High-Quality Dataset-Sharing and Trade Based on A Performance-Oriented Directed Graph Neural Network

Journal:	*IEEE Transactions on Automation Science and Engineering*
Manuscript ID	T-ASE-2024-3211
Manuscript Type:	Regular Paper (S1)
Date Submitted by the Author:	10-Oct-2024
Complete List of Authors:	Zeng, Yingyan; University of Cincinnati, Zhou, Xiaona; University of Illinois Urbana-Champaign, Department of Computer Science Chilukuri, Premith; Virginia Tech, Department of Computer Science Lourentzou, Ismini; University of Illinois Urbana-Champaign, School of Information Sciences Jin, Ran (2024); Virginia Polytechnic Institute and State University, Industrial and Systems Engineering
Key Words:	Dataset-Sharing, Data Trading, Dataset Valuation, Manufacturing Industrial Internet, Representation Learning

High-Quality Dataset-Sharing and Trade Based on A Performance-Oriented Directed Graph Neural Network

Yingyan Zeng, Xiaona Zhou, Premith Kumar Chilukuri, Ismini Lourentzou, and Ran Jin,

***Abstract*—The advancement of Artificial Intelligence (AI) models heavily relies on large high-quality datasets. However, in advanced manufacturing, collecting such data is time-consuming and labor-intensive for a single enterprise. Hence, it is important to establish a context-aware and privacy-preserving data sharing system to share small-but-high-quality datasets between trusted stakeholders. Existing data sharing approaches have explored privacy-preserving data distillation methods and focused on valuating individual samples tied to a specific AI model, limiting their flexibility across data modalities, AI tasks, and dataset ownership. In this work, we propose a performance-oriented representation learning (PORL) framework in a Directed Graph Neural Network (DiGNN). PORL distills raw datasets into privacy-preserving proxy datasets for sharing and learns compact meta data representations for each stakeholder locally. The meta data will then be used in DiGNN to forecast the AI model performance and guide the sharing via graph-level supervised learning. The effectiveness of the PORL-DiGNN is validated by two case studies: data sharing in the semiconducting manufacturing network between similar processes to create similar quality defect models; and data sharing in the design and manufacturing network of Microbial Fuel Cell anodes between upstream (design) and downstream (Additive Manufacturing) stages to create distinct but related AI models.**

***Note to Practitioners*— To accelerate AI adoption in advanced manufacturing, there is an urgent need for data sharing among manufacturing participants to efficiently prepare high-quality datasets and improve AI model performance. This work proposes a dataset-sharing framework that lays the foundation for future data exchange and trade. Current approaches lack the flexibility to support sharing across diverse context and may expose the value of the data prematurely. To address these challenges, we introduce a performance-oriented representation learning framework that generates data for sharing and valuation to secure the value and preserve the private information, and then utilizes graph-based supervised learning to guide the sharing decisions for data receivers. The framework's effectiveness and generalizability are demonstrated through two real-world manufacturing dataset-sharing case studies. Industrial participants can use this framework to rank datasets from others based on their predicted utility for specific downstream AI tasks.**

***Index Terms*—Dataset-Sharing, Data Trading, Dataset Valuation, Manufacturing Industrial Internet, Representation Learning**

This work was partially supported by the National Science Foundation under grants CMMI-2208864 and CMMI-2331985.

Yingyan Zeng is with the Department of Mechanical and Materials Engineering, University of Cincinnati, Cincinnati, OH 45242, USA (email:zengyy@ucmail.uc.edu); Xiaona Zhou is with the Department of Computer Science, and Ismini Lourentzou is with the School of Information Sciences, both in the University of Illinois Urbana-Champaign, Champaign, IL 61820, USA (email:xiaonaz2@illinois.edu; lourent2@illinois.edu); Premith Kumar Chilukuri is with the Department of Computer Science, and Ran Jin is with the Grado Department of Industrial and Systems Engineering, both in Virginia Tech, Blacksburg, VA24060, USA (email:cpremithkumar@vt.edu; jran5@vt.edu).

I. Introduction

ADVANCES in machine learning have improved data-driven decision-making in advanced manufacturing using Artificial Intelligence (AI) methods. For example, convolutional neural networks are widely used for visual inspection [1] and design optimization [2]. Despite the substantial research efforts in new AI models, their effectiveness has been heavily dependent on the availability of large, informative, and high-quality datasets [3]. A small dataset can limit the capacity of AI models to learn complex patterns, resulting in overfitting and poor accuracy [4]. Even with a large dataset, the class imbalance, discrepancy between training and testing data, and model dynamics may cause a lack of representativeness of the training data, which leads to the low prediction performance of AI models [5], [6]. In summary, without high-quality datasets, the AI models yield unreliable and inferior performance, which causes untrustworthy AI-guided decision-making and low adoption rates of AI systems in the industry.

Approaches have been investigated to improve the training datasets for AI models to prevent unreliable and inferior performance. A Manufacturing Industrial Internet (MII) connects manufacturing processes and systems via a sensor and actuator network [7], which enables the collection of high-speed and large-volume sensing data as training datasets for AI models. However, for a single manufacturing enterprise, it still takes a long time and intensive effort to collect sufficient data to train advanced AI model. Data sampling, synthesis, and generation methods have been investigated to reduce the time and cost of collecting high-quality training data for AI models at one stakeholder recently [8]–[10]. It significantly reduces the costs of data annotation and preparation, yet it does not utilize potential datasets from other stakeholders.

In this work, we aim to investigating a data sharing and trading system to accelerate the data preparation process for AI systems. This is not only to scale high-quality data preparation from a single data source to multiple data sources, but also to incentivize the substantial investments in MII infrastructure and justify the value and the Return on Investment of generation, collection, and storage of datasets. It takes a solid step towards putting datasets as digital assets in the balance sheet of enterprises, which stimulates the formation of a global/regional data exchange market. Such a trading system is envisioned to

be formed by a firm foundation, including a quantitative metric for datasets generated from MII to compare "data quality", informative to differentiate the contribution of a dataset for the target AI performance; a sharing mechanism that can share "high-quality" datasets across different stakeholders to accelerate AI model development; and a distillation process before sharing to protect the privacy information based on the data stakeholders' own protocols, standards, and methodologies. To unlock and utilize the value of data for AI systems, the data trading system should consider the unique characteristics of data as a commodity [11], including the data processing method to preserve privacy, the freshness of the dataset for one task [12], and its contribution to one task to enable context-aware, fine-grained, and highly automatic data trading. In this paper, we investigate a paradigm of representation learning and dataset-sharing without pricing and trading, establishing it as the foundation of the broader concept of data trade.

Such a paradigm is expected to allow stakeholders to make informed economic decisions in a privacy-preserving manner, which includes protecting confidential and proprietary information contained by the raw datasets, securing details about the target AI task, and controlling access to data prior to making any data sharing decisions. To secure the data privacy of stakeholders in critical applications (*i.e.*, healthcare or competing manufacturers [13], [14]), existing research efforts have been made in different aspects, including differential privacy, federated learning, and data distillation. Differential privacy focuses on the privacy of individual data points rather than the entire dataset, which is difficult to be scalable [15]. Federated learning struggles to accommodate heterogeneous model structures, which limits its practical application [16]. Data distillation generates privacy-preserving proxy datasets for sharing [17]. However, it requires access to either the target AI model or the generated proxy datasets before sharing. This hampers their practical applications since valuable information has already been exposed without authorization. Just as a buyer estimates the worth of a commodity with a sample or demo rather than the entire commodity, it is critical to develop a dataset-sharing paradigm that evaluates a dataset's contribution to specific tasks without direct access to the full datasets.

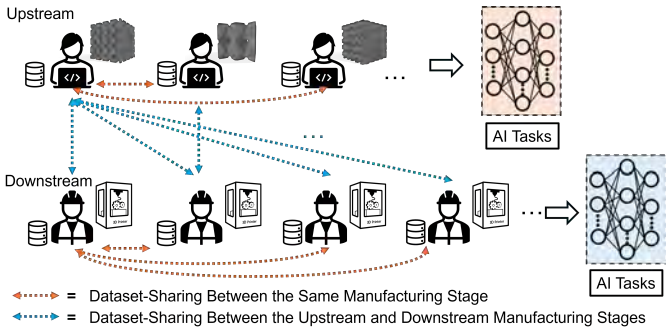

Fig. 1: Data sharing between manufacturing stages

Beyond privacy concerns, another challenge of data sharing and trade is how to share heterogeneous datasets from various stakeholders to enable on-demand sharing for context-aware AI enhancement. As illustrated in Fig. 1, within the same manufacturing system, data sharing between the same types of stakeholders (orange arrows) aims to improve the performance of similar AI tasks. For example, silicon ingots are manufactured by different manufacturers in similar furnaces. The different sensor network configurations, however, result in heterogeneous data formats based on different *in-situ* process data collected from furnaces. As a more generic scenario, data sharing between upstream and downstream stages (blue arrows) involves distinct AI tasks for data owners and data receivers. This is an example of the design and manufacturing of Microbial Fuel Cell (MFC) anodes in Additive Manufacturing (AM) [18]. The discrepancies between design feasibility and manufacturability may lead to suboptimal manufacturing process settings, resulting in reduced manufacturability of the purchased designs. When designers share additional information with manufacturers, such as design features and rules, manufacturers can better model the manufacturability with shared information as extra predictors. This helps them to optimize manufacturing processes and enhances the collaborative partnership between designers and manufacturers. However, heterogeneous data formats and distinct AI tasks make it difficult for data receivers to understand the value of other stakeholders' data and hinder their effective use in the data receivers' target AI tasks. On the other hand, while predicting design feasibility for the designers and predicting product manufacturability for the manufacturers employing different AI models, they are inherently connected. This motivates us to identify the similarity and interrelationship between the AI tasks, through which the stakeholders can be informed to achieve dataset sharing adaptive to the context (*i.e.*, AI models, data formats, and inherent dataset connections).

In summary, the research objective of this paper is to create a context-aware, effective, and privacy-preserving dataset-sharing framework that allows stakeholders make performance-oriented data sharing decisions to improve the target AI task(s). Here, the context-aware property refers to the ability to tailor dataset-sharing for different contexts, while performance-oriented property refers to recommending sharing decisions that significantly enhance the performance of the target AI task. We propose a performance-oriented representation learning (PORL) framework within a Directed Graph Neural Network (DiGNN). To preserve the private information of datasets prior to the sharing decision, the proposed PORL distills the dataset to generate the proxy dataset for sharing, securing the raw dataset. Simultaneously, it learns the meta data as highly compact and informative representation for each stakeholder. The meta data will be used to estimate the performance of AI tasks attained by dataset-sharing in the proposed DiGNN, thus securing the value of the proxy dataset. Specifically, each feature or modality in the dataset is encoded separately in the distillation process to create an independent latent space in consideration of the heterogeneity across datasets, in which the missing features or modalities are imputed. This aligns the latent space for proxy dataset generation while minimizing the assumption on the inter-feature or inter-modality dependency. Then, PORL employs the self-attention mechanism to capture the importance and dependency between latent features in the proxy

dataset coupled with the local AI task, where the weight matrix in the attention heads is treated as the representative meta data for each stakeholder. Thus, it characterizes each dataset with the stakeholder's local AI task. In brief, the PORL framework addresses the challenges for the dataset-sharing in rich context, establishing the foundation for context-aware sharing. Subsequently, we formulate the estimation of the performance for a target AI task with one dataset-sharing decision as a directed graph-level supervised learning problem. By treating each stakeholder (*i.e.*, dataset-AI task pair) as one node in a directed graph, we can conveniently learn the interaction between the stakeholders and distinguish the impact of dataset-sharing in different combinations compared to traditional supervised learning methods. Thus, the proposed framework provides context-aware sharing suggestions. In addition to effective sharing recommendations, DiGNN has good interpretability by the learned sharing pattern for each stakeholder. We further propose a positive ranking loss to enhance the positive information transfer with high-quality datasets between the stakeholders so that the performance of the target AI task can be improved after sharing.

The remainder of this work is organized as follows: Section II summarizes the related work. Section III introduces the proposed PORL-DiGNN framework. Section IV and Section V validate PORL-DiGNN via two real case studies. We conclude this work in Section VI with discussion of future directions.

II. Related Work

A. Data Valuation

Data valuation methods aim to quantify the usefulness of data sources and assign value to individual data points, guiding data selection strategies. In the literature, three primary approaches for data valuation are proposed: leave-one-out (LOO), which assess each sample based on the performance difference caused by its removal during the training of downstream AI models [19]; Shapley-value-based methods, which quantify marginal performance improvements through game theory [20]–[22]; and reinforcement learning-based methods, which learn to rank the importance of samples in conjunction with training the target AI model [23], [24]. However, these approaches typically focus on valuating data points closely tied to training one target AI model, making them inefficient and inflexible for multiple, different AI modeling needs from multiple stakeholders and heterogeneous datasets. To generalize the valuation, a model-agnostic framework named LAVA was proposed using class-wise Wasserstein distance to value datasets effectively [3]. However, LAVA does not consider contextualized task information and sharing direction between two parties. Additionally, all these methods do not address privacy concerns as they necessitate access to the entire dataset and the target AI task for valuation.

B. Data Sharing

Data sharing has gathered significant research attention in the big data era. Existing approaches employ blockchain technology to resolve the efficiency and privacy issues in data sharing [25]–[27]. For instance, edge computing was integrated with blockchain [28] to enable medical data sharing and improve data processing efficiency and security. However, these sharing systems cannot provide an efficient and accurate estimation of the performance for the target AI task prior to the sharing decision, thus being unable to support performance-oriented dataset-sharing.

Other techniques are also used in the literature. For instance, differential privacy was utilized via local perturbation of data at the workers' end before sharing [29]. A federated learning framework was proposed to develop the health prognostic models from heterogenous edge data by matching the feature similarity, but it is unable to address heterogenous model structure [16]. As a previous work, privacy-preserving data distillation was proposed to generate synthetic data representations and an attention mechanism was used to dynamically select data tailored for specific downstream machine learning tasks [17]. In this framework, data points from the datasets were accessed sequentially during the sharing process, which cannot be simply adapted to the sharing of entire datasets. Moreover, the value of the shared content cannot be protected before making the online sharing decision.

III. Methodology

A. Assumptions and Problem Definition

To create the dataset-sharing framework, we make the following assumptions: (1) In the data trading system, the stakeholders may never share the raw datasets to protect the confidential information. Instead, proxy datasets can be generated from the raw datasets for sharing after the sharing decision is made to secure their value, while the meta data can be shared freely to inform the data sharing decision. Each data owner has one dataset to share. This framework can be scaled to accommodate multiple datasets by treating each dataset as a node within the proposed DiGNN. Each data receiver can request multiple dataset from different stakeholders. (2) Each stakeholder is associated with a target AI task. AI tasks should share similar context (i.e., AI models, data formats, and connections to the datasets). (3) The dimension of the proxy dataset is predefined before sharing. (4) The performance of dataset-sharing decisions within a data trading system can be learned from historical data sharing decisions and the target AI model's performance using supervised learning.

Consider that we have a set N of stakeholders in the data trading system, where stakeholder $\mathcal{S}_i$ produces the local dataset $\mathcal{D}_i = \{\boldsymbol{X}_i, \boldsymbol{Y}_i\}, i \in N$. A target AI task is defined with the testing set $\mathcal{D}_i^T = \{\boldsymbol{X}_i^T, \boldsymbol{Y}_i^T\}, i \in N$, together with metrics to evaluate the performance (*e.g.*, F1 score for binary classification models). Among the stakeholders, let $\mathcal{S}_k$ denote the data receiver, who will request datasets from the other stakeholders. The local AI task of the data receiver $\mathcal{S}_k$ becomes the target AI task in the dataset-sharing. To enable more efficient dataset sharing, we allow the data receiver to simultaneously request datasets from multiple data owners to improve the performance of the target AI task. To consider the heterogeneity across different datasets, we denote the available feature or modality set of local dataset $\mathcal{D}_i$ as $\mathcal{F}_i$ with cardinal $|\mathcal{F}_i|, i \in N$, where each element corresponds to the index

of an available feature or modality in $\mathcal{D}_i$. Define the complete feature or modality set $\mathcal{F}$ as $\{1, 2, \cdots, F\}$, which is the union of the feature sets $\mathcal{F}_i$ across all stakeholders within this trading system. Let $\boldsymbol{x}_i^j, j \in \mathcal{F}_i$, $i \in N$ be the data vector for the feature indexed by j in the local dataset $\mathcal{D}_i$, and the sample size of $\mathcal{D}_i$ as n_i. We have $\boldsymbol{x}_i^j \in \mathbb{R}^{n_i \times d_j}$, where d_j represents the dimension of feature or modality j. Correspondingly, we have the input data $\boldsymbol{X}_i \in \mathbb{R}^{n_i \times \sum_{j \in \mathcal{F}_i} d_j}$, with the response $\boldsymbol{Y}_i \in \mathbb{R}^{n_i}, i \in N$ for each stakeholder.

B. Overview of the Proposed Methodology

As shown in Fig. 2, we propose a PORL-DiGNN framework consisting of two key components: performance-oriented representation learning (PORL) of proxy dataset and meta data, and a Directed Graph Neural Network (DiGNN) for learning the anticipated performance of dataset-sharing. Our key idea is to preserve the privacy and secure the value of data through a two-level representation learning approach. Before one stakeholder $\mathcal{S}_k$ makes the dataset sharing decision, all the stakeholders in the trading system will distill the raw dataset to generate proxy dataset $\boldsymbol{Z}_i, i \in N$ in the first level so that it will be used in sharing instead of the raw dataset. The heterogeneity in features (or modalities) is standardized with separate encoding and imputation in the latent space during the distillation process. Simultaneously, the meta data $\boldsymbol{E}_i, i \in N$ associated with the stakeholder's local task is generated via PORL in the second level to characterize each stakeholder. Then, the meta data is used to model the interrelationships between stakeholders in a graph neural network to estimate the performance of a sharing decision. With historical sharing information available, each sharing decision is modeled as a graph, where the stakeholders' meta data serve as node features, and the graph-level response is defined as the performance of the target task attained after executing the sharing decision. Therefore, the evaluation of a new sharing decision can be efficiently achieved through a single inference using the trained DiGNN. After evaluating all the potential sharing decisions of stakeholder $\mathcal{S}_k$, the optimal decision can be made by selecting the one that yields the highest performance.

The advantage of the proposed framework lies in three aspects: (1) The framework allows limited information sharing before making the sharing decision, thus protecting the data privacy and the target task information of both the data owners and receivers. (2) The separate encoding of features and modality "standardize" the heterogeneous datasets in the data distillation and imputation. Thus, it generalizes the framework to datasets with different features or modalities. (3) By modeling the sharing decision in the trading system as a graph, the interaction between stakeholders can be effectively learned via historical sharing decisions and outcomes, which can be applied to new stakeholders through the learned node embedding function in the "cold start" scenarios.

C. Performance-Oriented Representation Learning

To secure the privacy of the raw datasets, the proxy dataset are distilled from the raw datasets as privacy-preseving data representations. Then, to further secure the value of proxy

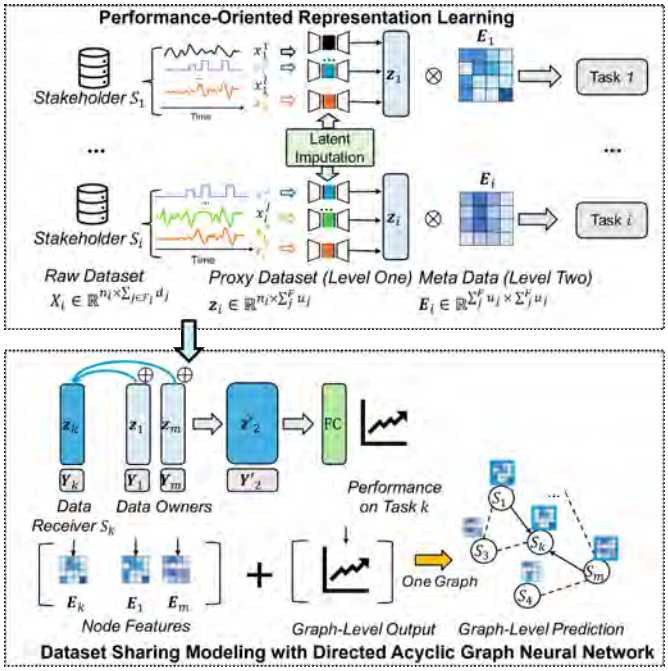

Fig. 2: Overview of the proposed PORL-DiGNN framework

datasets, we propose to generate meta data to characterize each stakeholder. We follow the idea in [17] to utilize Variational Autoencoder (VAE) in the proxy dataset distillation. Specifically, in the local training of stakeholder $\mathcal{S}_i$, a feature or modality-specific encoder $f_{\text{enc}}^j(\cdot), j \in \mathcal{F}_i$ projects the feature or modality vetcor $\boldsymbol{x}_i^j$ into an independent latent space as the latent variable vector $\boldsymbol{z}_i^j$ with dimension u_j, then a corresponding decoder $f_{\text{dec}}^j(\cdot)$ reconstructs the data on the original feature space. The latent variables are concatenated to form the joint latent vector, *i.e.*, the proxy dataset $\boldsymbol{Z}_i$ of dataset $\mathcal{D}_i$. To standardize the proxy dataset across stakeholders for sharing and then employing them in the target AI task modeling, we propose to impute the missing features as standard Gaussian noise in the latent space, *i.e.*, $\boldsymbol{z}_i^j \sim \mathcal{N}(0, 1), j \in \mathcal{F} \backslash \mathcal{F}_i, i \in N$. The Gaussian noise is commonly used as the prior distribution $p_\theta(\cdot)$ for the available latent variables $\boldsymbol{z}_i^j, j \in \mathcal{F}_i, i \in N$ [30]. To this end, the latent space can be aligned across datasets, where each latent dimension corresponds to one feature. The dimension of the proxy dataset is unified as $\sum_{j=1}^{F} u_j$ for all stakeholders. In the encoding and decoding, the VAE loss is applied on each available feature, which minimizes the reconstruction error and the Kullback-Leibler (KL) divergence:

$$\mathcal{L}_i^{\text{VAE}} = \sum_{j \in \mathcal{F}_i} \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})} \left[\log p_\theta(\boldsymbol{x}_i^j \mid \boldsymbol{z}_i^j) \right] - \lambda_1 D_{KL} \left(q_\phi(\boldsymbol{z}_i^j \mid \boldsymbol{x}_i^j) \| p_\theta(\boldsymbol{z}) \right), \tag{1}$$

where $q_\phi(\boldsymbol{z}_i^j | \boldsymbol{x}_i^j)$ is the approximate posterior for the prior distribution of latent variables $p_\theta(\boldsymbol{z})$, and λ_1 is the weight of the KL loss term. By transforming the raw data into a lower-dimensional representation in the latent space, the process obscures direct relationships and identifiable features in the raw data. The network architecture for the encoder can be selected

based on the specific feature. For example, a Long Short-Term Memory (LSTM) network [31] is used for time series data, a Convolutional Neural Network (CNN) is employed for image or tensor data, and a Multilayer Perceptron (MLP) is used for other tabular features in this work.

As shown in the upper panel of Fig. 2, the distilled proxy dataset is further used as the extracted feature for the local AI task. To characterize the proxy dataset associated with the local AI task in an informative, and compact manner, we propose to learn the task-specific feature importance and the inter-latent variable dependency as the meta data via the self-attention mechanism [32] during the distillation process. The meta data are generated in the local AI task via the following layer:

$$f_{\text{att}}(\boldsymbol{Z}_i) = \frac{1}{t} \oplus_t \left[\boldsymbol{Z}_i \odot \text{softmax}\left(\boldsymbol{W}^t_{f_{\text{att}}} \boldsymbol{Z}_i + \boldsymbol{b}^t_{f_{\text{att}}} \right) \right], \quad (2)$$

where t is the number of attention heads; the attention head weight matrix $\boldsymbol{W}^m_{f_{\text{att}}} \in \mathbb{R}^{\sum_{j=1}^{F} u_j \times \sum_{j=1}^{F} u_j}; m \in [t]$ represents the relations between the input latent variables in the proxy dataset learned in attention head m; $\boldsymbol{b}^m_{f_{\text{att}}}$ is the corresponding bias; $\odot$ refers to the Hadamard product; and $\oplus$ refers to the Hadamard summation across attention heads. Hence, the attention head weight captures how the latent variables are correlated in the learning process of the local AI task, encompassing information from both the data itself and the task-specific dependency. Afterwards, the output of this attention layer $f_{\text{att}}(\boldsymbol{Z}_i)$ is used as the input for the following two fully connected layers in the local AI task as $f_{\text{clf}}(\boldsymbol{Z}_i) = a_{l_2}\left(\boldsymbol{W}_2 \cdot (a_{l_1}\left(\boldsymbol{W}_1 \cdot f_{\text{att}}(\boldsymbol{Z}_i) + \boldsymbol{b}_{l_1}\right)) + \boldsymbol{b}_{l_2}\right)$, where $a_{l_1}(\cdot)$ and $a_{l_2}(\cdot)$ are the activation functions for layer l_1 and l_2, respectively. Based on the self-attention, the meta data for proxy dataset $\boldsymbol{Z}_i$ is defined as:

$$\boldsymbol{E}_i = \frac{1}{t} \oplus_t \boldsymbol{W}^t_{f_{\text{att}}}. \quad (3)$$

$\boldsymbol{E}_i$ represents how each latent variable is related to the response. Therefore, by identifying the patterns in the inter-latent variable dependency in the DiGNN framework, the difference in how the input data correlates with the response can be identified in its local AI task, indicating whether one dataset will be beneficial to an AI task of another stakeholder. The ability to capture task-specific dependencies is crucial for dataset sharing among stakeholders operating in heterogeneous streams with distinct tasks. Without using explicit descriptive information on the target task, the proposed approach also protects the privacy of target tasks in dataset-sharing.

To perform the representation learning in an end-to-end manner, suppose we have a classification problem as the local AI task for each stakeholder, then the training cross-entropy loss becomes $\mathcal{L}^{\text{Clf}} = -\sum_i y_i \log\left(\text{softmax}(f_{\text{clf}}(\boldsymbol{Z}_i)\right)$. The total loss of the proxy dataset distillation and representation learning process is the summation of the VAE loss and the classification loss: $\mathcal{L}^{\text{PORL}} = \mathcal{L}^{\text{VAE}} + \lambda_2 \mathcal{L}^{\text{Clf}}$, where λ_2 is the weight of the cross-entropy loss term. λ_1 and λ_2 are the tuning parameters to balance the impact of different loss terms during the training process. This framework can also be directly applied to other supervised learning tasks by replacing the cross-entropy loss with an appropriate loss function tailored to the specific local AI task.

D. Directed Graph Neural Network for Dataset-Sharing

Assume that all stakeholders in the trading system have generated proxy data sets and meta-data. We formulate a graph-level supervised learning problem to guide the data receiver $\mathcal{S}_k$ to make effective dataset-sharing decisions to improve the target AI model performance. Specifically, each data-sharing decision between two stakeholders is represented by a directed edge (i, j) between two nodes i and j in one directed graph $\mathcal{G}$, where the performance gain on the target testing set $\mathcal{D}^T_i, i \in N$ attained by the sharing is the graph-level response. The lower panel in Fig. 2 demonstrates one sharing decision, where stakeholder $\mathcal{S}_k$ requests the dataset from stakeholder $\mathcal{S}_1$ and $\mathcal{S}_m$. The direction of the edge indicates the direction of dataset transfer. This is critical as sharing the same datasets in different directions can lead to significantly different outcomes. By modeling the sharing decision as a directed graph, these differences are naturally encoded.

To address the graph-structured data, graph neural networks encode the node features as node representations and aggregate them based on the graph structure to produce a graph representation. In the widely used message-passing neural network (MPNN) [33], node representation is updated via iterative message passing between neighboring nodes. Instead of aggregating with neighbors, we propose a structure based on directed acyclic graph neural networks [34], where the aggregation and updating of node representation strictly follow the partial order defined by the directed edges.

Denote the number of layers in the DiGNN as L and the predecessor set of node i as $\mathcal{P}(i)$. For example, if node j shares the corresponding stakeholder's dataset with node i, generating edge (j, i), then j is the predecessor of node i in this graph. In the ℓ-th layer, the aggregated message m^ℓ_i of node i is obtained as a weighted combination of the representation of all its predecessors $\mathcal{P}(i)$ defined by the directed edges in the same layer:

$$\begin{aligned} m^\ell_i &= A^\ell\left(\left\{h^\ell_j \mid j \in \mathcal{P}(i)\right\}, h^{\ell-1}_i\right) \quad (4) \\ &= \frac{1}{|\mathcal{P}(i)|} \oplus_{j \in \mathcal{P}(i)} \underset{j \in \mathcal{P}(i)}{\text{softmax}}\left(h^{\ell-1}_i \cdot W^\ell_{ij} \cdot {h^\ell_{ju}}^T \right. \\ &\quad \left. + h^{\ell-1}_i \cdot W^\ell_{ij} \cdot \gamma(j,i)\ \right) h^\ell_j, \end{aligned}$$

where A^ℓ is the aggregate operator for the ℓ-th layer; $h^{\ell-1}_i$ is the aggregated representation of node i in the previous (*i.e.*, $(\ell-1)$-th) layer; and $\gamma(j, i)$ is the attribute of edge (j, i). In this way, the edge direction is precisely encoded in the aggregated messgae. Here, the attention mechanism is employed in A^ℓ. We treat W^ℓ_{ij} as the attention weight between node i and j in the ℓ-th layer, $h^{\ell-1}_i$ as the query, h^ℓ_j and $\gamma(j, i)$ as the key, and h^ℓ_j itself as the value. The representation of node i at the first layer is the meta data of the corresponding stakeholder: $h^1_i = \boldsymbol{E}_i, \forall i \in N$. Additionally, we employ trainable embeddings as edge attributes $\gamma(j, i)$ and set it as the same dimension as h^ℓ_i. We enforce that the two key contents h^ℓ_j and $\gamma(j, i)$ share the same attention weight W^ℓ_{ij} to reduce the parameters in the GNN. This approach allows the preference and relative importance of each stakeholder to be captured through the embeddings and the attention weights learned from historical sharing decisions. The stakeholder who owns the dataset that

enhances the performance of the target AI task will be given a higher preference by the data-receiving stakeholder.

To update the representation h_v^ℓ via the message m_i^ℓ, a recurrent architecture is employed: $h_i^\ell = f_{\text{GRU}}^\ell(h_i^{\ell-1}, m_i^\ell)$, where $h_i^{\ell-1}$ is the input; m_i^ℓ is the previous state; and h_i^ℓ represents the current state of a Gated Recurrent Unit (GRU) [35]. The GRU mechanism allows better learning of dependencies indicated by the order in the directed graph $h_{\mathcal{G}}$.

As the final step, the updated node representations in each layer are concatenated, followed by a max-pooling operation and a fully-connected (FC) layer, to generate the graph-level output: $h^{\mathcal{G}} = f_{\text{FC}}\left(\underset{i\in N}{\text{Max-Pool}}\left([h_i^1; ...; h_i^\ell; ...; h_i^L]\right)\right)$, where $[\cdot;\cdot]$ is the concatenation operator between vectors. With a binary classification problem as the target AI task, the F1 score evaluated on the target testing set can be adopted as the performance gain by the dataset sharing. Denote the ground-truth F1 score for one dataset-sharing decision as $y_k^{\mathcal{G}}$, where k is the index of one data-sharing decision among n sharing decision in the trading system. The loss for the graph-level regression problem is calculated as: $\mathcal{L}^{\mathcal{G}} = \frac{1}{n}\sum_{k=1}^{n}(y_k^{\mathcal{G}} - h_k^{\mathcal{G}})^2$.

During the DiGNN training, the mean squared error (MSE) loss ensures accurate F1 score predictions to recommend the "best" sharing decision. However, it is also critical to avoid negative impact of additional shared datasets to the data receiver, which can be achieved by accurately identifying the sharing decisions that would result in a decrease of the F1 score. Therefore, we propose the following ranking loss:

$$\mathcal{L}^R = \frac{1}{n}\sum_{k=1}^{n}\max\left(0, -y_{k\,\text{ind}}^{\mathcal{G}*}\cdot(h_k^{\mathcal{G}} - y_k^{\mathcal{G}*}) + \delta\right), \quad (5)$$

where $y_k^{\mathcal{G}*}$ is the original F1 score of the data receiver before this sharing decision; $y_{k\,\text{ind}}^{\mathcal{G}*} = 1$ if the predicted F1 score of the sharing decision should be higher than the original F1 score, and vice-versa for $y_{k\,\text{ind}}^{\mathcal{G}*} = -1$; δ is the margin that enforces a buffer zone between the predicted and the original F1 score, and the prediction falling out of the zone will be penalized. The ranking loss $\mathcal{L}^R$ encourages the DiGNN to correctly classify whether one sharing decision will cause the F1 score gain or loss. The training loss function of the DiGNN is proposed as: $\mathcal{L}^{\text{DiGNN}} = \mathcal{L}^{\mathcal{G}} + \lambda_3\mathcal{L}^R$, where λ_3 is the hyperparameter that tunes the balance between the MSE loss and the ranking loss. A low weight λ_3 is suitable for a trading system where the majority of stakeholders' datasets can improve each other's target task performance. In contrast, when there are substantial dataset sharing decisions that may cause performance decreases, λ_3 should be set to a higher value. The standard Adam optimizer is used for model estimation and hyperparameter fine tuning through cross-validation to achieve good model performance.

IV. Case Study: Dataset Sharing in the Same Manufacturing Stage

We apply the proposed PORL-DiGNN to the first motivation example, the dataset sharing between manufacturers in the semiconducting manufacturing network. As shown in Fig. 3, multiple ingots were manufactured in similar furnaces by different manufacturers in the CZ process. The process is sensitive to subtle changes in conditions. The failure results in the growth of defective polycrystalline ingots [36]. Prediction of the binary polycrystalline defect indicator is crucial in the crystal growth process, as it significantly impacts the initial product quality in semiconductor manufacturing. The setting variables and process variables collected from sensors during the manufacturing are used to train a supervised learning model to predict the polycrystalline defect [36]. Consider each manufacturer as one stakeholder who is associated with a binary classification problem for the defects as the target AI task. It is important to facilitate the sharing of datasets across manufacturers due to the extensive time required to collect adequate samples for training advanced AI models and the class imbalance caused by the limited number of defective samples compared to normal ones.

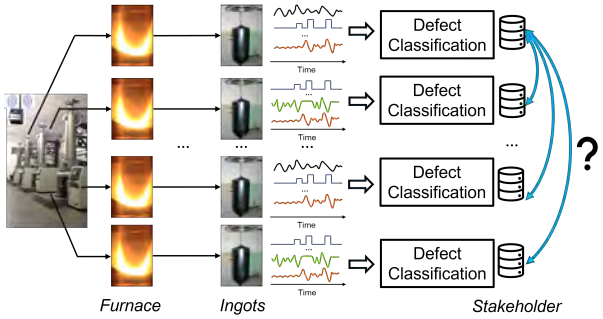

Fig. 3: The CZ process and collaborative dataset sharing (elements redraw from [17])

A. Datasets Information

This study involves three groups, each consisting of four manufacturers, indexed as $A1, A2, A3, A4, B1, \cdots$, *etc.* Originally, all manufacturers had identical sensing systems. We modified the raw data to simulate manufacturers with varied sensors, creating heterogeneous datasets for sharing. The complete set of *in-situ* variables includes six features: Gas Flow Rate (slpm), Main Chamber Pressure (Torr), Crystal Growth Rate (mm/hr), Main Heater Current (Amps), Liquid Surface Temperature (°C), and Crystal Pulling Speed (mm/hr). Besides, two setting variables are employed for the quality modeling: Crystal Diameter (mm), and Main Chamber Pressure (Torr) [36]. All variables are collected in time series format. Each ingot's measurements are segmented into 15-minute intervals, each constituting one sample, and binary quality labels are assigned by a domain expert. The dataset collected by each manufacturer is a tensor with dimension (n_i, l, m_i), where n_i is the sample size, m_i is the number of time series features which varies from 4 to 6, and l is the number of time stamps of each sample. 30% samples in each dataset are randomly selected as the testing set $\mathcal{D}_i^T, i \in \{1, \cdots, 12\}$. We assume that stakeholders are permitted to simultaneously purchase up to three datasets, yielding 231 potential sharing decisions per stakeholder. This total is derived by summing all the combinations of 1, 2, or 3 data owners with one data receiver (*i.e.*, $11 + 55 + 154 = 231$). Consequently,

within this data trading system, there are a total of 2,772 (*i.e.*, $231 \times 12 = 2772$) potential sharing decisions. The sharing of more datasets can be achieved by sequential sharing.

TABLE I: Dataset information, ground-truth sharing decisions, and outcomes of classification model for case study one

Manuf.	Local F1 Score	Num. of Positive Shar. Comb.	Sample Size	Ratio of Positive Class	Best F1 with Three Shar.	Best Comb. with Three Shar.
A1	0.800	131	257	0.109	1.000	[A2, A3, A4]
A2	0.727	77	854	0.069	0.848	[A1, B3, C4]
A3	0.753	6	1098	0.076	0.753	[A2, A4, B2]
A4	0.526	229	736	0.095	0.778	[B1, B2, C4]
B1	0.667	122	379	0.063	0.933	[A2, B2, C4]
B2	0.667	0	425	0.014	0.667	[A1, A2, A3]
B3	0.667	2	229	0.035	0.800	[B1, C2, C3]
B4	0.857	27	255	0.051	1.000	[A1, A2, C4]
C1	0.909	45	453	0.044	1.000	[A1, A3, C3]
C2	1.000	0	218	0.050	1.000	[A1, B4, C4]
C3	0.762	160	297	0.111	0.900	[A4, B1, C1]
C4	0.900	41	641	0.051	0.952	[A1, A2, C2]

Table I summarizes the local F1 score, the sample size, ratio of positive samples of each dataset and the ground-truth best sharing results for each manufacturer. The structure of the local AI model is defined by encoders and decoders (Sec. IV-B), followed by two fully-connected layers. The ground-truth is evaluated based on the sharing of proxy datasets across all 2772 combinations. The third column shows the number of sharing decisions that improve each stakeholder's local AI task performance. It is clear that most stakeholders, as data receivers, benefit from using datasets from other data owners. However, for $A3$, $B2$ and $C2$, the dataset-sharing does not yield performance improvements. For $B2$, the lack of improvement may be attributed to the severe class imbalance in the dataset. It can also be observed that a substantial number of sharing decisions could lead to performance degradation for stakeholders (*e.g.*,$B3$, $C1$, $C4$, *etc.*). This highlights the importance of accurately identifying positive and negative sharing decisions within this data trading system.

B. Hyperparameter Settings

In the representation learning process, a LSTM-based architecture with a two-layer encoder and a two-layer decoder is applied to each time series feature. Specifically, the latent dimension for each feature is $u_i = 8$, thus generating the proxy dataset with dimension $8 \times 8 = 64$, and the dimension of the meta data is $\boldsymbol{E}_i \in \mathbb{R}^{64\times 64}, i \in \{1, \cdots, 12\}$. The number of attention heads t is set to 2. In the proposed DiGNN, we use an encoder consisting of three convolutional layers to embed the node feature $\boldsymbol{E}_i$. The number of layers L is set to 2. For the whole framework, the weights in the loss are set as: $\lambda_1 = 1, \lambda_2 = 1, \lambda_3 = 1$ to balance the different loss terms. The margin δ for ranking loss is set to 0.1.

C. Benchmark Methods and Evaluation Metrics

We evaluate the proposed PORL-DiGNN framework in benchmark comparison. We adopt the following benchmark methods that partially address the challenges in dataset-sharing as state-of-the-art solutions: (1) **LAVA** [3]: Valuating the dataset based on the class-wise Wasserstein distance, where datasets closer to the data receiver's dataset contribute more to the target task. LAVA requires access to the proxy dataset. (2) **UtlPred** [37]: Estimating the performance on the target task by learning a parameterized function that takes the meta data as the input and the performance as the output. Here we generate the summary statistics of the proxy dataset as the meta data of a stakeholder (*i.e.*, sample size, number of features, mean of mean values, standard deviation of mean values, mean of kurtosis, mean of skewness, mean value of all responses, standard deviation of all responses, range of all responses, kurtosis of all response, and skewness of all responses), and then use it as the input for UtlPred. (3) **Graph Convolutional Neural Network (GCN)** [38]: GCN is employed to learn the performance of dataset-sharing between stakeholders. GCN considers the direction of edges in the adjacency matrix in the graph. The input for GCN is the summary statistic of the proxy dataset; (4) **PORL-UtlPred**: Using the meta data generated by PORL as the input for **UtlPred**; (5) **PORL-GCN**: Using the meta data generated by PORL as the node input for **GCN**. We include **PORL-DiGNN** with only the MSE loss and **PORL-DiGNN with Ranking Loss** as the proposed methods.

We create two scenarios for evaluation. In scenario one, 10%, 20%, and 30% of the historical dataset-sharing information of each stakeholder among all possible potential decisions is available as training data respectively, with the remainder as testing data. In scenario two (Leave-One-Out (LOO)), we simulate a real-world application where a completely new stakeholder requests datasets from the data trading system without historical sharing information.

Four performance metrics are employed in the evaluation: (1) **Normalized Discounted Cumulative Gain (NDCG)** of the top k^* datasets, which measures the ranking prediction accuracy for individual datasets, the higher the better. k^* is the number of combinations (*i.e.*, sharing decisions) that lead to a positive performance gain (*i.e.*, the third column in Table I); (2) the **Pearson Correlation** between the ground-truth and the predicted F1 score of the target AI model on the testing set, the higher the better; (3) the **Normalized-Root-Mean-Squared Error (NRMSE)** of the predicted F1 score to the ground-truth F1 score of the target AI model, the lower the better; (4) and the **F1 score** of the target AI model of each data receiver by executing the best dataset-sharing decision, which evaluates the effectiveness of the dataset-sharing framework. LAVA is an unsupervised approach that does not predict the F1 score and, thus, is not evaluated by the NRMSE.

D. Experimental Results for Scenario One

The results of scenario one are summarized in Table II. We use a small training ratio to validate the proposed method in a real-world application setting. First, the effectiveness of the proposed representation learning method is validated by comparing UtlPred with PORL-UtlPred, and GCN with PORL-GCN. Secondly, the proposed PORL-DiGNN with ranking loss outperforms all the benchmark methods, which indicates that the proposed method can achieve effective data-sharing with limited historical data. This superior performance is attributed to the explicit consideration of edge direction in node message aggregation in the DiGNN, which is vaguely expressed in the adjacency matric in GCN. UtlPred does not

TABLE II: The average values and standard errors of the performance in $10\%, 20\%, 30\%$ training ratio reported over 10 replications. Significant best results are highlighted in **bold** and the best benchmark performance is in blue. Blue arrows indicate relative improvement over the best benchmark method. All numbers in the table have been multiplied by 100.

Training Ratio	Method	Evaluation Metrics			
		NDCG (↑)	Pearson Correlation (↑)	NRMSE (↓)	F1 Score (↑)
10%	LAVA	87.8 ± 0.0	10.1 ± 0.0	-	77.3 ± 0.0
	UtlPred	88.1 ± 0.5	26.7 ± 0.0	18.6 ± 0.0	75.0 ± 1.9
	GCN	88.4 ± 0.8	64.6 ± 0.0	15.0 ± 0.0	75.3 ± 2.2
	PORL-UtlPred	88.3 ± 1.6	15.4 ± 9.3	17.6 ± 0.1	76.7 ± 3.6
	PORL-GCN	89.0 ± 0.5	70.4 ± 0.0	14.1 ± 0.0	76.5 ± 3.9
	PORL-DiGNN	89.5 ± 0.6	75.6 ± 0.0	14.0 ± 0.0	78.2 ± 2.9
	PORL-DiGNN with Ranking Loss	$\mathbf{90.2 \pm 0.7}$ ↑ **1.2**	$\mathbf{76.5 \pm 0.0}$ ↑ **6.1**	$\mathbf{13.1 \pm 0.0}$ ↓ **1.0**	$\mathbf{80.8 \pm 2.5}$ ↑ **4.1**
20%	LAVA	88.7 ± 0.0	8.1 ± 0.0	-	77.9 ± 0.0
	UtlPred	89.2 ± 0.4	26.6 ± 0.0	17.4 ± 0.0	77.6 ± 1.6
	GCN	89.2 ± 0.7	72.6 ± 0.0	13.1 ± 0.0	78.2 ± 2.3
	PORL-UtlPred	90.7 ± 0.8	15.7 ± 9.3	16.9 ± 0.1	80.8 ± 2.7
	PORL-GCN	90.2 ± 0.7	73.5 ± 0.0	12.8 ± 0.0	79.0 ± 2.8
	PORL-DiGNN	89.8 ± 1.9	76.6 ± 0.0	12.5 ± 0.0	81.2 ± 3.0
	PORL-DiGNN with Ranking Loss	$\mathbf{91.1 \pm 1.0}$ ↑ **0.4**	$\mathbf{76.8 \pm 0.0}$ ↑ **3.3**	$\mathbf{12.1 \pm 0.0}$ ↓ **0.7**	$\mathbf{83.4 \pm 2.9}$ ↑ **2.6**
30%	LAVA	89.7 ± 0.0	11.5 ± 0.0	-	78.1 ± 0.0
	UtlPred	90.8 ± 0.6	26.1 ± 0.2	15.3 ± 0.0	78.6 ± 2.3
	GCN	90.0 ± 1.9	71.1 ± 0.1	11.1 ± 0.0	78.2 ± 3.4
	PORL-UtlPred	91.9 ± 1.3	18.6 ± 8.8	14.5 ± 0.1	80.5 ± 2.6
	PORL-GCN	90.8 ± 1.7	72.9 ± 0.0	10.8 ± 0.0	79.2 ± 3.2
	PORL-DiGNN	92.0 ± 1.7	76.9 ± 0.0	10.5 ± 0.0	82.5 ± 3.0
	PORL-DiGNN with Ranking Loss	$\mathbf{92.5 \pm 1.2}$ ↑ **0.6**	$\mathbf{77.1 \pm 0.0}$ ↑ **4.2**	$\mathbf{10.0 \pm 0.0}$ ↓ **0.8**	$\mathbf{84.0 \pm 2.4}$ ↑ **3.5**

account for sharing direction, resulting in lower Pearson correlation and higher NRMSE. Despite using proxy dataset directly and violating the privacy-preserving requirements, LAVA does not outperform the proposed method. This shortfall is mainly due to its lack of task-specific contextualized information (*i.e.*, the dependency between input variables in the local AI task), which is effectively integrated from PORL into graph-level supervised learning. Additionally, because LAVA uses the Wasserstein distance as a proxy for estimating performance, the Pearson correlation is notably poor. The results also indicate that when a larger number of training samples is used (*i.e.*, historical dataset-sharing information), the performance of all the methods improves, where the improvement in NDCG, NRMSE, and F1 score is significant.

E. Experimental Results for Scenario Two

In this scenario, each of the twelve datasets is excluded in turn to serve as the testing set, while the sharing decisions from the remaining datasets are used as the training set. Table III listed the average F1 score after executing the estimated best sharing decision of each stakeholder over 10 replications. Coherently, most stakeholders can achieve greater improvement with the proposed dataset-sharing framework. In particular, by investigating the result of $B4$ and $C1$, the proposed method is the only one that enables positive information transfer, thereby validating the effectiveness of the ranking loss in scenarios where information about the new stakeholder is limited. It can also be found that the F1 score after dataset-sharing is further enhanced by the proposed ranking loss when we compare the two proposed methods.

V. Case Study: Dataset-Sharing from Upstream to Downstream Manufacturing Stage

The proposed PORL-DiGNN framework is also applied to the second motivation example, the dataset sharing from designers to manufacturers in design and manufacturing of MFC anodes. The design and manufacturing of MFC anodes involve intricate geometries and process parameters that significantly impact their performance and manufacturability [39]. Evaluating the manufacturability of a design before initiating production is crucial for additive manufacturing, which reduces production cost and control the quality defects. Additionally, assessing a design's manufacturability allows manufacturers to adjust the process settings proactively, thereby avoiding production failures. However, predicting manufacturability for a design is challenging for the manufacturers. The availability of data on finished manufacturing parts is limited compared to the large design space, resulting in poor manufacturability prediction performance. The competitive roles of other manufacturers who have similar machine configurations hinder the potential data sharing for performance improvement. Designers, with their extensive design data, are prime candidates of data source as they are collaborators with manufacturers.

Fig. 4 visualizes a MII consisting of six upstream designers and nine downstream manufacturers on the left panel and summarizes the modality and ownership of data during the design and manufacturing of MFCs on the right panel. In the design process, the designers define the design space with design variables (*i.e.*, cell type, cell count, volume fraction factor, layer thickness, X- and Y- axis rotation) and generate designs. The intermediate features (*i.e.*, minimum feature size, number of disjoint volume, thickness of cavity, and number of disjoint cavities) of each design are simulated by a design model [18] and evaluated by a set of design rules to determine the design's feasibility. Additionally, a 3D CAD model of each design is generated in STL format, providing a digital representation of the MFC anode geometry. During the manufacturing process, manufacturers recieve the STL file of a small subset of feasible designs from designers for anode production. Based on varying manufacturing process settings (*i.e.*, layer thickness, feed/flow ratio, and nozzle temperature), the chosen designs are printed and then subjected to a quality assessment process to evaluate their manufacturability. In summary, as shown in Fig. 4, the designers own the design data (*i.e.*, design variables, intermediate features, and design rules), the STL file, and the feasibility as the quality indicator,

TABLE III: The average F1 score gain in percentile in LOO reported over 10 replications. Significant best results are highlighted in **bold**. F1 score gain is colored with green while F1 score reduction is colored with red.

Furnace	A1	A2	A3	A4	B1	B2	B3	B4	C1	C2	C3	C4
LAVA	**14.22%**	5.49%	-9.16%	33.95%	0.52%	0.00%	-1.67%	-4.44%	-0.12%	-20.33%	12.16%	-4.18%
UtlPred	1.67%	-4.58%	-16.21%	29.18%	6.36%	0.00%	0.00%	-2.04%	-8.44%	-14.29%	2.74%	-0.47%
GCN	13.89%	6.50%	-17.47%	30.88%	0.00%	0.00%	0.00%	-1.39%	-7.12%	-21.32%	9.06%	-0.26%
PORL- UtlPred	8.33%	4.15%	-14.06%	32.93%	0.88%	0.00%	0.00%	-1.94%	-2.17%	-19.13%	7.11%	-1.65%
PORL- GCN	5.02%	7.22%	-8.07%	31.04%	2.68%	0.00%	0.00%	-4.37%	-2.58%	-11.54%	9.54%	-0.75%
PORL- DiGNN	6.27%	5.49%	-9.16%	33.95%	0.52%	0.00%	-1.67%	-4.44%	-0.12%	-10.33%	12.16%	-0.34%
PORL- DiGNN with Ranking Loss	7.52%	**8.24%**	**-7.17%**	**37.67%**	**8.69%**	0.00%	0.00%	**1.87%**	**1.92%**	**-7.70%**	**16.35%**	**3.84%**

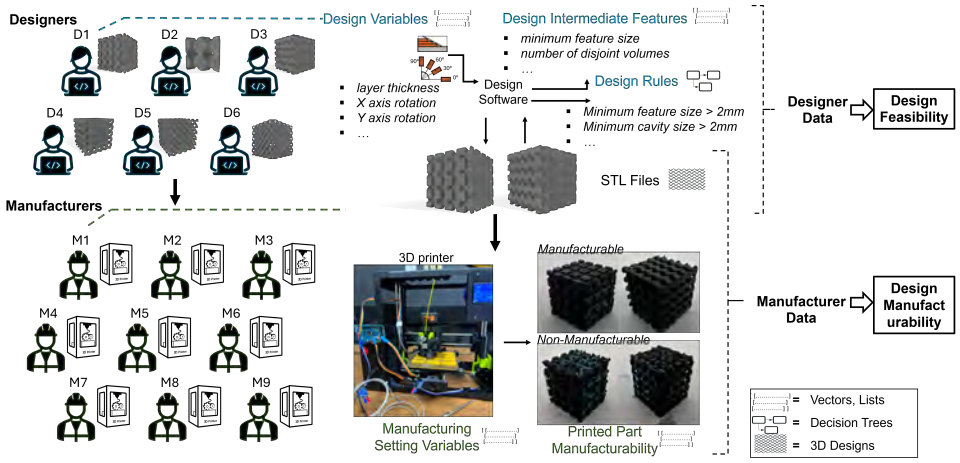

Fig. 4: The MII of designers and manufacturers (left); the data modality and ownership among upstream designers and downstream manufacturers in MFCs industry (right) (elements redraw from [18])

typically with a huge sample size. The manufacturers own the manufacturing data (*i.e.*, setting variables), the STL file, and the manufacturability as the quality indicator, typically with a limited sample size and highly imbalanced classes.

In the MII depicted in Fig. 4, designers and manufacturers are stakeholders in the data trading system, each possessing heterogeneous datasets and AI tasks. The target AI task for the designers is to classify the designs on feasibility and the target AI task for the manufacturers is to classify the designs on manufacturability. This leads to different data modalities with different-but-related target AI tasks, creating a challenging situation in data sharing. The AI tasks are related since the design feasibility are generated by design rules, which is an approximate and in many cases, conservative model for manufaturability [18]. Our proposed PORL-DiGNN framework addresses both challenges to enable effective dataset-sharing from the upstream designers to the downstream manufacturers. The intuition is to identify the designers whose variations in the design process demonstrate correlations similar to those observed in the manufacturing process, which can be captured by the learned meta data and the proposed DiGNN. Table IV summarizes the target tasks and predictors associated with designers and manufacturers. The architecture of the local AI model for manufacturability and design feasibility prediction are the same: defined by encoders and decoders (Sec. V-A), followed by two fully-connected layers. When applying the proposed dataset-sharing framework, the manufacturers need to impute their local data, while the designers need to impute their shared data in PORL for the proxy and meta data generation as shown in Table IV.

TABLE IV: Dataset and target AI task information for designers and manufacturers (shortened as Manuf.)

Stake-holder		Predictors			Response	
		Design Variables, Design Rules, Design Intermediate Features	STL Files	Manufact-uring Setting	Feasib-ility	Manufact-urability
Designers		✓			✓	
Manuf.			✓	✓		✓
Designers → Manuf.	Local Data	Latent Imputation	✓	✓		✓
	Shared Data	✓	✓	Latent Imputation	✓	

A. Dataset Information

This study includes design datasets from six designers and manufacturing datasets from nine manufacturers collected via experiments and simulations. By varying the design space and design rules, six design datasets of MFC anodes are generated, each of which contains 500 STL files created by a designer. We collected the first manufacturing dataset through experiments in real Additive Manufacturing. From each designer, three feasible designs with a large surface area are selected for production by the manufacturer. Then, by varying the values of the process setting variables in a fractional factorial design, 132 MFC anodes are printed in LulzBot Mini 2 3D Printer in the lab. The manufacturability of each printed anode is assessed and annotated by a set of predefined criteria, including irregularities on faces, deformities on cavities/pores, the base of the design, edge quality, and conformity of the printed pores to the CAD design. Using the first manufacturing dataset, we build a Deep Neural Network (DNN) baseline model to predict manufacturability, using STL files and process setting

TABLE V: Dataset information, ground-truth sharing decisions, and outcomes of classification model for case study two

Manuf.	Sample Size	Local F1 Score	Num. of Comb. For Positive Shar.	Best F1 with Shar.	Best Comb.	Average F1 across All Comb.
M1	132	0.809	62	0.920	D2	0.885
M2	96	0.816	62	0.980	D2, D3, D4	0.905
M3	96	0.846	62	1.000	D5	0.983
M4	72	0.766	62	0.941	D4, D5, D6	0.908
M5	72	0.843	61	1.000	D1, D2	0.959
M6	96	0.838	62	1.000	D1	0.994
M7	96	0.807	62	0.907	D1	0.905
M8	60	0.831	62	0.983	D2, D3	0.910
M9	96	0.862	62	1.000	D4	0.977

variables as inputs. The STL file is converted into a voxel representation, sliced into 64 layers, and arranged into an 8x8 grid to form a 512×512 image [40]. The DNN model has two encoders: a ResNet-based encoder for the images and a MLP-based encoder for process variables. To simulate eight manufacturers with different machine configurations, we perturb the parameters in the last two layers to generate their manufacturability labels.

In Table V, we summarize the local sample size, the F1 score of the local manufacturability AI model of each manufacturer, and the groud-truth performance after receiving shared datasets from designers (columns 4-7). Despite different AI tasks, the results show that the performance of most manufacturers' AI tasks can be significantly improved (*i.e.*, columns 4, 5, 7). This is primarily due to the large design space exploited by the design datasets and the other design information not accessible to manufacturers but has substantial impact on the actual manufacturability of one design. Additionally, variations in the upstream design stage propagate to the downstream manufacturing stage, allowing the information from the feasibility classification task to aid the manufacturability classification task.

B. Hyperparameters and Settings

During the representation learning process, a ResNet-based encoder is applied to the processed images while a MLP-based encoder is applied to the remaining design data (*i.e.*, design variables, the threshold of design rules, and the intermediate features). Another MLP-based encoder is applied to the remaining manufacturing data (*i.e.*, manufacturing setting variables). The latent dimension for the image data is set to 32, while the latent dimensions for both the remaining design data and manufacturing data are set to 16. The weight for the ranking loss λ_3 is set to be 0.1.

C. Experimental Results

With six designers available for the dataset sharing, there are 63 candidate sharing decisions available for each manufacturer by summing all the combinations of 1-6 designers with one manufacturer (*i.e.*, $6 + 15 + 20 + 15 + 6 + 1 = 63$), resulting in $63 \times 9 = 567$ sharing decisions in total. We use the same benchmark methods and evaluation metrics in Section IV and evaluate the proposed method in 10-fold cross validation.

From the results summarized in Table VI, we find the proposed PORL-DiGNN with ranking loss achieves the best

TABLE VI: The average values and standard errors of the performance in 10-fold cross validation reported over 10 replications. Significant best results are highlighted in **bold** and the best benchmark performance is in blue. Blue arrows indicate relative improvement over the best benchmark method. All numbers in the table have been multiplied by 100.

Method	Input Data	Evaluation Metrics			
		NDCG (↑)	Pearson Corr. (↑)	NRMSE (↓)	F1 Score (↑)
LAVA	Proxy	94.9 ± 0.2	40.6 ± 0.3	-	93.5 ± 0.2
UtlPred	Sum. Stat.	95.1 ± 0.0	19.3 ± 7.8	30.4 ± 0.1	91.6 ± 0.0
GCN	Sum. Stat.	95.8 ± 0.1	82.0 ± 1.1	17.6 ± 0.2	92.1 ± 0.0
PORL-UtlPred	Meta Data	95.2 ± 0.0	19.1 ± 7.0	30.1 ± 0.0	93.7 ± 0.0
PORL-GCN	Meta Data	95.8 ± 0.2	83.4 ± 0.2	15.6 ± 0.1	93.7 ± 0.0
PORL-DiGNN	Meta Data	99.4 ± 0.2	86.2 ± 0.1	$\mathbf{12.5} \pm 0.0$	$\mathbf{95.2} \pm 0.2$
PORL-DiGNN with Ranking Loss	Meta Data	$\mathbf{99.5 \pm 0.2}$ ↑ 3.7	$\mathbf{86.8 \pm 0.4}$ ↑ 3.4	$\mathbf{12.5 \pm 0.0}$ ↓ 3.1	$\mathbf{95.2 \pm 0.0}$ ↑ 1.5

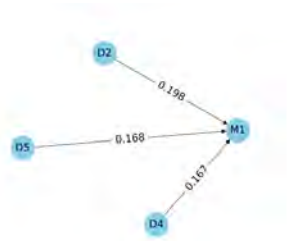

Fig. 5: The normalized attention weight for manufacturer 1

performance under most evaluation metrics. No significant differences were observed when comparing UtlPred with PORL-UtlPred and GCN with PORL-GCN. This might be due to the relatively low dimension of the input variables (*i.e.*, manufacturing setting, design variables), which results in a simpler feature representation in the proxy dataset and leads to the simpler correlation that can be featured well by summary statistics. We also found PORL-DiGNN with ranking loss exhibits similar performance to PORL-DiGNN. This is anticipated, as most sharing decisions result in positive information transfer (*i.e.*, column 4 in Table V). Hence, the ranking loss has a limited impact on performance enhancement.

Furthermore, we calculate the normalized attention weight in the aggregated message m_i^2 for $i \in N$ in the second layer in DiGNN based on Eq. 4. There is no attention mechanism applied to the first layer. In detail, we create a graph to present the sharing decision for each manufacturer. For each node i representing a manufacturer, we have:

$$\text{score}(j,i) = \underset{j \in \mathcal{P}(i)}{\text{softmax}} \left(h_i^1 \cdot W_{ij}^2 \cdot h_j^{2T} + h_i^1 \cdot W_{ij}^2 \cdot \gamma(j,i) \right),$$

where the predecessor set $\mathcal{P}(i)$ includes all six designers. Thus, $\text{score}(j,i)$ represents the normalized attention weight of the designer represented by node j to the manufacturer represented by node i. We visualize the normalized attention weight for manufacturer $M1$ and $M3$ in Fig. 5 and Fig. 6 as their sharing paths, respectively.

In Fig. 5 and Fig. 6, only designers with a score exceeding the threshold $\frac{1}{6}$, which represents the average score of all the six designers, is retained. The number on each edge represents the score of the designers, while the direction of the edge indicates the dataset-sharing direction. For example, Fig. 5

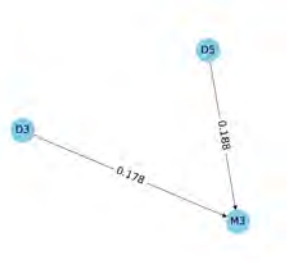

Fig. 6: The normalized attention weight for manufacturer 3

shows that Designer $D2$, $D4$ and $D5$ have a score higher than $1/6$, and the score of Designer $D2$ for $M1$ is 0.198. By comparing the column "Best Comb." in Table V with the results in the two figures, it can be observed that the normalized attention weight is consistent with the relative importance of the designer's datasets to the manufacturer's target AI task. Although the normalized attention scores do not deviate significantly from the average value of $\frac{1}{6}$, they effectively highlight the importance of a potential stakeholder when making dataset-sharing decisions, thus indicating the sharing pattern.

VI. Conclusions

High-quality data is key to AI model performance. To efficiently train advanced AI models for critical applications, a context-aware, fine-grained, and privacy-preserving data trading system is essential. This work focuses on a dataset-sharing framework as the foundation for data trade. Existing data sharing approaches cannot provide accurate estimation of the value of a dataset across different data modalities, target AI tasks, and dataset ownership in a privacy-preserving manner. We propose a performance-oriented representation learning framework within a Directed Graph Neural Network named PORL-DiGNN. With representation learning generating the proxy dataset for sharing and meta data for making the sharing decisions locally, the proposed method secures the privacy and value of datasets while effectively characterizing stakeholders by the attention mechanism. By graph-level supervised learning, the performance of the target AI task attained by dataset-sharing between stakeholders can be estimated accurately. Two real case studies demonstrate the advantage of PORL-DiGNN over benchmark methods in effectively guiding sharing decisions for stakeholders from the same manufacturing stage and from upstream to downstream stages.

The work leaves us with several directions for future research. Firstly, the proposed dataset-sharing framework can be generalized to the scenario of multiple datasets and AI tasks for each stakeholder by integrating the utility functions of each AI model. Secondly, we will investigate a dynamic graph neural network [41] to consider the changing distribution of data of each stakeholder in an online setting. Furthermore, we plan to investigate a pricing mechanism building on the existing dataset-sharing paradigm. We will utilize techniques from personalized differential privacy [42] to align privacy settings with different pricing tiers to adapt to stakeholder's protocols, standards, and methodologies.

References

[1] L. Wang, X. Chen, D. Henkel, and R. Jin, "Pyramid ensemble convolutional neural network for virtual computed tomography image prediction in a selective laser melting process," *Journal of Manufacturing Science and Engineering*, vol. 143, no. 12, p. 121003, 2021.

[2] W. Wei, C. Jiang, and Y. Huang, "A data-driven human–machine collaborative product design system toward intelligent manufacturing," *IEEE Transactions on Automation Science and Engineering*, 2023.

[3] H. A. Just, F. Kang, T. Wang, Y. Zeng, M. Ko, M. Jin, and R. Jia, "Lava: Data valuation without pre-specified learning algorithms," in *The Eleventh International Conference on Learning Representations*. OpenReview, 2023.

[4] Q. P. He and J. Wang, "Fault detection using the k-nearest neighbor rule for semiconductor manufacturing processes," *IEEE transactions on semiconductor manufacturing*, vol. 20, no. 4, pp. 345–354, 2007.

[5] N. Gupta, S. Mujumdar, H. Patel, S. Masuda, N. Panwar, S. Bandyopadhyay, S. Mehta, S. Guttula, S. Afzal, R. Sharma Mittal *et al.*, "Data quality for machine learning tasks," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2021, pp. 4040–4041.

[6] P. Branco, L. Torgo, and R. P. Ribeiro, "A survey of predictive modeling on imbalanced domains," *ACM Computing Surveys (CSUR)*, vol. 49, no. 2, pp. 1–50, 2016.

[7] S. Kang, R. Jin, X. Deng, and R. S. Kenett, "Challenges of modeling and analysis in cybermanufacturing: a review from a machine learning and computation perspective," *Journal of Intelligent Manufacturing*, pp. 1–14, 2023.

[8] Y. Zeng, X. Chen, and R. Jin, "Ensemble active learning by contextual bandits for ai incubation in manufacturing," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 1, pp. 1–26, 2023.

[9] Y. Zeng, P. Thiyagarajan, B. M. Chan, and R. Jin, "Synthetic data generation and sampling for online training of dnns in manufacturing supervised learning problems," in *2023 IEEE 19th International Conference on Automation Science and Engineering (CASE)*. IEEE, 2023, pp. 1–6.

[10] H. Chen, W. Wan, M. Matsushita, T. Kotaka, and K. Harada, "Automatically prepare training data for yolo using robotic in-hand observation and synthesis," *IEEE Transactions on Automation Science and Engineering*, 2023.

[11] B. Milletler, "Data economy: radical transformation or dystopia?" *Frontier Technology Quarterly*, 2019.

[12] N. Lu, B. Ji, and B. Li, "Age-based scheduling: Improving data freshness for wireless real-time traffic," in *Proceedings of the eighteenth ACM international symposium on mobile ad hoc networking and computing*, 2018, pp. 191–200.

[13] K. Gupta, D. Saxena, P. Rani, J. Kumar, A. Makkar, A. K. Singh, and C.-N. Lee, "An intelligent quantum cyber-security framework for healthcare data management," *IEEE Transactions on Automation Science and Engineering*, 2024.

[14] S. J. Shackelford, "Smart factories, dumb policy? managing cybersecurity and data privacy risks in the industrial internet of things," *Minn. JL Sci. & Tech.*, vol. 21, p. 1, 2019.

[15] M. U. Hassan, M. H. Rehmani, and J. Chen, "Differential privacy techniques for cyber physical systems: a survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 1, pp. 746–789, 2019.

[16] A. Arunan, Y. Qin, X. Li, and C. Yuen, "A federated learning-based industrial health prognostics for heterogeneous edge devices using matched feature extraction," *IEEE Transactions on Automation Science and Engineering*, 2023.

[17] P. Shojaee, Y. Zeng, M. Wahed, A. Seth, R. Jin, and I. Lourentzou, "Task-driven privacy-preserving data-sharing framework for the industrial internet," in *2022 IEEE International Conference on Big Data (Big Data)*. IEEE, 2022, pp. 1505–1514.

[18] S. Kang, X. Deng, and R. Jin, "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*, vol. 21, no. 6, p. 061008, 2021.

[19] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *International conference on machine learning*. PMLR, 2017, pp. 1885–1894.

[20] R. Jia, D. Dao, B. Wang, F. A. Hubis, N. Hynes, N. M. Gürel, B. Li, C. Zhang, D. Song, and C. J. Spanos, "Towards efficient data valuation based on the shapley value," in *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 1167–1176.

[21] S. Tang, A. Ghorbani, R. Yamashita, S. Rehman, J. A. Dunnmon, J. Zou, and D. L. Rubin, "Data valuation for medical imaging using shapley value and application to a large-scale chest x-ray dataset," *Scientific reports*, vol. 11, no. 1, p. 8366, 2021.
[22] Y. Kwon and J. Zou, "Beta shapley: a unified and noise-reduced data valuation framework for machine learning," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 8780–8802.
[23] J. Yoon, S. Arik, and T. Pfister, "Data valuation using reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2020, pp. 10 842–10 851.
[24] Y. Wang, J. Wang, F. Gao, and J. Song, "Unveiling value patterns via deep reinforcement learning in heterogeneous data analytics," *Patterns*, vol. 5, no. 5, 2024.
[25] J.-S. Lee, C.-J. Chew, J.-Y. Liu, Y.-C. Chen, and K.-Y. Tsai, "Medical blockchain: Data sharing and privacy preserving of ehr based on smart contract," *Journal of Information Security and Applications*, vol. 65, p. 103117, 2022.
[26] B.-K. Zheng, L.-H. Zhu, M. Shen, F. Gao, C. Zhang, Y.-D. Li, and J. Yang, "Scalable and privacy-preserving data sharing based on blockchain," *Journal of Computer Science and Technology*, vol. 33, pp. 557–567, 2018.
[27] F. A. Putra, H. Febriansyah, and R. F. Sari, "Blockchain-based data owner rating in medical record data sharing using ethereum," in *2022 20th International Conference on ICT and Knowledge Engineering (ICT&KE)*. IEEE, 2022, pp. 1–9.
[28] Z. Li, J. Zhang, J. Zhang, Y. Zheng, and X. Zong, "Integrated edge computing and blockchain: A general medical data sharing framework," *IEEE Transactions on Emerging Topics in Computing*, 2023.
[29] X. Zheng and Z. Cai, "Privacy-preserved data sharing towards multiple parties in industrial iots," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 5, pp. 968–979, 2020.
[30] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
[31] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[32] B. Škrlj, S. Džeroski, N. Lavrač, and M. Petkovič, "Feature importance estimation with self-attention networks," *arXiv preprint arXiv:2002.04464*, 2020.
[33] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," in *International conference on machine learning*. PMLR, 2017, pp. 1263–1272.
[34] V. Thost and J. Chen, "Directed acyclic graph neural networks," *arXiv preprint arXiv:2101.07965*, 2021.
[35] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.
[36] H. Sun, X. Deng, K. Wang, and R. Jin, "Logistic regression for crystal growth process modeling through hierarchical nonnegative garrote-based variable selection," *Iie Transactions*, vol. 48, no. 8, pp. 787–796, 2016.
[37] Y. Zeng, J. T. Wang, S. Chen, H. A. Just, R. Jin, and R. Jia, "Modelpred: A framework for predicting trained model from training data," in *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE, 2023, pp. 432–449.
[38] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.
[39] Y. Shi, Y. Zhang, S. Baek, W. De Backer, and R. Harik, "Manufacturability analysis for additive manufacturing using a novel feature recognition technique," *Computer-Aided Design and Applications*, vol. 15, no. 6, pp. 941–952, 2018.
[40] S. K. P. K. Chilukuri, B. Song and R. Jin, "Generating optimized 3d designs for manufacturing using a guided voxel diffusion model," in *Proc. ASME 2024 Int. Manuf. Sci. Eng. Conf., MSEC2024*, 2024.
[41] J. Skarding, B. Gabrys, and K. Musial, "Foundations and modeling of dynamic networks using dynamic graph neural networks: A survey," *iEEE Access*, vol. 9, pp. 79 143–79 168, 2021.
[42] H. Ebadi, D. Sands, and G. Schneider, "Differential privacy: Now it's getting personal," *Acm Sigplan Notices*, vol. 50, no. 1, pp. 69–81, 2015.

Yingyan Zeng is an assistant professor in the Department of Mechanical and Materials Engineering at the University of Cincinnati. She received her Ph.D. degree in Industrial and Systems Engineering, her Master's degree in Computer Engineering, both at Virginia Tech in 2024, and her bachelor's degree in Shanghai Jiao Tong University, Shanghai. Her research interest lies in the interface between data quality and machine learning in Manufacturing Industrial Internet, with an emphasis on the quality-assured multimodality data generation, data acquisition, privacy-preserving data valuation and sharing methods to enhance the AI modeling performance in advanced manufacturing.

Xiaona Zhou is a 2nd-year Ph.D. student in Computer Science at the University of Illinois Urbana-Champaign. Her research experience includes privacy-preserving machine learning, dataset selection through reinforcement learning, and cross-modal information retrieval. Alongside her academic work, she is a year-round research intern at Sandia National Laboratories, where she focuses on machine learning for anomaly detection across various scenarios.

Premith Kumar Chilukuri is pursuing a Master's in Computer Science at Virginia Tech. He specializes in developing cutting-edge AI technologies, including LLMs, computer vision, and NLP. Premith has authored papers accepted at top-tier conferences and journals, and holds patents in video inpainting and dynamic calibration systems. His research interests include generative AI, federated learning, and AI applications in robotics and manufacturing. He has internship experience as Machine Learning Engineer at ABB Research, Samsung R&D Institute, Supervue AI, and Zebi Data.

Ismini Lourentzou is an assistant professor in the School of Information Sciences at the University of Illinois Urbana-Champaign, where she leads the Perception and LANguage (PLAN) Lab. Her research focus is multimodal machine learning, primarily the intersection of vision and language in settings with limited supervision, and its applications in healthcare, embodied AI, and other fields. She has served on the NeurIPS organizing committees, and in editorial and area chair roles for PLOS Digital Health, ECCV, ACL, IEEE BigData and MICCAI. Her research has been supported by the National Science Foundation (NSF), Defense Advanced Research Projects Agency (DARPA), Commonwealth Cyber Initiative (CCI), and Amazon. Before joining iSchool, she was an assistant professor at Virginia Tech, Computer Science Department.

Ran Jin is an Associate Professor and the Director of Laboratory of Data Science and Visualization at the Grado Department of Industrial and Systems Engineering at Virginia Tech. He received his Ph.D. degree in Industrial Engineering from Georgia Tech, Atlanta, his Master's degrees in Industrial Engineering, and in Statistics, both from the University of Michigan, Ann Arbor, and his bachelor's degree in Electronic Engineering from Tsinghua University, Beijing. His research focuses on machine learning in Manufacturing Industrial Internet, human centered AI, human-AI collaboration, and data quality. His research outcomes have been broadly applied in additive manufacturing, thermal spray coating, broaching, semiconductor, printed electronics, optical fiber, and continuous fiber manufacturing industries.

Ran Jin
Associate Professor
Grado Department of Industrial and Systems Engineering
1185 Perry Street (0118)
Virginia Tech, Blacksburg, VA 24061

111 Durham Hall
1145 Perry Street, Blacksburg, VA 24061
Phone: (540) 231-5936
Email: jran5@vt.edu

Editor-in-Chief
IEEE Transactions on Automation Science and Engineering

Dear Editor-in-Chief,

We are submitting our manuscript in title "High-Quality Dataset-Sharing and Trade Based on A Performance-Oriented Directed Graph Neural Network" for possible publication in IEEE Transactions on Automation Science and Engineering.

Data quality is the key to the performance of Artificial Intelligence (AI) tasks in advanced manufacturing systems. In this manuscript, we proposed a performance-oriented representation learning (PORL) framework in a Directed Graph Neural Network (DiGNN) as the foundation for future data exchange and trade. By a two-level representation learning approach and formulating the estimation of target AI task performance attained by dataset-sharing as a graph-level supervised learning problem, the proposed framework enables the effective and privacy-preserving dataset sharing between data stakeholders with different contexts (*i.e.*, data formats, target AI tasks, and inherent connections between datasets) to improve the data receiver's AI task. The advantages of this method were demonstrated in two case studies: one in the semiconducting manufacturing network where datasets are shared between similar manufacturing processes, another in the design and manufacturing network of Microbial Fuel Cell anodes where datasets are shared between upstream (design) and downstream (Additive Manufacturing) stages. Our work shows the potential to be readily employed by many applications and research studies.

We believe this work will attract the readers of IEEE Transactions on Automation Science and Engineering, given its relevance to automation science and engineering for the AI development in advanced manufacturing systems.

The primary methodology of this paper is Artificial Intelligence and Machine Learning, and the primary application is Manufacturing.

Thank you for considering our submission. We look forward to hearing feedback from the editors and reviewers.

Sincerely,

Ran Jin
Associate Professor
Grado Department of Industrial and Systems Engineering
1185 Perry Street (0118)
Virginia Tech, Blacksburg, VA 24061
Phone: (540) 231-5936
Email: jran5@vt.edu