Article

Not peer-reviewed version

Federated Learning for Heterogeneous Data Integration and Privacy Protection

Chenwei Gong , Xuyang Zhang * , Yuzhen Lin , Hang Lu , Pei-Chiang Su , Jingwei Zhang

Posted Date: 28 March 2025

doi: 10.20944/preprints202503.2211.v1

Keywords: Federated Learning; Feature Alignment

 doi:10.20944/preprints202503.2211.v1

Article

Federated Learning for Heterogeneous Data Integration and Privacy Protection

Chenwei Gong, Xuyang Zhang *, Yuzhen Lin, Pei-Chiang Su, Hang Lu and Jingwei Zhang

[1] Henry Samueli School of Engineering, University of California, Los Angeles, Los Angeles, 90024, USA
[2] Rackham School, University of Michigan-Ann Arbor, Ann Arbor, MI, 48109, USA
[3] School of Information Systems and Management, Carnegie Mellon University, Jersey City, 07302, USA
[4] Department of Electrical and Computer Engineering, Carnegie Mellon University Pittsburgh, PA 15213, USA
[5] Muma college of Business, University of South Florida, Florida, 33620, USA
[6] ERP DSC Logistics ENG CN, SAP Labs China, Shanghai, 201203, China
* Correspondence: xuyangzh@umich.edu

Abstract: Federated learning (FL) represents a promising approach that enables the collaborative training of machine learning models without compromising data privacy. This approach is particularly advantageous when handling heterogeneous data dispersed across numerous institutions or devices, as centralized data aggregation is often constrained by privacy concerns and data regulations. In order to address the challenges posed by heterogeneous data, we have devised an adaptive data integration mechanism. This mechanism maps the features of disparate data sources to a unified feature space through the use of feature alignment technology, thereby facilitating the effective fusion of data. This fusion is achieved through the application of statistical alignment and multiperspective learning technology. Furthermore, in order to safeguard the confidentiality of data, we integrate differential privacy and homomorphic encryption techniques, thereby preventing the disclosure of information during model updates and data transfers. Furthermore, a multi-level privacy protection strategy is proposed, which employs de-identification, secure multi-party computation, and federated averaging technologies at the three stages of data preprocessing, model training, and result aggregation, respectively. This approach ensures data security and facilitates effective model updates. The experimental results demonstrate that the proposed framework exhibits enhanced model performance and robustness in comparison to traditional federated learning methods on a multitude of real-world heterogeneous datasets.

Keywords: Federated Learning; Feature Alignment

1. Introduction

In the contemporary era, data has emerged as a pivotal element of production and a strategic asset across a spectrum of societal domains. It is regarded as the "new oil" of the 21st century. The accelerated development and extensive implementation of information technology have resulted in the generation and aggregation of vast quantities of data at an unprecedented rate. The comprehensive analysis and utilization of these data sets have not only facilitated the transformation of conventional industrial development models but have also markedly enhanced production efficiency, thereby providing robust support for the sustained growth of the economy and the improvement of people's lives. In the field of scientific research, the utilization of data-driven scientific discovery is becoming a dominant paradigm, alongside the traditional three. Scientists regard data as the core object and tool of research. By analyzing and modeling large-scale data, scientists can reveal the laws behind complex phenomena, thereby guiding and implementing various scientific research projects [1]. To illustrate, the advancement of data acquisition, analysis, and processing capabilities has significantly facilitated the advancement of cutting-edge science in fields

 doi:10.20944/preprints202503.2211.v1

such as genomics, astronomy, and climate science. Data serves not only to validate existing theories but also to provide new avenues for developing new hypotheses and exploring previously uncharted territory [2].

Concurrently, the pervasive application of data presents a multitude of challenges. The exponential growth in data volume has created a pressing need to develop effective methods for storing, processing, and analyzing these data, while also ensuring the privacy and security of the data. In light of these considerations, the implementation of the national big data strategy is not only intended to facilitate technological advancement but also to guarantee that data [3], as a pivotal resource, can facilitate economic and social development through institutional innovation and policy guidance.

In the contemporary data-driven society, machine learning technology has been extensively employed across a range of domains, including healthcare, finance, and the Internet of Things. However, with the increasing prominence of data privacy protection and security issues, traditional centralized data collection and model training methods are facing significant challenges [4]. To address this challenge, Federated Learning (FL), a novel distributed machine learning approach, offers a promising solution to data privacy concerns. Federated learning enables multiple participants to collaboratively train a global model without sharing their local data, thereby avoiding direct exposure to sensitive data.

Federated learning represents a novel approach to addressing the issue of data silos. It enables joint modeling between disparate data holders while circumventing the potential risks associated with data sharing, such as the compromise of privacy. In the conventional federated learning model, the data is retained at the local level, with each participant generating model parameters through local training and subsequently transmitting these parameters to a central server for aggregation. Nevertheless, this model presents two significant shortcomings.

First and foremost, the current approach is not sufficiently versatile. The models and algorithms utilized by each participant frequently necessitate intricate tuning and transformation to align with the specific requirements of local training, particularly in the context of disparate machine learning algorithms [5]. This significantly constrains the extensive applicability of federated learning frameworks across diverse contexts. Secondly, the efficiency of the algorithm training is low. As the training process depends on regular communication between the central server and individual data nodes, each iteration necessitates a waiting period until all nodes have completed their training and uploaded the model parameters. This process not only requires a significant investment of time, particularly in the context of prolonged network communication delays, but is also constrained by the disparities in computing capabilities across individual nodes. In the event that the computing power of certain nodes is insufficient, the overall training process will be slowed down by these nodes, resulting in a reduction in the aggregation efficiency of the global model [6].

A variety of strategies can be employed to enhance these processes. As an illustration, the deployment of differential privacy technology enables the processing of data at the local level for the purpose of safeguarding its confidentiality. Subsequently, the processed data can be transmitted to a central server in a single operation. This approach has the additional benefit of reducing the communication overhead associated with model training while circumventing the complexities inherent to local training [7]. However, this approach also introduces novel challenges pertaining to the protection of data during transmission, necessitating more rigorous safeguards. The primary paradigms of traditional federated learning include horizontal federated learning and vertical federated learning (i.e., data is divided in the feature dimension, with each data holder possessing disparate features of the same sample) [8]. However, it is not always the case that real-world data can be divided in such a neat manner. In many practical applications, data holders frequently possess only a subset of features for a given sample, rendering traditional horizontal or vertical divisions inapplicable. To illustrate, in the context of Internet of Things (IoT) devices, some devices may only be capable of collecting a subset of the feature data for a given user. This presents a challenge in

 doi:10.20944/preprints202503.2211.v1

aligning with the requirements of traditional federated learning frameworks, which typically assume a more uniform distribution of data across all participants.

While federated learning can provide a certain degree of data privacy, practical applications, particularly those involving distributed data across diverse institutions, devices, or users, have encountered significant challenges due to the issue of data heterogeneity. The heterogeneity of data from different participants, in terms of format, feature space, and distribution, presents a significant challenge in the construction of a unified and effective federated learning model in such environments. Furthermore, as an increasing number of regulations and policies impose stringent requirements for data privacy (such as the General Data Protection Regulation (GDPR) in the European Union), the need to enhance data privacy protection in federated learning and prevent potential information leakage during model training and aggregation has also emerged as a pressing issue that requires immediate attention.

The majority of current research and applications concentrate on horizontally and vertically partitioned datasets. However, there is a paucity of joint modeling studies on heterogeneous data partitioning. In practice, the diversity among data holders and the incompleteness of datasets underscore the urgent need to develop federated learning algorithms that can handle heterogeneous data. Such algorithms must be capable of addressing structural imbalances in data and identifying an optimal balance between computational complexity, privacy protection, and communication costs. This not only enhances the applicability of federated learning but also facilitates its extensive deployment in heterogeneous data-rich contexts, including the Internet of Things and intelligent healthcare.

Differential privacy and homomorphic encryption are common privacy-preserving techniques in federated learning. However, existing methods often suffer from inefficiencies in dealing with heterogeneous data. By integrating these technologies into the FL framework, this study enables effective data privacy protection even in the case of uneven data distribution. At the same time, the combination of these technologies improves the computational efficiency of the model and the overall robustness of the system.

Centralized data processing methods are heavily constrained in traditional data collection and processing due to concerns over data privacy and security. For that reason, FL for Gau-M has the potential of a new paradigm shift that can be inspired out of its nature of heterogeneity across institutions or devices. But the current FL methods still suffer from inefficiency and limited flexibility in scenarios where data is heterogeneous. We propose a new framework to solve existing methods limitations on heterogeneous data applications with advanced privacy protection. Therefore, this paper proposes a novel framework designed to address the inefficiencies and challenges encountered in the application of FL to heterogeneous data scenarios, by incorporating advanced privacy protection mechanisms.

2. Related Work

Jakub et al. [10] have previously highlighted that the objective of federated learning is to develop high-quality models from data distributed across a vast number of clients while maintaining the data locally. Jakub's proposed training approach entails each client independently computing the model's update parameters based on its local data, which are then conveyed to a central server. On the central server, the updated parameters from the individual clients are aggregated in order to calculate the new global model. In a previous study, Sharma et al. [11] proposed a federated transfer learning algorithm that enables the sharing of model knowledge and facilitates knowledge transfer between different neural networks, while ensuring data privacy. In this approach, the knowledge of the source domain is transferred to the target domain through the construction of a cross-domain model, thereby enhancing the learning capacity of the target domain task. The essence of federated transfer learning lies in the integration of the strengths of both federated learning and transfer learning. This approach not only safeguards data from unauthorized access but also enhances the model's generalization capacity when there are discrepancies between the source and target domains.

 doi:10.20944/preprints202503.2211.v1

In their study, Chromiak et al. [12] put forth a data model for heterogeneous data integration that takes into account various forms of data partitioning, including horizontal, vertical, and hybrid partitioning scenarios. However, this model does not safeguard the privacy of data during the process of data integration, which may result in an increased risk of privacy violations in practical applications. While the model is effective in addressing the fusion of heterogeneous datasets, the absence of essential privacy protection mechanisms represents a significant limitation in scenarios where data privacy is a primary concern.

Madaan et al. [13] conducted a further analysis of the necessity for data integration in the context of the Internet of Things (IoT) and identified a distinctive privacy leakage threat in this scenario. The data collected by IoT devices is often scattered and derived from a multitude of heterogeneous sources, increasing the risk of information leakage during integration, particularly in the absence of robust privacy protection mechanisms. The protection of data privacy in the context of the Internet of Things is not solely a matter of safeguarding personal information; it also entails ensuring the trustworthiness of devices and the security of the system as a whole. Consequently, the need for robust privacy protection is of paramount importance.

In order to address these challenges, Clifton et al. [14] proposed a research topic on privacy-preserving data integration. This topic discussed how data fusion from multiple data sources could be achieved through third-party matching (Matching) and integration techniques without violating user privacy. The aim was to provide secure query results. The study demonstrates that the deployment of third-party privacy protection mechanisms can effectively mitigate the concerns of all parties involved in the process of data integration. By encrypting and anonymizing the data, third parties can securely match and integrate multiparty data without direct access to the original data.

In their seminal work, Kasiviswanathan et al. [15] introduced the concept of Local Differential Privacy (LDP), a groundbreaking approach that eliminates the reliance on trusted third parties and enables users to locally perturb their data, thereby markedly enhancing privacy protection. By introducing noise directly into the data at the user level, local differential privacy provides a heightened level of privacy for each user, rendering it impossible for even the data collector or central server to recover the original data. This model is particularly advantageous in a decentralized data environment, as it reduces the necessity for a global trusted third party and expands the scope of privacy-preserving technologies.

On this basis, Wei et al. [16] combined differential privacy with federated learning to propose an improved privacypreserving federated learning framework (FL). In this framework, each participant introduces differential privacy noise into the local training process, thereby rendering it challenging to disclose the sensitive information of the original data even if the locally generated model parameters are intercepted. Subsequently, the noise-disturbed local model parameters are uploaded to a central server for integration during the global aggregation phase. This approach demonstrably enhances the security of federated learning, guarantees the confidentiality of participants, and enhances the resilience of the model.

Yu et al. [17] discussed the security of federated learning (FL) in horizontal and vertical data partitioning, particularly how the model poisoning attacks to be dealt with. In this study, the authors proposed the Rogue Device Detection (MDD) to detect and prevent the attack of rogue devices from interferingthe learning process. Anees et al. [18] presented a vertical federated learning framework based on neural networks and improved its performance using server integration techniques. In vertical federated learning, different parties share data features rather than splitting data by sample dimensions as in horizontal federated learning. Our study focuses on how to coordinate heterogeneous actors, and how to optimize the complexity of the framework through an integrated server.

3. Methodologies

3.1. Differential Privacy

 doi:10.20944/preprints202503.2211.v1

Initially, we assume that the distributions of datasets X and X , respectively, are p(X) and p X , then aligning the features of these distributions directly may result in confusion regarding the features in the high-dimensional data. Therefore, we adopt an optimal transport (OT)-based approach to map data with different distributions into a common space, while considering preservation of local structure. Transmission problem can be represented by solving the following convex optimization problem, denoted as Equation 1.

$$\min_{\in\Gamma(\mathrm{p}(\ ),\mathrm{p}\ \ )} \ \in\ ,\ \in \gamma(x\ ,\ x\ )d\ (x\ ,\ x\ )\ , \qquad 1$$

where $d(x\ ,\ x\ )$ represents a distance metric (for example, Euclidean distance) defined on the data feature space. $\gamma(x\ ,\ x\ )$, on the other hand, denotes the mapping matrix, which describes the manner in which the features between X and X are aligned. Γ(p(X), p X)represents a joint distribution set that satisfies the edge constraints.

$$\Gamma\ \mathrm{p}(X\ ),\ \mathrm{p}\ X\ \ = \{\gamma \in \mathbb{R}\ \ \ : \gamma 1$$

$$= \mathrm{p}(X\ ),\ \gamma^{\intercal} 1\ \ = \mathrm{p}\ X\ \ \}, \qquad 2$$

By solving this optimal transmission problem, we can ascertain the optimal methodology for mapping the variables X and X to the common feature space Z, thereby achieving the alignment of heterogeneous data. In order to solve this problem in an efficient manner, the optimal transmission method of entropy regularisation is employed. This rewrites the optimal transmission problem in Equation 2 as follows in Equation 3.

$$\min_{\in\Gamma(\mathrm{p}(\ ),\mathrm{p}\ )\ ,} \ \gamma(x\ ,\ x\ )d\ (x\ ,\ x\ ) - \epsilon H(\gamma), \qquad 3$$

where the term $H(\gamma) = -\sum\ ,\ \gamma(x\ ,\ x\ )log\ \gamma(x\ ,\ x\ )$ represents the entropy regularisation, whereas ϵ controls the regularisation intensity. As the privacy budget increases, the amplitude of the noise decreases, improving the performance of the model. Conversely, when the privacy budget is small, the noise increases and the privacy protection increases, but the accuracy of the model decreases.

In order to further enhance the flexibility of feature alignment, we combine kernel methods in order to map the data into a high-dimensional Reproducing Kernel Hilbert Space (RKHS). Assuming that the feature map is $\emptyset(\cdot)$, a suitable kernel function $k(x, y)$ may be selected, denoted as Equation 4.

$$k\ (x, y) = \exp\left(-\frac{\|\ x - y\ \|^2}{2\sigma^2}\right). \qquad 4$$

The kernel function allows us to define the corresponding kernel matrices, K and K , which represent the similarity of X and X in high-dimensional space. The optimal transport problem is generalized to the kernel space with the objective of achieving feature alignment in a more flexible feature space. The objective of the nuclear method is to minimize the distance between the two nuclear matrices, represented as Equation 5

$$\min_{\emptyset\ ,\emptyset\ \ \ ,} \ \|\ K - K\ \|\ , \qquad 5$$

where $\|\cdot\|$ represents the Frobenius norm, which serves to align disparate datasets in the kernel space by minimizing the distance between them. This allows the features of each heterogeneous data source to be mapped to a common high-dimensional space, thereby achieving feature fusion.

On the basis of feature alignment, further fusion of heterogeneous data is achieved through multi-view learning. In multi-perspective learning, it is postulated that there are K distinct perspectives, designated as V , V , ..., V , each of which encompasses features derived from disparate

 doi:10.20944/preprints202503.2211.v1

data sources. The implementation of feature fusion is achieved through the utilization of a weighted federated learning objective function, denoted as Equation 6.

$$\min_{\alpha, W} \| W^{\top}X - Y \|^2 + \lambda \sum_{k=1} \| \alpha_k \|_2^2 \quad (6)$$

where X represents the data feature matrix of the initial k perspective, W denotes the corresponding feature weight, α signifies the perspective weight, and Y represents the global target. By optimizing both α and W concurrently, it is possible to identify the most effective combination of features across different perspectives and achieve an adaptive fusion of heterogeneous data sources. In order to achieve dynamic weight adjustment, the Lagrangian multiplier method is introduced to optimize the perspective weights, with the final objective function being, shown as Equation 7.

$$L(W, \alpha, \lambda) = \alpha \| W^{\top}X - Y \|^2 + \lambda \sum_{k=1} \| \alpha_k \|_2^2 - \eta \left(\alpha_1 + \cdots + \alpha_k - 1 \right), \quad (7)$$

where η represents the Lagrangian multiplier, which is employed to ensure the normalization of the weights associated with each perspective. By optimizing gradient descent for this objective function, the weight α and feature matrix W of each perspective can be adjusted dynamically, thus enabling the adaptive integration of heterogeneous data. Figure 1 shows framework diagram of proposed model.

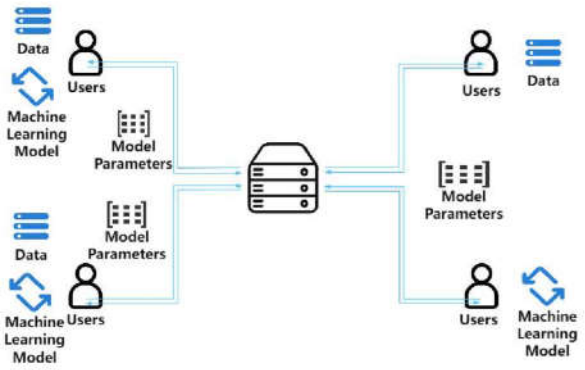


Figure 1. Model Architecture with Federated Learning.

3.2. Privacy Protection Mechanism

The protection of privacy is a pivotal concern in the context of federated learning frameworks, particularly in scenarios where multi-party collaboration occurs within heterogeneous data environments. In order to guarantee the confidentiality of data during the processes of model training and data transfer, we have integrated sophisticated privacy-preserving technologies, including differential privacy, homomorphic encryption and secure multi-party computation. In this section, we will undertake a detailed analysis of the complex mathematical and formulaic aspects underlying these privacy protection mechanisms.

The differential privacy mechanism conceals the individual data of the participants by introducing noise during the model update process. This ensures that even if an attacker gains access to information from the training process, it is not possible to derive the data of a single participant.

 doi:10.20944/preprints202503.2211.v1

Let us consider the scenario in which the gradient provided by participant i is updated to $\Delta\theta$ at each round of t model update. In order to protect the privacy of this gradient, we introduce the differential privacy noise $\eta \sim \mathcal{N}(0, \sigma)$, which is finally updated as Equation 8.

$$\Delta\theta = \Delta\theta + \eta . \quad (8)$$

The variability of the noise σ is determined by the privacy budget ϵ and the sensitivity Δf. Sensitivity Δf is defined as Equation 9.

$$\Delta f = \max \| \theta(X) - \theta X \| . \quad (9)$$

In accordance with the definition of ϵ -differential privacy, it is imperative to ensure that the gradient update subsequent to noise addition satisfies the following Equation 10.

$$\mathbb{P}\left(\mathcal{M}(X) = y\right) \leq e \cdot \mathbb{P}\left(\mathcal{M}\ X' = y\right) . \quad (10)$$

In the event that datasets X and X are adjacent, the perturbed model $\mathcal{M}$ is employed. The noise is selected to be Gaussian distributed, and its standard deviation, denoted as σ, is determined by the following Equation 11.

$$\sigma = \frac{\Delta f}{\epsilon}\sqrt{2\log\left(\frac{1.25}{\delta}\right)}, \quad (11)$$

where the term δ represents the probability of a privacy failure.

In addition to global differential privacy, local differential privacy can be employed to introduce noise to the data at the local level, rather than at the server side. Assuming that the local data of the participant's i is X , the data will first be perturbated through the Randomized Response mechanism. This ensures that each participant can add noise locally, thereby guaranteeing that even the data received on the server side cannot be directly recovered from the original data, denoted as Equation 12.

$$X' = X + \eta , \eta \sim \mathcal{N}\left(0, \sigma^2\right) . \quad (12)$$

Homomorphic encryption represents a highly sophisticated privacy-preserving technology, enabling direct computation of encrypted data. In federated learning, the use of homomorphic encryption enables the server to avoid decrypting the data when updating model parameters, thereby preventing any potential data leakage.

Let us consider the case where the local gradient of the participant i is updated to $\Delta\theta$. In this scenario, the public key encryption mechanism is employed to encrypt the gradient, with the encryption operation represented by $Enc(\cdot)$ .In homomorphic encryption, the following homomorphism is satisfies Equation 13.

$$Enc(\Delta\theta) + Enc\ \Delta\theta = Enc\ \Delta\theta + \Delta\theta . \quad (13)$$

The server is capable of directly assessing the relative importance of the participants in the encrypted state, denoted as Equation 14.

$$Enc(W) = \frac{1}{N} Enc(\Delta\theta) . \quad (14)$$

Subsequently, upon the necessity for decryption, the server is able to decrypt the result W utilizing the private key sk, calculated as Equation 15.

$$W = Dec\ Enc(W) . \quad (15)$$

To further reinforce the security measures in place, we employ the use of homomorphic encryption mechanisms with noise resilience, such as BFV (Brakerski-Fan-Vercauteren) or CKKS (Cheon-Kim-Kim-Song) encryption algorithms. The aforementioned algorithms guarantee that the

 doi:10.20944/preprints202503.2211.v1

results can be retrieved even in the event of noise amplification during the computation by incorporating a noise term into the encryption process. The accumulation of noise is a gradual process that occurs during the course of computation. Consequently, it is necessary to implement a process of fine-tuning in order to maintain the desired level of noise. Let us consider the encrypted gradient $\Delta\theta$, which contains an implicit noise term η . The objective is to ensure that the final noise η remains within the allowable threshold σ , described as Equation 16.

$$\| \eta \| \leq \sigma \quad . \qquad 16$$

After applying a weighted average, the cumulative form of noise can be expressed as Equation 17.

$$\eta = \frac{1}{N} \eta_1 . \qquad 17$$

In each encryption iteration, the variance σ of the noise is adjusted in order to control the final accumulated noise and ensure the accuracy of the resulting calculation.

3.3. Secure Multi-Party Computation

In light of the aforementioned considerations, we put forth a multi-level privacy protection strategy encompassing three stages: data preprocessing, model training, and result aggregation. In the initial phase of data processing, the elimination of personally identifiable information is achieved through the utilisation of de-identification technology. The hash function $h(\cdot)$ is employed to encode all potentially exposed data, described as Equation 18.

$$h(X) = hash(X), \qquad 18$$

This process ensures that any identifiable information is removed from the data set prior to its integration into the federated learning system. In the model training phase, we employ the use of Secure Multi-Party Computation technology. Each participant performs their own gradient update $\Delta\theta$ locally and then decomposes the updated value into multiple sub-parts, which are then sent to other participants through secret sharing, denoted as Equation 19.

$$\Delta\theta = \quad s , \qquad 19$$

Once each participant has received a portion of the other participant's data, local aggregation is conducted, and security protocols are employed to prevent the intermediate computation process from disclosing the data. In the result aggregation stage, the Federated Averaging technique is employed to update the global model, whereby the local model updates of each participant are weighted by Equation 20.

$$W = \frac{1}{N} \Delta\theta . \qquad 20$$

In this instance, the confidentiality and integrity of the aggregated data are assured by the use of homomorphic encryption and differential privacy.

4. Experiments

4.1. Experimental Setups

The experiment utilized the PUMS (Public Use Microdata Sample) dataset, a publicly accessible data set provided by the United States Census Bureau that contains comprehensive, anonymized

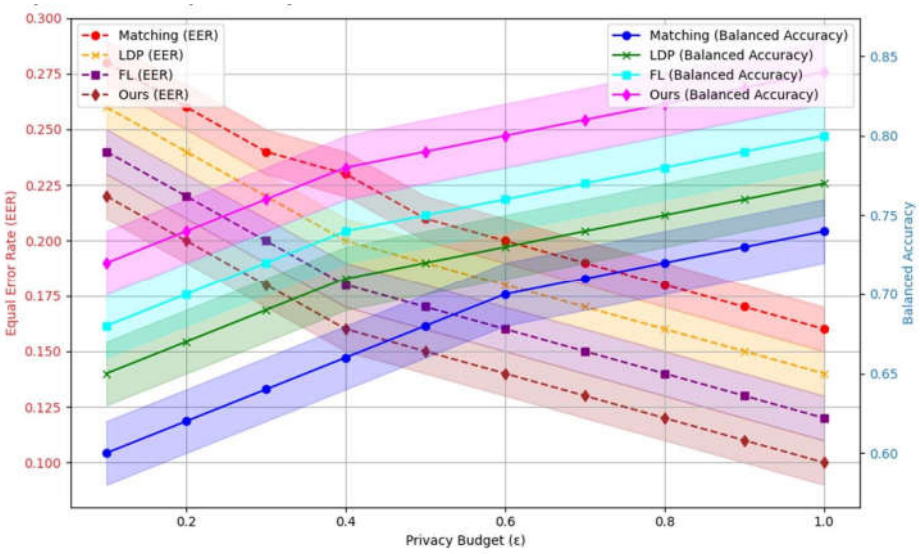
 doi:10.20944/preprints202503.2211.v1

information about the population and housing. The dataset encompasses a multitude of characteristics, including age, gender, educational attainment, occupation, income, family structure, and financial attributes such as credit scores, investment types, and loan types. The privacy budget is ε = 1.0, the noise variance is σ = 0.5, and the number of iterations is set to 100. The global model employs a two-layer, fully connected neural network, wherein each layer comprises 64 neurons, and the activation function utilizes the rectified linear unit function. The learning rate is set to 0.01, and the optimization algorithm utilizes Adam. The experiment was conducted in parallel on 10 different data nodes, each containing a heterogeneous data source, thereby simulating a real-world distributed data scenario.

4.2. Experimental Analysis

The Equal Error Rate (EER) is a pivotal metric for evaluating the efficacy of classifiers, particularly in the context of binary classification problems. The EER represents the error rate at which the false positive and false negative rates are equal. It provides a comprehensive assessment of the model's error rate between different classes by identifying the equilibrium between the two. A lower EER indicates that the model demonstrates superior performance in balancing positive and negative class errors, as evidenced by a lower error rate. Further, balanced accuracy, which is employed to address data imbalances, represents a variant of computational accuracy that balances the effects of category imbalances by averaging the accuracy of each category. By considering the recall rate of each class simultaneously, the equalization accuracy can more equitably assess the classifier's performance on an imbalanced dataset, which is particularly well-suited to scenarios where class distributions are uneven.

Figure 2 shows the experimental comparison of the Balanced Error Rate (EER) and Balanced Accuracy of the four methods Matching, LDP, FL and Ours under different privacy budgets on the PUMS dataset, and the error range is indicated by shading. The privacy budget is increased from 0.1 to 1, and the curves for EER and equalization precision correspond to the performance of different methods. It can be seen that with the increase of privacy budget, the equalization error rate of each method gradually decreases, while the equalization accuracy gradually increases, especially the Ours method shows the best comprehensive performance when the privacy budget is high. When the privacy budget is low, the noise added by the Local Differential Privacy (LDP) method increases, resulting in more serious data distortion, which affects the training effect of the model. In contrast, other methods, such as those proposed in this paper, are able to better balance privacy protection and model performance by combining global and local privacy protection mechanisms.



 doi:10.20944/preprints202503.2211.v1

Figure 2. Comparison of EER and Balanced Accuracy with Different Privacy Budgets on PUMS Dataset.

Resource utilization is an indicator that assesses the effective utilization of computing resources on disparate devices and nodes in the context of federated learning. By monitoring the utilization of CPU, memory, and network bandwidth on each node, one can analyze the balance and effectiveness of the model in resource allocation. The evaluation of resource utilization is beneficial for the optimization of the deployment strategy of the algorithm and the assurance that disparate devices are able to reasonably allocate tasks in accordance with their respective performance levels. This is particularly relevant in the context of heterogeneous computing environments, which encompass both high-computing power servers and low-computing edge devices. Figure 3 presents a comparative analysis of the resource utilization experiments conducted for the four methods, namely Matching, LDP, FL, and the proposed method, under varying privacy budgets. As the privacy budget increases, so does the resource utilization of each method. The Ours method demonstrates the highest resource utilization across all privacy budgets. This indicates that the proposed approach exhibits superior performance in terms of resource management and efficiency, particularly in the context of high privacy budgets.

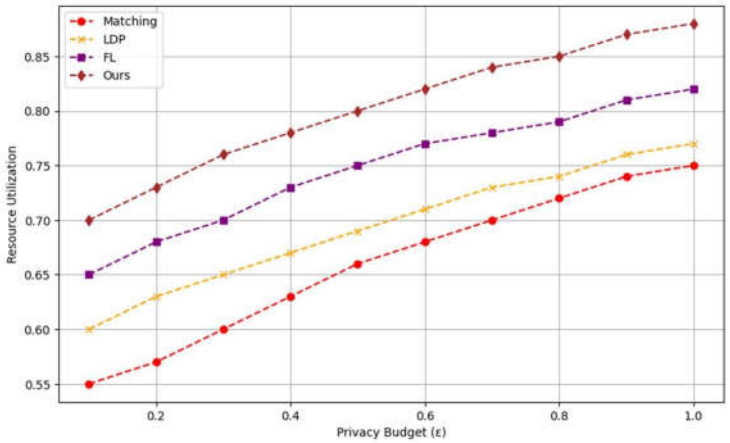


Figure 3. Comparison of Resource Utilization with Different Privacy Budgets on PUMS Dataset.

While privacy protection represents a fundamental prerequisite for federated learning, the extent to which users can trust that their privacy is adequately safeguarded also constitutes a crucial aspect of evaluation. The term "privacy transparency" is used to describe the extent to which users are able to comprehend and regulate the privacy mechanisms employed by the system. By comparing the interpretability and transparency of different privacy policies (such as the noise mechanism of differential privacy and the encryption scheme of homomorphic encryption), the extent to which users trust and accept the federated learning system can be evaluated. Figure 4 depicts user engagement on a scale of 0.1 to 1.0 and compares it by calculating privacy

 doi:10.20944/preprints202503.2211.v1

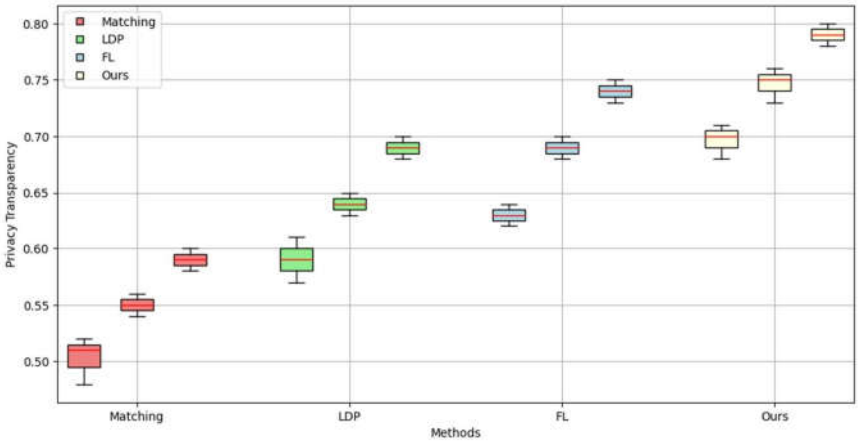


Figure 4. Comparison of Privacy Transparency at Different User Engagement Levels.

5. Conclusion

In conclusion, this work presents a comprehensive comparison of the privacy transparency of various data processing methods, including matching, LDP, FL, and our proposed approach, across different user engagement levels. The experimental results demonstrate that the various methods exhibit distinct characteristics with regard to their efficacy in safeguarding privacy, with the "Ours" method exhibiting optimal transparency and stability when user engagement is high. This demonstrates that the integration of heterogeneous data fusion and differential privacy technology can enhance the effective utilization of data while maintaining the confidentiality of the data subjects. In light of the ongoing expansion in data volume and the concomitant rise in data privacy requirements, the integration of federated learning with privacy protection will assume an increasingly pivotal role. Further research should concentrate on enhancing the computational efficiency of privacy-preserving mechanisms, strengthening the resilience of data integration, and addressing the complexities of heterogeneous data environments. transparency. As user engagement increases, the privacy transparency of various methods improves, with our method consistently performing the best. This suggests that our method offers superior transparency and user trust.

References

1. Jia, Bin, et al. "Blockchain-enabled federated learning data protection aggregation scheme with differential privacy and homomorphic encryption in IIoT." IEEE Transactions on Industrial Informatics 18.6 (2021): 4049-4058.
2. Rodríguez, Eva, Beatriz Otero, and Ramon Canal. "A survey of machine and deep learning methods for privacy protection in the internet of things." Sensors 23.3 (2023): 1252.
3. Akter, Mahmuda, et al. "Edge intelligence: Federated learning-based privacy protection framework for smart healthcare systems." IEEE Journal of Biomedical and Health Informatics 26.12 (2022): 5805-5816.
4. Lei, Moyixi, et al. "Integration of privacy protection and blockchainbased food safety traceability: Potential and challenges." Foods 11.15 (2022): 2262.
5. He, Zaobo, Lintao Wang, and Zhipeng Cai. "Clustered federated learning with adaptive local differential privacy on heterogeneous iot data." IEEE Internet of Things Journal (2023).
6. Phan, Thanh Chi, and Hung Chi Tran. "Consideration of data security and privacy using machine learning techniques." International Journal of Data Informatics and Intelligent Computing 2.4 (2023): 20-32.
7. Huang, Chenxi, et al. "A robust approach for privacy data protection: IoT security assurance using generative adversarial imitation learning." IEEE Internet of Things Journal 9.18 (2021): 17089-17097.
8. Bi, Hongliang, Jiajia Liu, and Nei Kato. "Deep learning-based privacy preservation and data analytics for IoT enabled healthcare." IEEE Transactions on Industrial Informatics 18.7 (2021): 4798-4807.
9. Gures, Emre, et al. "Machine learning-based load balancing algorithms in future heterogeneous networks: A survey." IEEE Access 10 (2022): 37689-37717.
10. Konečný, Jakub. "Federated Learning: Strategies for Improving Communication Efficiency." arXiv preprint arXiv:1610.05492 (2016).

 doi:10.20944/preprints202503.2211.v1

11. Sharma, Shreya, et al. "Secure and efficient federated transfer learning." 2019 IEEE international conference on big data (Big Data). IEEE, 2019.
12. Chromiak, Michał, and Krzysztof Stencel. "A data model for heterogeneous data integration architecture." Beyond Databases, Architectures, and Structures: 10th International Conference, BDAS 2014, Ustron, Poland, May 27-30, 2014. Proceedings 10. Springer International Publishing, 2014.
13. Madaan, Nishtha, Mohd Abdul Ahad, and Sunil M. Sastry. "Data integration in IoT ecosystem: Information linkage as a privacy threat." Computer law & security review 34.1 (2018): 125-133.
14. Clifton, Chris, et al. "Privacy-preserving data integration and sharing." Proceedings of the 9th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery. 2004.
15. Kasiviswanathan, Shiva Prasad, et al. "What can we learn privately?." SIAM Journal on Computing 40.3 (2011): 793-826.
16. Wei, Kang, et al. "Federated learning with differential privacy: Algorithms and performance analysis." IEEE transactions on information forensics and security 15 (2020): 3454-3469.
17. Yu, Chong, et al. "Secure and Efficient Federated Learning Against Model Poisoning Attacks in Horizontal and Vertical Data Partitioning." IEEE Transactions on Neural Networks and Learning Systems (2024).
18. Anees, Amir, Matthew Field, and Lois Holloway. "A neural networkbased vertical federated learning framework with server integration." Engineering Applications of Artificial Intelligence 138 (2024): 109276.