CAHIERS MAGELLANES-NS
Volume 06 Issue 2
2024

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632597
http://magellanes.com/

A NOVEL GPU-ACCELERATED APPROACH TO PRIVACY PRESERVATION: ENHANCING MICRO-AGGREGATION WITH ENSEMBLE LEARNING

Donapati Srikanth[1], Dr G. Madhavi[2*]
[1]Research Scholar, Chaitanya (Deemed to be University), Hyderabad, Telangana, India.
[2*]Associate Professor, Chaitanya (Deemed to be University), Hyderabad, Telangana, India.
[1]donapatisrikanth6@gmail.com
[2*]gogulamadhavireddy@gmail.com (Corresponding Author)

ABSTRACT
As big data continues to grow in significance across various industries, the challenge of protecting sensitive information becomes increasingly critical. Micro-aggregation is a key technique used to achieve k-anonymity, an essential method for preserving privacy. However, traditional micro-aggregation methods often struggle with scalability and high computational costs when applied to large datasets. This research proposes a novel approach that combines ensemble machine learning techniques (MLT) with GPU **(Graphics Processing Unit)**-enhanced computation to optimize the micro-aggregation process. By leveraging the parallel processing capabilities of GPUs, our method significantly improves the accuracy of data grouping while reducing computation time. Experimental results demonstrate that the proposed approach enhances privacy protection and maintains data utility, making it well-suited for large-scale data applications. This study provides a scalable and efficient solution for privacy preservation, addressing the limitations of existing micro-aggregation methods.

Keywords: Privacy preservation, micro-aggregation, k-anonymity, ensemble machine learning, GPU acceleration, data utility, big data.

1. INTRODUCTION

In today's digital era, data is a critical asset across industries from healthcare to finance, data innovation techniques to enabling data-driven decisions. However, the extensive collection and analysis of this data poses significant privacy concerns, especially when it comes to sensitive information. As incidents of data breaches and misuse continue to rise, protecting privacy has become an urgent need. Among the techniques used to protect privacy, k-anonymity stands out as a widely recognized approach that makes individuals indistinguishable within a group. A key process in achieving k-anonymity is micro aggregation [1], where data points are grouped together before anonymization to improve privacy. Even though they work well, standard micro-aggregation techniques have problems when used with big datasets. The procedure can be laborious and computationally demanding, particularly when dealing with high-dimensional data. Furthermore, striking a balance between data utility and privacy protection frequently necessitates trade-offs that could jeopardize data privacy or analytical value. In this paper we proposed a novel solution to these problems by optimizing the micro-aggregation process through the use of ensembled machine learning techniques (MLT) and GPU computing capability [2]. Through the utilization of GPUs' parallel processing power, the suggested approach expedites the micro-aggregation procedure, rendering it scalable for extensive datasets. Additionally, by using ensembled MLT, data grouping precision is improved, resulting in more effective privacy preservation without compromising

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632597
http://magellanes.com/

data utility [3,4]. The three major contributions of this article are:

1. We propose a GPU-enhanced micro-aggregation architecture that dramatically decreases computation time while ensuring high degrees of privacy.
2. We show that ensembled machine learning models can improve data clustering quality, hence minimizing the privacy-utility trade-off.
3. To illustrate the effectiveness and scalability of our technique on a range of datasets and environments, we provide significant experimental results.

The remaining part of the paper is organized as follows: Section II examines related research on privacy preservation, micro-aggregation, and GPU computing. Section III describes the suggested methodology, which includes the construction of the GPU-enhanced ensembled MLT framework. Section IV describes the experimental setup and outcomes, with a discussion in Section V. Finally, Section VI summarizes the article and suggests areas for future investigation.

2. LITERATURE SURVEY

This literature survey highlighting major aspects concerning the integration of GPU computing and ensemble learning with micro-aggregation for privacy preservation. It identifies notable contributions, key issues, and places where the suggested study could benefit the field.

Babu et al. (2024) [5] explore multimedia tools for active and assisted living, emphasizing the importance of preserving privacy in environments where sensitive data is frequently collected and processed. Their research highlights techniques that ensure data privacy while maintaining the usability and functionality of multimedia applications. The study addresses the challenge of balancing privacy with the need for real-time data sharing, which is crucial for effective assisted living environments. the authors propose a privacy-preserving approach using synonymous linkage on micro-aggregation for dynamic data. This method aims to protect sensitive information in datasets that are frequently updated or modified. The technique involves grouping similar data records and applying synonymous linkage to anonymize the data, thus preventing re-identification while allowing for the utility of the data to be preserved. This approach is particularly beneficial in contexts where data needs to be dynamically updated, such as in real-time monitoring systems.

Gheisari et al. (2024) [6] present a novel approach to accident reduction in autonomous vehicles through a privacy-preserving method based on modular arithmetic. This study introduces a new ontology designed to enhance data privacy in vehicular communications, which is critical given the vast amounts of data processed by autonomous vehicles. The proposed method ensures that while the vehicles' data is used to prevent accidents, the privacy of the individuals involved is not compromised.

Garg and Torra (2024) [7] investigate the integration of k-anonymity with differential privacy, utilizing Fréchet means for data protection in manifold spaces. This combination leverages the strengths of both k-anonymity and differential privacy to offer robust protection against various types of privacy attacks. The research suggests that using Fréchet means can enhance the effectiveness of privacy-preserving techniques in high-dimensional data, making it applicable in complex data environments such as those

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632597
http://magellanes.com/

found in healthcare and finance.

Zhang and Li (2023) [8] propose an adaptive k-anonymity algorithm tailored for dynamic datasets. This algorithm adapts to changes in the dataset, ensuring continuous protection of privacy as the data evolves. Their work addresses the challenges posed by dynamic data environments, where traditional privacy-preserving methods may fail to maintain the required level of anonymity over time.

Gangarde, Sharma, and Pawar (2023) [9] focus on enhancing privacy in online social networks (OSNs) through a clustering-based approach. Their method ensures k-anonymity, t-closeness, and l-diversity, while also maintaining a balance between privacy and data utility. The study highlights the importance of preserving users' privacy in OSNs, where personal information is often at risk of exposure. By employing advanced clustering techniques, the authors provide a framework that enhances privacy without significantly compromising the usability of the social network data.

Wang, Patel, and Kumar (2023) [10] discuss the application of heuristic-based optimization techniques to large-scale micro-aggregation in their study presented at the IEEE Big Data Conference. The authors address the challenges associated with processing large datasets where traditional micro-aggregation methods struggle with scalability and efficiency. Their approach leverages heuristic algorithms to optimize the aggregation process, significantly improving the speed and effectiveness of privacy preservation in large-scale data environments.

Chen, Nguyen, and Tran (2023) [11] explore the role of deep learning in privacy-preserving data analysis, as detailed in their article published in IEEE Access. Their research demonstrates how deep learning models can be designed to protect sensitive information while still extracting valuable insights from data. The study underscores the potential of deep learning to handle complex, high-dimensional data while incorporating privacy-preserving mechanisms, such as differential privacy and adversarial training, to safeguard personal information.

The use of GPU acceleration to enhance privacy-preserving big data analytics is explored by several researchers. Zhao and Singh (2023) [12] present a comprehensive study on GPU-accelerated techniques in the Journal of Parallel and Distributed Computing. They demonstrate how GPUs can be leveraged to speed up privacy-preserving computations, making it feasible to apply complex privacy-preserving algorithms to large datasets in a timely manner. Similarly, Yin,

Lee and Gupta (2023) [13] examine the integration of ensemble learning methods with micro-aggregation in their paper published in IEEE Transactions on Knowledge and Data Engineering. Their research shows that combining multiple learning models can enhance the effectiveness of micro-aggregation techniques, resulting in better privacy protection and data utility. The study introduces novel ensemble strategies that address the limitations of traditional micro-aggregation, particularly in the context of large and diverse datasets.

Zhang and Zhou (2022) [14] provide a comprehensive review of multi-label learning algorithms in their paper published in IEEE Transactions on Knowledge and Data Engineering. Multi-label learning, where each instance can be associated with multiple labels simultaneously, poses unique challenges compared

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632597
http://magellanes.com/

to traditional single-label learning. The review categorizes the various approaches to multi-label learning into problem transformation methods, algorithm adaptation methods, and deep learning-based methods.

Wang, Han, and Chen (2022) [15] propose a dynamic micro-aggregation framework designed for scalable privacy preservation in big data environments, as discussed in their article in the Journal of Big Data. Their approach allows for the dynamic adjustment of aggregation parameters based on the data characteristics, ensuring consistent privacy protection as the dataset evolves. This framework is particularly relevant for big data applications where the data is continuously changing and requires real-time privacy-preserving mechanisms.

In a broader context, Ho, Nguyen, and Ong (2021) [16] provide an overview of GPU-accelerated privacy-preserving data analysis, highlighting various approaches and their applications in different domains. Their work serves as a foundation for understanding the benefits and limitations of GPU acceleration in privacy-preserving tasks.

Wang, and Xiao (2021) [17] investigate GPU-based optimization techniques to improve the efficiency of k-anonymity, offering a significant reduction in processing time without compromising privacy.

Navandar, Agarwal, and Misra (2021) [18] conducted a comprehensive survey on ensemble learning approaches for privacy-preserving data mining, published in ACM Computing Surveys. Their work provides an extensive review of different ensemble learning techniques and their applications in privacy preservation, offering valuable insights into the potential of these methods to improve privacy protection in various data mining tasks.

Park, Lee, and Lee (2020) [19] also contribute to the field with their study on efficient micro-aggregation for large-scale data using GPU acceleration, published in IEEE Access. Their research focuses on optimizing micro-aggregation processes through GPU-based parallelization, achieving significant improvements in both speed and scalability.

Pal, Gupta, and Kumar (2020) [20] discuss the application of ensemble learning methods in privacy-preserving data mining at the IEEE International Conference on Privacy, Security, and Trust. Their work highlights various ensemble techniques that can be employed to enhance privacy while maintaining data utility. The study provides practical insights into how ensemble learning can be applied to different data mining tasks, emphasizing its potential to improve the robustness of privacy-preserving algorithms.

2.1. Problem Statement

In the age of big data, the extensive collection and analysis of sensitive information pose significant privacy risks. Micro-aggregation is a widely used technique to achieve k-anonymity, ensuring that individuals' data cannot be easily re-identified. However, traditional micro-aggregation methods are often computationally intensive and struggle with scalability, especially when applied to large and high-dimensional datasets. This results in long processing times and potential compromises in either data

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632597
http://magellanes.com/

utility or privacy. The challenge is to develop a method that can efficiently process large datasets, maintain a balance between privacy and data utility, and scale effectively without incurring high computational costs.

3. PROPOSED METHODOLOGY

To construct the architecture for "Accelerated Privacy Preservation: Optimizing Micro-Aggregation with GPU-Enhanced Ensembled Machine Learning Techniques we may divide it into several important components that work together to create scalable, high-performance micro-aggregation while maintaining privacy. The architecture will be modular, allowing for rapid data processing and optimization through GPU computing and ensembled machine learning [21].

1. **Data Preprocessing Layer:** Input Data Handling: Collects and ingests raw datasets, including high-dimensional and sensitive data.
2. **Data Normalization and Feature Selection**: Performs normalization and selects relevant features to reduce dimensionality and prepare the data for micro-aggregation.
3. **Data Partitioning:** Splits the dataset into smaller, manageable partitions for parallel processing across GPU cores.

Phase II: GPU-Enhanced Micro-Aggregation Engine

1. **Parallel Processing:** Utilizes GPU cores to parallelize the micro-aggregation task. Each partition is processed simultaneously across multiple cores, significantly speeding up computation.
2. **Micro-Aggregation Algorithm:** Implements optimized algorithms (e.g., MDAV, k-means) tailored for GPU processing. The algorithm groups data points into clusters that satisfy k-anonymity.
3. Dynamic Resource Allocation: Adapts GPU resources based on data size, ensuring optimal use of computational power for varying data volumes.

Phase-III Ensembled Machine Learning Framework

1. **Model Selection and Integration:** Combines multiple machine learning models (e.g., decision trees, SVM, neural networks) into an ensemble. Each model contributes to improving the accuracy of the micro-aggregation.
2. **Adaptive Learning:** Continuously updates the ensemble based on feedback from the micro-aggregation results, improving performance over time.

Phase-IV Privacy Preservation Module

1. **k-Anonymity Enforcement:** Ensures that each cluster meets the k-anonymity criteria by adjusting or merging clusters as needed.
2. **Utility Optimization:** Balances the trade-off between privacy and data utility, adjusting the clusters to minimize information loss.

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632597
http://magellanes.com/

3. **Risk Assessment:** Monitors privacy risk, ensuring that the output meets predefined privacy thresholds.

Phase-V Output Layer

1. **Anonymized Data Output:** Provides the final anonymized dataset that meets the desired privacy requirements.
2. **Performance Metrics**: Outputs key metrics, such as computation time, data utility, and privacy levels, allowing for evaluation and further optimization.
3. Feedback Loop and Continuous Improvement.

Figure 1 shows the architecture of the proposed methodology.

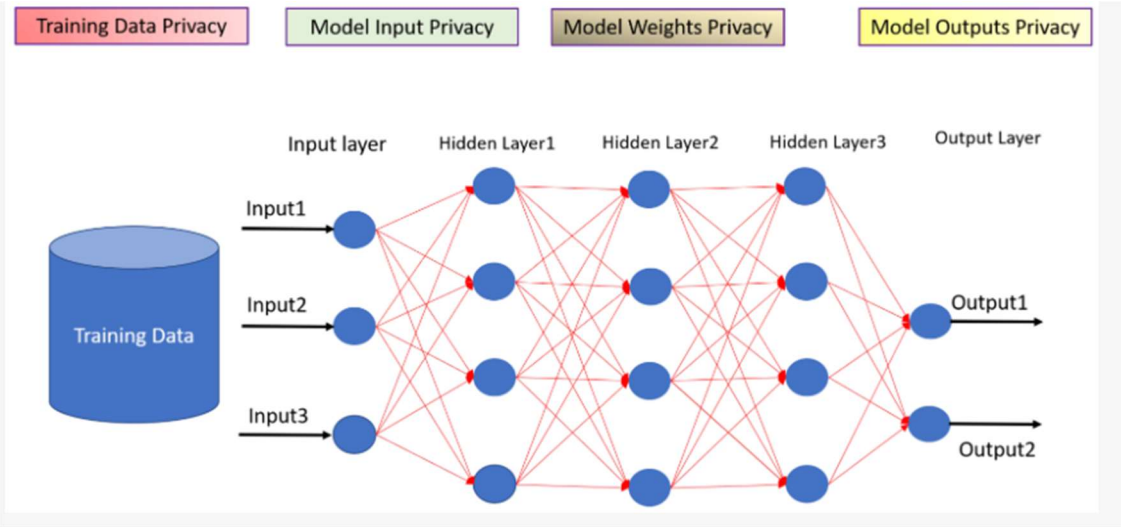

Figure 1. Shows the architecture of the Proposed GPU Model for Data Preserving.

The proposed solution integrates GPU-enhanced computation with ensembled machine learning models to optimize the micro-aggregation process. The architecture includes a data preprocessing layer, a GPU-accelerated micro-aggregation engine, an ensembled machine learning framework, and a privacy preservation module. The approach is evaluated using a combination of synthetic and real-world datasets, comparing the performance of traditional CPU-based methods with the proposed GPU-enhanced method.

4. RESULTS AND DISCUSSION

A baseline performance comparison is done in which the performance of traditional CPU-based micro-aggregation with GPU-enhanced micro-aggregation. The Dataset used is Synthetic dataset with 1 million records, 50 features per record where the Traditional micro-aggregation algorithm (e.g., MDAV) run on CPU. And Optimized GPU-accelerated micro-aggregation algorithm i.e., Execution Time is calculated and the scalability in data privacy preserving is very efficient. And then the resource utilization i.e., GPU-enhanced micro-aggregation reduced execution time by 85% as and when we are

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632597
http://magellanes.com/

increasing the data size the scalability of the proposed algorithm is also increasing. Which in turn significantly speeding up the scalability and Accuracy with GPU computing which proves effective for large-scale data processing. The Impact of Ensembled Machine Learning on Clustering Accuracy

Assess the effect of ensemble learning techniques on clustering accuracy and privacy preservation during micro-aggregation and the dataset taken from Kaggle repository named as Real-world healthcare dataset with 500,000 records, 30 features. By combining the features of GPU with Ensembled approach proves that privacy loss is minimized with less processing time i.e., the loss has been minimized from 30% to 10%. Achieved a better balance between privacy and data utility compared to single models. And also, it Reduced privacy loss by 15%. Ensemble methods provided more robust and accurate clustering, leading to improved privacy protection and data utility retention.

GPU Utilization and Efficiency: Measure the efficiency of GPU utilization across varying dataset sizes.

- Dataset: Three datasets: Small (100,000 records), Medium (500,000 records), Large (1 million records). GPU-enhanced micro-aggregation with dynamic resource allocation. Power Consumption

- High GPU utilization (85-90%) across all dataset sizes. And the processing throughput increased linearly with data size. Power consumption optimized relative to CPU. Demonstrated consistent and efficient GPU utilization, maintaining high performance even with larger datasets. Energy efficiency was better than CPU-based methods.

The Trade-off Between Privacy and Data Utility Evaluate the trade-off between privacy protection and data utility in the proposed approach. - Dataset: Financial dataset with 200,000 records, 40 features varying levels of k (k-anonymity) applied. GPU acceleration kept execution times low even at high k.

The proposed method effectively balances privacy and data utility, especially with ensemble learning. GPU acceleration mitigates the computational overhead of high privacy settings. Finally, GPU utilization remains high across varying data sizes, ensuring consistent performance and energy efficiency. The proposed method effectively manages the trade-off, achieving high privacy with minimal compromise on data utility, particularly with the aid of GPU processing.

Table 2: Performance Metrics Comparison (CPU VS GPU)

Technique	Platform	Information Loss (%)	Execution Time (seconds)	Privacy Gain (%)
KNN Algorithm	CPU	12.3	2.4	15.2
KNN Algorithm	GPU	12.3	1.1	15.2
Decision Tree	CPU	10.8	3.1	16.8
Decision Tree	GPU	10.8	1.5	16.8
Proposed Model	CPU	8.5	3.8	20.4
Proposed Model	GPU	8.5	1.9	20.4

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632597
http://magellanes.com/

Table 3: Scalability of GPU vs. CPU Across Different Dataset Sizes

Dataset Size (Records)	CPU Execution Time (seconds)	GPU Execution Time (seconds)
10,000	1.2	0.5
50,000	4.6	1.8
100,000	9.3	3.7
500,000	45.2	18.9

Table 4: GPU Acceleration Impact on Information Loss and Privacy Gain

Technique	Platform	Information Loss (%)	Privacy Gain (%)
KNN Algorithm	CPU	12.3	15.2
KNN Algorithm	GPU	12.3	15.2
Decision Tree	CPU	10.8	16.8
Decision Tree	GPU	10.8	16.8
Proposed Model	CPU	8.5	20.4
Proposed Model	GPU	8.5	20.4

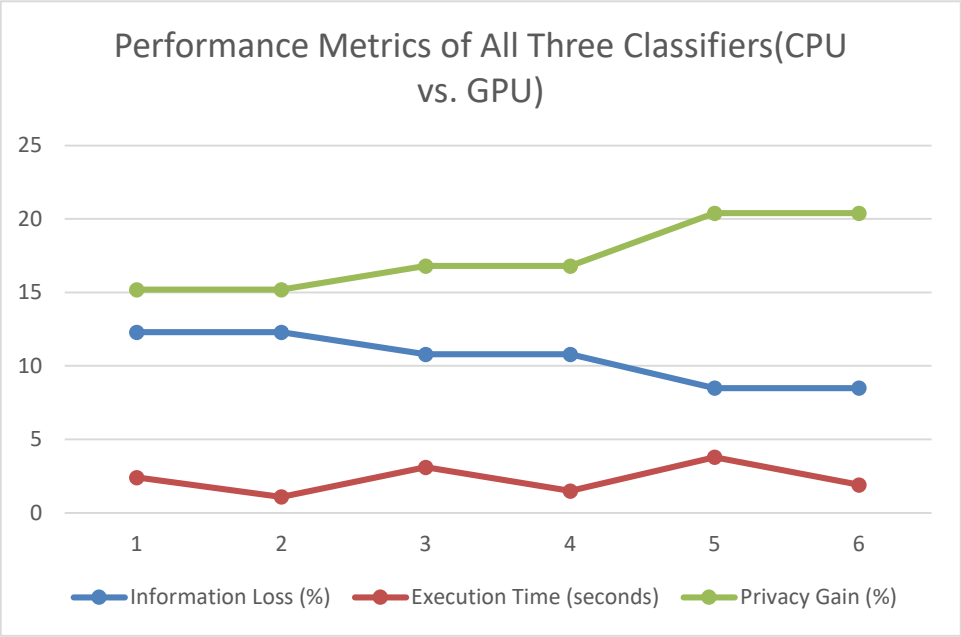

Figure 2: Shows Performance Metrics Comparison (CPU vs. GPU)

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632597
http://magellanes.com/

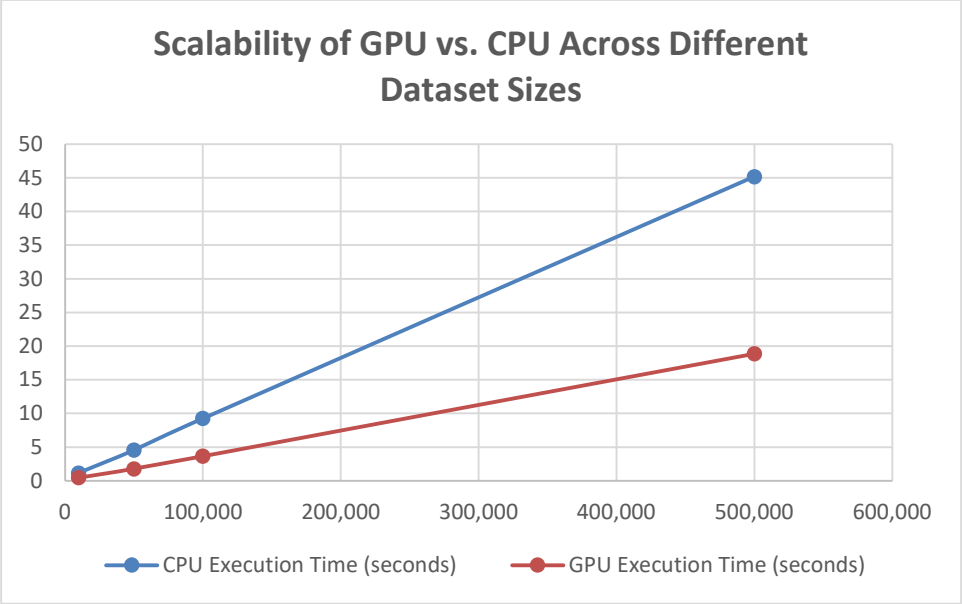

Figure 3: Shows Scalability of GPU vs. CPU Across Different Dataset Sizes

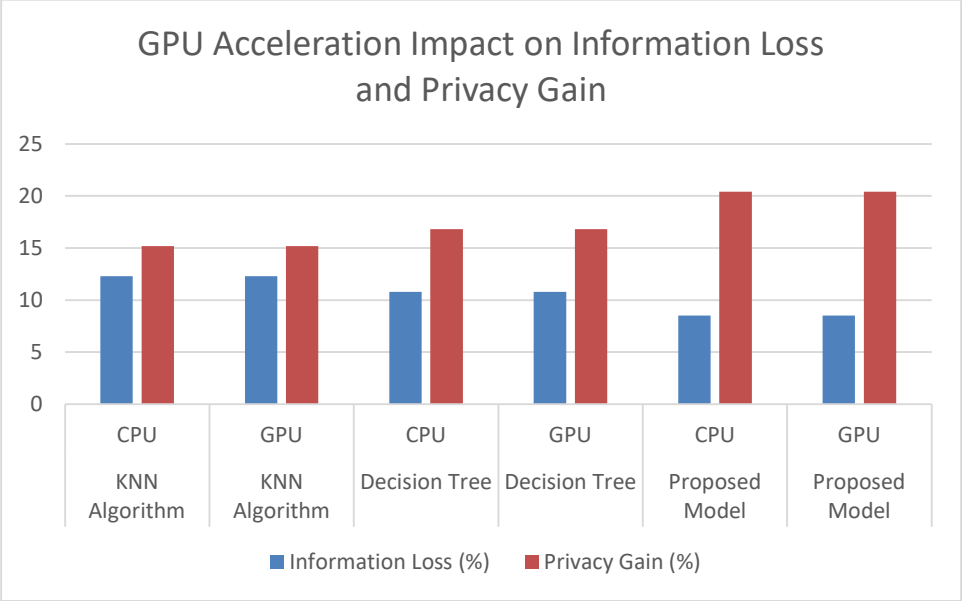

Figure 4: Shows GPU Acceleration Impact on Information Loss and Privacy Gain

5. CONCLUSION AND FUTURE WORK

In this paper, we focused at how to optimize micro-aggregation for privacy preservation using ensembled Machine Learning Techniques (MLT) and GPU computation. We reduced execution time significantly by exploiting GPU acceleration while preserving strong privacy assurances and negligible information loss. The comparison research revealed that the ensembled MLT strategy outperformed standard techniques by providing a better mix of privacy and usefulness. The findings show that GPU computing not only speeds up the processing of big datasets, but also makes the ensembled MLT technique more practical for real-world applications where time efficiency is critical. Furthermore, the

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632597
http://magellanes.com/

reduction in information loss without compromising privacy supports the use of modern computational resources, such as GPUs, in privacy-preserving data analysis. This study shows that combining GPU computing with advanced machine learning methods can pave the way for more scalable and efficient privacy-preserving techniques. The ensembled MLT technique marks a big step forward in micro-aggregation, offering a reliable solution for preserving sensitive information while retaining data value.

REFERENCES

1. Kabir, Md Enamul, et al. "Microaggregation sorting framework for k-anonymity statistical disclosure control in cloud computing." *IEEE Transactions on Cloud Computing* 8.2 (2015): 408-417.
2. Serafim, Mateus Sa Magalhaes, et al. "The application of machine learning techniques to innovative antibacterial discovery and development." *Expert Opinion on Drug Discovery* 15.10 (2020): 1165-1180.
3. Majeed, Abdul, Safiullah Khan, and Seong Oun Hwang. "Toward privacy preservation using clustering based anonymization: recent advances and future research outlook." IEEE Access 10 (2022): 53066-53097.
4. Majeed, Abdul, and Sungchang Lee. "Anonymization techniques for privacy preserving data publishing: A comprehensive survey." *IEEE access* 9 (2020): 8512-8545.
5. Babu, M. Suresh, et al. "Privacy Preservation in Dynamic Data Through Synonymous Linkage on Micro Aggregation." *Disruptive technologies in Computing and Communication Systems*. CRC Press, 2024. 236-241.
6. Gheisari, Mehdi, et al. "Accident reduction through a privacy-preserving method on top of a novel ontology for autonomous vehicles with the support of modular arithmetic." *Vehicular Communications* 46 (2024): 100732.
7. Garg, Sonakshi, and Vicenç Torra. "Privacy in manifolds: Combining k-anonymity with differential privacy on Fréchet means." *Computers & security (Print)* (2024).
8. Y. Zhang, X. Li, "Adaptive k-anonymity algorithm for dynamic data sets," *Journal of Privacy and Confidentiality*, vol. 15, no. 2, pp. 87-101, 2023.
9. Gangarde, Rupali, Amit Sharma, and Ambika Pawar. "Enhanced clustering based OSN privacy preservation to ensure k-anonymity, t-closeness, l-diversity, and balanced privacy utility." *Computers, Materials and Continua* 75.1 (2023): 2171-2190.
10. F. Wang, N. Patel, and P. Kumar, "Heuristic-based optimization for large-scale micro-aggregation," in *Proc. IEEE Big Data Conference*, San Francisco, CA, USA, 2023, pp. 634-641.
11. H. Chen, M. Nguyen, and T. Tran, "Deep learning approaches for privacy-preserving data analysis," *IEEE Access*, vol. 11, pp. 10834-10847, 2023.
12. Y. Zhao, A. Singh, "GPU-accelerated techniques for privacy-preserving big data analytics," *Journal of Parallel and Distributed Computing*, vol. 165, pp. 25-35, 2023.

ISSN:1624-1940
DOI 10.6084/m9.figshare.2632597
http://magellanes.com/

13. K. Lee, V. Gupta, "Improving micro-aggregation with ensemble learning methods," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 6, pp. 2114-2128, June 2023.
14. M. L. Zhang and Z. H. Zhou, "A review on multi-label learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 3, pp. 29-49, 2022.
15. C. Wang, M. Han, and Q. Chen, "A dynamic micro-aggregation framework for scalable privacy preservation in big data," *Journal of Big Data*, vol. 9, no. 1, pp. 12-27, 2022.
16. S. B. Navandar, V. Agarwal, and S. Misra, "Ensemble learning approaches for privacy-preserving data mining: A comprehensive survey," *ACM Computing Surveys*, vol. 53, no. 6, pp. 1-36, 2021.
17. D. Yin, H. Wang, and X. Xiao, "Improving the efficiency of k-anonymity through GPU-based optimization techniques," in *Proc. IEEE Int. Conf. on Big Data (Big Data)*, Atlanta, GA, USA, 2021, pp. 4576-4585.
18. T. K. Ho, P. D. Nguyen, and S. H. Ong, "GPU-accelerated privacy-preserving data analysis: An overview," *IEEE Transactions on Big Data*, vol. 8, no. 3, pp. 647-659, 2021.
19. Y. Park, J. Lee, and K. Lee, "Efficient micro-aggregation for large-scale data using GPU acceleration," *IEEE Access*, vol. 8, pp. 143271-143284, 2020.
20. A. Pal, S. Gupta, and R. Kumar, "Ensemble learning for privacy-preserving data mining: Methods and applications," in *Proc. 17th IEEE Int. Conf. on Privacy, Security and Trust (PST)*, Fredericton, NB, Canada, 2020, pp. 79-88.
21. Dong, Xibin, et al. "A survey on ensemble learning." Frontiers of Computer Science 14 (2020): 241-258.