size introduces an exploitable periodic structure. By resampling the cycle length independently at every cycle boundary, RC-CNA removes the predictability of that structure. The accuracy improvement observed with the privacy improvement is a practically important secondary benefit. It also suggests that the combination of a shifting cycle boundary, cumulative noise scaling helps prevent the classifier from adapting to fixed, repeating noise pattern, which fixed-cycle, hard-reset addition may otherwise allow to happen over a long-running stream.

It is important to note that RC-CNA changes four things relative to the base study's LRW baseline at once. The experiment demonstrates that this mechanism improves on the fixed-cycle baseline.

IX. CONCLUSION

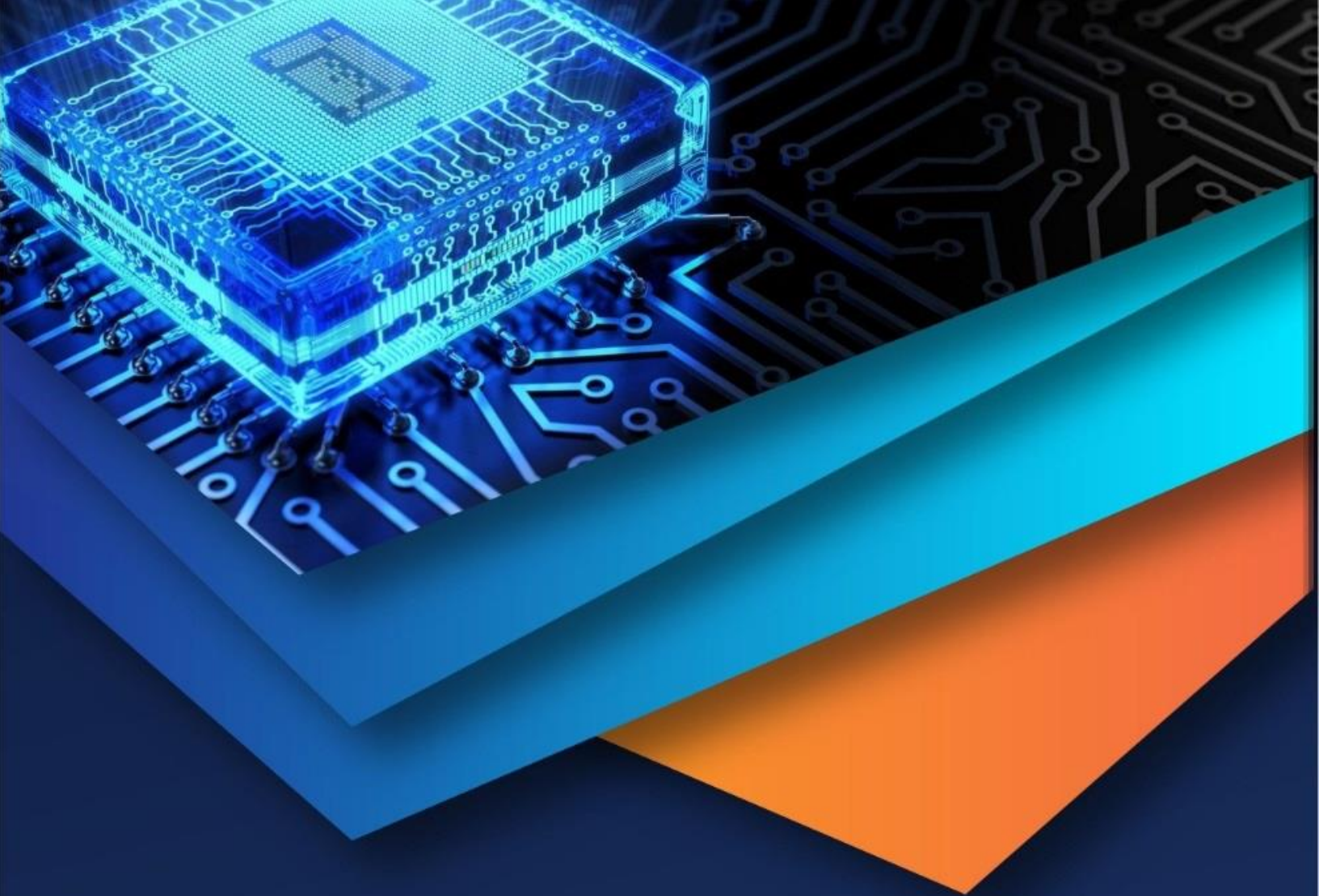
This paper proposed Random-Cycle Cumulative Noise Addition (RC-CNA), a privacy-preserving perturbation mechanism for data stream mining. It randomizes the cycle length used in cumulative noise addition instead of fix cycle size. RC-CNA was evaluated against the fixed-cycle using an Adaptive Random Forest classifier under prequential evaluation and a known input–output attack model. The proposed method improved classification accuracy on the perturbed stream, reduced relative reconstruction, reduced ϵ -privacy breach probability and reduced the overall Privacy–Accuracy Magnitude. These results confirm that randomizing cycle size is an effective and low overhead way to close the fixed-cycle vulnerability of cumulative noise addition while also improving classification utility.

REFERENCES

- [1] Domingos, P. and Hulten, G. 2000. Mining high-speed data streams. Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 71–80.
- [2] Gomes, H. M., Bifet, A., Read, J., Barddal, J. P., Enembreck, F., Pfahringer, B., Holmes, G. and Abdesslem, T. 2017. Adaptive random forests for evolving data stream classification. Machine Learning, 106(9-10), 1469–1495.
- [3] Denham, B., Pears, R. and Naeem, M. A. 2020. Enhancing random projection with independent and cumulative additive noise for privacy-preserving data stream mining. Expert Systems with Applications, 152.
- [4] Hewage, U. H. W. A., Pears, R. and Naeem, M. A. 2022. Optimizing the Trade-off Between Classification Accuracy and Data Privacy in the Area of Data Stream Mining. International Journal of Artificial Intelligence, 20(1), 147–167.
- [5] A. Cuzzocrea, Privacy-Preserving Big Data Stream Mining: Opportunities, Challenges, Directions, IEEE International Conference on Data Mining Workshops, 2017.
- [6] U. H. W. A. Hewage, R. Sinha, M. Asif Naeem, Privacy-preserving data (stream) mining techniques and their impact on data mining accuracy: a systematic literature review, Springer, 2023.
- [7] Liu, K., Kargupta, H. and Ryan, J. 2006. Random Projection-Based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining. IEEE Transactions on Knowledge and Data Engineering, 18, 92–106.
- [8] Liu, K. 2007. Multiplicative Data Perturbation for Privacy Preserving Data Mining. PhD thesis, University of Maryland, Baltimore County. Giannella, C., Kargupta, H. and Liu, K. 2008. A Survey of Attack Techniques on Privacy-Preserving Data Perturbation Methods.
- [9] Giannella, C., Kargupta, H. and Liu, K. 2008. A Survey of Attack Techniques on Privacy-Preserving Data Perturbation Methods.
- [10] Feifei Li, Jimeng Sun, Hiding in the Crowd: Privacy Preservation on Evolving Streams through Correlation Tracking, IEEE, 2007.
- [11] Ching-Ming Chao, Po-Zung Chen, Chu-Hao Sun, Privacy-Preserving Classification of Data Streams, Journal of Science and Engineering, 2009.
- [12] M. A. P. Chamikara, P. Bertok, D. Liu, S. Camtepe, I. Khalil, Efficient data perturbation for privacy preserving and accurate data stream mining, Pervasive and Mobile Computing, 48 (2018) 1–19.
- [13] Fatlawi, Hayder K. and Kiss, Attila, Differential privacy based classification model for mining medical data stream using adaptive random forest, Acta Universitatis Sapientiae, Informatica, vol. 13, no. 1, 2021, pp. 1–20.



- [14] Waruni Hewage, R. Sinha, M. Asif Naeem, An accuracy-privacy optimization framework considering user's privacy requirements for data stream mining, 2025
- [15] Jerzy Stefanowski, Dariusz Brzezinski, Stream Classification, in Encyclopedia of Machine Learning, Springer.
- [16] Albert Bifet, MOA data stream mining: A practical approach, vol 8, pp. 127-141, May 2011.
- [17] Bifet, A., Kirkby, R., Kranen, P. and Reutemann, P. 2012. Massive Online Analysis (MOA) Manual.
- [18] Kaggle. 2017. New York City Taxi Trip Duration. <https://www.kaggle.com/c/nyc-taxi-trip-duration>



10.22214/IJRASET



45.98



IMPACT FACTOR:
7.129



IMPACT FACTOR:
7.429



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089  (24*7 Support on Whatsapp)