

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 privacy and security of the client’s data by not 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 FL, a novel framework that allows users to define custom data readiness (DR) standards, metrics, rules, and remedies tailored to specific FL tasks. Our framework generates comprehensive DR reports based on the user-defined metrics, rules, and remedies to ensure datasets are optimally prepared for FL while preserving privacy. We demonstrate the framework’s practical application by integrating it into an existing PPFL framework. We conducted experiments across six diverse datasets, addressing seven different DR issues. The results illustrate the framework’s versatility and effectiveness 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 by identifying and addressing data-related issues before the training phase.

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

I. INTRODUCTION

Federated Learning (FL) [1], [2] is a decentralized machine learning approach allowing multiple 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 on local devices. However, new challenges emerge when privacy-preserving techniques are applied in FL. A recent study [3] (part of a series on Privacy-Preserving Federated Learning (PPFL) by NIST in collaboration with the UK government’s Responsible Technology Adoption Unit) highlights significant challenges, primarily due to its unique approach that prevents organizations from accessing training data. This restriction complicates essential pre-processing tasks like data cleaning and feature selection, as data scientists cannot view data across different sites. This may lead to potential inconsistencies and deployment failures. Many studies [4]–[6] in the literature have demonstrated that low-quality data directly impacts the model by significantly 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 in PPFL, we introduce Data Readiness for AI (DRAI) into the PPFL domain. Our recent survey [9] presented a comprehensive six-pillar taxonomy for assessing DRAI, focusing on data quality, organization, fairness, understandability, governance, and value. To operationalize these dimensions, we developed AIDRIN (AI Data Readiness Inspector) [10], a framework designed to evaluate the DRAI of datasets across these pillars. However, AIDRIN was initially designed for centralized AI training, where data is uploaded to a standalone platform for evaluation. In contrast, PPFL requires a decentralized approach to data readiness (DR) assessment including methods that preserve privacy and security while evaluating DR across distributed datasets. In addition AIDRIN was designed to visualize the metric evaluations that were predefined. This study aims to bridge the gap in meeting specific DR standards tailored to PPFL tasks by allowing users to define custom metrics and evaluation criteria to suit the FL context while not compromising privacy.

An extendable framework for supporting user-defined metrics, rules, and remedies is still unavailable. For example, in the healthcare industry, PPFL can be used to develop a predictive model for diagnosing a specific disease using MRI scans from multiple hospitals [11]. However, challenges such as data heterogeneity, data quality, and privacy concerns arise. Hospitals often use different MRI machines, which 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 we can allow stakeholders to define custom DR standards, metrics, rules, and remedies tailored to this FL task. Experts can establish standards for what constitutes an “AI-ready” MRI scan, such as specific format and resolution requirements, and implement metrics to evaluate data 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 data quality. It is required that each hospital to ensure independently that its data meets the necessary standards before participating in the FL process. This may result in

a consistent and high-quality dataset across all participants. Consequently, the reliability and performance of the predictive model are improved while maintaining patient privacy.

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 users to define custom DR standards, such as metrics, rules, and remedies tailored to specific FL tasks. In this context, “users” refers to the individuals or stakeholders who are involved in and responsible for a specific FL task, and who collaborate to establish the necessary standards, metrics, rules, and remedies. In the rest of the paper, “users” will be used to denote these key participants. This framework allows clients to locally execute these user-defined functions to dynamically ensure their data meets the necessary standards without compromising privacy. Clients can verify compliance with these rules and apply remedies to their data if necessary. The results of these metric evaluations are compiled into a DR report for user 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 users in the definition, validation, and refinement of DR standards.

This process ensures that only clients with qualified data participate in the FL system. By doing so, it maintains DR and integrity while preserving privacy. The 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. By this integration, we showcase how this approach can be used in existing PPFL workflows. The main contributions of this study are highlighted below.

- 1) We propose a novel framework that enables users within a PPFL system to define custom metrics, rules, and remedies. Our framework addresses the challenge of executing these user-defined standards by automating the process and ensuring that clients can locally apply these actions to meet required data standards while preserving privacy.
- 2) The framework generates comprehensive DR reports that evaluate the standards defined by users. This method ensures privacy is preserved by only including aggregated metric evaluations without exposing any raw data. Users can review these reports to assess whether clients have met the expected standards and gain insights into the data’s characteristics.
- 3) Integrating CADRE into the APPFL framework demonstrates its practical value and compatibility with existing PPFL workflows.

We evaluated CADRE using six datasets with various data modalities (e.g. 2D images, tabular data, 3D volumetric data)

and downstream tasks (such as classification, segmentation, and survival analysis). These datasets were either naturally occurring or intentionally constructed to reflect seven key DR challenges: noise, class imbalance, duplicate records, high memory consumption, bias, outliers, and insufficient anonymity. As our framework allows users within a PPFL system to define custom metrics, rules, and remedies, these issues were effectively addressed. Through these experiments, we demonstrated how datasets that initially failed to meet these standards could be systematically improved. This illustrates the versatility and effectiveness of our framework in real-world and heterogeneous data scenarios.

II. RELATED WORK

Users seeking to adopt AI rely on structured frameworks to evaluate their DR, with a focus on key 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. These tools focus on user-friendly interfaces, advanced profiling, duplicate removal, and other validation rules, enabling both technical and non-technical users to efficiently clean and standardize data.

Despite their strengths, these frameworks exhibit critical gaps when applied to modern, distributed AI environments. Most notably, they lack integration with FL architectures, which are increasingly important as organizations move toward decentralized data models. Existing frameworks generally assume a centralized data environment and provide little to no guidance on assessing DR across distributed clients with heterogeneous data distributions. They also fall short in addressing compliance challenges related to cross-border data flows, which are common in FL scenarios. FL has become a prominent decentralized machine learning approach, allowing multiple clients to collaboratively train models while maintaining data privacy. Apart from its advantages, FL faces key challenges such as robustness against malicious clients and heterogeneous data distributions [12]. Several frameworks have been developed to address these challenges with unique methodologies and trade-offs.

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 a two stage framework that first identifies clients with noisy labels and then applies label correction strategies to improve the overall robustness of the global model. However, it focuses solely on label noise and ignores other critical aspects of DR. FedDQA [21] introduces a metric to evaluate client data quality without additional computational cost, 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 user-specific 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 users 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 users to take control of DR.

III. DESIGN OVERVIEW

The objective of CADRE is to allow users of FL systems to define and utilize both foundational and customizable actions. To support this requirement, CADRE provides the following main components: *metrics*, *DR reports*, *rules*, and *remedies*. The functionality of each of these components will be discussed in details in the following subsections. In Figure 1, we show an outline of CADRE by illustrating its components.

A. Metrics Component

We divided the metrics component of CADRE into two main parts: standard metrics and customizable metrics. The standard metrics include a set of universally applicable measures such as evaluating sample sizes, data sparsity, and basic 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 module contains basic visualizations, such as bar charts and scatter plots, which are

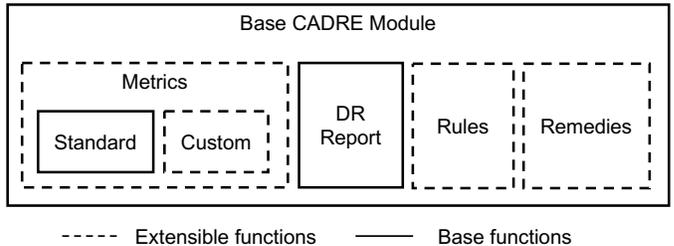


Fig. 1: An overview of CADRE framework for FL tasks. Metrics include commonly known standard DR evaluation measurements. The extensible functions are used to define custom DR metrics, rules, and remedies. The DR report consolidates the standard and custom metric evaluations with visualizations.

included in the DR reports to provide a visual representation of the client data’s characteristics.

The customizable metrics sub-module offers an extensible function that allows users to define custom metrics tailored to their unique FL task and evaluation needs. This flexibility ensures that users can assess client data according to the specific requirements of their projects. For example, if a particular task requires assessing the completeness or skewness of the data, users can define these metrics within CADRE. These standard and custom metric evaluations and visualizations allow users 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 Sub-Modules

Beyond metrics, CADRE includes a robust sub-module for defining rules and remedies. This sub-module allows users to establish custom rules that data must meet to be considered ready for the next stages of the FL pipeline. The users can also define custom remedies to improve the readiness of the data to meet the specified rules.

For instance, if users 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 users 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 highlight the heterogeneity among clients. These insights are compiled into an easily readable HTML report, allowing users to quickly 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, the 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. Figure 2 presents an actual HTML DR report generated for this specific example during an FL experiment. 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.

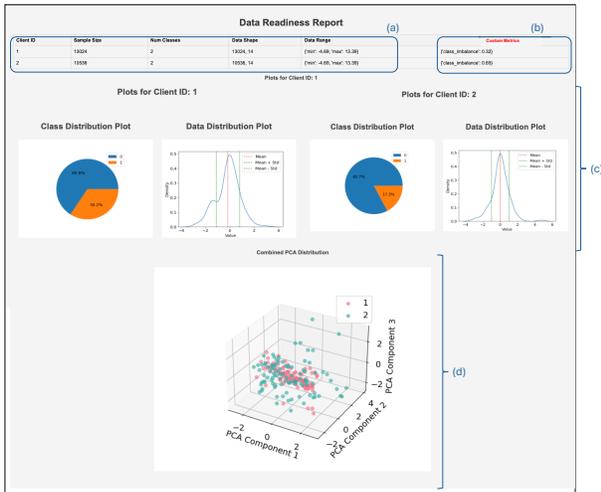


Fig. 2: The figure illustrates an example DR report from an FL experiment featuring: (a) Standard metrics, (b) Custom metrics in CADRE for this specific FL task, (c) Individual client plots, and (d) Combined data plots.

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 users within CADRE will be applied 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 as needed while preserving privacy.

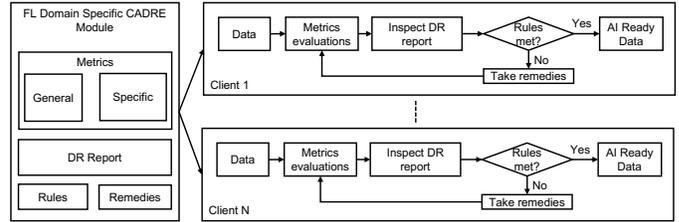


Fig. 3: The figure illustrates the iterative data evaluation and remediation process within the clients involved in FL framework. It outlines how clients use the CADRE’s functions to assess DR, apply custom rules, and implement remedies as needed.

By integrating these sub-modules, CADRE provides a comprehensive and flexible framework for ensuring DR in FL systems. This framework allows users 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 federated learning systems. It allows researchers to implement, test, and deploy federated learning experiments across distributed clients while ensuring data privacy.

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 federated learning 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 federated learning 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. Users 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. Figure 4 provides a visual representation of its implementation within the APPFL framework.

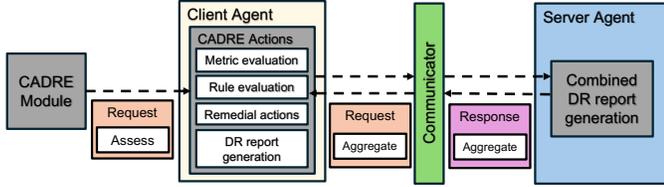


Fig. 4: The figure illustrates the integration of CADRE within the APPFL framework.

Configuring CADRE for specific FL tasks is a straightforward process that allows users 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, users 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, users 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, users 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 users to define evaluation metrics, rules, and remedies specific to their needs.

With the integration of CADRE into the APPFL framework, users 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 users with timely interventions, allowing them to focus on strategic decisions rather than manual data remediation tasks.

Additionally, the comprehensive DR reports offer transparency and accountability. These reports provide users 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, users 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 online [32]

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

V. EVALUATIONS

In this section, we will explore the datasets, experimental setups, and custom CADRE modules used to demonstrate how we can effectively address DR related challenges across various dimensions, including data quality, privacy, fairness, and more. Since most datasets in our study do not naturally exhibit these issues, we used various data pollution techniques, which we will also discuss here. Finally, we will illustrate how our custom DR standards are achieved by utilizing the tailored metrics, rules, and remedies within the custom CADRE modules.

A. Datasets and Experimental Setup

In this study, we utilized six diverse datasets spanning both standard benchmarks and 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-READI (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-READI 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

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.

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-READI 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-READI 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 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 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 its metric. The 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, the quasi-identifiers were “work-class,” “race,” and “gender.” We selected these features 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.”

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

CADRE Module ID	MNIST	CIFAR-10	Adult Income	Flamby TCGA-BCRA	Flamby IXI Tiny	AI-READI
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)

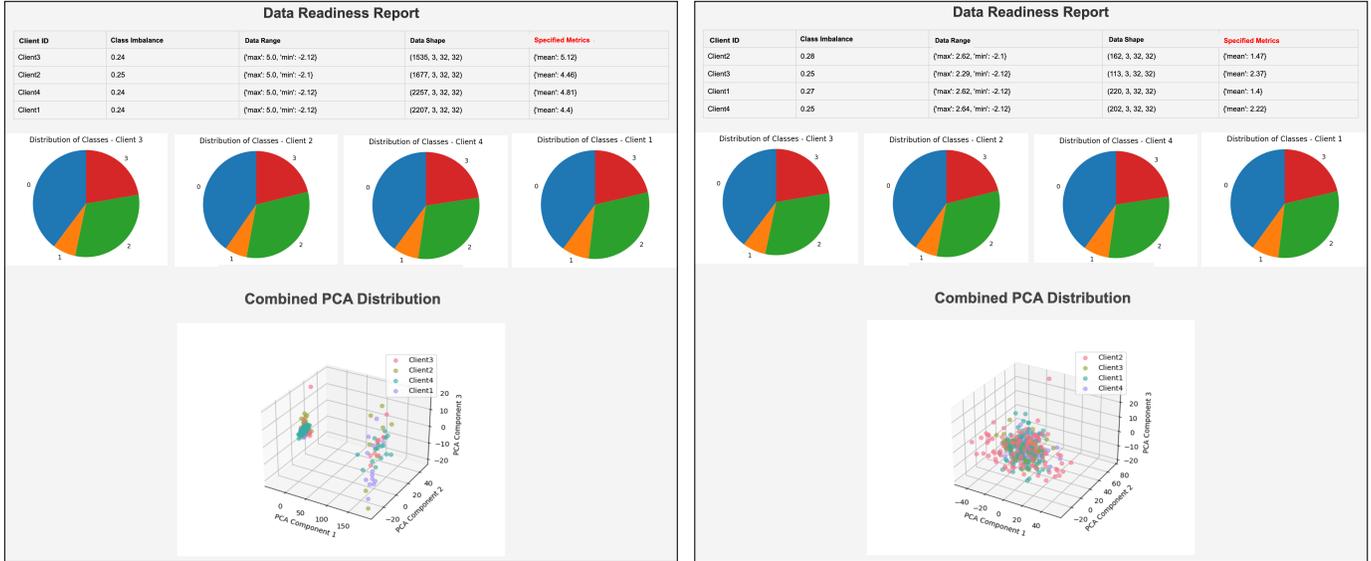


Fig. 6: Example DR reports generated before (left) and after (right) applying CADRE module 1 show an improvement in the average mean after removing noisy data. Results are shown in the table’s rightmost column. The combined PCA plot at the bottom right confirms that noise-related anomalies in the data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

These attributes were chosen due to their potential to link individuals across datasets and may pose privacy concerns if not anonymized.

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 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.

Since this work precedes the training phase of the FL task, it provides valuable insights for users to analyze the data and assess its readiness for training. This pre-training

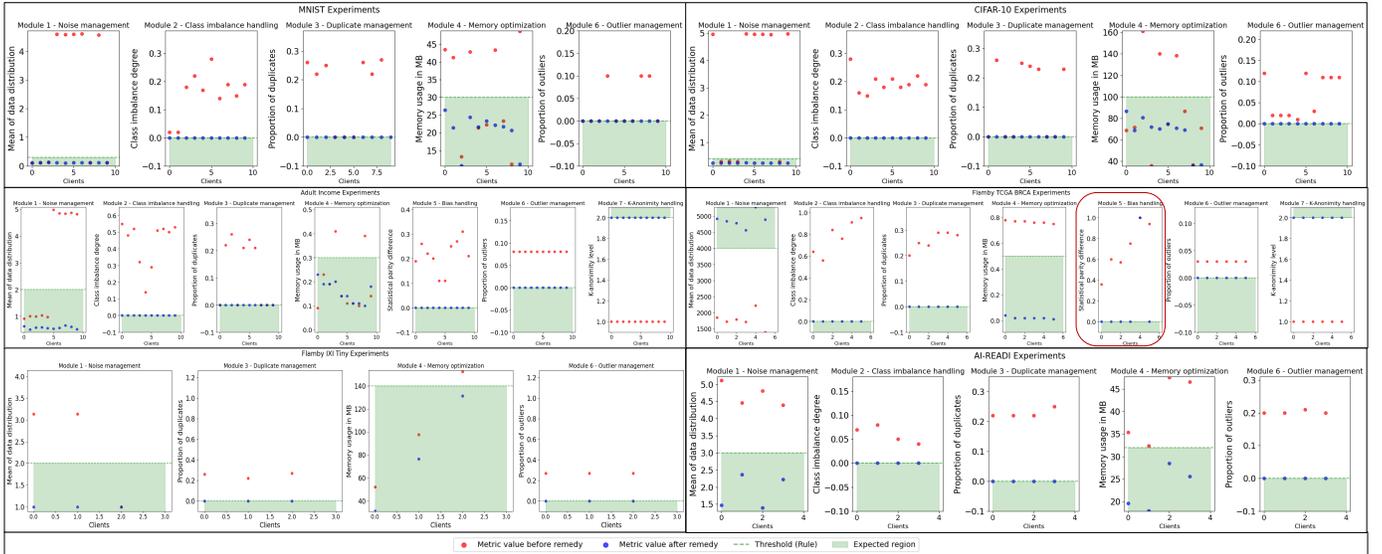


Fig. 7: Evaluation of custom metrics for each CADRE module, before and after remedy application. Threshold lines indicate predefined rule criteria. The red box highlights one case where a client’s post-remedy metric remains above the threshold.

evaluation helps determine whether proceeding with training is worthwhile, potentially saving significant resources. While we do not report the impact on overall model performance due to not entering the training phase, existing literature suggests that many remedies applied by these modules can influence model outcomes. For instance, high-quality, noise-free data significantly enhances model performance [4], [22], while balanced class distributions can reduce bias and mitigate model drifting issues [44], [45].

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 [46], though it may compromise overall performance and accuracy. Similarly, as we increase the privacy budget of the data, model accuracy tend to decrease [47]. However, users 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 [48]. 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 users 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 users 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 generates comprehensive DR reports that include evaluations from user-defined and standard metrics, along with visualizations that provide insights into data characteristics and DR levels. It enables users to make informed decisions before training while preserving data privacy. Experiments across six diverse datasets and seven distinct DR issues demonstrate the framework’s versatility and effectiveness. The integration of the framework into the APPFL framework highlights its practical applicability. By addressing DR issues before training, our approach conserves resources and enhances model outcomes. However, there can be requirements that necessitate compromising model performance. For instance, achieving perfect fairness may affect accuracy, and increasing privacy budgets can decrease precision. These trade-offs highlight the importance of prioritizing data standards based on specific FL task requirements. Our framework allows users to set realistic expectations for training, optimize resource utilization, and lay the groundwork for reliable and equitable FL results.

For future work, we plan to expand its applicability to a broader range of datasets and explore automated methods for CADRE to further streamline the DR process. Additionally, we will investigate more computationally intensive tasks, and explore adding custom privacy-preserved modules to CADRE for user-controlled privacy protection. This will enhance the framework’s adaptability to evolving privacy standards in PPFL environments.

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. Pentylala, "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. Grafberger, "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://dataplatfom.cloud.ibm.com/docs/content/wsj/quality/dq-sla-compliance.html?context=cpdaas&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-ready 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] C. Xiao and S. Wang, "An experimental study of class imbalance in federated learning," *arXiv preprint arXiv:2109.04094*, 2022.
- [45] 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/>
- [46] 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.
- [47] 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.
- [48] 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.