

CADRE: Customizable Assurance of Data Readiness in Privacy-Preserving Federated Learning

Kaveen Hiniduma^{*†}, Zilinghan Li[†], Aditya Sinha^{†‡}, Ravi Madduri[†], Suren Byna^{*}

^{*}The Ohio State University [†]Argonne National Laboratory [‡]University of Illinois Urbana-Champaign
{hiniduma.1, byna.1}@osu.edu, {zilinghan.li, madduri}@anl.gov, aditya47@illinois.edu

Abstract—Privacy-Preserving Federated Learning (PPFL) is a decentralized machine learning approach where multiple clients train a model collaboratively. PPFL preserves the privacy and security of a client’s data without exchanging it. However, ensuring that data at each client is of high quality and ready for federated learning (FL) is a challenge due to restricted data access. In this paper, we introduce CADRE (Customizable Assurance of Data REadiness) for federated learning (FL), a novel framework that allows users to define custom data readiness (DR) metrics, rules, and remedies tailored to specific FL tasks. CADRE generates comprehensive DR reports based on the user-defined metrics, rules, and remedies to ensure datasets are prepared for FL while preserving privacy. We demonstrate a practical application of CADRE by integrating it into an existing PPFL framework. We conducted experiments across six datasets and addressed seven different DR issues. The results illustrate the versatility and effectiveness of CADRE in ensuring DR across various dimensions, including data quality, privacy, and fairness. This approach enhances the performance and reliability of FL models as well as utilizes valuable resources.

Index Terms—Data readiness for AI, Data quality assessment, Federated learning.

I. INTRODUCTION

Federated Learning (FL) [1], [2] allows multiple decentralized participants to train a model collaboratively without sharing their raw data. Rather than centralizing data, FL allows each participant to locally train a model on their data and transmit only the model updates to a central server. This method enhances privacy and security by keeping sensitive data locally. However, new challenges emerge when privacy-preserving techniques are applied in FL. A recent study on Privacy-Preserving Federated Learning (PPFL) led by NIST [3] highlights significant challenges, primarily due to the lack of access to training data. Data cleaning and feature selection are complicated because data scientists cannot view data across different clients. This may lead to inconsistencies and deployment failures. Many studies [4]–[6] have demonstrated that low-quality data directly impacts the model by lowering the performance and robustness. Additionally, PPFL’s privacy protections make it difficult to detect poor-quality or maliciously crafted data, which may lead to degrading the final model’s quality. While recent research is beginning to address these issues with techniques like secure input validation and adaptations of data poisoning defenses [7], [8], these solutions are not yet widely implemented in practical PPFL libraries.

In our efforts to address these challenges, we introduce Data Readiness for AI (DRAI) into the PPFL domain. Our recent survey [9] presented a comprehensive taxonomy for assessing

DRAI, focusing on data quality, organization, fairness, understandability, governance, and value. We developed AIDRIN (AI Data Readiness Inspector) [10] framework to evaluate the DRAI of datasets across these dimensions. However, AIDRIN was initially designed for centralized AI training, where data is uploaded to a standalone platform for evaluation. In contrast, PPFL requires decentralized data readiness (DR) assessment, including methods that preserve privacy and security.

A framework for supporting user-defined metrics, rules, and remedies to allow data stewards and FL administrators to define custom metrics and evaluation criteria while preserving privacy is needed. However, such a framework targeting either PPFL or centralized data readiness assessment is still unavailable. For example, in healthcare, PPFL can be used to develop a model for diagnosing a specific disease using MRI scans from multiple hospitals [11]. However, challenges such as data heterogeneity, quality, and privacy concerns arise because hospitals often use different MRI machines that leads to variations in image quality, resolution, and file format due to differences in hardware, software, and imaging protocols. In addition, some datasets may contain noisy or incomplete images, caused not only by machine differences but also by factors such as scanning artifacts, acquisition errors, or data corruption. To address these challenges, data owners should define custom DR standards, metrics, rules, and remedies tailored to their FL tasks. Data owners can establish standards for what constitutes an “AI-ready” MRI scan, such as data format and resolution requirements, and implement metrics to evaluate quality. Rules can be set to automatically flag images that do not meet these standards, and remedies, such as pre-processing techniques, can be applied to improve quality. It is required that each hospital ensure independently that its data meets the standards before participating in the FL process.

To meet these challenging requirements in preparing and ensuring DR in PPFL, we propose a novel framework, called CADRE (Customizable Assurance of Data REadiness). This framework allows FL “administrators” to define custom data readiness (DR) standards, including metrics, rules, and remedies tailored to specific FL tasks. Here, *administrators* refers to the individuals or stakeholders responsible for a given FL task who collaborate to establish these data readiness standards. CADRE allows clients to locally execute these custom functions to ensure their data meets the necessary standards at run time without compromising privacy. Clients can verify compliance with these rules and apply remedies to

their data as necessary. The results of these metric evaluations are compiled into a DR report for administrators' inspection. The report includes evaluations based on the custom readiness standards, along with standard metrics and visualizations of client data statistics. This framework brings a human-in-the-loop approach to FL by involving administrators in the definition, validation, and refinement of DR standards. CADRE can include predefined techniques and rules to showcase its capabilities, but its primary functionality lies in its customizability. The framework enables administrators to define a wide range of DRAI evaluation metrics, rules, and remedies tailored to specific FL tasks. This flexibility makes CADRE adaptable and practical for diverse PPFL scenarios, which enhances its usefulness across applications.

DR evaluation ensures that only clients with qualified data participate in the FL system. The CADRE framework is designed to be generalizable, applicable to any FL task, and adaptable to various domains. To demonstrate its practical application, we have developed an extensible module for the APPFL (Advanced Privacy-Preserving Federated Learning) framework [12], [13], an open-source software framework that enables researchers and developers to implement, test, and validate various PPFL techniques. With this integration, we showcase usage of CADRE in existing PPFL workflows. The main contributions of this study are:

- We propose a novel framework that enables FL administrators within a PPFL system to define custom metrics, rules, and remedies. CADRE addresses the execution of these custom standards by automating the process and ensuring that clients can locally apply these actions to meet required data standards while preserving privacy.
- We generate comprehensive DR reports in CADRE that evaluate the metrics defined by FL administrators. This ensures privacy preservation by only including aggregated metric evaluations without exposing any raw data. Administrators can review these reports to assess whether clients have met the expected standards and gain insights into the data's characteristics.
- We integrate CADRE into APPFL, demonstrating compatibility with existing PPFL workflows.

We evaluated CADRE using six datasets with a variety of data modalities (e.g., 2D images, tabular data, 3D volumetric data) and downstream tasks (such as classification, segmentation, and survival analysis). In some cases, we polluted the datasets to add noise, class imbalance, duplicate records, high memory consumption, bias, outliers, and insufficient anonymity. CADRE allows administrators within a PPFL system to define custom metrics, rules, and remedies, showing that the issues caused by our pollution were effectively addressed. We also demonstrate CADRE's impact further using an example where resolving DR challenges leads to improvements in model performance. In the remainder of the paper, we describe related work (§II), CADRE design (§III), its integration into APPFL (§IV), and its evaluation (§V).

II. RELATED WORK

A few frameworks evaluate data with a focus on aspects such as data quality, governance, and infrastructure. Existing frameworks [14]–[17] primarily assess data availability, volume, quality, governance, and ethics. A wide range of data cleansing tools [17]–[19] are available today, each offering unique features to ensure the accuracy, reliability, and trustworthiness of data. However, most users prefer manual cleaning of the data and decide on AI readiness themselves or skip these tools entirely.

Despite their strengths, these frameworks exhibit critical gaps when applied to modern, distributed AI environments. They lack integration with FL architectures. Existing frameworks generally assume a centralized data environment. They also fall short in addressing compliance challenges related to cross-border data flows, which are common in FL scenarios.

Ensuring the integrity of model updates is critical in FL, as malicious clients can degrade the quality of the global model. FLTrust [8] addresses this by establishing a root of trust using a clean dataset to assign trust scores to client updates. However, its reliance on a single trusted dataset introduces a vulnerability if that dataset is compromised. EIFFeL [7] enhances integrity while preserving privacy through secure aggregation and verification of client updates. It effectively filters out malicious contributions. However, it does not address the challenge of data heterogeneity, which can affect convergence and overall model performance.

The performance of FL models is often affected due to heterogeneous and noisy data distributions. In FL, where data is distributed between multiple clients, label noise refers to incorrect or inconsistent labels in the training data, which can significantly reduce model performance. To address this issue, FedELC [20] proposes to identify clients with noisy labels and apply label correction strategies to refine the labels. However, it ignores other critical aspects of DR. FedDQA [21] introduces a metric to evaluate client data quality, allowing the selection of higher-quality clients for training. Although effective in minimizing the influence of noisy data, this approach risks introducing selection bias and does not actively improve the underlying data. In the domain of PPFL, methods such as lazy influence approximation [22] and FedDQC [23] offer quality assessments that preserve privacy using influence scores and relevance alignment, respectively. Although these approaches maintain confidentiality, they have computational overhead and suffer from reduced data resolution under strict privacy constraints.

Another key limitation of existing FL frameworks is their lack of flexibility in supporting custom DR metrics and remediation workflows. Most rely on static, predefined evaluation criteria, making it difficult to accommodate domain-specific requirements. Remediation processes are often rigid and lack support for custom operations such as federated anonymization or edge-device preprocessing. Even unified data platforms rarely allow integration of custom rules or remedies. To address these gaps, our proposed framework enables FL administrators to define customized metrics, rules,

and remedies aligned with the needs of specific FL systems. This flexibility helps manage data heterogeneity by enforcing consistent standards across clients, all while preserving privacy. The framework integrates seamlessly with existing PPFL workflows and supports DR evaluation before initiating resource-intensive training. Moreover, it aligns with the vision of Industry 5.0 [24], emphasizing human-centric, privacy-aware, and adaptable AI systems that empower administrators to take control of DR.

III. DESIGN OVERVIEW

The objective of CADRE is to allow administrators of FL systems to define and utilize both foundational and customizable actions. To support this, CADRE provides the following main components: *metrics*, *DR reports*, *rules*, and *remedies*. In Figure 1, we show an outline of CADRE with its components, including metrics that are standard and custom (i.e., administrator-defined). Rules and remedies are also administrator-defined functions.

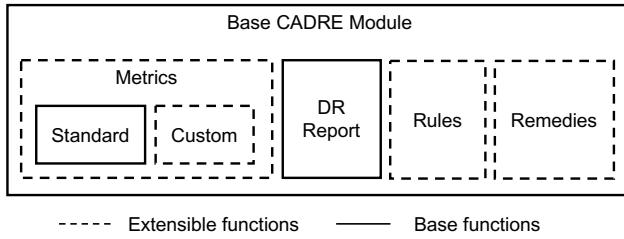


Fig. 1: An overview of CADRE framework for FL tasks. Metrics include commonly known standard DR evaluation measurements. The extensible functions are used to define custom DR metrics, rules, and remedies. The DR report provides standard and custom metric evaluations with visualizations.

A. Metrics Component

We divided the metrics component of CADRE into two main parts: standard metrics and custom metrics. The standard metrics include a set of DR metrics defined in AIDRIN [10] including quality metrics such as evaluating sample sizes, data sparsity, and statistical measures like mean, median, and standard deviation of the client's data distribution. These metrics serve as a baseline for assessing DR of clients' data across any FL task. Additionally, the standard metrics component contains basic visualizations, such as bar charts and scatter plots, which are included in the DR reports to provide a visual representation of the client data characteristics.

The custom metrics component is an extensible capability that allows FL administrators to provide custom metrics tailored to their unique FL task and evaluation needs. This flexibility ensures that administrators can assess client data according to the specific requirements of their projects. For example, if a task requires assessing the completeness or skewness of the data, administrators can define these metrics within CADRE. These standard and custom metric evaluations and visualizations allow administrators to quickly grasp the readiness of clients' data and identify potential issues that may lead to unexpected behavior in downstream FL tasks.

B. Rules and Remedies Components

Besides metrics, CADRE also includes sub-modules for defining rules and remedies. These sub-modules allow administrators to establish custom rules that their custom metric must meet to be considered ready for the next stages of the FL pipeline. The administrators can also define custom remedies to improve the readiness of the data to meet the specified rules. For instance, if administrators need to assess noise levels in the data, they can use metrics such as the standard deviation of the data distribution to quantify noise. A high standard deviation may indicate excessive variability and suggest the presence of noise. The administrators can then establish a rule where the standard deviation must not exceed a predefined threshold. If this threshold is surpassed, remedies could be implemented, such as filtering out extreme values or including only a subset of the affected client's data in the analysis.

C. DR Reporting Module

CADRE generates detailed DR reports by aggregating metric evaluations and visualizations produced by individual clients. It also includes principal component analysis (PCA) [25] graphs, to illustrate the combined data distribution and heterogeneity among clients. These insights are compiled into an easily readable HTML report, allowing administrators to assess whether clients meet specified standards while ensuring data privacy. This feature is essential to maintain transparency and accountability throughout the DR process.

For instance, for a given FL task, a custom metric could involve measuring class imbalance within each client's dataset in the FL system. Identifying class imbalance is important because it can bias the learning process, especially in classification tasks where underrepresented classes may be poorly learned [26]. In this scenario, a rule would be to flag any client datasets where the class distribution significantly deviates from a defined threshold of balance. If a client is flagged, the remedy might involve data augmentation or re-sampling techniques to mitigate the imbalance until the metric indicates an acceptable distribution. The resulting report will display the class distribution statistics for each client's dataset, making it easy to identify and address any flagged issues. In Figure 2, we present a DR report generated for this specific example. The report includes evaluations of both standard and custom metrics, visualizations for each of the two clients involved in the experiment, and combined plots. The visualizations include standard plots such as class distribution and data distribution charts, while the combined plot is a PCA visualization of a sample of the data from the clients. For this example, we used the Adult Income dataset [27], and CADRE is integrated into the APPFL framework. More details about this integration and the experiments can be found in sections IV and V.

Clients participating in the FL framework use custom metrics within CADRE to locally evaluate their data and generate DR reports. If the client data meets the specified rules, the data will proceed to the subsequent stages of the FL pipeline. Conversely, if the data does not meet the rules, remedies defined by the administrators within CADRE will be applied

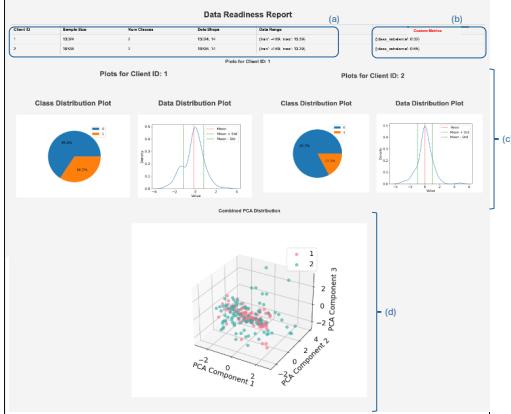


Fig. 2: The figure illustrates an example DR report from an FL experiment featuring: (a) Standard metrics, (b) Custom metrics in CADRE for this specific FL task, (c) Individual client plots, and (d) Combined data plots.

to improve the DR. This process will iterate until the data complies with the established rules. This will ensure DR for the next stages of the FL pipeline. Figure 3 provides a visual representation of this iterative approach by illustrating how clients use CADRE’s functions to assess DR, apply custom rules, and implement remedies while preserving privacy.

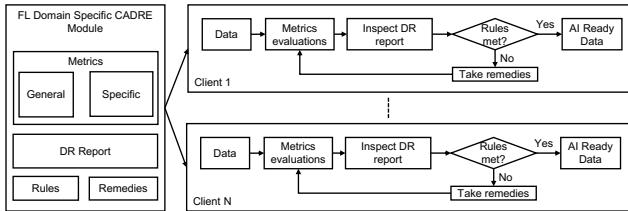


Fig. 3: An iterative data evaluation and remediation process where clients are involved in an FL framework. It outlines how clients use the CADRE’s functions to assess DR, apply custom rules, and implement remedies as needed.

By integrating metrics, rules, and remedies components into DR frameworks, CADRE provides a comprehensive and flexible framework for ensuring DR in FL systems. This framework allows administrators to tailor the DR process to the specific needs of their projects while maintaining high standards of DR and privacy.

IV. INTEGRATION INTO EXISTING PPFL FRAMEWORKS

In this study, we utilize the APPFL framework to demonstrate the practical application of CADRE. APPFL is an open-source framework designed to enhance privacy and security in FL systems. It allows researchers to implement, test, and deploy FL experiments across distributed clients while ensuring data privacy. We chose APPFL as the testbed for CADRE due to its modular and extensible architecture, which aligns well with CADRE’s design principles and integration goals. Its built-in support for differential privacy, asynchronous and synchronous training algorithms, and flexible customization of core FL components makes it a suitable and practical

platform for evaluating CADRE’s capabilities in real-world PPFL scenarios.

APPFL consists of six key components: an aggregator, scheduler, trainer, privacy module, communicator, and compressor. These components work together to tackle challenges such as computational disparities and security concerns in distributed machine learning, while also enabling enhanced privacy protection, supporting flexible model training on decentralized data, simulating various FL algorithms, implementing lossy compression for efficient data transfer, and providing a highly extensible framework for customizing aggregation algorithms, server scheduling strategies, and client local trainers. The framework supports various popular synchronous and asynchronous FL algorithms such as FedAvg [1], FedAvgM [28], FedBuff [29], and FedCompass [30], and incorporates differential privacy techniques [31].

CADRE will be integrated into the APPFL framework as an extensible module. Administrators can use the extensible nature of CADRE to define the metrics, rules, and remedies for a specific FL task. This allows clients to use its functions locally. This integration enables clients to evaluate data using custom metrics and apply custom remedies if the rules are not satisfied. After evaluating the data, the client agent will compile a DR report of the evaluations. These evaluations are then aggregated by the communicator within APPFL to combine the results from all clients for review. This integration demonstrates CADRE’s ease of use and versatility within existing PPFL frameworks. In Figure 4, we provide an overview of its implementation within the APPFL framework.

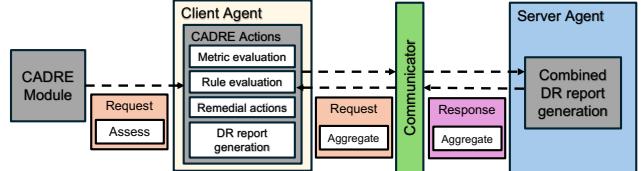


Fig. 4: An integration of CADRE in the APPFL framework.

Configuring CADRE for specific FL tasks is a straightforward process that allows administrators to tailor its extensible functionality to meet the unique requirements of each task. The process begins with the utilization of the base CADRE module. The base CADRE module serves as a foundational template with extensible functions. By using this template, administrators can create a specialized CADRE module that incorporates the necessary evaluation metrics, rules, and remedies specific to their task.

Once the custom CADRE module is configured, it is seamlessly integrated into the APPFL framework by uploading it. The framework is designed to accommodate such modular additions, making the integration process smooth and efficient. To activate the newly created CADRE module, administrators simply update the configuration file within the APPFL framework. This involves specifying the path to the custom CADRE module file that will allow the framework to recognize and utilize it appropriately. Additionally, administrators can pass other relevant arguments specific to the CADRE module by

defining them in the configuration file. For instance, a CADRE module may require additional inputs, such as feature indices and other identifiers, for various DR-related tasks. Figure 5 illustrates an example of this configuration, showcasing the YAML-based setup used to define a custom CADRE module.

```
cadre_configs:
  cadre_path: path/to/custom/cadre/module
  cadre_name: CustomCADREModule
  remedy_action: true
  cadre_kwargs:
    kwargs1: value1
    kwargs2: value2
```

Fig. 5: YAML configuration for customizing a CADRE module in FL tasks, allowing administrators to define evaluation metrics, rules, and remedies specific to their needs.

With the integration of CADRE into the APPFL framework, administrators gain significant advantages that aid in making informed decisions before entering the costly training phase. As data flows through the system, CADRE automatically executes defined actions, ensuring that DR issues are addressed promptly and consistently. This automation provides administrators with timely interventions, allowing them to focus on strategic decisions rather than manual data remediation tasks.

Additionally, the DR reports offer transparency and accountability. These reports provide administrators with a clear overview of the DR actions taken and allow effective assessment of DR compliance. By reviewing the detailed evaluations without exposing any raw data, while maintaining privacy and security, administrators can ensure that only clean and compliant data is used. Overall, this streamlined approach highlights how easily CADRE can be adapted for different FL tasks and data modalities. This concept will enhance the flexibility and effectiveness of PPFL. The documentation and code for this integration are available as part of APPFL [32].

Integration of CADRE into APPFL leads to improved model performance, as AI-ready data reduces the risk of errors and noise affecting the training process. CADRE also supports scalability by allowing the system to efficiently handle large datasets. This allows administrators to make better-informed decisions, optimizing resource allocation and minimizing risks before committing to the next phases in FL.

V. EVALUATIONS

To demonstrate the effectiveness of CADRE in evaluating data quality, privacy, and fairness, we use multiple datasets, experimental setups, and custom CADRE modules. Since most of the publicly available datasets that have been used to develop FL models are relatively clean and preprocessed, we used various data pollution techniques to evaluate with CADRE. We will illustrate how our custom DR standards are achieved by utilizing the tailored metrics, rules, and remedies within the custom CADRE modules. We will present an example illustrating the performance improvement of the final FL model on the downstream task when using CADRE, compared to without using it.

A. Datasets and Experimental Setup

In this study, we used six datasets spanning both standard benchmarks and those from real-world medical research. The benchmark datasets include MNIST [39], a collection of handwritten digit images widely used for image classification; CIFAR-10 [40], which comprises color images across ten classes for object recognition tasks; and Adult Income [27], a tabular dataset from the UCI repository used to predict whether an individual's income exceeds \$50K based on census data.

In addition to these, we used three datasets derived from real-world medical research. TCGA-BRCA from the Flamby collection [41] contains clinical data from breast cancer patients and is used for survival analysis. The IXI Tiny dataset, also from Flamby, consists of 3D brain MRI scans and serves as a benchmark for medical image segmentation tasks. Both of these datasets are naturally partitioned among clients, such as different hospitals or research centers, and are widely used in FL research. Finally, the AI-READY (Artificial Intelligence Ready and Equitable Atlas for Diabetes Insights) dataset [42] is a new comprehensive and ethically sourced collection designed to advance AI research in Type 2 Diabetes Mellitus (DM2), consisting over 15 data modalities, such as vitals, retinal imaging, electrocardiograms, and other health-related measurements, all aimed at exploring salutogenic pathways to health. For our research, we utilized color fundus photography (CFP) images from the AI-READY collection to classify the severity of diabetes by analyzing the retinal health using the CFP images. To simulate real-world heterogeneity, we divided the dataset among four clients based on the imaging devices used: iCare Eidon, Optomed Aurora, Topcon Maestro2, and Topcon Triton. By considering these datasets from various modalities and with different downstream tasks, we demonstrate the versatility of our proposed framework, which is not constrained by data modality or task.

To facilitate the evaluation of class imbalance, we transformed MNIST, CIFAR-10, and the AI-READY data into binary classification tasks. In MNIST, digits 0–4 were grouped into one class, while digits 5–9 formed another. In CIFAR-10, images with class indices 0–4 were assigned to one class, while images with class indices 5–9 were categorized as the other. For the AI-READY dataset, we categorized the classes as follows: the “pre-diabetes (lifestyle controlled)” and “oral medication and/or non-insulin injectable medication controlled” classes were combined into one group, while the “healthy” and “insulin-dependent” classes formed the other group. This transformation simplifies the evaluation process and improves understandability. The Adult Income dataset is inherently a binary classification task, so no further modifications were necessary.

As discussed in section IV, we employed APPFL to integrate CADRE and conduct the experiments. We consistently used FedAvg [1] as the primary FL algorithm across all experiments. Since MNIST, CIFAR-10, and Adult Income are not inherently FL datasets, we applied non-independent and identically distributed (non-IID) partitioning to ensure data heterogeneity. For these three datasets, we partitioned the data

TABLE I: Overview of custom CADRE modules used in experiments.

CADRE Module ID	Category	Metric	Rule	Remedy
1	Noise Management	Mean magnitude of the data (image intensities or feature values)	Applied remedy when the data distribution mean exceeded a threshold (e.g., > 0.37 for MNIST).	Data points with noisy indices were removed.
2	Class Imbalance Handling	Class imbalance degree [33]	Applied when imbalance degree > 0 .	SMOTE [34] was used to oversample the minority class.
3	Duplicate Management	Proportion of duplicates	Applied when duplicates proportion > 0 .	Duplicates were identified and removed.
4	Memory Optimization	Memory usage in megabytes (MB) to store the client's data	Applied when memory usage was excessively high.	Data types were optimized or duplicates removed depending on the dataset's pollution method.
5	Bias Handling	Statistical parity difference [35] for Adult Income dataset and representative rate difference for TCGA-BRCA dataset	Applied when metric value > 0 .	Stratified resampling [36] to balance sensitive groups and labels in the Adult Income dataset, while SMOTE to oversample the minority group in the TCGA-BRCA dataset.
6	Outlier Management	Proportion of outliers using Interquartile range (IQR) method [37]	Applied when outliers proportion > 0 .	Outliers were clipped at IQR bounds.
7	K-anonymity Handling	K-anonymity level [38]	Applied when anonymity level ≤ 1 .	Data records with low anonymity levels were suppressed to ensure the desired level of anonymity.

into 10 clients per experiment and ran the experiments for 10 global epochs. On the other hand, TCGA-BRCA and IXI Tiny datasets are genuine FL datasets, already partitioned into 6 and 3 clients, respectively. As previously mentioned, the AI-READI dataset was partitioned based on the imaging device used, resulting in four clients corresponding to the four devices. CADRE operates before the actual training phase, so FL training related configurations do not impact CADRE's execution. However, to ensure the completeness of our experiments and to validate integration in FL tasks, we reported these configurations. For the AI-READI dataset, we utilized a single node with 64GB RAM and one NVIDIA A40 GPU on the Delta supercomputer at NCSA [43]. The rest of the experiments were conducted on an Apple M2 Max MacBook Pro with 32GB unified memory.

B. Custom CADRE Modules

In this study, we used seven custom CADRE modules, each designed to address a specific DR issue. These modules incorporate tailored metrics, rules, and remedies to ensure that the client's data meets the expected standards. The selection of modules covers a broad spectrum of DR challenges, as identified in the [9] study, including data quality, fairness, privacy, and structure. Table I provides a detailed overview of these custom modules by outlining the metrics, rules, and remedies each module uses to evaluate and enhance the data's readiness for specific AI tasks.

As seen in Table I, for module 5, we measured statistical parity difference in the Adult Income dataset and representation rates in the TCGA-BRCA dataset. Statistical parity involves assessing class labels and sensitive groups, making it suitable for the Adult Income dataset, which deals with classification tasks. However, for the TCGA-BRCA dataset, which is used for survival analysis, measuring statistical parity is not feasible. Instead, we evaluate the representation rates of sensitive attributes and balance them as a remedy. For the Adult Income dataset, "gender" was selected as the sensitive feature for analysis by the module. This feature contains two categories: "male" and "female." In contrast, for the TCGA-BRCA dataset, "race_white" was identified as the sensitive feature, represented as a binary attribute where "1" indicates that the race is white, and "0" signifies otherwise.

Module 7 uses k-anonymity level as a metric. A remedy is applied when the anonymity level is less than or equal to 1 by ensuring that each entity remains identical from at least $k - 1$ others based on quasi-identifiers [38]. Quasi-identifiers are attributes that are not unique identifiers on their own but can be combined to identify individuals. For the Adult Income dataset, quasi-identifiers were "workclass," "race," and "gender." We selected these as the quasi-identifiers because they are commonly available in public records and, when combined, could increase re-identification risk. Similarly, for the TCGA-BRCA dataset, the quasi-identifiers included demographic and self-reported characteristics such as "age_at_index," "ethnicity_not hispanic or latino," "ethnicity_not reported," "race_asian," "race_black or african american," "race_not reported," and "race_white." These attributes were chosen due to their potential to link individuals across datasets and may pose privacy concerns if identified.

C. Data Pollution

To fully demonstrate the remedies provided by our custom CADRE modules, it was essential to ensure that the datasets used in our study exhibited the relevant issues. Some datasets naturally contained issues such as class imbalance, which was present in all classification tasks due to non-IID partitioning. Other issues were intentionally introduced through data pollution techniques. Table II provides detailed information on the pollution methods applied to each dataset. By polluting data, it enables the activation of rule and remedy actions in the custom CADRE modules in every experiment.

Figure 6 presents two DR report samples from an experiment conducted before and after meeting a CADRE module's standards. These reports illustrate how easily data-related issues can be identified and addressed, ensuring that standards defined by the custom CADRE modules are met. For this sample, we used the AI-READI dataset's before-and-after DR reports from the experiment conducted for CADRE module 1.

D. Results

After conducting experiments across all datasets and custom CADRE modules, as detailed in Tables I and II, we observed that nearly all client data met the required standards defined by each custom CADRE module. The process generated DR

TABLE II: Dataset-specific data pollution methods applied for each CADRE module.

CADRE Module ID	MINIST	CIFAR-10	Adult Income	Flamby TCGA-BRCA	Flamby IXI Tiny	AI-READY
1	Added Gaussian noise (std. dev. = 2) to 90% of the data	Added Gaussian noise (std. dev. = 2) to 90% of the data	Added Gaussian noise (std. dev. = 2) to 90% of the data	Added Gaussian noise (std. dev. = 2) to 90% of the data	Added Gaussian noise (std. dev. = 2) to 90% of the data	Added Gaussian noise (std. dev. = 2) to 90% of the data
2	Imbalanced class distribution due to non-IID partitioning	Imbalanced class distribution due to non-IID partitioning	Imbalanced class distribution due to non-IID partitioning	Not applicable (survival analysis task)	Not applicable (segmentation task)	Device-based partitioning inherently resulted in an imbalanced class distribution
3	20% of data was randomly duplicated	20% of data was randomly duplicated	20% of data was randomly duplicated	20% of data was randomly duplicated	20% of data was randomly duplicated	20% of data was randomly duplicated
4	Converted feature values to higher precision (float32 to float64)	Converted feature values to higher precision (float32 to float64)	Converted feature values to higher precision (float32 to float64)	Duplicates added to increase memory usage	Duplicates added to increase memory usage	Duplicates added to increase memory usage
5	Not applicable (image data has no sensitive features)	Not applicable (image data has no sensitive features)	Statistical parity differences were inherent	Representative rate differences were inherent	Not applicable (image data has no sensitive features)	Not applicable (image data has no sensitive features)
6	Added random gaussian noise (std. dev. = 2) to the data to simulate outliers	Added random gaussian noise (std. dev. = 2) to the data to simulate outliers	Added random gaussian noise (std. dev. = 2) to the data to simulate outliers	Features inherently contained outliers	Added random gaussian noise (std. dev. = 2) to the data to simulate outliers	Added random gaussian noise (std. dev. = 2) to the data to simulate outliers
7	Not applicable (no quasi-identifiers in image data)	Not applicable (no quasi-identifiers in image data)	Quasi-identifiers already contained low levels of anonymity	Quasi-identifiers already contained low levels of anonymity	Not applicable (no quasi-identifiers in image data)	Not applicable (no quasi-identifiers in image data)

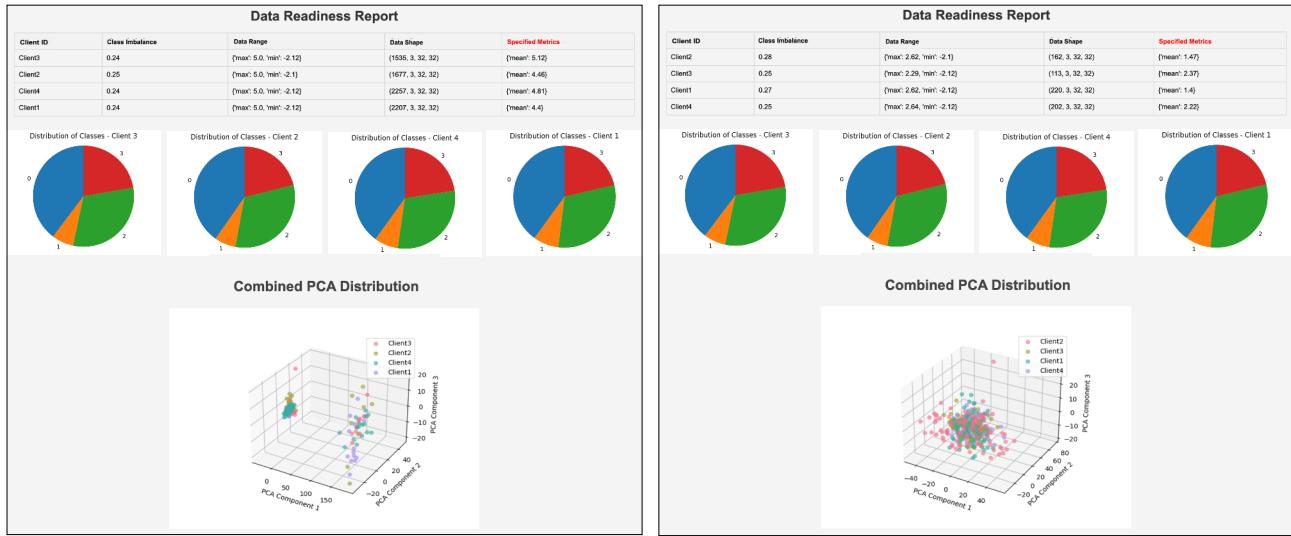


Fig. 6: Example DR reports generated before (left) and after (right) applying CADRE module 1 show an improvement in the average mean after removing noisy data. Results are shown in the table's rightmost column. The combined PCA plot at the bottom right confirms that noise-related anomalies in the data distribution have been resolved.

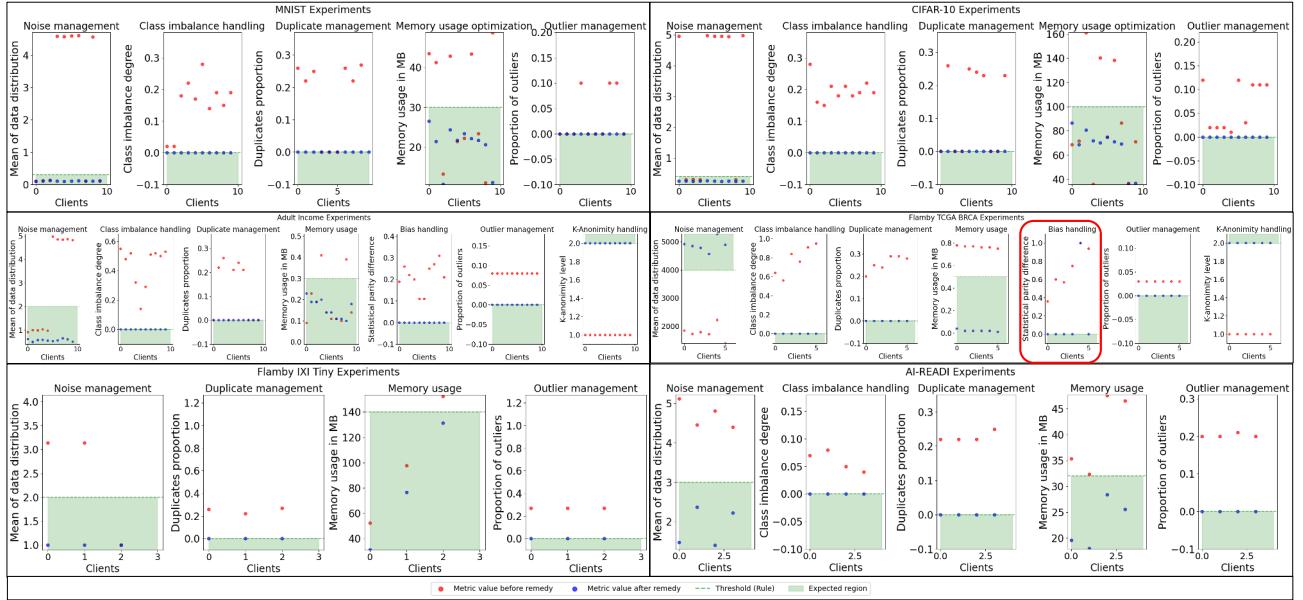


Fig. 7: Evaluation of custom metrics for each CADRE module, before and after remedy application. Threshold lines indicate predefined rule criteria. The red box highlights one case where a client's post-remedy metric remains above the threshold.

reports that reflected these metric evaluations, along with standard metrics and visualizations, as depicted in the Figure 6. Figure 7 illustrates the metric values before and after applying the remedies of custom CADRE modules, with threshold values indicating the rules set for each experiment. As shown in the figure, almost all post-remedy data points fall within the expected range. However, there is only one exception, observed in the figure that is boxed in red, where one client's post-remedy metric value remains above the threshold. The DR report's representative rates plot of the sensitive feature helped identify that this particular client contained only one ethnic group, preventing the remedy action from balancing the feature due to the absence of a second group. This example highlights the importance of DR reports in understanding the DR levels of clients before proceeding to the training phase.

Although this work focuses on the pre-training phase of the FL pipeline, it offers important insights into the quality and readiness of data before initiating costly training procedures. By evaluating and improving DR early on, administrators can make informed decisions about whether to proceed with training. This will ultimately help conserving computational and organizational resources.

To demonstrate the downstream impact of CADRE, we conducted an experiment using the IXI Tiny dataset from the Flamby benchmark suite. This dataset consists of 3D brain MRI scans and is commonly used for medical image segmentation tasks, where performance is typically measured using the Dice score [44]. Figure 8 presents the average Dice scores between clients during 10 rounds of FL training. The blue curve shows performance before applying the CADRE noise management module, while the green curve reflects results after CADRE was used to remove noisy indices. The shaded regions represent the standard deviation between clients to capture the variability in performance.

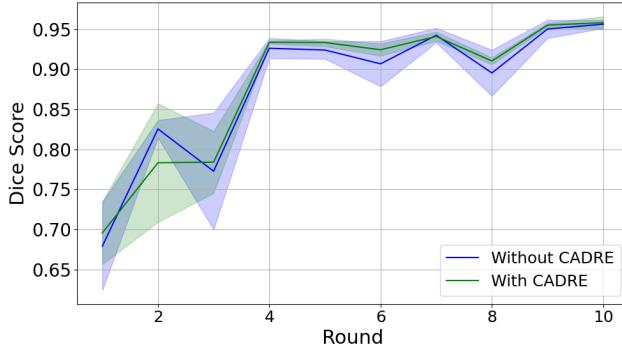


Fig. 8: The figure presents a comparative analysis of FL performance on the Flamby IXI Tiny dataset by evaluating the impact of CADRE's noise handling module through Dice scores over ten training rounds.

The results indicate that applying CADRE leads to consistently improved and more stable Dice scores, particularly from round 4 onward. This suggests that CADRE's early noise-handling improves DR and reduces inter-client variability. These factors are critical for achieving better generalization in

FL models. These findings align with prior research showing that high-quality, noise-free data improves model robustness and accuracy [4], [22], and that addressing other DR dimensions, such as class imbalance, can mitigate bias and reduce model drift [45], [46].

However, other factors, such as achieving perfect fairness and optimal anonymity levels, may affect different aspects of model performance. A dataset with minimal statistical parity can improve model fairness [47], though it may compromise overall performance and accuracy. Similarly, as we increase the privacy budget of the data, model accuracy tend to decrease [48]. However, administrators might choose to prioritize data fairness, and privacy standards over model performance. Also, memory usage optimization is crucial for FL clients, as resource-constrained edge devices have limited computational and memory capacity [49]. Efficient optimization helps maintain training efficiency while preventing performance degradation. Overall, these results demonstrate that our framework can be effectively integrated into PPFL systems to meet DR-related standards before training to conserve valuable resources and funds. Moreover, the informative DR reports simplify the process for administrators by providing a clear understanding of the data's condition for the FL task and setting expectations for the training phase.

VI. CONCLUSION AND FUTURE WORK

In this study, we introduced a novel framework to enhance DR in PPFL systems. The framework allows FL administrators to define CADRE modules tailored to address diverse DR challenges across various downstream tasks and data modalities. By specifying custom metrics, rules, and remedies, these modules allow clients to execute processes locally and to ensure that their data meets the necessary standards while preserving privacy. CADRE allows administrators to set realistic expectations for training, optimize resource utilization, and lay the groundwork for reliable and equitable FL results.

In our future work, we will expand CADRE's applicability to a broader range of usecases and explore automated methods to streamline the DR process. Additionally, we will investigate computationally intensive tasks and explore adding custom privacy-preserving modules to CADRE for user-controlled privacy protection. This will enhance the adaptability of CADRE to evolving privacy standards in PPFL frameworks.

ACKNOWLEDGMENT

This work is supported in part by the U.S. Department of Energy, Office of Science, under contract numbers DE-AC02-06CH11357, DE-AC02-05CH11231, and subcontracts GR138942 and GR130493 at OSU. This research uses computing resources provided by the National Artificial Intelligence Research Resource (NAIRR) Pilot, supported by award NAIRR240008.

REFERENCES

- [1] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," *Artificial intelligence and statistics*, pp. 1273–1282, 2017.

- [2] 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.
- [3] X. Huang, Y. Dong, and S. Pentyala, "Data pipeline challenges in privacy-preserving federated learning," <https://www.nist.gov/blogs/cybersecurity-insights/data-pipeline-challenges-privacy-preserving-federated-learning>, February 2024, nIST Cybersecurity Insights Blog Post. Part of a series on privacy-preserving federated learning in collaboration with the UK government's Responsible Technology Adoption Unit (RTA).
- [4] G. Nilsson, "The impact of data quality on federated versus centralized learning," Master of Science in Engineering: AI and Machine Learning, Blekinge Institute of Technology, 371 79 Karlskrona, Sweden, July 2024.
- [5] G. Nilsson, M. Boldt, and S. Alawadi, "The role of the data quality on model efficiency: An exploratory study on centralised and federated learning," in *2024 9th International Conference on Fog and Mobile Edge Computing (FMEC)*, 2024, pp. 253–260.
- [6] W. Zhao, Y. Du, N. D. Lane, S. Chen, and Y. Wang, "Enhancing data quality in federated fine-tuning of foundation models," *arXiv preprint arXiv:2403.04529*, Mar 2024.
- [7] A. R. Chowdhury, C. Guo, S. Jha, and L. van der Maaten, "Eiffel: Ensuring integrity for federated learning," 2022. [Online]. Available: <https://arxiv.org/abs/2112.12727>
- [8] X. Cao, M. Fang, J. Liu, and N. Z. Gong, "Fltrust: Byzantine-robust federated learning via trust bootstrapping," 2022. [Online]. Available: <https://arxiv.org/abs/2012.13995>
- [9] K. Hiniduma, S. Byna, and J. L. Bez, "Data readiness for ai: A 360-degree survey," *ACM Comput. Surv.*, Mar. 2025, just Accepted. [Online]. Available: <https://doi.org/10.1145/3722214>
- [10] K. Hiniduma, S. Byna, J. L. Bez, and R. Madduri, "Ai data readiness inspector (aidrin) for quantitative assessment of data readiness for ai," in *Proceedings of the 36th International Conference on Scientific and Statistical Database Management*, ser. SSDBM '24. New York, NY, USA: Association for Computing Machinery, 2024. [Online]. Available: <https://doi.org/10.1145/3676288.3676296>
- [11] T.-H. Hoang, J. Fuhrman, M. Klarqvist, M. Li, P. Chaturvedi, Z. Li, K. Kim, M. Ryu, R. Chard, E. A. Huerta *et al.*, "Enabling end-to-end secure federated learning in biomedical research on heterogeneous computing environments with appfl," *Computational and Structural Biotechnology Journal*, vol. 28, pp. 29–39, 2025.
- [12] Z. Li, S. He, Z. Yang, M. Ryu, K. Kim, and R. Madduri, "Advances in appfl: A comprehensive and extensible federated learning framework," *arXiv preprint arXiv:2409.11585*, 2024.
- [13] M. Ryu, Y. Kim, K. Kim, and R. K. Madduri, "Appfl: open-source software framework for privacy-preserving federated learning," in *2022 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. IEEE, 2022, pp. 1074–1083.
- [14] S. Shrivastava *et al.*, "Dqlearn: A toolkit for structured data quality learning," in *Proceedings of the IEEE International Conference on Big Data (Big Data)*, 2020, pp. 1644–1653.
- [15] N. Gupta, H. Patel *et al.*, "Data quality toolkit: Automatic assessment of data quality and remediation for machine learning datasets," *arXiv preprint arXiv:2108.05935*, 2021.
- [16] S. Afzal, C. Rajmohan, M. Kesarwani, S. Mehta, and H. Patel, "Data readiness report," in *Proceedings of the IEEE International Conference on Smart Data Services (SMDS)*, 2020, pp. 42–51.
- [17] S. Schelter, D. Lange, P. Schmidt, M. Celikel, F. Biessmann, and A. Graefberger, "Automating large-scale data quality verification," *Proceedings of the VLDB Endowment*, vol. 11, no. 12, pp. 1781–1794, August 2018.
- [18] T. Rekatsinas, X. Chu, I. F. Ilyas, and C. Ré, "Holoclean: holistic data repairs with probabilistic inference," *Proc. VLDB Endow.*, vol. 10, no. 11, p. 1190–1201, Aug. 2017. [Online]. Available: <https://doi.org/10.14778/3137628.3137631>
- [19] IBM, "Data quality sla rule compliance and remediation," https://dataplatform.cloud.ibm.com/docs/content/ws_j/quality/dq-sla-compliance.html?context=cpdas&audience=wdp, 2015, accessed: 2025-04-30.
- [20] X. Jiang, S. Sun, J. Li, J. Xue, R. Li, Z. Wu, G. Xu, Y. Wang, and M. Liu, "Tackling noisy clients in federated learning with end-to-end label correction," in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, ser. CIKM '24. ACM, Oct. 2024, p. 1015–1026. [Online]. Available: <http://dx.doi.org/10.1145/3627673.3679550>
- [21] Z. Zhang, G. Chen, Y. Xu, L. Huang, C. Zhang, and S. Xiao, "Feddq: A novel regularization-based deep learning method for data quality assessment in federated learning," *Decision Support Systems*, vol. 180, p. 114183, 2024. [Online]. Available: <https://doi.org/10.1016/j.dss.2024.114183>
- [22] L. Rokvic, P. Danassis, S. P. Karimireddy, and B. Faltings, "Lia: Privacy-preserving data quality evaluation in federated learning using a lazy influence approximation," 2024. [Online]. Available: <https://arxiv.org/abs/2205.11518>
- [23] Y. Du, R. Ye, F. Yuchi, W. Zhao, J. Qu, Y. Wang, and S. Chen, "Data quality control in federated instruction-tuning of large language models," 2025. [Online]. Available: <https://arxiv.org/abs/2410.11540>
- [24] European Commission, "Industry 5.0," 2021, accessed: 2025-04-14. [Online]. Available: https://research-and-innovation.ec.europa.eu/research-area/industrial-research-and-innovation/industry_50_en
- [25] K. Pearson, "On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.
- [26] J. Zhang, C. Li, J. Qi, and J. He, "A survey on class imbalance in federated learning," 2023. [Online]. Available: <https://arxiv.org/abs/2303.11673>
- [27] R. Kohavi, "Adult," UCI Machine Learning Repository, 1996, doi: 10.24432/C5GP7S. [Online]. Available: <https://doi.org/10.24432/C5GP7S>
- [28] T.-M. H. Hsu, H. Qi, and M. Brown, "Measuring the effects of non-identical data distribution for federated visual classification," *arXiv preprint arXiv:1909.06335*, 2019.
- [29] J. Nguyen, K. Malik, H. Zhan, A. Yousefpour, M. Rabbat, M. Malek, and D. Huba, "Federated learning with buffered asynchronous aggregation," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 3581–3607.
- [30] Z. Li, P. Chaturvedi, S. He, H. Chen, G. Singh, V. Kindratenko, E. A. Huerta, K. Kim, and R. Madduri, "FedCompass: efficient cross-silo federated learning on heterogeneous client devices using a computing power aware scheduler," *arXiv preprint arXiv:2309.14675*, 2023.
- [31] C. Dwork, "Differential privacy," in *International colloquium on automata, languages, and programming*. Springer, 2006, pp. 1–12.
- [32] APPFL Contributors, "Data readiness assurance framework in appfl," https://appfl.ai/en/latest/tutorials/examples_dr_integration.html, 2025.
- [33] C. Xiao and S. Wang, "An experimental study of class imbalance in federated learning," in *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2021.
- [34] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.
- [35] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226, 2012.
- [36] E. Liberty, Z. Karnin, B. Xiang, L. Rouesnel, B. Coskun, R. Nallapati, J. Delgado, A. Sadoughi, Y. Astashonok, P. Das *et al.*, "Stratified sampling meets machine learning," in *Proceedings of The 33rd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 48. PMLR, 2016, pp. 2320–2329.
- [37] J. W. Tukey, "Exploratory data analysis," 1977.
- [38] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," *Technical report, SRI International*, 1998.
- [39] Y. LeCun and C. Cortes, "Mnist handwritten digit database," 2010. [Online]. Available: <http://yann.lecun.com/exdb/mnist/>
- [40] A. Krizhevsky, "Learning multiple layers of features from tiny images," Tech. Rep., 2009.
- [41] C. He, S. Rasouli, I. Zachariah, P. Tiwari, P. Bacon, Y. Shen, A. Kotti, O. Marfoq, H. Benali, T. Clozel *et al.*, "Flamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings," *arXiv preprint arXiv:2210.04620*, 2022.
- [42] A.-R. Consortium, "Flagship dataset of type 2 diabetes from the ai-readi project (1.0.0)," 2024. [Online]. Available: <https://doi.org/10.60775/fairhub.1>
- [43] W. Gropp, T. Boerner, B. Bode, and G. Bauer, "Delta: Balancing gpu performance with advanced system interfaces," National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign,

- Technical Report, 2023, funded by National Science Foundation (award OAC 2005572).
- [44] Owkin, “FLamby IXI Dataset,” https://github.com/owkin/FLamby/blob/main/flamby/datasets/fed_ixi/README.md#prediction-task, 2021, accessed: 2025-07-14.
 - [45] C. Xiao and S. Wang, “An experimental study of class imbalance in federated learning,” *arXiv preprint arXiv:2109.04094*, 2022.
 - [46] R. Labs. (2023) Understanding the impact of class imbalance in federated learning. [Online]. Available: <https://risingwave.com/blog/understanding-the-impact-of-class-imbalance-in-federated-learning/>
 - [47] W. Huang, T. Li, D. Wang, S. Du, and J. Zhang, “Fairness and accuracy in federated learning,” *Information Sciences*, vol. 589, pp. 170–185, 2022.
 - [48] M. Fisichella, G. Lax, and A. Russo, “Partially-federated learning: A new approach to achieving privacy and effectiveness,” *Information Sciences*, vol. 610, pp. 1–18, 2022.
 - [49] H. Huang, W. Zhuang, C. Chen, and L. Lyu, “Fedmef: Towards memory-efficient federated dynamic pruning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Computer Vision Foundation, 2024.