International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 November 2023*

Modified EPPXGBOOST for Effective Data Stream Mining in Cloud

Aniket Patel[1]*, Dr. Kiran Amin[2]
[1]Department of Computer Science and Engineering
Ganpat University
Mehsana,India
aniketpatel.it@gmail.com

[2]Department of Computer Engineering
Ganpat University
Mehsana,India
kiran.amin@ganpatuniversity.ac.in

*Corresponding Author:aniketpatel.it@gmail.com

Abstract— In today's technology-driven landscape, the perva- sive use of online services across diverse domains has led to the generation of vast datasets, necessitating advanced data mining techniques for meaningful insights. The advent of data streams, characterized by continuous and dynamic data flows, presents a significant challenge, prompting the evolution of data stream mining. This field addresses issues such as rapid changes in streaming data and the need for quick algorithms. To tackle these challenges, an innovative approach named (Effective Privacy Preserving eXtreme Gradient Boosting) EPPXGBOOST is proposed, combining Adaptive XGBOOST for continuous learning from evolving data streams with PPXGBOOST for privacy preservation.

Keywords- Data mining, Privacy preservation, Adaptive XGBOOST, Data privacy, Machine learning, EPPXGBOOST

I. INTRODUCTION

In today's technology-driven era, where the digitization of various aspects of life is prevalent, the significance of privacy preservation in data mining and machine learning cannot be overstated [1]. As vast amounts of personal, sensitive, and business-related information are processed and analyzed, the need to safeguard individuals' privacy has become a paramount concern. Ensuring the confidentiality and integrity of data is not only an ethical imperative but also a legal requirement in many jurisdictions [2]. Moreover, the escalating frequency and sophistication of cyber threats under- score the urgency of implementing robust privacy-preserving techniques. However, integrating effective privacy measures becomes particularly challenging when dealing with dynamic and continuous data streams. The rapid evolution of information in scenarios such as online transactions, healthcare records, and communication logs necessitates innovative solutions to protect privacy while extracting meaningful insights [3].

The primary justification for choosing Adaptive XGBOOST lies in its effectiveness for data streaming [4]. An emerging trend in Machine Learning (ML) involves adapting to evolving data streams, offering a compelling alternative to traditional batch learning across various scenarios. For instance, consider the application of fraud detection in online banking, where model training involves massive datasets. The critical factor in such cases is the runtime, as a prolonged training period may allow potential frauds to go undetected. Similarly, in the analysis of communication logs for security, storing all logs becomes impractical and often unnecessary. This constraint poses a significant limitation for methods requiring multiple passes over the data. Therefore, the selection of Adaptive XGBOOST is driven by its ability to address the challenges posed by dynamic data streams in real-time applications [5]. Dealing with data streaming in ML poses specific chal- lenges. First, models have a single opportunity to access the data and must process it continuously as new information arrives [6]. This real-time processing requirement is crucial to keep up with the dynamic nature of streaming data. Sec- ond, models need to be capable of providing predictions at any given moment, emphasizing the need for efficiency and responsiveness. Lastly, the presence of concept drift introduces another layer of complexity. Concept drift refers to the po tential change in the relationship between features and learn- ing targets, a common occurrence in real-world applications that aim to model dynamic systems [6].

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 November 2023*

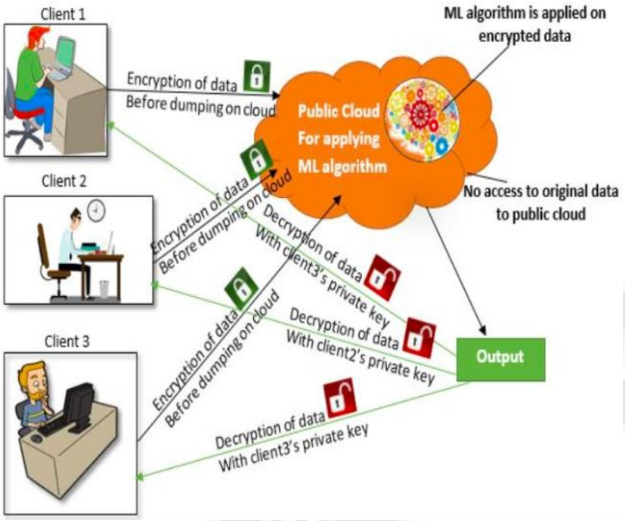

Fig. 1. PPML Concept

To address this challenge, a common strategy involves replacing the model when concept drift is detected, typically indicated by the decline in performance of a batch model. However, this approach demands substantial resources for data collection, processing, model training, and validation. In contrast, stream models offer a more efficient solution as they are continuously updated and can dynamically adapt to evolving concepts without the need for a complete model overhaul. This consideration has led to the adoption of Adaptive XGBOOST to effectively handle the dynamic nature of data streams and mitigate the resource-intensive drawbacks associated with traditional model replacement approaches [7].

In response to these challenges, the utilization of ML models, and specifically, the XGBOOST algorithm, has emerged as a powerful tool in privacy preservation. XGBOOST's capacity for handling large datasets, feature selection, and its capability to address complex relationships within the data make it a popular choice. However, the static nature of traditional XGBOOST poses limitations in scenarios where data streams continuously evolve. This leads to the introduction of EPPXGBoost, an acronym for Effective Privacy Preserving eXtreme Gradient Boosting. EPPXGBoost integrates the adaptability of Adaptive XGBOOST with privacy-preserving mechanisms to navigate the unique challenges posed by dynamic data streams. By combining the strengths of both adaptability and privacy preservation, EPPXGBoost aims to provide a comprehensive solution for secure and efficient data analysis in scenarios where preserving privacy is of paramount importance.

A. Privacy Preserving Machine Learning (PPML)

Privacy-Preserving Machine Learning (PPML) is a system-static approach designed to prevent data leakage within machine learning algorithms. This strategy facilitates the implementation of various privacy-enhancing techniques, allowing multiple input sources to collaboratively train machine learning models without exposing their raw private data [8]. While the accumulation of substantial datasets is crucial for advancing Artificial Intelligence (AI), particularly in Machine Learning (ML), the process of data collection differs significantly from its utilization in predicting behaviors. This distinction poses notable challenges to the privacy of individuals and organizations, introducing risks such as breaches of data privacy that can result in both financial losses and damage to reputation.

A significant portion of privacy-sensitive data analysis, including search algorithms, recommender systems, and advertising technology networks, relies on machine learning techniques. The primary goal of privacy-preserving machine learning is to bridge the gap between safeguarding privacy and leveraging the benefits offered by machine learning methodologies. This is crucial for ensuring the secure handling of acquired data and compliance with data privacy regulations. The overarching concept behind PPML is outlined in Figure 1.

B. Motivation And Contribution

The motivation for this paper stems from the critical need to address the challenges posed by dynamic data streams in the realm of machine learning, particularly in scenarios where privacy preservation is paramount. The ever-increasing reliance on online services and the generation of vast datasets demand innovative solutions that not only effectively mine valuable insights but also ensure the privacy of sensitive information. The paper aims to contribute by proposing a novel approach, EPPXGBoost (Effective Privacy Preserving eXtreme Gradient Boosting), which combines the adaptability of Adaptive XGBOOST with privacy-preserving techniques. The central motivation is to offer a comprehensive solution that efficiently handles data stream mining, providing timely and accurate results while safeguarding privacy. The proposed algorithm addresses the limitations of traditional models, especially in scenarios such as fraud detection in online banking and security analysis of communication logs, where waiting for model training on massive datasets is impractical and potential threats may go undetected. The contribution of this paper lies in introducing an integrated approach that not only adapts to the dynamic nature of data streams through Adaptive XGBOOST but also prioritizes privacy preservation through the incorporation of privacy- preserving techniques. By striking a balance between accuracy, adaptability, and privacy, the proposed EPPXGBoost algorithm aims to open new horizons for secure and efficient data analysis in domains where data privacy is a critical concern, such as healthcare, financial analysis, and government sectors. The paper is organized as follows: Section 2 provides a review of related work in the domain of privacy preservation. Section 3 details the proposed approach. Section 4 conducts an analysis of experimental results, and finally, Section 5 concludes the paper.

II. RELATED WORK

In the context of rapid advancements in mobile technology, there is a growing prevalence of continuous streams of spatio-temporal data, often automatically collected by various data holders. The study presented in [9] introduces an innovative technique aimed at anonymizing trajectories generated by individuals in these data streams. This approach prioritizes the timely release of anonymized trajectories while incorporating a freshness element. The study addresses two distinct privacy threats through a formalized approach and introduces an algorithm designed to incrementally anonymize dynamically-updated sliding windows within the sequence. The window structure efficiently accommodates substantial data volumes.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 November 2023*

Through evaluations using both simulated and real-life datasets, the study compares the performance of the proposed method with existing approaches, highlighting its effectiveness in anonymizing high-volume trajectory streams and demonstrating superior results [9]. In the domain of distributed mobile phone networks, the pa- per [10] introduces a method for privacy-preserving data sharing. This approach, termed shadow coding, employs shadow matrix computation to efficiently address the challenges associated with privacy concerns. The research demonstrates the feasibility of the proposed method through experiments with real-life datasets and the implementation of a pilot system within a city for distributed mobile phone data collection [10]. Future efforts in this research endeavor include the exploration of attack models, asynchronous distributed environments, and diverse variations of shadow matrices to cater to various privacy requirements. The paper also emphasizes the importance of enhancing formal security analysis and delving into data demander anonymity in the context of distributed data collection as promising directions for further research [10]. In [11], a privacy-preserving scheme designed to efficiently handle large-scale streaming categorical data is introduced. This scheme utilizes an innovative anonymization method for both candidates and categorical information, ensuring rapid processing and minimal communication overhead. The approach effectively protects data privacy against potential tampering by adversaries, as validated through empirical evaluation [11]. Looking ahead, the intention is to explore enhancements to the technique. One avenue of exploration involves investigating methods to reduce the number of dummy votes while maintaining privacy levels, with the aim of improving communication performance [11]. Additionally, plans include the development of algorithms capable of distinguishing tampered votes from dummy votes, enabling the identification of instances of information tampering and facilitating timely actions. The current design evenly distributes transmission workload across different channels to the central decoder, and future efforts will focus on optimizing this distribution based on channel risk assessments to enhance transmission efficiency [11]. In the current era dominated by advanced smart devices, the substantial increase in data volumes has triggered significant apprehensions regarding privacy and security. The act of sharing sensitive personal data with external entities exposes it to potential misuse, necessitating the development of robust solutions. To tackle this challenge, the paper introduces a blockchain-based system that leverages Hyperledger Iroha, aiming to provide users with data ownership, as outlined in [12]. In this innovative framework, users willingly share their personal information with service providers and, in return, receive coins as incentives. Crucially, the collected data is shared only based on permissions granted by the users. The mobility data, encompassing elements such as location, time, and name, has promising potential for trajectory mining tasks. Specifically, the paper focuses on privacy-preserving group mobility trajectory mining, involving the preprocessing of trajectory data into spatiotemporal regions and the extraction of frequent trajectories [12]. Comprehensive experiments affirm the scalability and relevance of the proposed system. Despite these achievements, various avenues for further exploration exist. For example, conducting a comparative analysis between decentralized and centralized systems under similar conditions could offer valuable insights. Additionally, exploring the delicate balance between privacy preservation and system performance remains an area for potential investigation [12]. In [13], the paper introduces an approach that applies Singular Value Decomposition (SVD) and 3D Rotation Data Perturbation (RDP) to simultaneously achieve the dual objectives of maintaining substantial data utility while preserving the privacy of the dataset. The central process involves utilizing SVD for perturbation, employing matrix decomposition to extract sensitive attributes and eliminate irrelevant information.

Following this, the iterative application of 3D RDP ensures the skewing of all sensitive features across different axes, thereby enhancing privacy protection. The perturbed and original data undergo classification using distinct classifiers, and accuracy is measured as the ratio of correctly classified instances to the total instances in the dataset [13]. Comparative analysis against existing methods demonstrates that the proposed approach offers a better balance between data privacy and utility. For future research endeavors, there is a potential avenue in exploring alternative perturbation techniques to further enhance classification accuracy [13].

The paper in [14] addresses the challenge of preserving privacy in data mining operations, emphasizing the limitations of traditional randomization-based strategies. It advocates for perturbation-based techniques as an effective approach to address privacy concerns that might not be adequately covered by randomization. The effectiveness of this approach is demonstrated through experiments involving diverse data types. Additionally, the research introduces a novel Base encoding technique, enhancing data privacy algorithms [14]. By incorporating both randomness and perturbation, the method modifies data to safeguard personal and sensitive information, showcasing its effectiveness on datasets featuring both continuous and categorical variables. The resulting data transformation allows for meaningful data mining while upholding data integrity [14]. In summary, the proposed method proves successful in preserving both data privacy and quality, offering valuable insights applicable to various domains such as purchase behavior, criminal records, medical history, and credit reports [14].

III. PROPOSED METHODOLOGY

A. Problem Formulation

The issue we tackle revolves around the classification of evolving data streams, where the association between features and classes may change over time due to concept drift. Traditional batch models prove insufficient in handling such scenarios, often requiring significant resources for replacement. Our proposed algorithm employs an ensemble approach to address this challenge. Figure 2 illustrates the identification of the problem.

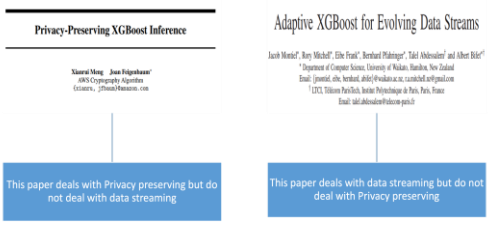

Fig. 2. Problem Identification

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 November 2023*

Boosting stands out as a prominent ensemble method, sequentially combining base models to achieve improved predictive accuracy. Specifically, eXtreme Gradient Boosting (XGB) has gained popularity as a learning algorithm [15]. Our proposal involves an adaptation of XGB designed for the classification of evolving data streams. In this context, where new data continuously emerges, potentially leading to concept drift altering the connection between class and features, our method generates new ensemble members from mini-batches of incoming data while maintaining a fixed maximum ensemble size. Importantly, learning is an ongoing process even after reaching this size, ensuring the ensemble continually updates with new data to align with the prevailing concept. Additionally, we explore the integration of privacy preservation throughout this entire process. The adoption of Adaptive XGBOOST provides distinct ad- vantages, particularly in its compatibility with data streaming. This approach adeptly responds to the nuances of learning from evolving data streams, offering an attractive alternative to traditional batch learning in various scenarios. For instance, it proves invaluable in domains like fraud detection within online banking operations, where prompt model training is imperative to prevent potential frauds from going undetected. Similarly, in the security analysis of communication logs, where retaining all logs is impractical and unnecessary, stream models overcome the limitations of methods requiring multiple passes over the data. Managing data streams in machine learning poses several challenges:

- ML models have one-time access to data, necessitating realtime processing as new data continuously streams in [16].
- Immediate predictions are demanded from machine learning models at any given moment.
- The potential occurrence of concept drift, wherein the relationship between features and learning targets evolves, presents a formidable challenge. This phenomenon is prevalent in real-world applications aimed at modeling dynamic systems.

The proposed algorithm was implemented using the Python programming language, utilizing Spyder as the Integrated Development Environment (IDE). Spyder, an open-source sci- entific environment tailored for Python, proved to be a robust choice, offering advanced editing, analysis, debugging, and profiling functionalities essential for scientific endeavors. The unique amalgamation of features in Spyder, including data exploration, interactive code execution, variable inspection, and visualization creation, made it particularly suitable for researchers and professionals in various fields. The imple- mentation process involved translating algorithmic concepts into Python code snippets, leveraging a variety of Python libraries for tasks such as data manipulation, machine learning, encryption, and visualization. Python's versatility and user- friendly syntax, coupled with Spyder's optimization for scientific tasks, facilitated the coding process. The debugging and profiling features of Spyder ensured the accuracy and reliability of the algorithm by identifying and resolving code issues. In summary, the implementation centered on Python libraries within the Spyder environment, aligning the code with the proposed methodology for privacy-preserving data stream analysis.

In executing our privacy-preserving algorithm, we employed three diverse datasets to illustrate its effectiveness across vari- ous contexts and data formats. The chosen datasets encompass the Amazon Product Reviews dataset1, the Titanic dataset2 real-time processing as new data continuously streams in [16].

- Immediate predictions are demanded from machine learning models at any given moment.
- The potential occurrence of concept drift, wherein the relationship between features and learning targets evolves, presents a formidable challenge. This phenomenon is

sourced from the Kaggle competition" Titanic: Machine Learning from Disaster," and the extensive US Census dataset3 provided by the United States Census Bureau. Each dataset presents distinct characteristics and complexities, affording us the opportunity to showcase the algorithm's capability in preserving sensitive information while successfully conducting crucial data stream mining and analysis activities prevalent in real-world applications aimed at modeling dynamic systems.

In handling concept drift, traditional practice involves replacing a model with a new one once degradation becomes evident, entailing substantial resource investments for data collection, processing, model training, and validation. In contrast, stream models exhibit continuous adaptation by updating to accommodate the evolving concept. This key attribute underscores our choice to work with Adaptive XGBOOST.

In summary, our proposed adaptation of XGB addresses the demands of evolving data streams, tackling challenges posed by continuous data influx, concept drift, and privacy preservation. Leveraging the strengths of Adaptive XGBOOST and stream modeling, our approach offers effective and dynamic solutions in privacy-conscious data stream scenarios.

The problem at the core of our study revolves around the classification of evolving data streams, where the dynamic nature of the relationship between features and classes is subject to change due to concept drift. Traditional batch models struggle to cope with such scenarios, frequently necessitating significant resources for replacement. In response to this challenge, our proposed algorithm adopts an ensemble approach as a strategic solution.

Algorithm 1 Modified Privacy-Preserving EPPXGBOOST Algorithm

Require: Data from various clients
Ensure: Results after Modified EPPXGBOOST
1: Aggregate data from various clients
2: Encrypt the aggregated data with Modified Homomorphic Encryption
3: Perform Adaptive XGBOOST Machine Learning Algo- rithm on the encrypted data
4: Note the results
5: Decrypt the data at the client end when necessary
=0

IV. RESULT ANALYSIS

This section offers a detailed analysis of the results obtained from the proposed approaches. First, the performance of Adap- tive XGBOOST is presented across three diverse datasets:

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 November 2023*

[1]https://www.kaggle.com/datasets/yasserh/amazon-product-reviews-dataset
[2]https://www.kaggle.com /competitions/titanic/data
[3]https://www.census.gov/data/datasets.html

Amazon, Titanic, and US Census. Each dataset possesses dis- tinct characteristics, including varying feature counts (ranging from 9 to 36), diverse class categories (ranging from 2 to 5 classes), and ensemble sizes set at either 8 or 16 within a single ensemble in Table I. Although the maximum and minimum weight parameters (Wmax and Wmin) are specified for each dataset, their contextual explanations remain unclear. Notably, the datasets demonstrate commendable classification accuracy, with rates spanning from 94% to 96%, underscoring the effectiveness of the applied models or classifiers. The graphical representation of these results is depicted in Figure 3.

TABLE I
ADAPTIVE XGBOOST PERFORMANCE

Dataset	Amazon	Titanic	US Census
Features	9	11	36
Classes	5	2	2
Ensemble size (E(s)), here s=1	16	8	8
Wmax	1024	256	512
Wmin	8	4	4
Accuracy	95%	96%	94%

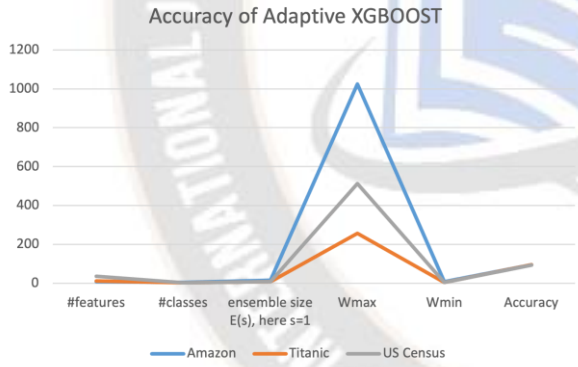

Fig. 3. Adaptive XGBOOST Result Analysis

Table II details the performance of PPXGBOOST across three datasets, featuring varying numbers of features (ranging from 9 to 36) and classes (ranging from 2 to 5). The analysis involves an ensemble approach with an ensemble size (E) set at 8 or 16, while s=1 implies a single ensemble. Notably, parameters denoted as Wmax and Wmin, presumed to be weight-related factors, lack specific contextual explanations. Despite these variations, all three datasets exhibit remarkable accuracy, ranging from 97% to 98%. This underscores the ex- ceptional performance of the employed PPXGBOOST models or classifiers in terms of predictive accuracy across diverse datasets. Figure 4 visually presents these results.

Analyzing the Titanic dataset revealed consistent patterns. The algorithm maintained a maximum decision tree depth of 5 and a learning rate of 0.01, initiating with an ensemble size of 8. Effective management of streaming data dynamics was achieved with sliding window sizes set at Wmax = 256

TABLE II
PPXGBOOST PERFORMANCE ANALYSIS

Dataset	Amazon	Titanic	US Census
Features	9	11	36
Classes	5	2	2
Ensemble size (E(s)), here s=1	16	8	8
Wmax	1024	256	512
Wmin	8	4	4
Accuracy	97%	98%	98%

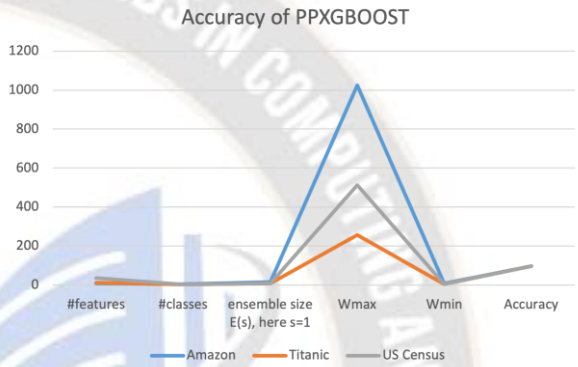

Fig. 4. PPXGBOOST Performance

and Wmin = 4. Impressively, the algorithm exhibited high efficiency, with a query time of 0.52 seconds. Despite the expansion in size to 12 KB compared to the conventional model's 3 KB, the EPPXGBOOST model did not compromise on accuracy. Demonstrating its competence, the algorithm achieved an impressive accuracy rate of 98% in privacy-preserving predictive tasks on the Titanic dataset.

Similarly, the performance trends observed in the US Census dataset were consistently notable. With a maximum decision tree depth of 5, a learning rate of 0.01, and an initial ensemble size of 8, the algorithm effectively managed streaming data input using sliding window sizes of Wmax= 512 and Wmin = 4. The algorithm showcased agility with a query time of approximately 0.81 seconds. While the EPPXGBOOST model exhibited a larger size of 2.5 MB com- pared to the conventional model's 210 KB, this trade-off was made for the sake of privacy preservation. Nevertheless, the algorithm maintained a robust accuracy of 97%, reinforcing its potential applicability across diverse datasets. The results presented in Table IV offer a comprehensive overview of the algorithm's performance across different datasets. Notably, on the Amazon dataset, the algorithm demonstrated effectiveness with a decision tree depth of 5, a learning rate of 0.01, and an initial ensemble size of 16. Efficient handling of streaming data was achieved with sliding window sizes set at Wmax = 1024 and Wmin = 8, resulting in a responsive query time of 0.83 seconds. Despite the EPPXGBOOST model's larger size of 4.2 MB, compared to the conventional model's 506 KB, accuracy remained impressive at 97%.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 November 2023*

TABLE III
VARIOUS PARAMETERS FOR EPPXGBOOST (DURING RUN 1)

Dataset	Amazon	Titanic	US Census
Max Depth	3	3	3
Learning Rate	0.01	0.01	0.01
Ensemble Size E(s), here s=1	16	8	8
Wmax	1024	256	512
Wmin	8	4	4
Query time (sec)	0.83	0.52	0.81
Model Size	506 KB	3 KB	210 KB
EPPXGBOOST Model Size	4.2 MB	12KB	2.5 MB
Accuracy	97%	98%	97%

Similarly, on the Titanic dataset, the algorithm maintained optimal performance with a decision tree depth of 5, a learning rate of 0.01, and an initial ensemble size of 8. Streamlined handling of evolving data was facilitated with sliding window sizes of Wmax = 256 and Wmin = 4, resulting in an efficient query time of 0.52 seconds. Despite the EPPXGBOOST model's expanded size to 12 KB, compared to theconventional model's 3 KB, accuracy remained high at 98%, demonstrating the algorithm's competence in privacy-preserving predictive tasks.

Lastly, the US Census dataset exhibited consistent performance trends, with a decision tree depth of 5, a learning rate of 0.01, and an initial ensemble size of 8. Effective management of streaming data was ensured with sliding window sizes of Wmax = 512 and Wmin = 4, resulting in an agile query time of approximately 0.81 seconds. Despite the EPPXGBOOST model's larger size of 2.5 MB, compared to the conventional model's 210 KB, the algorithm maintained a robust accuracy of 97%, emphasizing its potential across diverse datasets.

The heightened accuracy and adaptability of PPXGBOOST are complemented by its privacy-preserving capabilities. The algorithm incorporates advanced cryptographic methods, such as homomorphic encryption and secure multiparty computation, ensuring that sensitive information remains confidential during rigorous data analysis. This emphasis on privacy makes PPXGBOOST particularly well-suited for industries dealing with sensitive data, including healthcare, finance, and government sectors.

TABLE:IV
VARIOUS PARAMETERS FOR EPPXGBOOST
(DURING RUN 2)

Dataset	Amazon	Titanic	US Census
Max Depth	5	5	5
Learning Rate	0.05	0.05	0.05
Ensemble Size E(s), here s=1	16	16	32
Wmax	2048	512	2048
Wmin	8	8	16
Query time (sec)	0.79	0.47	0.87
Model Size	506 KB	3 KB	210 KB
EPPXGBOOST Model Size	4.2 MB	12KB	2.5 MB
Accuracy	98%	98%	97%

The trend persisted on the Titanic dataset, where the al- gorithm, with a decision tree depth of 7, a learning rate of 0.1, and an ensemble of 16 base models, efficiently processed streaming data with sliding window sizes of Wmax = 1024 and Wmin = 8. Notably, the algorithm achieved a rapid query time of 0.49 seconds. Despite the increased EPPXGBOOST model size to 12 KB, as opposed to the conventional 3 KB, the algorithm sustained a consistent accuracy rate of 97%.

Similarly, on the US Census dataset, employing a decision tree depth of 7, a learning rate of 0.1, and an ensemble size of 16 base models ensured effective data stream handling with sliding window sizes of Wmax = 1024 and Wmin =8. The algorithm maintained a stable query time of 0.79 seconds. Despite the privacy-preserving measures leading to an expanded EPPXGBOOST model size of 2.5 MB, compared to the original 210 KB, the accuracy rate remained resilient at 96%. This showcases the algorithm's versatility and reliability across different datasets.

V. CONCLUSION

In conclusion, this research has presented and evaluated privacy-preserving algorithms, specifically Adaptive XG-BOOST and PPXGBOOST, in the context of data stream mining. Through comprehensive analysis and implementation on three distinct datasets – Amazon Product Reviews, Kaggle's Titanic, and the US Census dataset – the proposed approaches have demonstrated notable capabilities in safeguarding sensitive information while effectively performing data stream mining tasks.

The performance assessment of Adaptive XGBOOST revealed commendable classification accuracy ranging from 94% to 96% across datasets with varying features, classes, ensemble sizes, and weight parameters. The transposed presentation of results offered valuable insights into the impact of these factors on algorithmic outcomes. Subsequently, the study extended to PPXGBOOST, showcasing its robust performance with accuracy rates between 97% and 98%. The PPXGBOOST exhibited robustness in handling diverse datasets, featuring varying numbers of features and classes.

REFERENCES

[1] Cuzzocrea, Alfredo. "Privacy-preserving big data stream mining: Opportunities, challenges, directions." 2017 ieee international conference on data mining workshops (icdmw). IEEE, 2017
[2] Hewage, U. H. W. A., 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 (2023): 1-38
[3] Mendes, Ricardo, and João P. Vilela. "Privacy-preserving data mining: methods, metrics, and applications." IEEE Access 5 (2017): 10562-10582.
[4] Xie, Lunchen, et al. "An efficient learning framework for federated xgboost using secret sharing and distributed optimization." ACM Trans- actions on Intelligent Systems and Technology (TIST) 13.5 (2022): 1-28.
[5] Montiel, Jacob, et al. "Adaptive xgboost for evolving data streams." 2020 International Joint Conference on Neural Networks (IJCNN). IEEE, 2020.
[6] Mohammadi, Mehdi, et al. "Deep learning for IoT big data and streaming analytics: A survey." IEEE Communications Surveys Tutorials 20.4 (2018): 2923-2960.
[7] Angbera, Ature, and Huah Yong Chan. "An adaptive XGBoost-based optimized sliding window for concept drift handling in non-stationary spatiotemporal data streams classifications." The Journal of Supercomputing (2023): 1-31.
[8] Li, Jing, et al. "Privacy preservation for machine learning training and classification based on homomorphic encryption schemes."

Information Sciences 526 (2020): 166-179.
[9] Al-Hussaeni, Khalil, Benjamin CM Fung, and William K. Cheung. ”Privacy-preserving trajectory stream publishing.” Data Knowledge Engineering 94 (2014): 89-109.
[10] Liu, Siyuan, et al. ”SMC: A practical schema for privacy-preserved data sharing over distributed data streams.” IEEE Transactions on Big Data 1.2 (2015): 68-81.
[11] Zhang, Ji, et al. ”On efficient and robust anonymization for privacy protection on massive streaming categorical information.” IEEE Transactions on Dependable and Secure Computing 14.5 (2015): 507-520.
[12] Talat, Romana, et al. ”A decentralised approach to privacy preserving trajectory mining.” Future generation computer systems 102 (2020): 382- 392.
[13] Kousika, N., and K. Premalatha. ”An improved privacy-preserving data mining technique using singular value decomposition with three-dimensional rotation data perturbation.” The Journal of Supercomputing 77 (2021): 10003-10011.
[14] Murugeshwari, B., S. Rajalakshmi, and K. Sudharson. ”Hybrid Approach for Privacy Enhancement in Data Mining Using Arbitrariness and Perturbation.” Computer Systems Science Engineering 44.3 (2023).
[15] Liu, Yang, et al. ”Boosting privately: Federated extreme gradient boosting for mobile crowdsensing.” 2020 IEEE 40th international conference on distributed computing systems (ICDCS). IEEE, 2020.
[16] Bifet, Albert, et al. Machine learning for data streams: with practical examples in MOA. MIT press, 2023.