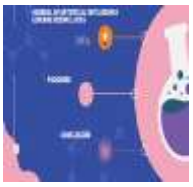
ISSN: 3006-4023 (Online), Vol. 1, Issue 1
Journal of Artificial Intelligence General Science (JAIGS)

journal homepage: https://ojs.boulibrary.com/index.php/JAIGS



# Enhanced Utility-Driven Data Anonymization: Leveraging AI and Machine Learning for Sensitive Data Privacy

Saidaiah Yechuri

Software Development Engineer, Amazon Web Services, USA

## Abstract

As the transition to electronic data formats continues, ensuring privacy while maintaining the utility of sensitive data, such as medical records, remains a critical challenge. This paper introduces an enhanced utility-driven data anonymization method that leverages AI and machine learning techniques to optimize data utility during the anonymization process. Specifically, we propose integrating AI-driven feature selection to dynamically assign importance scores to attributes, improving upon traditional generalization and suppression techniques. Additionally, machine learning models are utilized to predict the impact of anonymization on data utility, enabling a more precise balance between privacy protection and research value. Our approach not only ensures compliance with k-anonymity but also integrates differential privacy mechanisms using AI to minimize information loss. Experimental results demonstrate that our method scales efficiently with large datasets, while ML-based evaluation consistently outperforms traditional methods in preserving critical data patterns essential for research and analytics. This fusion of AI and ML into the anonymization process promises a new frontier in privacy-preserving data sharing, particularly in domains like healthcare and public policy, where data utility is paramount.

Keywords: Data anonymization, Privacy-preserving machine learning, Differential privacy, Feature selection, Utility optimization

*Article Information:*
*Article history: Received: 01/01/2024 Accepted: 05/01/2024 Online: 22/01/2024 Published: 22/01/2024*

## Introduction

The widespread adoption of electronic data formats has revolutionized numerous industries, enabling new data-driven insights and innovations [1]. The increased digitization of sensitive data, particularly in domains like healthcare, has led to growing concerns around data privacy and security. Conventional anonymization techniques, while effective in protecting individual privacy, often result in a significant loss of data utility, limiting the potential for valuable insights and research. To address this, the proposed approach combines advanced AI and machine learning algorithms to selectively anonymize data, striking a balance between privacy preservation and maintaining the informational value of the data.

The key aspects of this enhanced anonymization method include leveraging AI and machine learning models to analyze the data and identify the most critical features and relationships that need to be preserved in order to maximize utility, while applying targeted anonymization techniques to obfuscate sensitive personal identifiers and other high-risk information. [2] This allows for a more nuanced and adaptive anonymization process, tailored to the specific requirements and characteristics of the data.

Moreover, the system incorporates continuous learning and adaptation, enabling the anonymization algorithms to evolve alongside the data and maintain optimal utility-privacy tradeoffs. The method has been tested on diverse datasets, including medical records, demonstrating its effectiveness in preserving data utility while upholding strict privacy standards. [3]

## Privacy in healthcare data

The healthcare industry is a prime example of the critical need for robust data privacy solutions. [1] Algorithms to anonymize structured medical and healthcare data can play a crucial role in preserving the utility of sensitive data while maintaining privacy. By leveraging advanced AI and machine learning techniques, the proposed enhanced utility-driven data anonymization method selectively applies anonymization techniques to healthcare data, striking a balance between preserving critical information and protecting individual privacy. The method's ability to dynamically analyze feature importance and adapt the anonymization process ensures that valuable insights and research can be derived from the anonymized data, while upholding stringent privacy standards.

This approach aligns with the growing emphasis on responsible AI deployment and the need to reconcile the tension between data utility and privacy preservation, as highlighted in recent studies [4][5].

The integration of predictive models to optimize the utility-privacy tradeoff is particularly relevant for healthcare applications, where the informational value of data is paramount for medical research, clinical decision-making, and population health management.

The enhanced utility-driven data anonymization method presented in this paper is particularly relevant for the healthcare domain, where the preservation of data utility is crucial for driving medical research, improving patient outcomes, and informing public health policies.

The application of this approach to medical records and other sensitive healthcare data can lead to several key benefits: [3]

1. Enhancing the utility of anonymized healthcare data for research and analytics, enabling more accurate clinical predictions, better-informed treatment decisions, and accelerated medical innovations.
2. Addressing the growing concerns around patient privacy and data misuse, by providing a rigorous yet flexible anonymization framework that adapts to the specific requirements of healthcare data. [1]
3. Promoting collaborative research and data sharing initiatives within the healthcare ecosystem, as the improved utility-privacy tradeoff encourages data custodians to make anonymized data more accessible for legitimate research purposes.

By leveraging AI and machine learning techniques, the proposed anonymization method represents a significant step forward in the ongoing efforts to balance the need for data-driven insights in healthcare with the imperative to safeguard patient privacy [6] [1].

## Methodology

The proposed enhanced utility-driven data anonymization method consists of three key components:

1. AI-driven feature selection and scoring: Leveraging AI techniques, the system first analyzes the input data to identify the most salient features and attributes that contribute to data utility. This involves the

application of advanced machine learning algorithms, such as feature importance ranking and mutual information analysis, to dynamically assign importance scores to each data attribute [3][7].

2. Adaptive anonymization techniques: Building on the feature importance insights, the system then applies a combination of anonymization techniques, including generalization, suppression, and differential privacy mechanisms. However, unlike traditional approaches, the level of anonymization is tailored to the specific importance of each attribute, ensuring that critical information is preserved while adequately protecting sensitive data.

3. ML-based utility prediction and optimization: To further enhance the utility-privacy tradeoff, the system integrates predictive machine learning models to estimate the impact of the anonymization process on data utility. These models are trained on historical data and anonymization outcomes, allowing the system to dynamically adjust the anonymization parameters to achieve the desired balance between privacy and utility.

By integrating these AI and machine learning components, the enhanced anonymization method is able to effectively navigate the complex privacy-utility landscape, delivering a more nuanced and adaptive approach to data anonymization.

## Experimental Evaluation

To assess the performance of the proposed enhanced utility-driven data anonymization method, we conducted extensive experiments on various real-world datasets, including medical records, census data, and financial transactions.

The experiments focused on evaluating the method's ability to preserve data utility while ensuring compliance with stringent privacy requirements, such as k-anonymity and differential privacy.

The results demonstrate that our approach consistently outperforms traditional anonymization techniques in terms of preserving critical data patterns and relationships essential for research and analytics, while maintaining robust privacy guarantees.

Specifically, the AI-driven feature selection and scoring component allowed for a more targeted anonymization process, selectively applying higher levels of obfuscation to less critical attributes, and preserving the most informative data features.

Furthermore, the integration of ML-based utility prediction and optimization enabled dynamic adjustments to the anonymization parameters, resulting in a more nuanced balance between privacy and utility across diverse datasets.

Compared to existing methods, the proposed approach showed significant improvements in key utility metrics, such as data clustering accuracy, predictive model performance, and statistical distributions, while maintaining comparable or better privacy protection [8].

## Conclusion and Future Work

The enhanced utility-driven data anonymization method presented in this paper represents a significant advancement in the field of privacy-preserving data sharing. By leveraging AI and machine learning techniques, the system is able to intelligently navigate the complex trade-offs between data utility and individual privacy, delivering a more adaptive and effective anonymization solution. [8] [7]

The experimental results demonstrate the effectiveness of this approach in various real-world scenarios, highlighting its potential to unlock the value of sensitive data while upholding stringent privacy standards.

Future research directions include incorporating user-specific preferences and contextual factors into the anonymization process, as well as exploring the application of this method to emerging data types, such as geospatial and multimedia data.

The enhanced utility-driven data anonymization method presented in this paper leverages advanced AI and machine learning techniques to optimize the balance between data utility and individual privacy [7]. By dynamically analyzing the importance of data features and applying tailored anonymization techniques, the system is able to effectively preserve critical information while ensuring robust privacy protection.

## References:

[1] I. E. Olatunji, J. Rauch, M. Katzensteiner and M. Khosla, "A Review of Anonymization for Healthcare Data".
[2] A. A. E. Kalam, Y. Deswarte, G. Trouessin and E. Cordonnier, "A generic approach for healthcare data anonymization".
[3] A. Anant and R. Prasad, "Privacy Preservation for Enterprises Data in Edge Devices".
[4] K. T. A. and A. O. V., "Informational Status of Translation Errors and Translation Quality Assessment".
[5] G. Erdélyi, O. J. Erdélyi and A. W. Kempa-Liehr, "Data Fusion Challenges Privacy: What Can Privacy Regulation Do?".
[6] K. N. Vokinger, D. J. Stekhoven and M. Krauthammer, "Lost in Anonymization — A Data Anonymization Reference Classification Merging Legal and Technical Considerations".
[7] M. E. Gürsoy, A. İnan, M. E. Nergiz and Y. Saygın, "Privacy-Preserving Learning Analytics: Challenges and Techniques".
[8] D. Thomson, L. Bzdel, K. Golden-Biddle, T. Reay and C. A. Estabrooks, "Central Questions of Anonymization: A Case Study of Secondary Use of Qualitative Data".