A Task-Driven Privacy-Preserving Data-Sharing Framework for the Industrial Internet

Parshin Shojaee
Grado Department of Industrial and Systems Engineering
Virginia Tech
parshinshojaee@vt.edu

Yingyan Zeng
Grado Department of Industrial and Systems Engineering
Virginia Tech
yingyanzeng@vt.edu

Muntasir Wahed
Department of Computer Science
Virginia Tech
mwahed@vt.edu

Avi Seth
Department of Computer Science
Virginia Tech
aviseth@vt.edu

Ran Jin
Grado Department of Industrial and Systems Engineering
Virginia Tech
jran5@vt.edu

Ismini Lourentzou
Department of Computer Science
Virginia Tech
ilourentzou@vt.edu

***Abstract*—Industrial Internet provides a collaborative computational platform for participating enterprises, allowing the collection of big data for machine learning tasks. Despite the promise of training and deployment acceleration, and the potential to optimize decision-making processes through data-sharing, the adoption of such technologies is impacted by the increasing concerns about information privacy. As enterprises prefer to keep data private, this limits interoperability. While prior work has largely explored privacy-preserving mechanisms, the proposed methods naively average or randomly sample data shared from all participants instead of selecting the most well-suited subsets for a particular downstream learning task. Motivated by the lack of effective data-sharing mechanisms for heterogeneous machine learning tasks in Industrial Internet, we propose PriED, a task-driven data-sharing framework that selectively fuses shared data and local data from participants to improve supervised learning performance. PriED utilizes privacy-preserving data distillation to facilitate data exchange, and dynamic data selection to optimize downstream machine learning tasks. We demonstrate performance improvements on a real semiconductor manufacturing case study.**

***Index Terms*—Attention, Data-Sharing, Data Distillation, Industrial Internet, Privacy-Preserving, Reinforcement Learning, Task-Driven Data Selection.**

I. Introduction

Deep learning models have significantly improved performance on various supervised tasks [1, 2] but require large-scale labeled datasets to extract valuable knowledge [3, 4]. In industrial settings, collecting such annotated datasets is often infeasible, prohibitively expensive, or impractical for a single entity. Consequently, this calls for multiple participants to collaborate with their data. Lots of existing useful proprietary or sensitive data are distributed among data owners, yet privacy concerns prohibit sharing, particularly in critical applications, such as healthcare, or among participants with competing interests, such as manufacturing. The bottleneck, therefore, lies in how to share data while protecting data privacy and ensuring scalability in centralized model training.

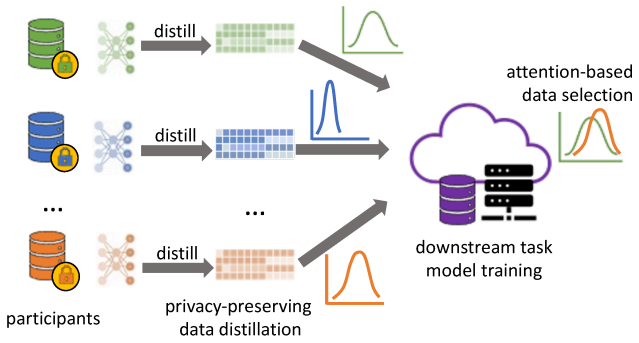

Fig. 1: Overview of the proposed PriED framework. Each participant can share distilled privacy-preserving data representations, by training their own generative data distillation model and desensitizing their private data. Given a specific downstream task, the server performs task-driven data selection to fit a model on the most suitable subset of distilled data from different participants.

To solve this problem, previous works have relied on different approaches, including but not limited to differential privacy (DP), compressive privacy, federated learning, *etc.* [5, 6, 7, 8, 9, 10]. Due to the large number of participating heterogeneous clients, these methods pose several challenges. For example, directly averaging parameters requires participants to have the same model structure. As the number of participants grows, it becomes infeasible in practice to collect and average model weights. In addition, requiring model weight or gradient exchange not only remains computationally expensive but comes with possible security risks and privacy concerns [11, 12, 13]. Model distillation has recently been applied to this task, where random subsets of clients are distilled on a server ensemble (student) that is trained by averaging the clients' (teachers') prediction logits on unlabeled, public, or artificially generated data [14, 15, 16]. These frameworks

978-1-6654-8045-1/22/$31.00 ©2022 IEEE

eliminate some of the vulnerabilities but rely on uniformly averaging model predictions or randomly sampling subsets of models at each distillation step, instead of selecting the most suitable data for the downstream task at hand. As such, randomly acquiring information from all participants results in suboptimal trained models and often may result in negative information transfer.

Beyond preserving privacy, another challenge thus lies in selecting which dataset or data instances are useful for a specific downstream task. Since participants in a data-sharing ecosystem might be fairly diverse in terms of the underlying company processes and variable relationships, a crucial component in improving model performance is the ability to retrieve participants that are similar. For example, in manufacturing data-sharing, these similar participants can produce similar products with the required specifications or use similar manufacturing processes, recipes, and equipment. As a result, some recent works focus on learning retrieval mechanisms for data selection [17, 18, 19]. Combining task-driven data-sharing frameworks and privacy-preserving generative methods would enable a more intelligent data-sharing ecosystem with built-in incentives and privacy protocols. However, this line of work has gained fairly limited attention.

To address these challenges, in this work, we propose **PriED**, a **Pri**vacy-pr**E**serving **D**ata-sharing framework that allows participants to share distilled synthetic data for downstream tasks while preserving data privacy. In contrast to parameter-based sharing or prediction-based sharing methods, PriED first learns intermediate privacy-enabled data representations for each participant. These distilled data representations are then fused with an attention mechanism that captures similarity among participants towards the intended task, resulting in a task-driven data-sharing framework. The learned attention-based similarity can effectively select data points from multiple data owners, conditioned on the corresponding data receiver. The model is trained end-to-end with reinforcement learning (policy gradient), assigning rewards based on the prediction accuracy for the downstream task. While in this paper, we focus on supervised learning downstream tasks, PriED can be easily adapted for various other tasks and domains, such as semi-supervised learning with corresponding discrete performance metrics. Most importantly, PriED mitigates privacy risks with similar computation costs as baseline global and local models and allows flexible task-driven data aggregation from heterogeneous data owners with varying data types, dimensionalities, and preprocessing mechanisms.

Our contributions can be summarized as follows: **(1)** We introduce a privacy-preserving data-sharing paradigm for industrial data that enables synergistic partnerships among participants and reduces data collection and annotation efforts for each participant. **(2)** We propose PriED, an end-to-end model that consists of privacy-preserving data distillation and attention-based data selection phases. Experimental evaluation on a real semiconductor manufacturing case study showcases the effectiveness of the proposed method. **(3)** We show that PriED progressively learns a task-driven similarity that can capture inherent engineering similarities among participants.

II. Related Work

A. Privacy-Preserving Data Sharing

Methods such as differential privacy enable sharing of useful data patterns while protecting individual aspects of the data that cannot be publicly shared. To achieve differential privacy, Xie et al. [20] propose DP-GAN, a Generative Adversarial Network (GAN) [21] model that is trained with injected gradient noise. Frigerio et al. [22] extend DP-GAN by incorporating clipping decay optimization [23], while Triastcyn and Faltings [24] propose a differentially private critic to ensure that synthetic data created by the generator do not appear in the training samples. Torkzadehmahani et al. [25] clip the gradients of the real and synthetic data separately to have better control over the sensitivity towards real data.

A separate line of work focuses on teacher-student ensemble models. For example, Private Aggregation of Teacher Ensembles (PATE) [26] utilizes multiple "teacher" models trained on disjoint datasets to train a single "student" model via noisy voting. Jordon et al. [27] replace the discriminator with a set of teacher and student discriminators, where the student discriminators learn from the data generated by the teacher discriminator. Long et al. [28] replace the discriminator in GAN with the teachers from the PATE model and train the student generator using a private gradient aggregation mechanism to produce synthetic samples.

Specifically in IoT settings, Du et al. [29] propose a communication efficient privacy-preserving protocol, where a differentially private approximate mechanism facilitates the distributed training. In our work, we consider sharing latent representations rather than the raw data itself. Most relevant to our setting, Li et al. [30] propose a robust privacy-preserving method to obscure important latent presentations for text preprocessing tasks. In contrast to prior work, and apart from learning latent representations that are invariant to sensitive features, we propose to combine distilled synthetic data in a task-driven manner instead of directly averaging or uniformly sampling, and thus improving data selection and downstream performance, while also learning inherent similarity mechanisms that are useful for grouping or retrieving participants with similar processes.

B. Learn to Select Data

For real-world tasks, the scarcity of annotations prevents the application of supervised deep learning methods that rely on large datasets. Several research works focus on addressing data collection and data selection challenges from various perspectives. One such direction is active learning (AL), which aims at incorporating targeted human annotations into the model training process, by iteratively selecting the most informative data points to pass to a subject matter expert for annotation. Commonly used criteria are uncertainty sampling [31], density-weighted uncertainty sampling [32, 33], diversity [34], QUIRE [35], extreme learning methods [36] and bayesian

active learning methods such as BALD (Bayesian Active Learning by Disagreement) [37].

Other works focus on interleaving processes to reduce annotator waiting time in batch active learning [38], and active learning domain adaptation settings by clustering uncertainty-weighted embeddings [39] or by utilizing reinforcement learning [40], Bayesian Optimization [41], and domain similarity metrics [42]. Recent works formulate active learning as a multi-armed bandit problem and select data from a set of candidates in each round [43, 44, 45]. Vu et al. [18] learn an active learning query strategy using reinforcement learning and feedback from human annotators, while Liu et al. [17] leverage imitation learning in a similar setup.

Nevertheless, most of these works target specific natural language processing and image classification problems and do not deal with privacy concerns. The benefit of privacy-preserving data-sharing models is that they enable the participants to share the data without violating any privacy constraints, and thus collaboratively utilize larger amounts of data for downstream tasks. Our work introduces a framework for learning to select data from a set of participants, where desensitized distilled data are shared in lieu of raw data that involve confidential and proprietary information.

III. Methodology

A. Problem Formulation

Let $\mathcal{M}$ be the set of participants (data owners) in the data-sharing ecosystem, each associated with a *private* multivariate time series data source $\{\boldsymbol{X}_i^t \in \mathbb{R}^F\}_{i=1}^M$. Here, $\boldsymbol{X}_i^t$ denotes all F features of i-th data owner at time t. Multivariate time series (MVTS) data collected from interconnected sensors and actuators are the most common types of data available in the Industrial Internet of Things (IIoT) and smart manufacturing [46, 47, 48, 49]. Let k be the data receiver, *i.e.*, the participant that wishes to request data for the downstream task. Subsequently, $\mathcal{M}_{-k}$ denotes the rest of the data owners, *i.e.*, excluding the data receiver k.

To enable privacy-preserving effective data sharing among the participants (*i.e.*, data owners and receivers), we introduce the PriED framework that consists of two key components, data distillation and data selection. Similar to prior works, we consider a subset of sensitive features, among the set of all features, that need to be kept private. For example, these sensitive features might reveal the setting of the underlying manufacturing process, that is typically private proprietary information. To alleviate privacy restrictions w.r.t. sensitive features, we propose a local privacy-preserving deep generative model (DGM) for the participants. The DGM model takes the private data as input and learns to approximate them in order to generate additional data points $\hat{\boldsymbol{X}}_i^t$ that resemble the learned data distribution (details in subsection III-B).

An attention-based model (described in subsection III-C) supports the dynamic data selection by progressively learning to retrieve data from participants based on data contributions to the supervised learning downstream task performance. Due to the sequential nature of the problem, *i.e.*, selecting the next data candidate based on what has been selected so far, we formulate the data selection task as a Markov Decision Process (MDP) and train with policy-gradient (subsection III-D). An overview of the PriED is presented in Fig. 2.

B. Privacy-preserving Data Distillation

To distill data from each data owner participant, we utilize privacy-preserving Variational Autoencoder Long Short-term Memory (VAE-LSTM) deep generative models [50]. Each data owner i trains their own private data distillation network.

The encoder is an LSTM network such that for the sequential data of each data owner i, the state $\boldsymbol{h}_i^t$ is calculated based on the previous state $\boldsymbol{h}_i^{t-1}$ and the input $\boldsymbol{X}_i^t$ of the current time step. The distribution of the VAE latent representation is obtained from the last state of the LSTM:

$$\boldsymbol{h}_i^t = f_{enc}(\boldsymbol{h}_i^{t-1}, \boldsymbol{X}_i^t), \tag{1}$$

$$\mu_{\boldsymbol{z}_i} = \boldsymbol{W}_\mu^\top h_i^{end} + b_\mu, \tag{2}$$

$$\log(\sigma_{\boldsymbol{z}_i}) = \boldsymbol{W}_\sigma^\top h_i^{end} + b_\sigma, \tag{3}$$

where $\boldsymbol{W}_\mu, \boldsymbol{W}_\sigma$ are learnable weight matrices, and b_μ, b_σ are bias terms. In addition, $\mu_{\boldsymbol{z}_i}$ and $\sigma_{\boldsymbol{z}_i}$ correspond to the mean and variance of the learned latent Gaussian distribution $\mathcal{N}(\mu_{\boldsymbol{z}_i}, \sigma_{\boldsymbol{z}_i})$. By using the reparameterization technique, a latent representation is sampled from the encoding distribution, *i.e.*, $\boldsymbol{z}_i \sim \mathcal{N}(\mu_{\boldsymbol{z}_i}, \sigma_{\boldsymbol{z}_i})$. The initial state $\boldsymbol{h'}_i^0$ of the LSTM-based decoder model is computed via

$$\boldsymbol{h'}_i^0 = \boldsymbol{W}_z^\top \boldsymbol{z}_i + b_z, \tag{4}$$

where $\boldsymbol{W}_z$ and b_z are the learnable weight matrix and bias vector, respectively. Thereafter, we can map the hidden state of each decoded time step into the multi-variate dimension of the input and reconstruct it as $\hat{\boldsymbol{X}}_i^t$

$$\boldsymbol{h'}_i^t = f_{dec}(\boldsymbol{h'}_i^{t-1}, \boldsymbol{X}_i^t), \tag{5}$$

$$\hat{\boldsymbol{X}}_i^t = \boldsymbol{W}_o^\top \boldsymbol{h'}_i^t + b_o, \tag{6}$$

where $\boldsymbol{W}_o$ is a learnable weight matrix, and b_o is a bias vector. The joint VAE-LSTM, denoted as $g_\phi(\boldsymbol{X}_i)$, will be locally trained for each data owner i, based on a hybrid loss function that combines the VAE loss with an adversarial loss for the target sensitive features that need to be kept private in the latent representations. In other words, we seek to learn latent representations that reconstruct the input very well, while being a poor representation for the reconstruction of target-sensitive features. This adds another layer of data protection. Note that our framework is fairly general and other privacy-preserving methods can be applied, *i.e.*, the model choice for generating distilled data is orthogonal to the proposed method.

Denoting the set of all features as $\mathcal{F}$ and the set of few sensitive target features as Ω, the hybrid loss is written as

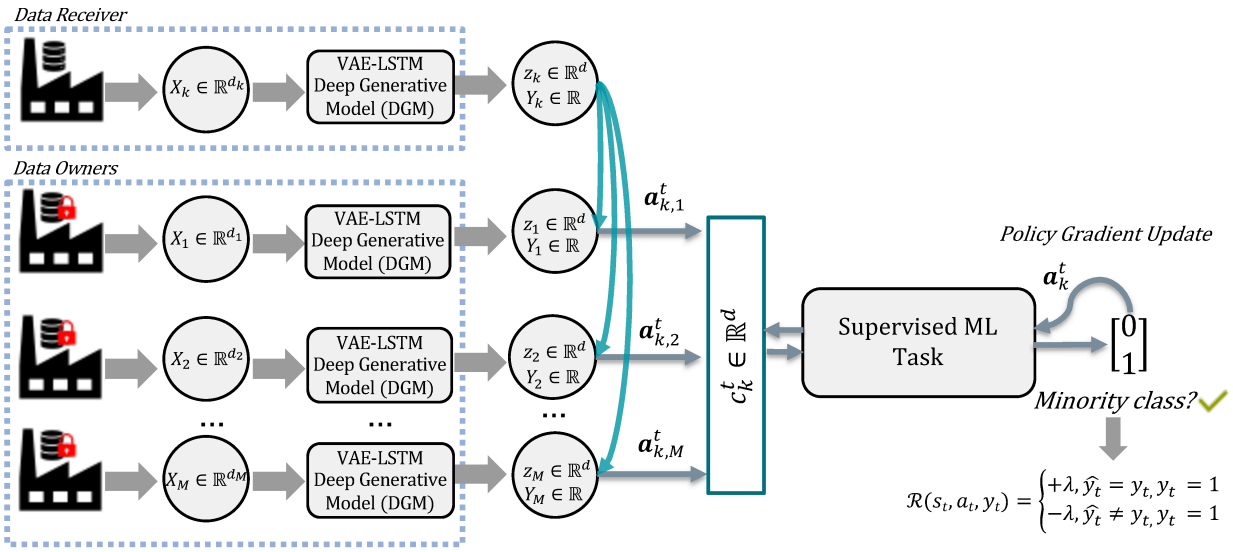

Fig. 2: Pictorial overview of PriED. Data from multiple participants (data owners) are distilled into low-dimensional vector representations that are invariant to sensitive features and hence preserve data privacy. An attention-based dynamic data selection mechanism progressively learns to retrieve data from participants based on their data contributions to the downstream task performance for the respective data receiver. PriED is trained end-to-end with reinforcement learning to allow the incorporation of flexible reward mechanisms, *e.g.*, incentivizing correct predictions for the minority class in an imbalanced classification task.

follows.

$$\mathcal{L} = \min_{\phi} \Big\{ \|\{\boldsymbol{X}_i - g_\phi(\boldsymbol{X}_i)\}_{w\in\mathcal{F}\setminus\Omega}\|^2 + \lambda_1 D_{KL}(\mu_{\boldsymbol{z}_i}, \sigma_{\boldsymbol{z}_i} || \mathcal{N}(0,1)) - \lambda_2 \|\{\boldsymbol{X}_i - g_\phi(\boldsymbol{X}_i)\}_{w\in\Omega}\|^2 \Big\}, \tag{7}$$

where the first two terms of the loss refer to the VAE loss, which consists of the input reconstruction squared error on non-sensitive features $w \in \mathcal{F} \setminus \Omega$ and the KL-divergence minimization between the learned distribution $(\mu_{\boldsymbol{z}_i}, \sigma_{\boldsymbol{z}_i})$ and the latent distribution, that is assumed to be a standard Gaussian distribution $\mathcal{N}(0,1)$. Such Gaussian latent priors are commonly used in the literature due to allowing reparameterization [51, 52, 53]. The negative sign of the third term referred to as the adversarial loss can be implemented by the gradient reversal layer during backpropagation [54]. In other words, the gradients are reversed in sign for the reconstruction of sensitive features $w \in \Omega$ since we want the VAE latent representations to be maximally poor for the reconstruction of these features. In this objective function, λ_1 and λ_2 are hyperparameters that balance the impact of different loss terms.

C. Data Selection Mechanism

Once each participant can share distilled data in the form of latent low-dimensional representations $\boldsymbol{z}_i \in \mathbb{R}^d, \forall i \in \mathcal{M}_{-k}$, the data receiver k has to select which data sources to query, conditioned on their local data and the contribution of the selected data to the downstream task. To facilitate task-driven data retrieval, we utilize attention. More specifically, we perform a cross-correlation similarity operation that compares data receiver k and data owner i via a bi-linear attention

$$sim\,(\boldsymbol{z}_k, \boldsymbol{z}_i) = \boldsymbol{z}_k{}^\top \boldsymbol{W}_a^t \boldsymbol{z}_i, \tag{8}$$

where $\boldsymbol{W}_a$ is a learnable weight matrix and $\boldsymbol{z}_k = g_\psi(\boldsymbol{X}_k)$ is an LSTM-based low-dimensional local representation for the data receiver k, to project the raw data from the data receiver to the same d-dimensional embedding space as the distilled data from data owners. This similarity is then normalized and transformed into a probability distribution:

$$\boldsymbol{a}_{k,i}^t = \frac{sim(\boldsymbol{z}_k, \boldsymbol{z}_i)}{\sum_{j\in\mathcal{M}_{-k}} sim(\boldsymbol{z}_k, \boldsymbol{z}_j)}, \tag{9}$$

where $\boldsymbol{a}_{k,i}^t$ refers to the attention weights between data receiver k and data owner i at time step t. After learning these attention weights for a specific task, we can quantify the task-driven similarity of each data receiver k with all the other data owners $i \in \mathcal{M}_{-k}$. Finally, we can compute the attention output $\boldsymbol{c}_k^t$ as the aggregated representation of data receiver k at time step t based on the weighted sum of attention weights and distilled data from different data owners as

$$\boldsymbol{c}_k^t = \sum_{i\in\mathcal{M}_{-k}} \boldsymbol{a}_{k,i}^t \boldsymbol{z}_i. \tag{10}$$

Here, the attention weights $\boldsymbol{a}_{k,i}^t$ indicate the preference of the data receiver k and determine the selection among the data owners. As our objective is to enable task-driven data-sharing between the participating data receiver and data owners, the attention weights can capture any inherent participant similarities and can be used to query data points based on performance improvements on the downstream task.

D. Policy Gradient

Due to the sequential nature of data selection, we frame the learning task as a Markov Decision Process (MDP) and utilize reinforcement learning. The MDP is formulated as follows:

State $\mathcal{S}$: The state of the environment is determined by the context vectors at time t, *i.e.*, $s_t = \boldsymbol{c}_k^t, s_t \in \mathcal{S}$.

Action $\mathcal{A}$: We consider the label set as the action set since the model prediction is w.r.t. the given downstream task, that is we train the model end-to-end. For a binary classification problem, the action set is $\mathcal{A} = \{0, 1\}$, *i.e.*, we denote the model prediction at time step t as $a_t \in \mathcal{A}$.

Reward $\mathcal{R}$: To address the potential imbalance in the binary classification task, following prior work [55], we define the reward as follows:

$$\mathcal{R}(s_t, a_t, y_t) = \begin{cases} +1, a_t = y_t \text{ and } y_t = 0, \\ -1, a_t \neq y_t \text{ and } y_t = 0, \\ +\lambda, a_t = y_t \text{ and } y_t = 1, \\ -\lambda, a_t \neq y_t \text{ and } y_t = 1, \end{cases} \tag{11}$$

where y_t is the label for the data sample observed at state s_t, $\lambda > 1$ and $y_t = 1$ denotes the minority class. This reward function formulation has been shown to outperform several imbalanced classification algorithms [55].

Policy $\pi_{\boldsymbol{\theta}}(a_t|s_t)$: the policy network is parameterized by the downstream supervised learning model that is conditioned on context vectors $\boldsymbol{c}_k^t$. The model is a fully-connected layer with sigmoid activations, *i.e.*, $\pi_{\boldsymbol{\theta}}(a_t|s_t) = f(\boldsymbol{W}_\theta^\top \boldsymbol{c}_k^t + b_\theta)$, where f is the sigmoid function, $\boldsymbol{W}_\theta$ are learnable weights, and b_θ is a bias term.

The objective is to maximize the expected cumulative reward over time, *i.e.*,

$$\arg\max_{\boldsymbol{\theta}} \mathcal{L}_{\boldsymbol{\theta}} = \mathbb{E}_{\pi_{\boldsymbol{\theta}}} \left[\sum_{t=0}^{T} \mathcal{R}(s_t, a_t, y_t) \right]. \tag{12}$$

We adopt policy gradients for training. Therefore, updating the policy parameters can be derived as:

$$\nabla_{\boldsymbol{\theta}_t} \mathcal{L}_{\boldsymbol{\theta}} = \mathbb{E}_{\pi_{\boldsymbol{\theta}}} \left[\left(\sum_{t=0}^{T} \mathcal{R}(s_t, a_t, y_t) \right) \left(\sum_{t=0}^{T} \nabla_{\boldsymbol{\theta}_t} log \pi_{\boldsymbol{\theta}_t}(a_t|s_t) \right) \right]. \tag{13}$$

IV. Experiments

A. A Real Manufacturing Test Case

Our research is motivated by Czochralski crystal growth processes (CZ processes) to manufacture ingots, considered as manufacturing participants. A pictorial overview of the CZ process is shown in Fig. 3. The CZ process determines the initial product quality in the semiconductor manufacturing industry. Therefore, it is extremely important to model and control product quality at this stage. To do so, we can model the reworks, *i.e.*, restarting a segment of the CZ process. [56, 57, 58]. There are different reasons for rework. For example, the normal conditions of the CZ process will produce single crystalline silicon ingots; and the abnormal conditions may lead to mono-crystalline defects. In quality inspection, data collected are often used to train a supervised model to predict rework due to defects based on CZ process variables.

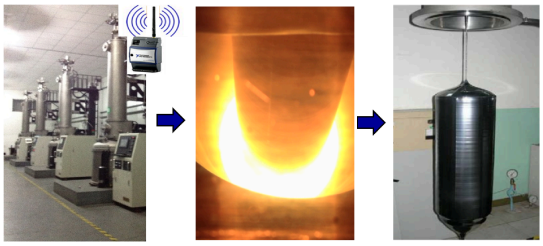

Fig. 3: CZ manufacturing process (redrawn from [59]).

One of the common issues encountered in quality modeling is imbalanced class distributions, *e.g.*, the number of defective samples (minority class) is disproportionate to the number of normal samples. It may take several months to collect sufficient data to train a neural network model for quality modeling. Since the sample size depends on the number of furnaces present in the manufacturing system, small manufacturers face considerably greater difficulties. Due to these difficulties, exchanging manufacturing data from related processes is crucial for the effectiveness of the modeling since it allows for the collection of more defective samples. Directly sharing raw data may cause the disclosure of manufacturing recipes, which can be easily utilized by competitors, resulting in substantial economic loss. In general, data-sharing offers a productive means to promote high-quality CZ process modeling, but it can also lower costs for data collection, storage, registration, and annotation for several manufacturing tasks. Most importantly, data-sharing facilitates collaborative partnerships, as participants can synergistically strengthen their collaboration, if being able to identify participants with similar underlying manufacturing processes and therefore share valuable data with each other. Driven by these practical challenges, we apply the proposed PriED for privacy-preserving data-sharing. We continue with a description of our data collection and experimental evaluation.

B. Dataset

Our case study involves data collected from three groups of furnaces that vary in terms of machine configurations, recipes, and product designs. We denote these three (3) groups as $G1$, $G2$, and $G3$. In this case study, we define twelve (12) data owners (D1-D12) as referring to manufacturers who produce particular ingots by their own furnaces. Multiple ingots are assigned to each furnace, and for each ingot, we collect multivariate time series data, obtained by various sensors in the CZ process. The defective samples are labeled by the rework actions recorded in the dataset, where, as mentioned, rework refers to redoing a segment of the process.

Table I represents each data owner (D1-D12) with their associated furnace group, the sample size of each class, and the imbalance ratio. From this table, it can be observed that for each group $G1$, $G2$, and $G3$, we have 5, 4, and 3 associated data owners, respectively. Each data owner's multivariate time

TABLE I: Data distribution per data owner

Data Owner ID	Furnace Group	Sample Size		Imbalance Ratio
		Normal	Abnormal	
D1	G1	2,082	641	3.25
D2	G1	1,509	1,183	1.28
D3	G1	1,507	1,425	1.06
D4	G1	1,354	1,542	0.88
D5	G1	2,330	815	2.86
D6	G2	3,089	859	3.60
D7	G2	3,982	1,388	2.87
D8	G2	3,575	601	5.95
D9	G2	4,837	512	9.45
D10	G3	1,438	1,435	1.00
D11	G3	1,890	1,017	1.86
D12	G3	1,448	1,491	0.97

series data has a tensor structure (m, l, d), where m refers to the sample size, l refers to the time length of each sample, and d refers to the multi-variate dimension, *i.e.*, number of process variables. Some examples of these process variables collected over time are pressure of the main chamber, main heater current, main heater power, pulling speed, thermal field temperature, diameter, main thermal field resistance, *etc.* In this study, the time length l is considered to be 20 minutes for all data owners with the data collection frequency of 1 minute.

The data imbalance for normal and abnormal events can be observed from the Table I sample size column. Based on the data distributions presented, it can be observed that data owner D9 has the most imbalance data with 9.45 imbalance ratio for the minority class. D6 and D8 also contain data with high imbalance. In addition, each data owner from furnace type $G1$ has 12 process variables and each data owner from the *G2* and *G3* groups has 35 process variables. Therefore, data from different data owners may have varying dimensions. Notably, there exist several important common features among all the groups, which are related to the setting of the furnace and need to be kept private.

C. Experimental Setup

In all experiments, we consider a $80/10/10$ split ratio for training, validation, and testing. We first generate distilled data for each data owner based on the privacy-preserving deep generative models described in subsection III-B. We consider the main heater power (*i.e.*, power supplied to the furnace to change the temperature gradient in the furnace) and thermal field SP value (*i.e.*, temperature set points measured by a thermocouple near the heater) as the sensitive target features which are desired to be kept private in the distilled data. Based on Sun et al. [56], these two features are found to be the most important for the quality prediction downstream task. Moreover, to ensure that input data can be represented with a tensor structure, instead of zero padding, we perform weighted sampling (*i.e.*, considering class proportion) on data owners with smaller sample sizes.

TABLE II: Model capacity and training times

Model	# Trainable Parameters	Training Time (seconds/epoch)
GlbR	130,413	14.8s
AttR	127, 873	9.8s
CosD	145,982	11.5s
GlbD	145,982	11.4s
PriED	127,502	12.0s

D. Baselines and Evaluation Metrics

To evaluate the proposed PriED data-sharing framework, we perform comparisons to demonstrate: (1) the effectiveness of the privacy-preserving data distillation in the framework, and (2) the effectiveness of the data selection mechanism. Specifically, we compare model variants under a combination of *data sharing* and *data selection* settings. For data-sharing, we consider: (1) raw data: using raw data without any privacy-preserving method and (2) distilled data: using the proposed privacy-preserving data distillation. Furthermore, we vary the data selection methods as follows: (1) local model: models are trained on the local data from each data owner without data-sharing; (2) global model: a global model is trained on (raw or distilled) data from all participants concatenated together; and (3) cosine similarity: data from data owners are selected based on their corresponding cosine similarity with the data from the data receiver.

Therefore, we conduct experiments with five baselines in total: (1) **GlbR:** raw data with a global model, (2) **LocR:** raw data with local models, (3) **AttR:** raw data with the proposed attention-based data selection mechanism, (4) **CosR:** distilled data with the cosine similarity criterion, (5) **GlbD:** distilled data with a global model, and (6) **PriED** our proposed framework. Finally, since the downstream task is to predict the rework event during the CZ process, *i.e.*, imbalanced binary classification, we evaluate with recall, precision, and F1 score. We report macro-averaged evaluation metrics (recall, precision, F1 score) computed via 5-fold cross-validation.

E. Hyper-parameter Details

To ensure fair comparison as much as possible, hyperparameters are set such that model capacity (*i.e.*, the number of learnable parameters) is approximately the same across various baselines and the proposed PriED. An overview of model capacity and training times per epoch for each baseline can be found in Table II. We use the same LSTM-based model architecture as the local model of all data owners. In detail, the VAE-LSTM consists of a 2-layer LSTM encoder and a 2-layer LSTM decoder with 64 neurons per layer. We set $\lambda_1 = 1$, $\lambda_2 = 1$, and $d = 16$ as the dimension for the learned VAE latent presentations $\boldsymbol{z}_i \in \mathbb{R}^d$, $\forall i \in \mathcal{M}$. For the data selection, PriED consists of a bi-linear attention layer and a fully-connected layer with ReLU activation functions, both with 16 neurons per layer. LocR (local model operating on raw data) is an LSTM architecture that consists of 2 layers with 16 neurons per layer. The final data representations for the local models are based on an element-wise max-pooling

TABLE III: Results for all data receivers by groups. Mean and standard deviation reported over 5 experimental trials. Best performance is highlighted in **blue** for models operating on raw data and **green** for models trained on distilled data. Overall best performance across all baselines is highlighted with grey cell background. Green arrows indicate relative gains over the next best method operating on distilled data. Best average performance (column-wise, AVG column) in **bold**.

Metric	Method			Data Group ID			AVG
	Name	Privacy-Preserving Method	Data Selection Method	G1	G2	G3	
Precision	GlbR	raw data	global model	0.60±0.02	0.38±0.01	0.78±0.02	0.57±0.01
	LocR	raw data	local model	0.82±0.06	0.00±0.00	0.76±0.03	0.53±0.04
	AttR	raw data	attention	**0.91±0.02**	**0.45±0.16**	**0.88±0.01**	0.75±0.07
	CosD	distilled data	cosine similarity	**0.91±0.14**$^{\uparrow 0.04}$	0.22±0.12	0.83±0.03	0.66±0.10
	GlbD	distilled data	global model	0.76±0.06	0.41±0.16	0.73±0.06	0.64±0.08
	PriED	distilled data	attention	0.87±0.04	**0.71±0.11**$^{\uparrow 0.30}$	**0.86±0.02**$^{\uparrow 0.03}$	**0.81±0.06**
Recall	GlbR	raw data	global model	0.39±0.02	0.17±0.02	0.56±0.02	0.36±0.01
	LocR	raw data	local model	0.53±0.04	0.00±0.00	0.72±0.04	0.40±0.03
	AttR	raw data	proposed	**0.92±0.02**	**0.34±0.13**	**0.86±0.02**	0.71±0.06
	CosD	distilled data	cosine similarity	0.56±0.03	0.01±0.01	0.64±0.03	0.39±0.02
	GlbD	distilled data	global model	0.43±0.04	0.03±0.02	0.20±0.06	0.24±0.04
	PriED	distilled data	attention	**0.84±0.02**$^{\uparrow 0.28}$	**0.56±0.10**$^{\uparrow 0.53}$	**0.86±0.01**$^{\uparrow 0.22}$	**0.75±0.05**
F1 score	GlbR	raw data	global model	0.46±0.02	0.24±0.02	0.65±0.02	0.43±0.02
	LocR	raw data	local model	0.56±0.05	0.00±0.00	0.74±0.03	0.42±0.03
	AttR	raw data	proposed	**0.91±0.02**	**0.37±0.11**	**0.87±0.01**	0.72±0.05
	CosD	distilled data	cosine similarity	0.63±0.03	0.01±0.01	0.71±0.02	0.45±0.02
	GlbD	distilled data	global model	0.55±0.04	0.07±0.03	0.31±0.07	0.32±0.04
	PriED	distilled data	attention	**0.86±0.03**$^{\uparrow 0.23}$	**0.62±0.10**$^{\uparrow 0.55}$	**0.86±0.01**$^{\uparrow 0.15}$	**0.78±0.05**

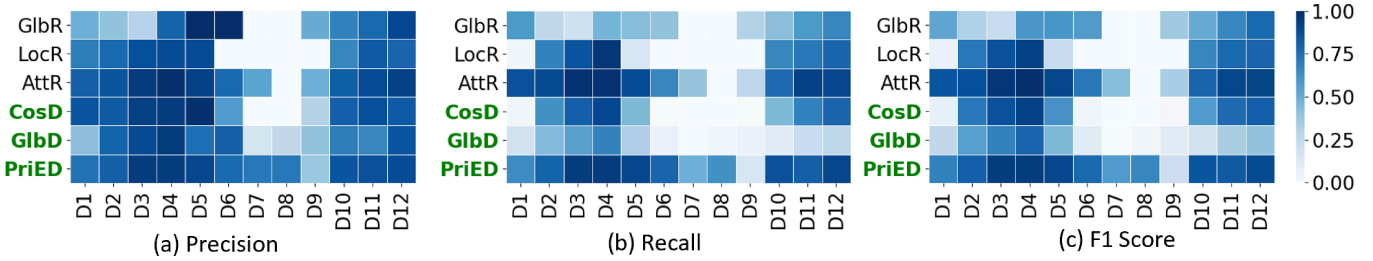

Fig. 4: Heatmaps of the (a) precision, (b) recall, and (c) F1 scores for all data receivers. The top-3 rows correspond to baselines operating on raw data (GlbR, LocR, AttR), and the bottom-3 rows to baselines operating on distilled data (CosD, GlbD, PriED).

of all hidden states. GlbR (global model operating on raw data) consists of 3 fully-connected layer with {256, 128, 32} neurons and ReLU activations, accompanied by a sigmoid output layer. To apply the proposed data selection mechanism to raw data with different feature dimensions, MLP projection heads are utilized to project data to the same dimensionality. GlbD (global model operating on distilled data) employs an LSTM network with 128 neurons with a sigmoid output layer. We train all models with Adam optimizer, 10^{-3} learning rate, and batch size of 64.

F. Experimental Results

In Table III, we summarize results for data receivers by different groups. We present mean and standard deviation over 5 experimental trials. To showcase the effectiveness of the proposed PriED, we compare with global and local models operating on either raw or distilled data. Note, however, that sharing raw data is an infeasible option in practice due to the need to preserve private information. These models are adopted as references to anchor the potentially best model performance. Overall, PriED yields the best average precision, recall, and F1 score across all groups (Table III, last column).

By investigating the performance of the attention-based data selection method, we find that AttR achieves the best performance with raw data (*i.e.*,, GlbR, LocR, and AttR) while PriED achieves the best performance with distilled data (*i.e.*,, CosD, GlbD, and PriED). These results validate that the data selected by the task-driven data selection method are more informative and contribute to the downstream task of the target data receiver. Moreover, the comparison between AttR and GlbR shows that most data receivers achieve better performance on the downstream task by sharing informative data from similar manufacturing processes, rather than utilizing data from all participants without performing data selection.

By comparing AttR and PriED, we find that using the distilled data may not achieve the best performance compared to using raw data (*e.g.*, for groups $G1$ and $G3$). However, the proposed method is closer to the best-performing raw data method, as compared to other baselines. Moreover, results clearly show that distilled data can produce superior performance for $G2$, where the most severe class imbalance is observed (Table I). This illustrates that the proposed VAE-LSTM learns informative data representations that improve the

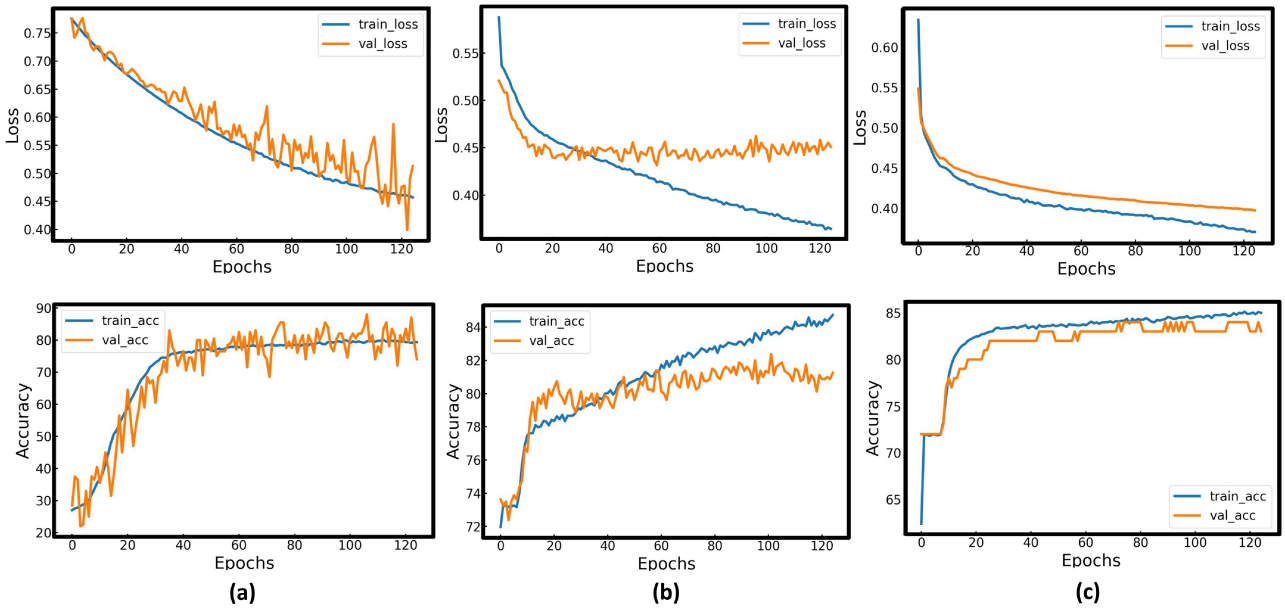

Fig. 5: An example of accuracy and loss improvements over training epochs for data owner D2. (a) Local model trained only on the D2 data, (b) PriED trained with cross-entropy, and (c) PriED trained with the policy gradient (RL).

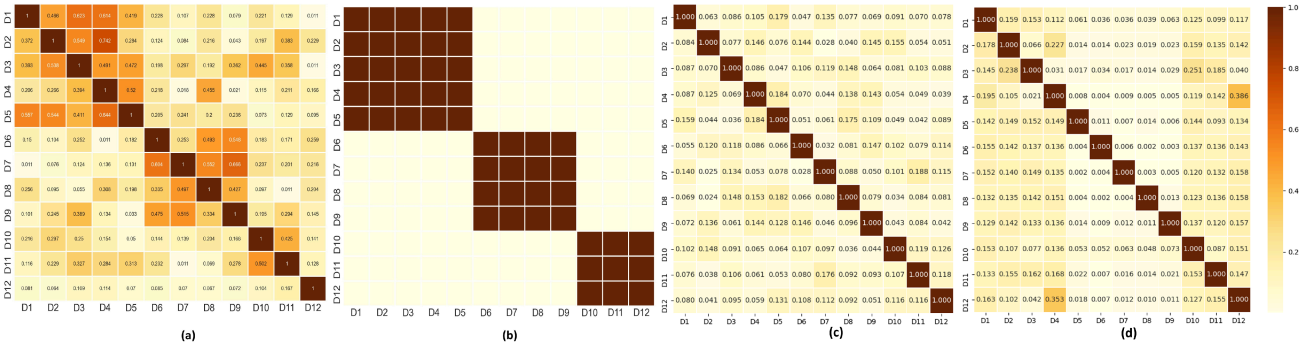

Fig. 6: (a) Learned attention weights between different data owners by PriED, (b) Intrinsic similarity of data owners based on the ingot furnace type, (c) Cosine similarity between different data owners, and (d) Learned attention weights between different data owners by raw data.

modeling performance on imbalance classification tasks.

In Figure 4, we visualize the precision, recall, and F1 score for all data receivers, respectively. To demonstrate the possible best performance (*i.e.*, upper bound), we present results for baselines employing raw data. As expected, the patterns of AttR and PriED are similar in the presented heatmaps. Although the darker color indicates AttR achieves better performance for most data receivers, PriED outperforms AttR on D7 and D8, which illustrates its advantage in data with high data imbalance ratios. As far as models trained on distilled data (bottom-3 rows), the proposed PriED demonstrates a dominant advantage over other baselines by achieving better performance across all data owners.

G. Ablation Study on RL Loss

Figure 5 illustrates the accuracy and loss improvements during training (across epochs) for a randomly selected data owner (D2). In general, we observe much smoother learning trends on the validation data for PriED trained with an RL loss, as shown in Figure 5(c), when compared to PriED trained with cross-entropy (shown in Figure 5(b)) and a local model with no data-sharing (shown in Figure 5(a)). We observe similar benefits across all data owners. Therefore, PriED with RL can reach comparable performance with a lower number of epochs and more stable training (fewer loss oscillations), as compared to other optimization objectives.

H. Qualitative Analysis

Figure 6 presents the PriED attention weights (which encode similarity between different data owners), in comparison to the inherent similarities between all data owners shown in Figure 6(b), the cosine similarity between different data owners (*i.e.*, CosD in Figure 6(c)), and the learned attention-based similarity matrix on raw data in (*i.e.*, AttR in Figure 6(d)).

Figure 6 shows that the patterns learned from CosD and AttR (Fig. 6 (b),(c)) are less clear than PriED (Fig. 6 (a)). The patterns calculated by cosine similarity seem more random since cosine only measures the element-wise distance between data points from different owners. Figure 6(d) shows that the attention-based data selection mechanism learns the similarity between the D1-D4 and D10-D12. However, the other patterns are poorly learned, which is most likely due to the information lost during the projection of the raw data. This indicates the importance of privacy-preserving data distillation.

V. Conclusion

Data-sharing mechanisms can accelerate machine learning training and deployment and improve data-driven decision-making. Yet, privacy concerns significantly limit data-sharing operations in Industrial Internet domains. On the other hand, utilizing shared data with poor information utility for a specific downstream task may increase the computational burden or even hamper modeling performance. In this work, we propose PriED as a privacy-preserving data-sharing framework that enables effective data-sharing to improve the performance of supervised downstream tasks with privacy guarantees. Experimental results on a real semiconductor manufacturing case study demonstrate the effectiveness of the proposed method. PriED can be generally applied to other supervised learning tasks. In our future work, we hope to further investigate the effectiveness of different method configurations through comprehensive ablation studies on a variety of downstream tasks. Furthermore, we aim to extend the data-sharing framework to multiple downstream tasks.

VI. Acknowledgements

The authors acknowledge the funding support from the National Science Foundation (Grant No. 2208864).

References

[1] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[2] H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and T. Zhao, "Smart: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization," *arXiv preprint arXiv:1911.03437*, 2019.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

[4] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *IEEE International Conference on Computer Vision*, 2017.

[5] C. Dwork, "Differential privacy: A survey of results," in *International Conference on Theory and Applications of Models of Computation*. Springer, 2008.

[6] P. Martins, L. Sousa, and A. Mariano, "A survey on fully homomorphic encryption: An engineering perspective," *ACM Computing Surveys (CSUR)*, vol. 50, no. 6, pp. 1–33, 2017.

[7] J. Li, X. Kuang, S. Lin, X. Ma, and Y. Tang, "Privacy preservation for machine learning training and classification based on homomorphic encryption schemes," *Information Sciences*, vol. 526, pp. 166–179, 2020.

[8] B. Jiang, J. Li, G. Yue, and H. Song, "Differential privacy for industrial internet of things: opportunities, applications and challenges," *IEEE Internet of Things Journal*, 2021.

[9] K. Bonawitz, H. Eichner, W. Grieskamp, D. Huba, A. Ingerman, V. Ivanov, C. Kiddon, J. Konečnỳ, S. Mazzocchi, B. McMahan *et al.*, "Towards federated learning at scale: System design," *Machine Learning and Systems*, vol. 1, pp. 374–388, 2019.

[10] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[11] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, no. 3, pp. 50–60, 2020.

[12] L. Lyu, H. Yu, and Q. Yang, "Threats to federated learning: A survey," *arXiv preprint arXiv:2003.02133*, 2020.

[13] J. Sun, A. Li, B. Wang, H. Yang, H. Li, and Y. Chen, "Soteria: Provable defense against privacy leakage in federated learning from representation perspective," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 9311–9319.

[14] T. Lin, L. Kong, S. U. Stich, and M. Jaggi, "Ensemble distillation for robust model fusion in federated learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2351–2363, 2020.

[15] X. Gong, A. Sharma, S. Karanam, Z. Wu, T. Chen, D. Doermann, and A. Innanje, "Ensemble attention distillation for privacy-preserving federated learning," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 15 076–15 086.

[16] Y. Zhou, G. Pu, X. Ma, X. Li, and D. Wu, "Distilled one-shot federated learning," *arXiv preprint arXiv:2009.07999*, 2020.

[17] M. Liu, W. Buntine, and G. Haffari, "Learning how to actively learn: A deep imitation learning approach," in *Annual Meeting of the Association for Computational Linguistics*, 2018.

[18] T. Vu, M. Liu, D. Phung, and G. Haffari, "Learning how to active learn by dreaming," in *Annual Meeting of the Association for Computational Linguistics*, 2019.

[19] J. Kreutzer, D. V. Torres, and A. Sokolov, "Bandits don't follow rules: Balancing multi-facet machine translation with multi-armed bandits," in *EMNLP Findings*, 2021.

[20] L. Xie, K. Lin, S. Wang, F. Wang, and J. Zhou, "Differentially private generative adversarial network," *arXiv preprint arXiv:1802.06739*, 2018.

[21] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, 2014.

[22] L. Frigerio, A. S. d. Oliveira, L. Gomez, and P. Duverger, "Differentially private generative adversarial networks for time series, continuous, and discrete open data," in *IFIP International Conference on ICT Systems Security and Privacy Protection*. Springer, 2019.

[23] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in *ACM SIGSAC Conference on Computer and Communications Security*, 2016.

[24] A. Triastcyn and B. Faltings, "Generating artificial data for private deep learning," *arXiv preprint arXiv:1803.03148*, 2018.

[25] R. Torkzadehmahani, P. Kairouz, and B. Paten, "Dp-cgan: Differentially private synthetic data and label generation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2019.

[26] N. Papernot, M. Abadi, U. Erlingsson, I. Goodfellow,

and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data," *arXiv preprint arXiv:1610.05755*, 2016.
[27] J. Jordon, J. Yoon, and M. Van Der Schaar, "Pate-gan: Generating synthetic data with differential privacy guarantees," in *International Conference on Learning Representations (ICLR)*, 2018.
[28] Y. Long, S. Lin, Z. Yang, C. A. Gunter, H. Liu, and B. Li, "Scalable differentially private data generation via private aggregation of teacher ensembles," 2020.
[29] W. Du, A. Li, P. Zhou, Z. Xu, X. Wang, H. Jiang, and D. Wu, "Approximate to be great: Communication efficient and privacy-preserving large-scale distributed deep learning in internet of things," *IEEE Internet of Things Journal*, vol. 7, no. 12, pp. 11 678–11 692, 2020.
[30] Y. Li, T. Baldwin, and T. Cohn, "Towards robust and privacy-preserving text representations," *arXiv preprint arXiv:1805.06093*, 2018.
[31] D. D. Lewis and J. Catlett, "Heterogeneous uncertainty sampling for supervised learning," in *International Conference on Machine Learning (ICML)*, 1994.
[32] P. Donmez, J. Carbonell, and P. Bennett, "Dual strategy active learning," in *European Conference on Machine Learning (ICML)*. Springer, 2007.
[33] H. T. Nguyen and A. Smeulders, "Active learning using pre-clustering," in *International Conference on Machine Learning (ICML)*. ACM, 2004.
[34] K. Brinker, "Incorporating diversity in active learning with support vector machines," in *International Conference on Machine Learning*, 2003.
[35] S.-J. Huang, R. Jin, and Z.-H. Zhou, "Active learning by querying informative and representative examples," in *Advances in Neural Information Processing Systems*, 2010.
[36] J. Qin, C. Wang, Q. Zou, Y. Sun, and B. Chen, "Active learning with extreme learning machine for online imbalanced multiclass classification," *Knowledge-Based Systems*, vol. 231, p. 107385, 2021.
[37] Y. Gal, R. Islam, and Z. Ghahramani, "Deep Bayesian Active Learning with Image Data," in *International Conference on Machine Learning*, 2017.
[38] I. Lourentzou, D. Gruhl, and S. Welch, "Exploring the efficiency of batch active learning for human-in-the-loop relation extraction," in *Companion Proceedings of The Web Conference 2018*, 2018, pp. 1131–1138.
[39] V. Prabhu, A. Chandrasekaran, K. Saenko, and J. Hoffman, "Active domain adaptation via clustering uncertainty-weighted embeddings," in *IEEE/CVF International Conference on Computer Vision (CVPR)*, 2021.
[40] M. Liu, Y. Song, H. Zou, and T. Zhang, "Reinforced training data selection for domain adaptation," in *Association for Computational Linguistics*, 2019.
[41] S. Ruder and B. Plank, "Learning to select data for transfer learning with bayesian optimization," *arXiv preprint arXiv:1707.05246*, 2017.
[42] S. Ruder, P. Ghaffari, and J. G. Breslin, "Data selection strategies for multi-domain sentiment analysis," *arXiv preprint arXiv:1702.02426*, 2017.
[43] Y. Zeng, P. Shojaee, S. H. Akhter Faruqui, A. Alaeddini, and R. Jin, "Contextual bandit guided data farming for deep neural networks in manufacturing industrial internet," in *2022 IEEE 5th International Conference on Industrial Cyber-Physical Systems (ICPS)*, 2022, pp. 1–6.
[44] W.-N. Hsu and H.-T. Lin, "Active learning by learning," in *AAAI Conference on Artificial Intelligence*, 2015.
[45] I. Lourentzou, A. Alba, A. Coden, A. L. Gentile, D. Gruhl, and S. Welch, "Mining relations from unstructured content," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2018, pp. 363–375.
[46] P. Shojaee, Y. Zeng, X. Chen, R. Jin, X. Deng, and C. Zhang, "Deep neural network pipelines for multivariate time series classification in smart manufacturing," in *2021 4th IEEE International Conference on Industrial Cyber-Physical Systems (ICPS)*, 2021, pp. 98–103.
[47] J. Astill, R. A. Dara, E. D. Fraser, B. Roberts, and S. Sharif, "Smart poultry management: Smart sensors, big data, and the internet of things," *Computers and Electronics in Agriculture*, vol. 170, p. 105291, 2020.
[48] F. Tao, Q. Qi, L. Wang, and A. Nee, "Digital twins and cyber–physical systems toward smart manufacturing and industry 4.0: Correlation and comparison," *Engineering*, vol. 5, no. 4, pp. 653–661, 2019.
[49] X. Yao, J. Zhou, Y. Lin, Y. Li, H. Yu, and Y. Liu, "Smart manufacturing based on cyber-physical systems and beyond," *Journal of Intelligent Manufacturing*, vol. 30, no. 8, pp. 2805–2817, 2019.
[50] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International Conference on Machine Learning (ICML)*, 2014.
[51] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.
[52] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *International Conference on Machine Learning (ICML)*, 2014.
[53] C. P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, and A. Lerchner, "Understanding disentangling in β-vae," *NeurIPS Workshop on Learning Disentangled Representations*, 2017.
[54] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," 2015.
[55] E. Lin, Q. Chen, and X. Qi, "Deep reinforcement learning for imbalanced classification," *Applied Intelligence*, vol. 50, no. 8, pp. 2488–2502, 2020.
[56] 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.
[57] G. Dhanaraj, K. Byrappa, V. Prasad, and M. Dudley, "Springer handbook of crystal growth," 2010.
[58] G. Fisher, M. R. Seacrist, and R. W. Standley, "Silicon crystal growth and wafer technologies," *Proceedings of the IEEE*, vol. 100, no. Special Centennial Issue, pp. 1454–1474, 2012.
[59] X. Chen, Y. Zeng, S. Kang, and R. Jin, "Inn: An interpretable neural network for ai incubation in manufacturing," *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2022.