SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V13I11P113

Volume 13 Issue 11, 221-231, November 2026
© 2026 Seventh Sense Research Group®

Original Article

A Context-Aware Cumulative Noise Framework for Privacy-Preserving Data Stream Classification

Anita A. Parmar[1], Chirag A. Patel[2]

[1]*Gujarat Technological University, Ahmedabad, Gujarat, India.*
[2]*Lukhdhirji Engineering College, Morbi, Gujarat, India.*

[2]*Corresponding Author: prof.capatel@gmail.com*

Received: 29 September 2026 Revised: 29 October 2026 Accepted: 05 Novmber 2026 Published: 27 November 2026

Abstract *- Privacy-Preserving Data Stream Mining (PPDSM) focuses on extraction of relevant patterns from large-volume data streams while protecting confidential information simultaneously. In streaming systems such as healthcare and IoT, data is continuously generated, which makes privacy preservation more challenging than with static datasets. In this paper, focus is on privacy-preserving classification of data streams. Most of the existing studies focus on either privacy or accuracy alone. They are missing trade-off evaluation. Additionally, existing methods treat all attributes equally and ignore feature contribution in classification process. Due to this, important features suffer unnecessary utility loss. To address this gap, this paper presents novel Context-Aware Cumulative Noise (CACN), a perturbation framework that perturbs features based on their importance. CACN allocates noise inversely proportional to feature importance. CACN is implemented and evaluated against baseline cumulative noise addition on two real-world streams using an Adaptive Random Forest classifier and a known input–output attack model. Experimental results confirm that CACN improves the privacy accuracy trade-off. These results show that context-aware noise allocation is an effective, low-overhead extension to cumulative noise addition, while also identifying feature correlation as an important factor that future privacy-aware noise mechanisms must account for.*

Keywords *- Adaptive Random Forest, Context-Aware noise addition, Cumulative noise, Privacy-Preserving data stream mining, Stream classification.*

1. Introduction

Applications of data mining are using a different criteria to analyze data, which include classification, association, forecasting and clustering. With the growing applications of data mining, substantial quantity of personal information is being analyzed. This information serves as resource to both business organizations and governments in their decision-making processes [1]. However, analysis of such data can pose threats to individuals' privacy. So, Privacy-Preserving Data Mining (PPDM) has emerged as an important research area. The primary objective of PPDM is to obtain useful information from large datasets while simultaneously preserving the privacy of sensitive information. Therefore, fundamental goal of PPDM is to design algorithms that transform or modify datasets in a manner that preserves data utility while preventing the disclosure of sensitive data.

Since the field emerged around 2000, a wide range of PPDM techniques has been developed for different data types and threat models [1, 2]. Early research introduced heuristic techniques to reduce the risk of sensitive information being revealed through data mining algorithms, and subsequent studies introduced frameworks for defining and measuring the degree of privacy preserved [3, 4]. Several surveys consolidate this body of work and provide the background for the more specific techniques discussed later in this paper [5-7].

Data stream mining extends traditional data mining by addressing environments in which data is generated continuously and needs to be processed incrementally rather than collected and analyzed as static batch [8, 9]. It has become applicable to various domains including networking, telecommunications, education, banking, weather forecasting, and financial markets, and privacy preservation in this streaming setting has correspondingly attracted growing research attention [7, 10]. Privacy-preserving algorithms mainly hide sensitive data by modifying sensitive attributes of individuals, while allowing data mining operations on stream. However, original distribution of the data may change due to such modification, and utility will decrease for data mining approaches. Hence, the proper balance between privacy protection and data stream remains a significant challenge. Moreover, stream mining algorithms are designed to provide quick and adapt changes in data distribution, and privacy-preserving algorithms may give additional overhead and delays. These effects can make it


This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/)

more difficult to detect concept drift and prevent mining method to adapt properly.

Privacy-preserving data stream mining techniques can be categorized by the mining task they support: Classification [11-17], Association rule mining [18, 19], Clustering [20] and Data stream publication [21, 24].

This paper concerns privacy-preserving data stream classification, in which data streams are classified, and privacy of underlying records is protected.

Among perturbation methods for stream classification, cumulative noise addition [15] and its cycle-based variants [25] give a good balance between privacy and accuracy, because noise added to one record is carried forward to later records, which makes reconstruction by an attacker harder. There are mainly two gaps. First, Most studies optimize either privacy or accuracy, and only a few studies combine the privacy-accuracy trade-off [15, 25]. Second, all existing cumulative noise addition methods add the same amount of noise to every feature. Features that are important for classification task are distorted and affect classification accuracy without a matching gain in privacy.

To address this gap, the paper proposes Context-Aware Cumulative Noise (CACN). The novelty of CACN lies in combining two ideas that have so far been studied separately: Cumulative Noise Addition For Streams and Feature-Important (context-aware) noise allocation. The main contributions are as follows. (i) A perturbation framework is designed that scores each attribute by its mutual information with the class label and sets the noise level of the attribute in inverse proportion to that score. (ii) This allocation is integrated with cumulative Gaussian noise injection and a linear cycle-based modulation that limits the long-run growth of noise. (iii) CACN is compared with the SRW baseline on two real streams using classification error, breach probability under a known input–output attack, and PAM. (iv) The analysis shows CACN improves all three measures on weakly correlated data, while strong inter-feature correlation increases breach probability.

The rest of the paper is organized as follows—section 2 reviews related work. Section 3 gives the preliminaries, and Section 4 formulates the problem. Section 5 describes CACN, Section 6 the experimental setup and Section 7 the results and discussion. Section 8 states limitations and future work, and Section 9 concludes the paper.

2. Related Work

2.1. Privacy-Preserving Stream Classification

Chao et al. [11] proposed PCDS, a two-stage privacy-preserving data stream classification method that first perturbs the stream and then applies stream mining. They achieved higher security and higher average squared distance, indicating both stronger privacy and lower data error. A related two-stage approach [17] combined a sliding-window perturbation algorithm with rotation perturbation [28], classifying the resulting perturbed stream with a Hoeffding tree. Their results showed data privacy could be achieved with minimal information loss for numeric attributes. Earlier studies applied additive, multiplicative, geometric and heuristic perturbations to streams before classification [12, 33-36], and Prasada and Hirwarkar [37] proposed an efficient model for privacy-preserving data stream classification.

Chamikara et al. [14] introduced P2RoCAl, a data stream perturbation algorithm reporting higher accuracy, efficiency, and attack resilience than rotation perturbation and data condensation. Gayathri Devi et al. [38] combined rotation-based perturbation with condensation to protect big data streams generated by Internet of Things devices. Kotecha and Garg [13] proposed Diverse HOeffding Tree (DAHOT), which combines a ℓ-diversity and k-anonymity with the Hoeffding Tree algorithm to preserve output privacy in data stream classification. DAHOT takes a data stream and an anonymity/diversity requirement specified by user as input and produces a privacy-preserving decision tree classifier. DAHOT was shown to be efficient on massive streams while offering higher interpretability, lower training time, and minimal information loss.

Fatlawi and Kiss [16] proposed a classification model for medical data streams based on differential privacy [32], built on Adaptive Random Forest (ARF), combining a medical-data preparation stage with an ensemble of fast decision trees and evaluating performance on real streaming data. Hoeffding tree variants, surveyed in [39], are the most common base learners in these studies, while ensemble learners such as ARF [40] adapt better to concept drift.

2.2. Cumulative Noise Addition

Denham et al. [15] introduced cumulative noise addition for data streams, in which noise added to a record also carries forward and accumulates into subsequent records rather than being generated each time independently, combined with random projection [29] and translation. Based on this, Hewage et al. [25] studied seven cumulative noise strategies, linear and logistic noise schedules, cycle-wise addition, and analyzed the resulting privacy–accuracy trade-off using the Privacy–Accuracy Magnitude (PAM) score. Their results showed that a cumulative noise model can outperform traditional static-noise approaches for stream classification. They showed Logistic Random Walk without Resetting (SRWR) variant achieves the best overall PAM. However, all variants apply noise uniformly across features and do not consider data context or feature importance.

More recently, Hewage et al. [26] 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 perturbation mechanism does not adapt the noise level to individual feature importance.

2.3. Context-Aware and Correlation-Aware Privacy Mechanisms

Li et al. [22] showed that, for evolving streams, perturbation noise should follow the correlation structure of the data. Their method tracks correlations online, but it targets stream publication only. Caruccio et al. [41] studied how privacy can be traded against classification accuracy in machine learning processes, and Calvi et al. [42] examined the trade-offs between privacy-enhancing technologies and fairness.

Clustering-based anonymization has been applied to data from wearable IoT devices [20], and the publication of transactional and health data streams has been studied in [21, 23, 24]. These studies show the value of adapting a privacy mechanism to the data and the task. None of them targets online stream classification, and none integrates cumulative noise addition, the mechanism designed to control noise growth in a continuously arriving stream.

2.4. Research Gap

The literature shows that PPDSM has attracted substantial research attention because streaming data is continuous and often sensitive, and that a good privacy-accuracy balance remains a central open problem. Two gaps remain unaddressed. First, most studies optimize accuracy or privacy in isolation rather than jointly, and only a few report a combined trade-off metric such as PAM [10, 15].

Second, existing cumulative noise methods, including the state-of-the-art SRW variant [25], treat every feature identically, so noise is not allocated according to a feature's importance to the classification task, and important features suffer unnecessary utility loss. No prior work found in the literature combines cumulative noise addition with context-aware noise allocation for privacy-preserving stream classification. This gap motivates the framework proposed in this paper.

3. Preliminaries

3.1. Stream Classification Process

In stream classification, data arrives continuously and is processed with incremental learning techniques that update the model progressively rather than retraining from scratch. To simulate this behavior, large data sets are divided into smaller chunks. The model first predicts on an incoming chunk and is then updated using that same chunk, followed by a prequential protocol. This paper uses the Adaptive Random Forest algorithm, an ensemble method for data stream classification, throughout this work.

3.2. Privacy Accuracy Trade-off Optimization

The ideal outcome is zero classification error together with zero privacy breach [43], which is unattainable in practice; a realistic objective is instead to jointly minimize classification error and breach probability. Following Denham et al. [15], this paper formalizes trade-off with the Privacy Accuracy Magnitude (PAM) score:

$$\text{PAM} = (\text{error})^2 + \left(\text{P}(\varepsilon - \text{privacy breach})\right)^2 \quad (1)$$

Breach probability quantifies the likelihood that how well an attacker can recover an original record successfully from its perturbed counterpart using a known Input /Output (I/O) attack [30]. Relative error is the magnitude of the difference vector between an original record x_i and its reconstructed version $\hat{x}_i$, normalized by the magnitude of the original record vector [25]. P(ε–privacy breach) is the fraction of attacked records whose relative error falls below a threshold ε. A lower PAM value indicates a better privacy-accuracy balance. If PAM falls below a given threshold, the trade-off is considered optimized.

4. Problem Formulation

The problem addressed in this paper is to design a perturbation method that preserves privacy while maintaining classification accuracy in a streaming environment, i.e, to optimize the privacy-accuracy trade-off. Specific objectives are to:

- Develop a privacy-preserving technique for data stream classification using context-aware cumulative noise.
- Achieve high classification accuracy on the perturbed data stream.
- Optimize the privacy-accuracy trade-off by minimizing classification error and breach probability.

Existing research addresses either cumulative noise without contextual adaptation or context-aware privacy without a cumulative noise framework, and no prior work found in the literature combines the two for privacy-preserving data stream classification. CACN fills this gap by adjusting the perturbation level of each feature according to its relevance to the class, which further improves the privacy–accuracy trade-off.

5. Proposed Method: Context-Aware Cumulative Noise (CACN)

5.1. Framework Overview

Figure 1 summarizes the CACN pipeline. The incoming data stream is first analyzed to estimate the importance of each feature via mutual information. This importance score parameterizes a context-aware noise model that determines per feature noise to inject. The context-aware noise injection stage then perturbs each record before it reaches the stream classifier (Adaptive Random Forest), and the resulting

predictions are evaluated jointly for accuracy, privacy, and PAM.

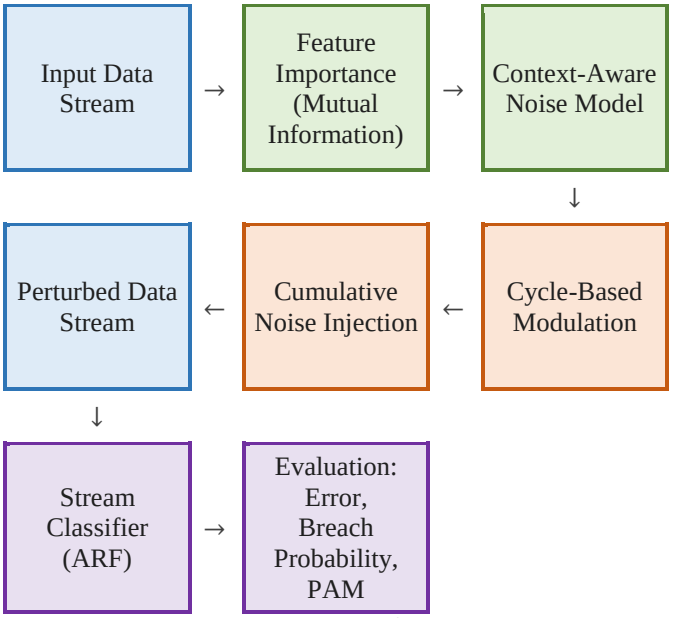

Fig. 1 Proposed CACN framework

5.2. Feature Importance Estimation

CACN allocates noise to each attribute according to its importance to the classification task. Feature importance is estimated using Mutual Information (MI) that measures the extent to which knowledge of a feature X reduces uncertainty regarding the target class Y. The higher mutual information, the more informative it is for classification. Formally,

$$MI(Xi,Y) = \sum_{x \in X,} \sum_{y \in Y,} P(X,Y) \log\left(\frac{P(x,y)}{P(x)P(y)}\right) \quad (2)$$

Noise is then allocated inversely proportional to a feature's importance. It means highly important features receive lower noise, and less important features receive higher noise. So that discriminative attributes are preserved for accurate classification while less-relevant attributes are distorted heavily for privacy.

5.3. Context-Aware Noise Model

This strategy retains the most important information for accurate classification while concentrating distortion on less important features to maintain privacy. As a result, the proposed method aims to preserve data utility with greater robustness against reconstruction attacks, improving the accuracy–privacy trade-off relative to uniform noise allocation.

5.4. Cycle-based Modulation

As in cycle-wise cumulative noise addition, a linear modulation function adaptively adjusts perturbation strength within each cycle. The modulation factor increases linearly from zero to a maximum value across the first half of the cycle, then falls linearly back to zero across the second half. This behavior balances privacy and utility by varying the noise level instead of applying a constant amount of noise throughout the cycle:

$$\lambda(p_t) = \begin{cases} \left(\frac{2\lambda_{max}}{C}\right) p_t, & 0 \le p_t < \frac{C}{2} \\ 2\lambda_{max}\left(1 - \frac{p_t}{C}\right), & \frac{C}{2} \le p_t < C \end{cases} \quad (3)$$

Where $p_t = t \bmod C$ is the position of record t within the current cycle, C is the cycle size, and λ_{max} is the maximum noise level. The time-dependent noise scale of feature i at record t is then given by Equation 4.

$$\sigma_{t,i} = \lambda(p_t)\sigma_i \quad (4)$$

5.5. Cumulative Noise Injection

Noise is added cumulatively, incrementally sampling from a Gaussian distribution to emulate a streaming environment:

$$Y = X + N(\mu, \sigma^2) \quad (5)$$

Where N(μ, σ²) is Gaussian noise with mean μ and standard deviation σ, X is the original data set, and Y is the perturbed data set.

5.6. Algorithm

CACN first computes feature importance and uses it to assign a per-feature noise variance. For each incoming stream record, cycle-based modulation then controls the scale of the noise actually injected. The perturbed stream is classified under prequential evaluation, privacy is measured with a known input–output attack, and PAM is computed to summarize the resulting trade-off. The full procedure is given in Algorithm 1.

6. Experimental Setup

6.1. Data Sets

Two real-world streaming data sets from the UCI repository were used in the experiments:

Table 1. Dataset

Dataset	#Attributes	#Instances	#Classes
AReM	6	≈41,000	5
Taxi	7	≈50,000	3

Activity Recognition system based on Multisensor data (AReM) consists of multivariate time series data which is fetched from the three wireless sensors corresponding to human activities. AReM is commonly used to evaluate activity-recognition systems in streaming and privacy-preserving settings. The Taxi data set contains taxi trip duration records.

Algorithm 1. Context-Aware Cumulative Noise Addition (CACN)

Input: original data stream D, cycle size C
Output: classification error, breach probability, PAM
1. For each feature $i = 1, \ldots, m$:
   compute importance s_i (mutual information with the target class)
2. Compute the noise standard deviation of each feature:
   $\sigma_i = \sigma \times (1 - s_i)$
3. For each incoming record X_t in stream D:
   cycle position: p_t = record_count mod C
   modulation factor $\lambda(p_t)$:
   $\lambda(p_t) = (2\lambda_{max} / C) \times p_t$ if $0 \le p_t < C/2$
   $\lambda(p_t) = 2\lambda_{max} \times (1 - p_t/C)$ if $C/2 \le p_t < C$
   noise scale: $\sigma_{t,i} = \lambda_{t,i} \times \sigma_i$
   Gaussian noise: $n_i = N(0, \sigma_i)$
   cumulative noise: cumulative_noise += mean(n_i)
   perturbed record: $Y_t = X_t$ + cumulative_noise

4. Repeat for entire stream
5. Output perturbed data stream

6.2. Classification Algorithm

Adaptive Random Forest (ARF) [40] is an ensemble-based learning algorithm designed for data stream classification. ARF is used under prequential evaluation. ARF predicts incoming stream by majority vote across all trees, improving accuracy and robustness.

6.3. Evaluation Metrics

Performance is evaluated with three metrics. Classification error is the measurement of the proportion of incorrectly classified instances. Breach probability is computed via a known input–output attack, which refers to probability that a privacy breach will occur described in Section 3.2.

PAM combines classification error and breach probability into a single trade-off score. PAM with lower values indicates a better balance between privacy and accuracy.

6.4. Parameters

Noise is generated from a Gaussian distribution, the standard choice in privacy-preserving data mining owing to its well-understood statistical properties. Experiments were conducted at four cycle sizes: 300, 600, 1000, and 2000 records.

7. Results and Discussion

7.1. Overall Performance

Tables 2 and 3 show CACN's classification error, breach probability, and PAM across all four cycle sizes for the AReM and Taxi data sets, respectively, also visualized in Figures 2 and 3.

Table 2. Overall performance of CACN on the AReM data set

Cycle Size	Classification Error	Breach Probability	PAM
300	0.119	0.463	0.220
600	0.115	0.463	0.228
1000	0.115	0.420	0.190
2000	0.121	0.470	0.238

Table 3. Overall performance of CACN on the Taxi dataset

Cycle Size	Classification Error	Breach Probability	PAM
300	0.315	0.593	0.483
600	0.315	0.635	0.502
1000	0.313	0.658	0.534
2000	0.313	0.482	0.362

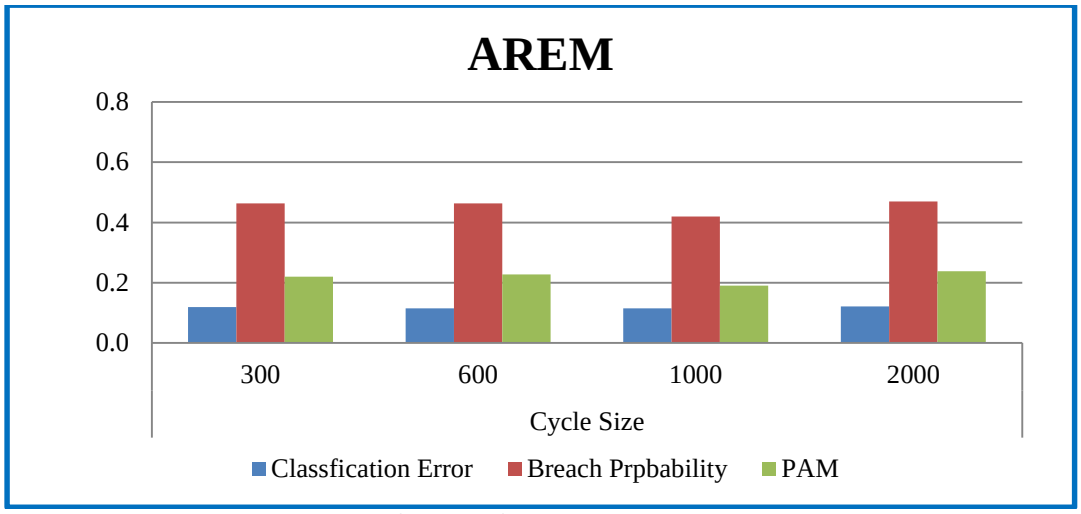

Fig. 2 Performance of CACN on the AReM data set

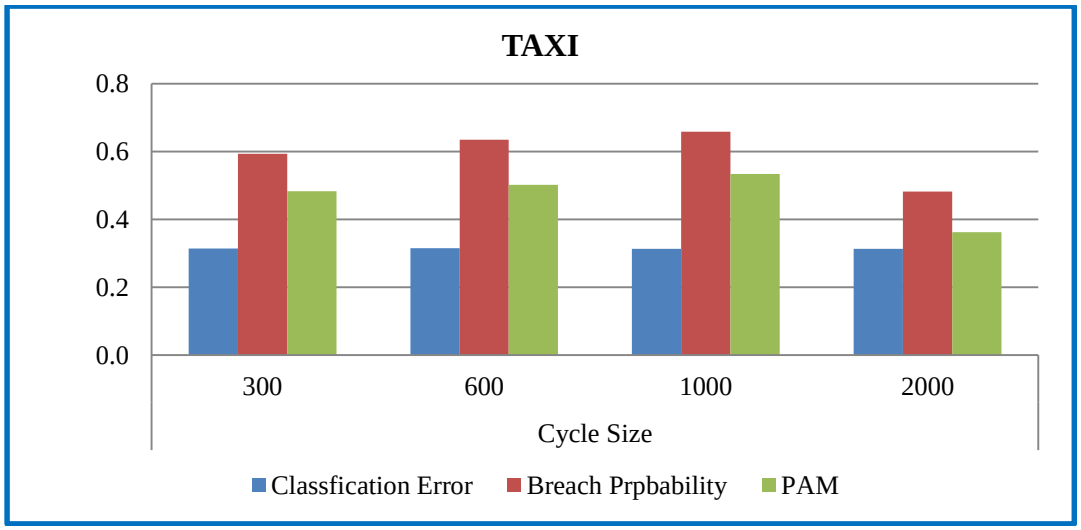

Fig. 3 Performance of CACN on the Taxi data set

On AReM, CACN reduces classification error and keeps breach probability comparable to the baseline, yielding a lower PAM.

On Taxi, CACN still reduces classification error, but breach probability rises relative to the baseline. Section 7.3 examines this contrast in detail.

7.2. Ablation Study: CACN vs. Baseline Cumulative Noise

To isolate the contribution of context-aware noise allocation, CACN is compared directly with the baseline cumulative noise addition method, the SRW (Logistic Random Walk without Resetting) variant identified as the best-performing method by Hewage et al. [25], under the same experimental configuration. Table 5 and Table 6 report this ablation for AReM and Taxi, respectively.

Table 4. Ablation study: CACN vs. baseline cumulative noise addition (AReM data set)

Cycle Size	Method	Classification Error	Breach Probability	PAM
300	Baseline (SRW)	0.161	0.523	0.308
	CACN	0.119	0.463	0.220
600	Baseline (SRW)	0.181	0.650	0.475
	CACN	0.115	0.463	0.134
1000	Baseline (SRW)	0.166	0.573	0.400
	CACN	0.115	0.420	0.190
2000	Baseline (SRW)	0.170	0.534	0.317
	CACN	0.121	0.470	0.238

Table 5. Ablation study: CACN vs. baseline cumulative noise addition (Taxi dataset)

Cycle Size	Method	Classification Error	Breach Probability	PAM
300	Baseline (SRW)	0.401	0.481	0.408
	CACN	0.315	0.593	0.483
600	Baseline (SRW)	0.406	0.397	0.375
	CACN	0.315	0.635	0.502
1000	Baseline (SRW)	0.386	0.288	0.244
	CACN	0.313	0.658	0.534
2000	Baseline (SRW)	0.382	0.335	0.269
	CACN	0.313	0.482	0.362

On AReM (Table 4), CACN reduces classification error at every cycle size relative to SRW at cycle sizes 300, 600, 1000 and 2000. Breach probability also falls at every cycle size, and PAM is lower throughout. When features are only weakly correlated, giving less noise to important features and more noise to less important ones therefore improves utility without weakening privacy, because an attacker cannot use one feature to infer another.

On Taxi (Table 5), CACN again reduces classification error at every cycle size relative to SRW, but breach probability is higher than that of the baseline, so PAM is higher for CACN than for the uniform baseline on this data set. This is attributed to strong correlation among the Taxi features. CACN places the least noise on the most important features. Because the heavily perturbed features are correlated with the lightly perturbed ones, an attacker can use this to reconstruct the data, and records are recovered more successfully. This agrees with earlier findings that correlation among attributes makes the reconstruction of perturbed data easier [4, 22, 27].

7.3. Comparison with SRW across Cycle Sizes

Figures 4 and 5 visualize the ablation results in Table 4 and Table 5 across all four cycle sizes, comparing (a) classification error, (b) breach probability, (c) PAM, and (d) the resulting privacy–accuracy trade-off for the AReM and Taxi data sets, respectively.

Overall, CACN achieves lower classification error and an improved trade-off on AReM, while on Taxi it improves utility at the cost of increased breach probability, underscoring that dataset characteristics materially affect whether context-aware noise allocation helps or hurts privacy.

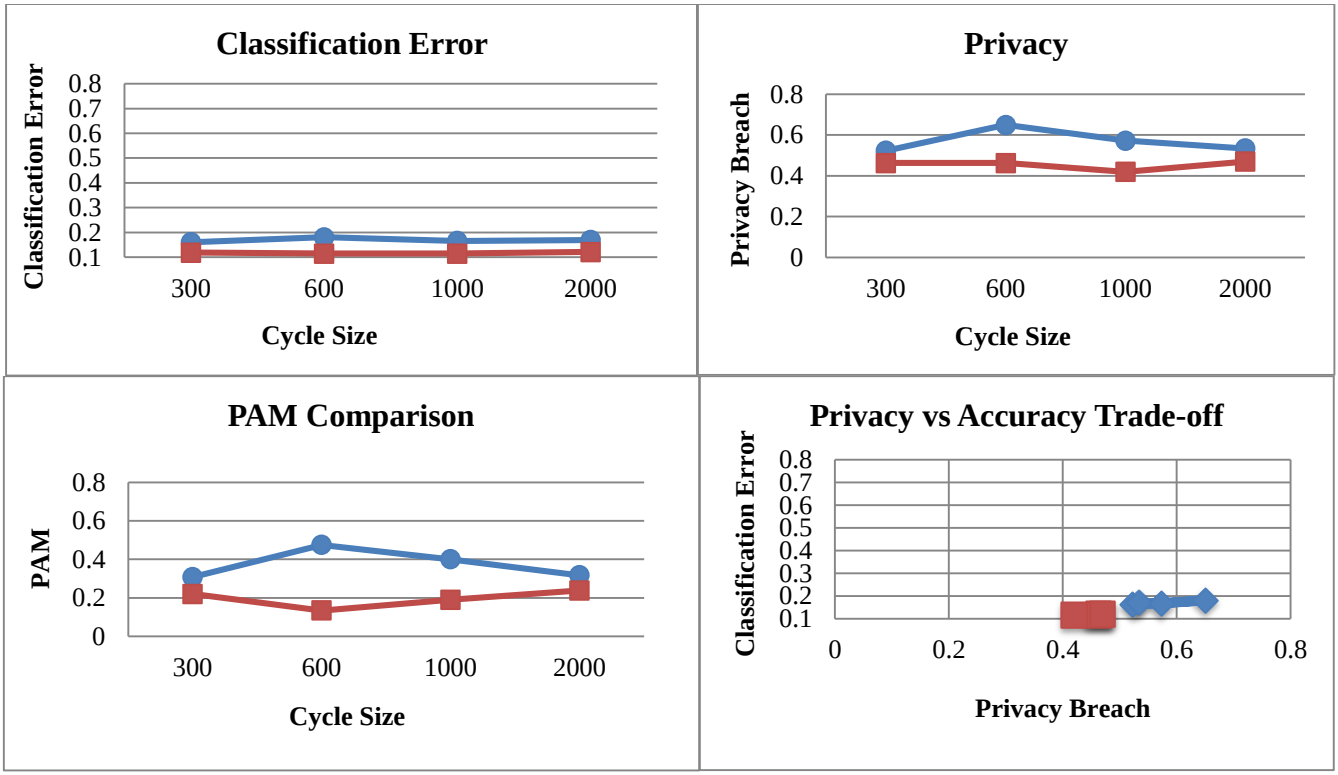

Fig. 4 Comparison of (a) Classification error, (b) Privacy breach, (c) PAM, and (d) Privacy–Accuracy trade-off (AReM dataset).

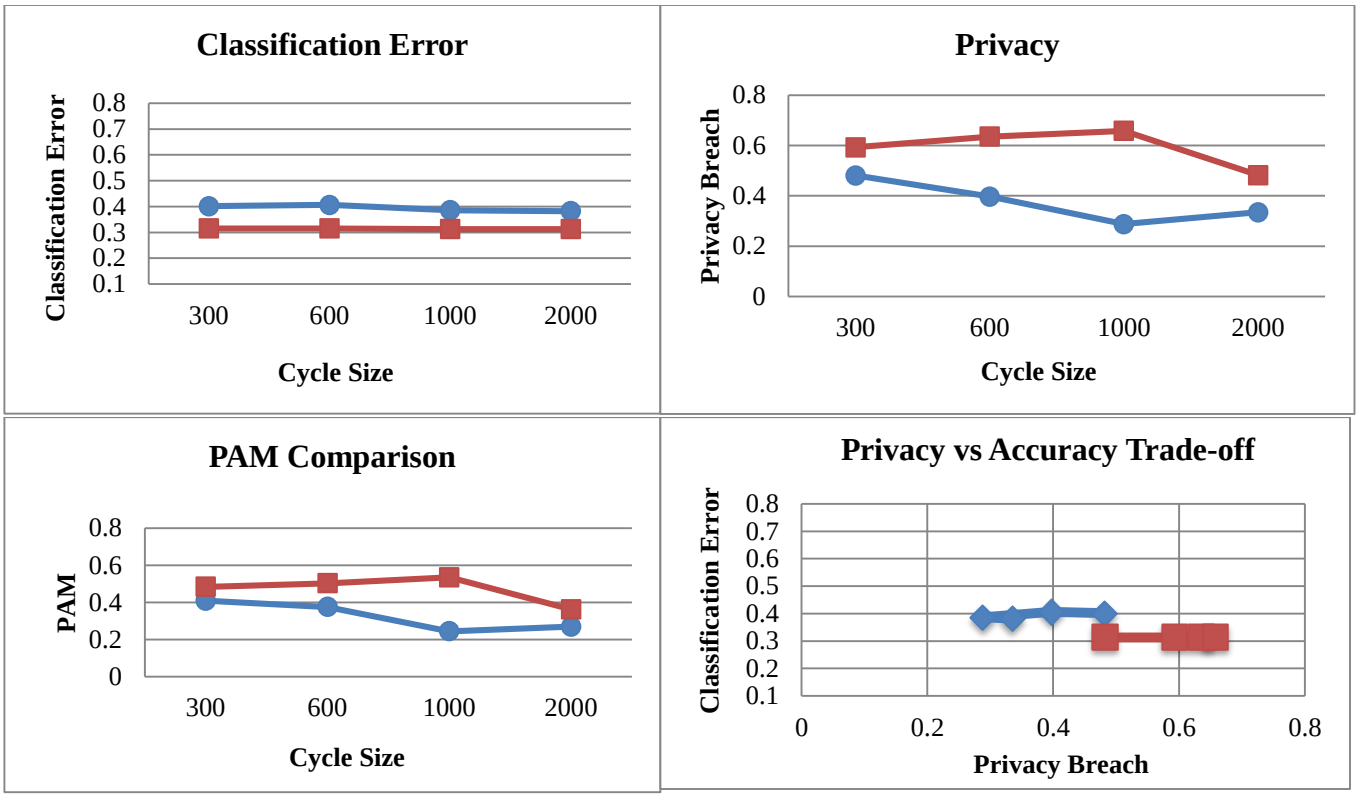

Fig. 5 Comparison of (a) Classification error, (b) Privacy breach, (c) PAM, and (d) Privacy–Accuracy trade-off (Taxi data set).

Table 6. Comparison of CACN with related cumulative-noise methods

Year	Method	Techniques	Objective	Strength	Limitation
2020	Random projection with independent and cumulative noise [15]	Random projection + translation + additive noise	Preserve privacy while maintaining classification accuracy in data streams	Better privacy–accuracy trade-off than basic perturbation	Noise accumulates indefinitely in long streams and can dominate real data
2022	Cumulative noise variations, incl. SRW (baseline) [25]	Random projection + cumulative noise (linear, logistic, cycle-wise)	Control noise growth and improve privacy–accuracy balance	Seven noise-addition variants with cycle-based noise control;	Noise is applied uniformly and is not adaptive to feature importance
2025	Accuracy-Privacy Optimization Framework (APOF) [26]	Logistic cumulative noise + regression	Let users set a target privacy level and predict achievable accuracy	Framework for tuning the privacy–accuracy trade-off	Requires parameter tuning and added computation; perturbation itself remains context-independent
Proposed	Context-Aware Cumulative Noise (CACN)	Cumulative noise + mutual-information-based context-aware adaptation	Allocate noise by feature importance to improve the privacy–accuracy balance	Improved accuracy on both data sets	Breach probability increases on data sets with strongly correlated features

7.4. How Balance is Achieved

Three properties explain the gains of CACN over SRW and the other cumulative noise methods in Table 6. First, uniform noise distorts every feature equally, regardless of its contribution to classification. Under the same amount of noise, the features that carry most of the class information lose as much signal as uninformative ones. CACN moves noise away from these features, so the split decisions of the

Hoeffding trees inside ARF are made on attributes that remain close to their original values.

This is why classification error falls on both data sets. Second, privacy against a known input–output attack depends on how well the attacker can get original attributes. When attributes are weakly correlated, the heavily perturbed low-importance attributes cannot be predicted from the lightly perturbed ones, so their larger noise adds to privacy and breach probability drops.

Third, the cumulative mechanism with cycle-based modulation is kept unchanged, so CACN gives same protection as cumulative noise against independent noise filtering [15], while changing only how the noise is distributed across attributes.

7.5. Comparison with State-of-the-Art Methods

Table 6 shows CACN relative to the cumulative-noise line of work it builds on, summarizing each method's technique, objective, strengths, and limitations, on the same experimental protocol. Earlier methods reduced the cost of noise by changing how much noise accumulates over time [15, 25] or by predicting the accuracy that a privacy target allows [26], and all of them treat features identically.

CACN changes where the noise is placed rather than how much is added over time, which is why it reduces classification error relative to the SRW baseline under the same experimental protocol

7.6. Discussion

The proposed approach has several advantages for privacy preserving stream classification. First, by allocating noise in a context aware manner while preserving data utility, important attributes retain their discriminative property while privacy is still protected.

Second, the mechanism is well-suited to real-time streams, since feature importance and noise allocation are computed incrementally. Third, CACN is fully compatible with standard stream learning algorithms such as ARF and integrates easily into existing stream-mining pipelines.

However, the approach has some limitations. It is sensitive to correlation among features. In the Taxi dataset, strong correlation allowed an attacker to partially reconstruct heavily perturbed attributes from lightly perturbed ones, increasing breach probability even as classification error improved.

These observations suggest that future noise-allocation mechanisms for stream classification should consider inter-feature correlation, not just individual feature importance, to achieve a robust privacy-accuracy balance across data sets with different correlation structures.

8. Limitations and Future Work

- CACN's noise allocation calculates importance of each feature independently and does not model inter-feature correlation. The Taxi results show this can increase breach probability on data sets with strongly correlated attributes. Extending the framework to allocate noise based on joint (correlation-aware) feature relevance is identified as future work.
- Results are reported for a linear, cycle-based noise modulation. Incorporating the logistic modulation shown to perform best for uniform cumulative noise addition into the context-aware setting is planned as the next step.

9. Conclusion

This paper proposed a Context-Aware Cumulative Noise (CACN), a framework for privacy preserving data stream classification in which noise is adaptively allocated according to feature importance rather than applied uniformly across all attributes. CACN estimates feature importance via mutual information and applies cumulative Gaussian noise with cycle-based modulation. Cumulative noise bounds long-run noise growth.

Experimental results show that CACN improves the privacy-accuracy trade-off relative to baseline cumulative noise addition on the AReM dataset, consistently reducing both classification error and PAM across all cycle sizes tested. On the Taxi data set, CACN still reduces classification error but increases breach probability because of strong feature correlation, which highlights that inter feature correlation is an important factor for privacy preserving stream mining. Future work will focus on incorporating logistic noise modulation and correlation-aware noise allocation to address this limitation.

Data Availability Statement

The AReM data set is publicly available from the UCI Machine Learning Repository (https://archive.ics.uci.edu/dataset/366/activity+recognition+system+based+on+multisensor+data+fusion+arem, doi: 10.24432/C5SS33).

The Taxi dataset is publicly available from the UCI Machine Learning Repository (https://archive.ics.uci.edu/dataset/339/taxi+service+trajectory+prediction+challenge+ecml+pkdd+2015,doi: 10.24432/C55W25).

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

[1] Liming Li, and Qishan Zhang, "A Privacy Preserving Clustering Technique using Hybrid Data Transformation Method," *2009 IEEE International Conference on Grey Systems and Intelligent Services (GSIS 2009)*, Nanjing, China, pp. 1502-1506, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[2] Rakesh Agrawal, and Ramakrishnan Srikant, "Privacy-Preserving Data Mining," *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 439-450, 2000. [CrossRef] [Google Scholar] [Publisher Link]

[3] S. Shimona, "Survey on Privacy Preservation Technique," *2020 International Conference on Inventive Computation Technologies (ICICT)*, Coimbatore, India, pp. 64-68, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[4] H. Kargupta et al., "On the Privacy Preserving Properties of Random Data Perturbation Techniques," *Third IEEE International Conference on Data Mining*, Melbourne, FL, USA, pp. 99-106, 2003. [CrossRef] [Google Scholar] [Publisher Link]

[5] Xueyun Li, Zheng Yan, and Peng Zhang, "A Review on Privacy-Preserving Data Mining," *2014 IEEE International Conference on Computer and Information Technology*, Xi'an, China, pp. 769-774, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[6] P. Ram Mohan Rao, S. Murali Krishna, and A.P. Siva Kumar, "Privacy Preservation Techniques in Big Data Analytics: A Survey," *Journal of Big Data*, vol. 5, pp. 1-12, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[7] Alfredo Cuzzocrea, "Privacy-Preserving Big Data Stream Mining: Opportunities, Challenges, Directions," *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, New Orleans, LA, USA, pp. 992-994, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[8] Jerzy Stefanowski, and Dariusz Brzezinski, "Stream Classification," *Encyclopedia of Machine Learning and Data Mining*, pp. 1-9, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[9] João Gama et al., "A Survey on Concept Drift Adaptation," *ACM Computing Surveys*, vol. 46, no. 4, pp. 1-37, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[10] U. H. W. A. Hewage, R. Sinha, and M. Asif Naeem, "Privacy-Preserving Data (Stream) Mining Techniques and their Impact on Data Mining Accuracy: A Systematic Literature Review," *Artificial Intelligence Review*, vol. 56, pp. 10427-10464, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[11] Ching-Ming Chao, Po-Zung Chen, and Chu-Hao Sun, "Privacy-Preserving Classification of Data Streams," *Tamkang Journal of Science and Engineering*, vol. 12, no. 3, pp. 321-330, 2009. [Google Scholar] [Publisher Link]

[12] Kiran Patel, Hitesh Patel, Parin Patel, "Privacy Preserving Data Stream Classification using Data Perturbation Techniques," *Proceedings of the International Conference on Emerging Trends in Electrical, Communication and Information Technologies (ICECIT)*, vol. 2, no. 2, pp. 2711-2714, 2014. [Google Scholar] [Publisher Link]

[13] Radhika Kotecha, and Sanjay Garg, "Preserving Output-Privacy in Data Stream Classification," *Progress in Artificial Intelligence*, vol. 6, no. 2, pp. 87-104, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[14] M.A.P. Chamikara et al., "Efficient Data Perturbation for Privacy Preserving and Accurate Data Stream Mining," *Pervasive and Mobile Computing*, vol. 48, pp. 1-19, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[15] Benjamin Denham, Russel Pears, and M. Asif Naeem, "Enhancing Random Projection with Independent and Cumulative Additive Noise for Privacy-Preserving Data Stream Mining," *Expert Systems with Applications*, vol. 152, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[16] Hayder K. Fatlawi, and Attila Kiss, "Differential Privacy based Classification Model for Mining Medical Data Stream using Adaptive Random Forest," *Acta Universitatis Sapientiae, Informatica*, vol. 13, no. 1, pp. 1-20, 2021. [Google Scholar] [Publisher Link]

[17] Ankit Jasoliya, and Tejal Patel, "Privacy Preserving in Data Stream using Sliding Window Method," *International Education and Research Journal*, vol. 3, no. 5, pp. 1-3, 2017. [Publisher Link]

[18] Nikunj H. Domadiya, and Udai Pratap Rao, "A Hybrid Technique for Hiding Sensitive Association Rules and Maintaining Database Quality," *Proceedings of the First International Conference on Information and Communication Technology for Intelligent Systems*, vol. 51, pp. 359-367, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[19] Jinyan Wang et al., "A Three-Phase Approach to Differentially Private Crucial Patterns Mining over Data Streams," *Computers & Security*, vol. 82, pp. 30-48, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[20] Fang Liu, and Tong Li, "A Clustering-Anonymity Privacy-Preserving Method for Wearable IoT Devices," *Security and Communication Networks*, vol. 2018, no. 1, pp. 1-8, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[21] Jinyan Wang, Chaoji Deng, and Xianxian Li, "Two Privacy-Preserving Approaches for Publishing Transactional Data Streams," *IEEE Access*, vol. 6, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[22] Feifei Li et al., "Hiding in the Crowd: Privacy Preservation on Evolving Streams through Correlation Tracking," *2007 IEEE 23rd International Conference on Data Engineering*, Istanbul, Turkey, pp. 686-695, 2007. [CrossRef] [Google Scholar] [Publisher Link]

[23] Ganesh Dagadu Puri, and D. Haritha, "A Novel Method for Privacy Preservation of Health Data Stream Publishing," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 4, pp. 4959-4963, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[24] Bin Zhou et al., "Continuous Privacy Preserving Publishing of Data Streams," *Proceedings of the 12th International Conference on Extending Database Technology*, pp. 648-659, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[25] Ullusu Hewage Waruni Amali Hewage, Russel Pears, and M. Asif Naeem, "Optimizing the Trade-off between Classification Accuracy and Data Privacy in the Area of Data Stream Mining," *International Journal of Artificial Intelligence*, vol. 20, no. 1, pp. 147-167, 2022. [Google Scholar] [Publisher Link]

[26] Waruni Hewage, R. Sinha, and M. Asif Naeem, "An Accuracy-Privacy Optimization Framework Considering User's Privacy Requirements for Data Stream Mining," *Journal of Big Data*, vol. 12, pp. 1-32, 2025. [CrossRef] [Google Scholar] [Publisher Link]

[27] Zhengli Huang, Wenliang (Kevin) Du, and Biao Chen, "Deriving Private Information from Randomized Data," *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pp. 37-48, 2005. [CrossRef] [Google Scholar] [Publisher Link]

[28] Keke Chen, and Ling Liu "Privacy Preserving Data Classification with Rotation Perturbation," *Fifth IEEE International Conference on Data Mining (ICDM'05)*, Houston, TX, USA, pp. 589-592, 2005. [CrossRef] [Google Scholar] [Publisher Link]

[29] Kun Liu, H. Kargupta, and J. Ryan, "Random Projection-based Multiplicative Data Perturbation for Privacy Preserving Distributed Data Mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 18, no. 1, pp. 92-106, 2006. [CrossRef] [Google Scholar] [Publisher Link]

[30] Kun Liu, Chris Giannella, and Hillol Kargupta, "An Attacker's View of Distance Preserving Maps for Privacy Preserving Data Mining," *Proceedings of the European Conference on Principles of Data Mining and Knowledge Discovery (PKDD)*, Berlin, Germany, vol. 4213, pp. 297-308, 2006. [CrossRef] [Google Scholar] [Publisher Link]

[31] Mahbubur Rahman, Mahit Kumar Paul, and A.H.M. Sarowar Sattar, "Efficient Perturbation Techniques for Preserving Privacy of Multivariate Sensitive Data Mining," *Array*, vol. 20, pp. 1-15, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[32] Cynthia Dwork et al., "Calibrating Noise to Sensitivity in Private Data Analysis," *Theory of Cryptography*, vol. 3876, pp. 265-284, 2006. [CrossRef] [Google Scholar] [Publisher Link]

[33] Kiran Patel, Hitesh Patel, and Parin Patel, "Privacy Preserving in Data Stream Classification using Different Proposed Perturbation Methods," *International Journal of Engineering Development and Research*, vol. 2, no. 2, pp. 2711-2714, 2014. [CrossRef] [Publisher Link]

[34] Tusharkumar J. Trambadiya, and Praveen Bhanodia, "A Heuristic Approach to Preserve Privacy in Stream Data with Classification," *International Journal of Engineering Research and Applications*, vol. 3, no. 1, pp. 1096-1103, 2013. [Google Scholar] [Publisher Link]

[35] Sanket P. Modi, and Ashil R. Patel, "Privacy Preserving Data Stream Mining using Two Phase Geometric Data Perturbation," *International Journal for Scientific Research and Development*, vol. 3, no. 1, pp. 1-4, 2015. [Google Scholar] [Publisher Link]

[36] Neha Gupta, and Indrjeet Rajput, "Preserving Privacy using Data Perturbation in Data Stream," *International Journal of Advanced Research in Computer Engineering and Technology*, 2013. [Google Scholar]

[37] P. Rajendra Prasada, and Tryambak Hirwarkar, "Efficient Model for Privacy Preserving Classification of Data Streams," *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 2, pp. 1475-1481, 2021. [Google Scholar] [Publisher Link]

[38] Gayathri Devi N, and Manikandan K, "Improved Perturbation Technique: Privacy-Preserving Rotation-based Condensation Algorithm for Privacy Preserving in Big Data Stream Using Internet of Things," *Transactions on Emerging Telecommunications Technologies*, vol. 31, no. 12, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[39] Arvind Kumar, Parminder Kaur, and Pratibha Sharma, "A Survey on Hoeffding Tree Stream Data Classification Algorithms," *CPUH-Research Journal*, 2015. [Google Scholar]

[40] Heitor M. Gomes et al., "Adaptive Random Forests for Evolving Data Stream Classification," *Machine Learning*, vol. 106, no. 9, pp. 1469-1495, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[41] Loredana Caruccio et al., "An Approach to Trade-off Privacy and Classification Accuracy in Machine Learning Processes," *CEUR Workshop Proceedings*, pp. 420-429, 2023. [Google Scholar] [Publisher Link]

[42] Alessandra Calvi, Gianclaudio Malgieri, and Dimitris Kotzinos, "The Unfair Side of Privacy Enhancing Technologies: Addressing the Trade-offs between PETs and Fairness," *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*, pp. 2047-2059, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[43] Kun Liu, "*Multiplicative Data Perturbation for Privacy Preserving Data Mining*," Ph.D. Thesis, University of Maryland, Baltimore County, 2007. [Google Scholar] [Publisher Link]