Proceedings of the IISE Annual Conference & Expo 2025
E. Gentry, F. Ju, X. Liu, eds.

High-Quality Privacy Preserving Text Data Sharing for Manufacturing Machine Learning Data Market

Hui Liu[1], Yingyan Zeng[3], Premith Kumar Chilukuri[2], Ran Jin[1]
[1] Grado Department of Industrial and Systems Engineering Virginia Tech
[2] Department of Computer Science, Virginia Tech
[3] Department of Mechanical and Materials Engineering, University of Cincinnati

Abstract

We investigated a multimodal dataset sharing method to share valuable text data from technical documents and narratives to improve Artificial Intelligence (AI) models in manufacturing. High-quality and informative datasets are essential for AI model training and deployment performances. The sharing of privacy-preserving proxy datasets, distilled from numerical raw data owned by other manufacturing stakeholders, can augment the local datasets and has proven to improve the AI model performance. However, it is challenging to share manufacturing text data under privacy preserving constraints, which is critical to protect know-how and IP information. The text data come from technical narratives about manufacturing. There is limited approach to share them due to limited access to such text data. In this paper, we modeled manufacturing domain knowledge and perceptions by employing multiagent-based large language models (LLMs) to generate high-quality, personalized text data. Then we integrate Multimodal-Aligned Variational Autoencoder (MAVAE) to fuse both text and numerical datasets to achieve privacy-preserving data sharing. We validated the proposed method based on microbial fuel cell (MFC) anode design problem with a focus to use text data to improve the design feasibility prediction by AI models. Different LLM agents are tuned to simulate different design styles, such as design space preferences and design rule configurations. The MAVAE encodes both numerical and text features into a shared latent space and predicts the post-sharing AI model performances for data-sharing decision-making. This method achieves an average F1 score of 0.928, outperforming baseline approaches such as Differential Privacy (0.898). This method is expected to be adopted by practitioners to share both numerical and text data in a data market.

Keywords
Data market, Large language models, Manufacturing design, Privacy preserving, Text data sharing

1. Introduction
With the advancement of smart manufacturing and Industry 4.0, artificial intelligence (AI) plays a crucial role in optimizing production efficiency, improving product quality, enhancing the product design flexibility, enabling the supply chain resilience, and reducing manufacturing costs. However, the effectiveness of AI models depends on high-quality multimodal data collaboration, requiring solutions for cross-modal knowledge fusion under privacy constraints and efficient heterogeneous data representation [1, 2].

In manufacturing multimodal data sharing, federated learning (FL) and differential privacy (DP) have advanced privacy protection and analysis for numerical data, yet privacy-preserving text data sharing remains challenging [1]. Current frameworks use numerical data (e.g., design variables, process parameters, *in situ* process data) for AI model training, but they fall short in capturing the contextual depth of design intent and rules [2]. While text data can supplement this implicit knowledge, its issues with missing information, unstructured formats, and privacy risks complicate meaningful representation learning; simple noise perturbation undermines semantic integrity [3, 4]. There is a lack of federated semantic modeling and privacy-enhanced text representation learning, which limits data security and sharing efficiency. Moreover, integrating numerical and text data is challenging due to structural and semantic differences [2, 5]. Although variational autoencoders (VAEs) and contrastive learning have been applied to image-numerical data fusion [6], they are less effective in text-numerical fusion, often failing to capture complex text semantics and leading to information redundancy [4].

To address these challenges, this paper proposes a privacy-preserving multimodal data-sharing framework that enables effective data fusion and secure sharing through a Multimodal-Aligned Variational Autoencoder (MAVAE) and multi-agent large language models (LLMs). The key innovation lies in MAVAE's dual-encoder structure, which integrates

numerical and textual data to generate proxy datasets, preserving essential design information while ensuring privacy protection. This approach allows secure data sharing without exposing sensitive process details. Additionally, KL divergence regularization and cross-modal alignment loss enhance modality consistency, improving heterogeneous data fusion. Multi-agent LLMs further refine text generation by simulating domain-specific design styles, enhancing semantic adaptability. Compared to flexible dataset distillation (FDD) [4], the proposed method enables privacy-aware multimodal fusion while maintaining high data utility for feasibility prediction. In this study, Privacy-preserving means protecting confidential data, securing target AI task details, and controlling access before data sharing [2]. Utility refers to the effectiveness of these shared proxies in supporting downstream AI tasks. The objective of this study is to maximize data utility while adhering to strict privacy constraints. In other words, our evaluation metric focuses solely on model performance.

The framework is validated using microbial fuel cell (MFC) anode design as a case study. Results demonstrate significant improvements in feasibility prediction accuracy and enhanced adaptability in cross-designer collaborative optimization. The remainder of this paper is organized as follows: Section 2 details the proposed method; Section 3 presents experimental validation; and Section 4 concludes the study with future research directions.

2. Methodology

2.1 Overview of the Proposed Methodology

This study explores a multi-designer data-sharing problem in MFC, where diverse design styles create variations in variable spaces, criteria, and combinatorial logic, resulting in a binary feasibility indicator [7]. The objective is to improve feasibility modeling via privacy-preserving data sharing.Table 1 presents an example dataset.

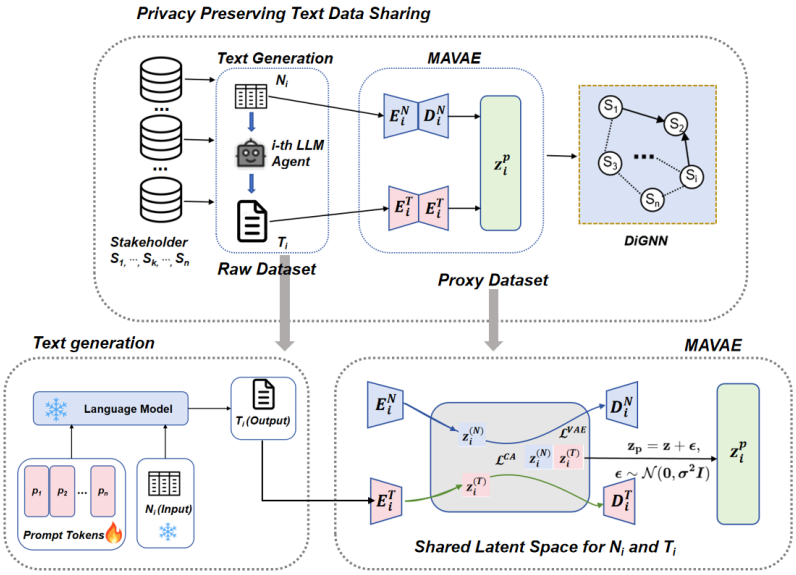

Figure 1: Overview of the Proposed Method Framework (Redrawn from [2] with authors' permission)

Figure 1 illustrates the LLM-based multimodal data-sharing framework. The process consists of three key components: (1) Text generation via LLMs: Multi-agent LLMs generate text descriptions of design styles, enriching the representation of domain-specific knowledge in manufacturing. (2) Multimodal fusion with MAVAE: MAVAE employs a modality-separated encoder to learn aligned representations of numerical and text data, integrating them into a shared latent space. Generative models further refine the multimodal representation Z_i, reducing heterogeneity while retaining critical task-related information. (3) Privacy-preserving data sharing via DiGNN: A DiGNN encodes dataset-sharing decisions among stakeholders, leveraging Z_i as node features to efficiently encode sharing decisions and predict post-sharing AI performance. The key innovation of our paper is the structured fusion of textual and numerical data for privacy-preserving data sharing in manufacturing. Compared with our earlier work [2], we employed multiagent-based large language models (LLMs) to generate high-quality, personalized text data and integrated a Multimodal-Aligned Variational Autoencoder (MAVAE) to enhance multimodal learning while protecting sensitive information.

The proposed method offers two major advantages: (1) Multimodal data fusion: MAVAE enables effective integration

of high-quality LLM-generated text with numerical data, ensuring a unified, low-dimensional representation while reducing heterogeneity. (2) Privacy-preserving sharing: Instead of direct raw data exchange, proxy data sharing protects sensitive information while balancing data utility and privacy.

Table 1: Design Samples with Numerical and Text Data

Sample	Numerical Data	Text Data	Design	Feasibility {0,1}
1	cell_type = SchwarzPrimitive cell_count = 4 ...	"Based on the design input provided, the feasibility analysis indicates that..."		1
2	cell_type = BCC cell_count = 3 ...	"The design input includes a "BCC" cell type, a cell count of ..."		0
		...		

2.2 Text Generation

LLM agents are trained to generate text data reflecting different design styles for each design sample from various stakeholders (i.e., designers in this paper). Each stakeholder S_i, associated with a numerical dataset N_i, follows goal definition, data type and text style analysis, and customized output prioritization to generate text data for different design scenarios. An optimization process ensures that the generated text aligns with specific requirements.

Figure 1 illustrates the text generation workflow, where soft prompt tuning is applied before LLM inference to optimize input prompts. It refines text generation by training soft prompt embeddings, guiding the LLM without modifying its core parameters [8]. These trainable vector prompts replace manual templates, adapting automatically to different manufacturing and design contexts.

In soft prompt tuning, a sequence of learnable embeddings is defined as $P = (p_1, p_2, ..., p_n)$, where each p_i is a trainable vector in the LLM's token embedding space. The modified input is given by $X_{\text{input}}^{\text{soft}} = (P, X_{\text{input}})$. The LLM generates text T_i, and its output $\hat{y}_t$ is optimized using cross-entropy loss: $L_{\text{text}} = -\sum_{t=1}^{T} y_t \log \hat{y}_t$, where only P is updated via gradient descent, preserving pre-trained LLM parameters. Soft prompting enables the LLM to efficiently adapt to new tasks, improve contextual understanding, and reduce the dependency on manual prompt engineering.

To improve tuning performance, we curate high-quality training data, pairing numerical inputs with annotated text reflecting stakeholder requirements, industry terminology, and stylistic nuances, ensuring context-aware and consistent text generation. Due to space limitations, details are omitted.

2.3 Multimodal-Aligned Variational Autoencoder

To ensure privacy while creating a universal embedding for high-dimensional multimodal inputs, MAVAE processes numerical data N_i, $i \in N$ and text data T_i by utilizing proxy datasets, which serve as privacy-preserving representations of the raw data. MAVAE was selected over other methods because it enables cross-modal alignment and latent-space regularization, which are essential for fusing heterogeneous modalities while preventing exposure of sensitive information [6]. MAVAE projects the modality of input data into a latent space z_i using modality-specific encoders E_i^N and E_i^T. The encoded representations are then reconstructed using decoders D_i^N and D_i^T, ensuring that the data remains informative while preventing direct access to raw values.

To train the MAVAE effectively, we employ a variational autoencoder (VAE) loss, which consists of reconstruction losses for both numerical and text data, as well as a KL divergence regularization term[6]

$$L_{\text{VAE}} = \sum_i \left[-\log p_\theta(N_i|z_i^{(N)}) - \log p_\theta(T_i|z_i^{(T)}) + \lambda D_{\text{KL}}(q_\phi(z_i^{(N)}, z_i^{(T)}|x_i) \| p_\theta(z)) \right] \tag{1}$$

where $z_i^{(N)} = E_i^N(N_i)$ is the latent representation of numerical data, $z_i^{(T)} = E_i^T(T_i)$ is the latent representation of text data, $D_i^N(z_i^{(N)})$ reconstructs numerical data from the latent space, $D_i^T(z_i^{(T)})$ reconstructs text data from the latent space, and $D_{\text{KL}}(q_\phi \| p_\theta)$ ensures the latent space follows a prior distribution.

Additionally, we introduce a cross-modality alignment loss to encourage consistency between the numerical and text

latent representations as

$$L_{\text{CA}} = \sum_i \|z_i^{(N)} - z_i^{(T)}\|^2. \tag{2}$$

We concatenate the latent representations of numerical and text data to form the combined latent variable as $z_i = [z_i^{(N)}; z_i^{(T)}]$. To incorporate AI task-specific objectives such as the design feasibility selection in this paper, a classification loss is added when the task is classification-based. This loss is designed to optimize the latent representations z_i for downstream tasks, which is defined as

$$L_{\text{clf}} = -\sum_i [y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)] \tag{3}$$

where y_i represents the ground truth label, and $f_{\text{clf}}(z_i)$ denotes the classifier's prediction logit derived from the latent space representation z_i. These losses ensure that the shared latent space z_i effectively captures the multimodal relationships between numerical and text data while optimizing for privacy-preserving, high-performance downstream tasks. Once trained, MAVAE enables the generation of proxy datasets by sampling latent variables from the learned distribution as $z_i^p = z_i + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. This ensures that the generated proxy dataset z_i^p retains essential information while preserving privacy. The sampled latent variables can then be used for downstream tasks such as feasibility assessment and design optimization.

2.4 Directed Graph Neural Network for Dataset-Sharing Decisions

To facilitate dataset-sharing decisions and enhance AI model performance, we formulate the problem as a graph-based learning task. Each stakeholder, with a dataset containing numerical and text data, is represented as a node $v \in \mathcal{V}$. Directed edges $(u,v) \in \mathcal{E}$ denote dataset-sharing actions from data owner to receiver, forming a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The edge direction reflects dataset transfer, and performance gains on the target testing set D_i^T serve as evaluation metrics.

Graph neural networks (GNNs) encode node features and aggregate representations [2, 9]. We adopt a directed acyclic graph (DAG)-based structure where each node representation is iteratively updated using message passing. At the t-th layer, the aggregated message for node v is computed as $h^G = f_{\text{FC}}\left(\text{Max-Pool}_{v\in\mathcal{V}}\left([h_v^1, \ldots, h_v^L]\right)\right)$. where MLP is a multi-layer perceptron that processes the aggregated information from neighboring nodes and edge features to learn higher-level representations, h_u^{t-1} is the previous layer's representation of node u, and $\gamma(u,v)$ encodes edge features. The node representation is updated via a Gated Recurrent Unit (GRU) as $h_v^t = f_{\text{GRU}}(h_v^{t-1}, m_v^t)$. The final graph representation is obtained by max-pooling all node embeddings

$$h^G = f_{\text{FC}}\left(\text{Max-Pool}_{v\in\mathcal{V}}\left([h_v^1, \ldots, h_v^L]\right)\right). \tag{4}$$

Dataset-sharing performance is evaluated using the F1 score on the target testing set at the data receiver. The graph-level loss for a single decision is defined as $L^G = \frac{1}{n}\sum_{j=1}^n (y_j^G - h_j^G)^2$, where y_j^G is the ground truth F1 score and h_j^G is the predicted score. Thus, DiGNN predicts post-sharing AI performance based on the latent representations of each stakeholder's dataset as nodes and sharing decisions as directed edges.

3. Case Study

The proposed method evaluates privacy-preserving data sharing among six MFC designers to predict binary design feasibility. MFCs generate electricity via microbial metabolism, offering a sustainable energy solution [7]. Assessing feasibility before production is crucial for optimizing design, reducing lead time, and increasing manufacturing success. However, the vast design space and diverse styles limit individual designers' predictive capabilities due to insufficient data, motivating the need for privacy-preserving data sharing. To validate the design feasibility prediction, we adopt a DNN-based classification model using both numerical and text features [2].

Data sharing enables collaboration among designers with similar rules and parameters, improving feasibility prediction accuracy by integrating diverse expertise.

3.1 Datasets Information

During the design phase, designers define the design space using variables such as unit type, unit quantity, volume fraction factor, layer thickness, and rotation angles. Intermediate 3D properties (e.g., minimum feature size, cavity thickness, and number of discontinuous volumes) are simulated and evaluated for feasibility based on predefined design rules [7]. Differences in design spaces and rules lead to distinct design styles, categorized as passionate, traditional, and eco-friendly. Passionate designers push design limits with extreme values, traditional designers favor mid-range stable parameters, and eco-friendly designers prioritize material efficiency to reduce environmental impact.

To model these styles, we developed three LLM agents to generate style-specific text descriptions. By encoding design rules and parameter characteristics, the agents captured stylistic preferences and integrated them into text generation. The synthesized data accurately reflects both design constraints and designer philosophies, aiding decision-making and collaborative optimization. This study used datasets from six designers (two per style) modifying design spaces and rules to create six MFC design datasets. Each dataset includes numerical design variables, LLM-generated text descriptions, and binary feasibility labels.

3.2 Hyperparameter Settings and Benchmark Methods

In the text generation process, the model used is specified as model="gpt-3.5-turbo", with max_tokens=1000 and temperature=0.7. For the MAVAE process, The dimension of z_i is set as 32, resulting in a shared space with 64 dimensions. The KL-weight λ is tuned to be 0.1 based on empirical validation using cross-validation and grid search. We evaluate the proposed framework in benchmark comparison. Although there is no approach that can address all the challenges mentioned in the literature, we adopt the following benchmark methods that partially address the challenges in dataset sharing as state-of-the-art solutions: (1) Differential Privacy (DP): Protects data privacy by adding carefully calibrated noise to data or model updates, ensuring that individual data points cannot be distinguished and preventing information leakage during dataset sharing [1]. (2) OptimShare: Enables privacy-preserving data sharing and collaborative model optimization through federated learning, allowing multiple parties to jointly train machine learning models without exposing their raw data [10].

3.3 Experimental Results

In a data market with six designers, each designer can share data in 31 possible ways with the remaining five designers ($5+10+10+5+1=31$), resulting in $31\times 6=186$ dataset-sharing decisions. The proposed method is evaluated using 10-fold cross-validation to train DiGNN and assess feasibility classification. Table 2 compares classification performance under different data-sharing strategies, highlighting the impact of text data integration and cross-designer sharing. Each column (D1-D6) represents a designer, with Row 1 showing the varying sample sizes of local datasets. The results indicate that incorporating text data (Rows 2-3) improves F1 scores, demonstrating its importance in feasibility prediction. Additionally, the results show that data sharing improves F1 scores compared to local datasets(Rows 3-6), and DiGNN (Rows 8-9) achieves performance close to the best dataset combination (Rows 6-7) while being more efficient than exhaustive search.

Data sharing consistently improved model performance by increasing parametric and stylistic diversity. Cross-designer integration expanded design space coverage, enriched pattern recognition, and provided complementary solutions to design challenges. This combinatorial approach enhanced model adaptability for complex tasks and improved prediction accuracy, particularly for extreme parameter configurations.

Table 2: F1 Score Comparison of Different Data Sharing Strategies

Row		D1	D2	D3	D4	D5	D6
1	Sample Size from Each Designer	154	179	100	96	253	224
2	Local F1 Score using Numerical Data	0.668	0.764	0.731	0.715	0.730	0.652
3	Local F1 Score using both Numerical and Text Data	0.691	0.835	0.816	0.741	0.935	0.818
4	Best F1 from One Shared Proxy Dataset	0.759	0.875	1	0.881	0.975	1
5	Best One Shared Proxy Dataset	D5	D5	D1	D6	D2	D1
6	Best F1 from All Comb. of Shared Proxy Datasets	0.798	0.913	1	0.937	1	1
7	Best Comb. of Shared Proxy Datasets	D3, D5	D3, D5, D6	D4, D5	D2, D3, D5, D6	D1, D2, D3, D4, D6	D3, D5
8	Our Method (F1 using DiGNN from 10-Fold CV)	0.798	0.890	1	0.881	1	1
9	Our Method (Shared Datasets)	D3, D5	D4, D5, D6	D4, D5	D3, D5, D6	D1, D2, D3, D4, D6	D3, D5

The proposed method, calculated across six datasets, achieves an average F1 score of **0.928**, which is better than Differential Privacy (0.898) and OptimShare (0.902) over 10-fold CV. This improvement stems from the MAVAE framework, which integrates numerical and textual data through a dual-encoder structure, generating privacy-preserving proxy datasets that maintain essential design information. By leveraging KL divergence regularization and cross-modal alignment loss, MAVAE ensures effective modality fusion while preserving data privacy. Additionally, the method optimizes dataset-sharing decisions by encoding stakeholder relationships using DiGNN, leading to better generalization and improved feasibility prediction.

4. Conclusion

This study presents a privacy-preserving multimodal data-sharing framework integrating multi-agent large language models (LLMs) and a Multimodal-Aligned Variational Autoencoder (MAVAE) to enhance manufacturing machine learning models. By leveraging LLMs, the proposed method generates high-fidelity text data, capturing diverse design styles and improving design feasibility predictions. MAVAE fuses numerical and text data while ensuring privacy protection through latent space transformations, preventing direct exposure of sensitive information. The framework was validated using MFC design as a case study, demonstrating significant improvements in feasibility prediction accuracy after privacy-preserving data-sharing decisions. Experimental results show that data-sharing strategies among designers enhance predictive performance by integrating knowledge across different design schemes. Compared to existing methods, the proposed approach achieves superior F1 scores while maintaining strong privacy constraint.

In future research, there are several promising directions to perform further investigations: (1) the paper assumes one dataset for each stakeholder, which can be relaxed for multiple datasets with similar statistical distributions for data sharing; (2) there is only one AI modeling task for each stakeholder, which will be extended to multiple AI tasks based on the same source of datasets; and (3) larger scale validation will be performed, especially based on different types of stakeholders with different objectives in AI tasks.

Acknowledgement

The authors acknowledge the funding support from National Science Foundation CMMI-2331985.

References

[1] Ebadi, H., Sands, D., and Schneider, G., 2015, "Differential Privacy: Now It's Getting Personal," ACM SIGPLAN Notices, vol. 50, no. 1, pp. 69–81.

[2] Zeng, Y., Zhou, X., Chilukuri, P. K., Lourentzou, I., and Jin, R., 2025, "High-Quality Dataset-Sharing and Trade Based on A Performance-Oriented Directed Graph Neural Network," IEEE Transactions on Automation Science and Engineering (Accepted).

[3] Chen, X., and Jin, R., 2024, "Lori: Local Low-Rank Response Imputation for Automatic Configuration of Contextized Artificial Intelligence," IEEE Transactions on Industrial Informatics, Major Revision.

[4] Bohdal, O., Yang, Y., and Hospedales, T., 2020, "Flexible Dataset Distillation: Learn Labels Instead of Images," arXiv preprint arXiv:2006.08572.

[5] Shojaee, P., Zeng, Y., Wahed, M., Seth, A., Jin, R., and Lourentzou, I., 2022, "Task-Driven Privacy-Preserving Data-Sharing Framework for the Industrial Internet," Proc. of the 2022 IEEE International Conference on Big Data (Big Data), IEEE, pp. 1505–1514.

[6] Khattar, D., Goud, J. S., Gupta, M., and Varma, V., 2019, "MVAE: Multimodal Variational Autoencoder for Fake News Detection," Proc. of The World Wide Web Conference, 2915–2921.

[7] Kang, S., Deng, X., and Jin, R., 2021, "A Cost-Efficient Data-Driven Approach to Design Space Exploration for Personalized Geometric Design in Additive Manufacturing," Journal of Computing and Information Science in Engineering, 21(6), 061008.

[8] Wang, C., Yang, Y., Gao, C., Peng, Y., Zhang, H., and Lyu, M. R., 2022, "No More Fine-Tuning? An Experimental Evaluation of Prompt Tuning in Code Intelligence," Proc. of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 382–394.

[9] Bi, W., Du, L., Fu, Q., Wang, Y., Han, S., and Zhang, D., 2024, "Make Heterophilic Graphs Better Fit GNN: A Graph Rewiring Approach," IEEE Transactions on Knowledge and Data Engineering.

[10] Chamikara, M. A. P., Jang, S. I., Oppermann, I., Liu, D., Musotto, R., Ruj, S., Pal, A., Mohammady, M., Camtepe, S., Young, S., Dorrian, C., and David, N., 2023, "OptimShare: A Unified Framework for Privacy Preserving Data Sharing – Towards the Practical Utility of Data with Privacy," arXiv preprint arXiv:2306.03379. Available: https://arxiv.org/abs/2306.03379.