

AIDRIN 2.0: A Framework to Assess Data Readiness for AI

Kaveen Hiniduma
The Ohio State University
hiniduma.1@osu.edu

Dylan Ryan
The Ohio State University
ryan.1989@osu.edu

Suren Byna
The Ohio State University
byna.1@osu.edu

Jean Luca Bez
Lawrence Berkeley National
Laboratory
jlbez@lbl.gov

Ravi Madduri
Argonne National Laboratory
madduri@anl.gov

Abstract

AI Data Readiness Inspector (AIDRIN) is a framework to evaluate and improve data preparedness for AI applications. It addresses critical data readiness dimensions such as data quality, bias, fairness, and privacy. This paper details enhancements to AIDRIN by focusing on user interface improvements and integration with a privacy-preserving federated learning (PPFL) framework. By refining the UI and enabling smooth integration with decentralized AI pipelines, AIDRIN becomes more accessible and practical for users with varying technical expertise. Integrating with an existing PPFL framework ensures that data readiness and privacy are prioritized in federated learning environments. A case study involving a real-world dataset demonstrates AIDRIN's practical value in identifying data readiness issues that impact AI model performance.

1 Introduction

The AI Data Readiness Inspector (AIDRIN) [3] has become an essential framework for assessing and improving data preparedness for AI applications. As organizations increasingly depend on AI-driven decision-making, ensuring high-quality, AI-ready data is critical. AIDRIN addresses this need by offering a comprehensive framework for evaluating data readiness across key dimensions, including data quality, bias and fairness, and privacy. In a previous study [2], we extensively examined the metrics across the pillars of data readiness. These pillars include *Data Quality*, *Understandability & Usability*, *Structure & Organization*, *Governance*, *Impact on AI*, and *Fairness*. AIDRIN categorizes the evaluation metrics within these pillars to provide users with a personalized assessment.

This paper presents enhancements to AIDRIN by mainly focusing on user interface (UI) improvements and integration to an existing privacy-preserving federated learning (PPFL) framework. Our objective is to make AIDRIN more intuitive, efficient, and aligned with the latest research in AI data readiness. By refining the UI, we aim to increase AIDRIN's accessibility and effectiveness for both data professionals and domain experts who may not have extensive technical backgrounds in data science or AI. Additionally, integrating AIDRIN into a PPFL framework strengthens its scalability and usability by allowing smooth operation across both centralized and decentralized AI pipelines.

2 UI Enhancements

The latest version of AIDRIN introduces UI enhancements designed to improve usability and provide clearer insights. One of the key improvements is the reorganization of evaluation metrics according

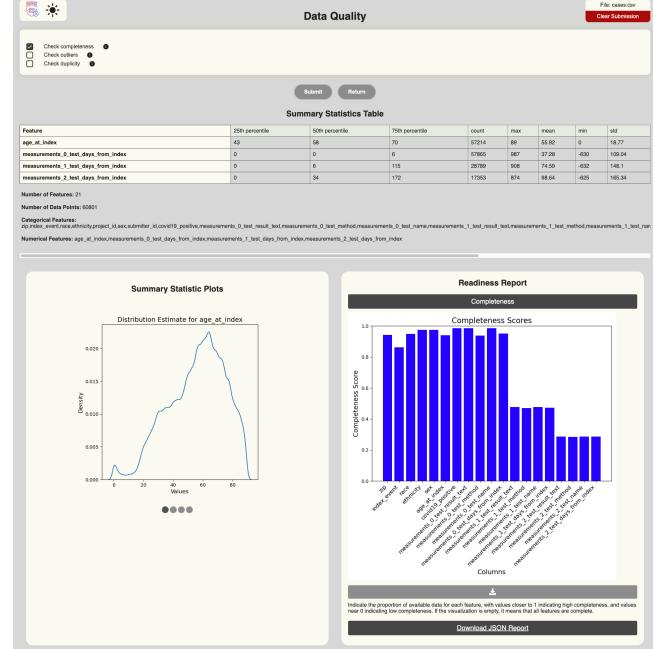


Figure 1: The figure displays the data quality page with real evaluation results from the MIDRIC cases dataset [6]

to the six pillars of AI data readiness, as established in our previous study [2]. These pillars offer a structured and comprehensive approach to assessing data readiness.

Each pillar now includes specific assessment criteria to ensure a more precise and meaningful evaluation. *Data Quality* examines completeness, duplicates, and outliers to ensure the datasets meet fundamental integrity standards. *Understandability & Usability* assesses adherence to FAIR principles [8] to improve data accessibility and reusability. *Impact on AI* evaluates feature relevance and correlations to help users understand how their data influences model performance. *Governance* analyzes data re-identification risks when quasi-identifiers are provided to address privacy concerns. Finally, *Fairness* examines class imbalance, statistical parity, and representational rates of sensitive features, ensuring equitable AI outcomes.

Additionally, AIDRIN presents a table with summary statistics displaying key metrics such as mean, median, and mode, alongside visualizations of data distributions. Based on user-selected metrics, AIDRIN dynamically generates quantitative evaluations and

visualizations to create a more intuitive and insightful assessment experience. These enhancements make AIDRIN a more powerful and user-friendly tool to ensure high-quality, AI-ready data.

Figure 1 illustrates these improvements in the UI, with the screenshot displaying the data quality page with real evaluation results generated from the MIDRC (Medical Imaging and Data Resource Center) cases dataset. The MIDRC dataset is a large, publicly available collection of de-identified medical imaging data focused on COVID-19-related chest imaging.

3 AIDRIN Integration to APPFL Framework

The integration of AIDRIN with APPFLx (Advanced Privacy Preserving Federated Learning as a Service) [1] marks a significant advancement in the field of PPFL. APPFLx provides a user-friendly platform for conducting cross-silo PPFL experiments by enabling secure and decentralized AI model training across multiple organizations. We selected APPFLx due to its open-source nature, modular architecture, and scalability, making it a good foundation for integrating AIDRIN's data evaluation features.

AIDRIN enables edge systems to locally evaluate data and use the communication strategies in the APPFLx framework to transmit evaluation results to the server. The server then generates HTML reports using AIDRIN by combining the results from all the users involved in the FL system. These reports provide stakeholders with clear, visually organized insights that are easy to interpret and review. This approach preserves privacy by ensuring that data remains on the edge system, with only evaluation results being transmitted to the server.

By integrating AIDRIN into APPFLx [1, 5], the system enables edge devices to locally evaluate data readiness before investing computational resources into training. Leveraging APPFLx’s communication framework, evaluation results are transmitted to a central server, which uses AIDRIN to generate aggregated HTML reports. These reports offer stakeholders clear, data readiness insights into data suitability across distributed clients.

4 Evaluations

AIDRIN has proven to be valuable in multiple use cases since its development, particularly in Privacy-Preserving Federated Learning (PPFL) settings. One such case involved using AIDRIN to analyze data quality issues in the Flamby Heart Disease [4, 7] dataset, a real-world federated dataset collected from four different hospitals.

Before training, AIDRIN's analysis revealed that one client's dataset contained only a single class, meaning all samples belonged to the same category. Additionally, the feature distribution in this client's data was highly sparse, with one feature consisting entirely of zeros, making it impossible to compute feature correlations for that feature. These issues likely led to model bias and poor generalization, ultimately degrading overall performance.

To address this, we conducted experiments where we excluded this problematic client from training, leading to a notable improvement in accuracy, from 70.6% to 74.7%. The initial performance drop was likely due to the imbalanced and uninformative data from the outlier client negatively influencing the global model, while removing it allowed for a more representative and effective learning process. Figure 2 presents visualizations generated by AIDRIN by

highlighting the problematic client's data characteristics, including a single-class distribution and sparse feature correlations.

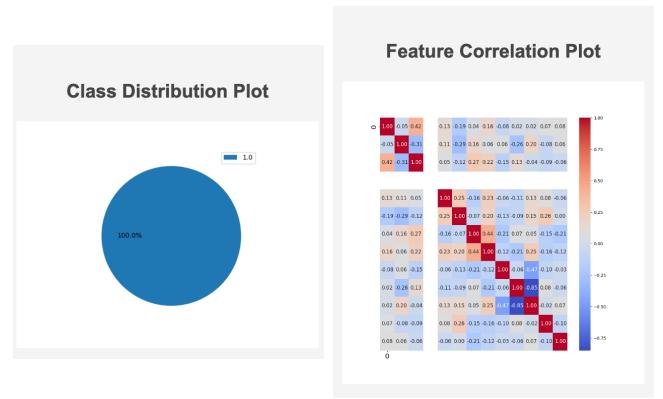


Figure 2: Visualizations from a client in the FLeamy Heart Disease dataset. The left plot shows a class distribution where all samples belong to a single class, indicating complete class imbalance. The right plot presents a feature correlation matrix, where one feature could not generate meaningful feature relevances due to zero values.

5 Conclusion

The enhancements made to AIDRIN improve its usability and effectiveness in AI-driven environments, particularly within PPFL setting. By refining the user interface and enabling integration into decentralized AI pipelines, AIDRIN is more accessible to users of varying technical backgrounds, including those with limited expertise. The evaluation results, particularly the case study involving the Flamby Heart Disease dataset, highlight AIDRIN's practical value in identifying critical data readiness issues that directly impact model performance. Overall, this study validates AIDRIN as a vital framework for promoting reliable, fair, and trustworthy AI development.

References

- [1] Li et al. 2023. APPFLx: Providing Privacy-Preserving Cross-Silo Federated Learning as a Service. In *2023 IEEE 19th International Conference on e-Science (e-Science)*. IEEE, 1–4. doi:10.1109/e-Science58273.2023.10254842
 - [2] Kaveen Hiniduma, Suren Byna, and Jean Luca Bez. 2025. Data Readiness for AI: A 360-Degree Survey. *ACM Comput. Surv.* (March 2025). doi:10.1145/3722214
 - [3] Kaveen Hiniduma et al. 2024. AI Data Readiness Inspector (AIDRIN) for Quantitative Assessment of Data Readiness for AI (*SSDBM '24*). doi:10.1145/3676288.3676296
 - [4] Andras Janosi, William Steinbrunn, Matthias Pfisterer, and Robert Detrano. 1988. Heart Disease. <https://archive.ics.uci.edu/dataset/45/heart+disease>. UCI Machine Learning Repository.
 - [5] Zilinghan Li et al. 2025. Advances in APPFL: A Comprehensive and Extensible Federated Learning Framework. In *2025 IEEE 25th International Symposium on Cluster, Cloud and Internet Computing (CCGrid)*. IEEE.
 - [6] Medical Imaging and Data Resource Center (MIDRC). [n. d.]. The Medical Imaging and Data Resource Center (MIDRC). <https://www.midrc.org/>. [n.d.] Accessed: 2025-04-03.
 - [7] Jean Ogier du Terrain and Samy-Safwan et al. Ayed. 2022. FLamby: Datasets and Benchmarks for Cross-Silo Federated Learning in Realistic Healthcare Settings. In *Advances in Neural Information Processing Systems*, Vol. 35. 5315–5334.
 - [8] Mark D Wilkinson et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* 3 (2016), 160018. doi:10.1038/sdata.2016.18

AIDRIN 2.0: A Framework to Assess Data Readiness for AI

 THE OHIO STATE UNIVERSITY
Kaveen Hiniduma¹, Dylan Ryan¹, Suren Byna¹, Jean Luca Bez², Ravi Madduri³
¹The Ohio State University, ²Lawrence Berkeley National Laboratory, ³Argonne National Laboratory
(hiniduma.1, ryan.1989, byna.1)@osu.edu, jlbez@lbl.gov, madduri@anl.gov

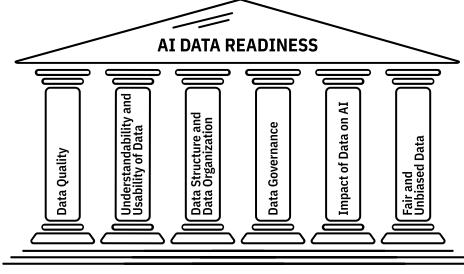
 BERKELEY LAB
 Argonne NATIONAL LABORATORY

HIGHLIGHT

- Solves the GIGO Problem (Garbage In, Garbage Out)**
Unready data leads to poor AI outcomes. AIDRIN prevents this by assessing data across six critical pillars: quality, fairness, usability, structure, AI impact, and privacy.
- Centralized & Decentralized, Privacy-Preserving by Design**
Supports data readiness assessments in both centralized and decentralized settings, integrating seamlessly with the APPFL^[3] framework for secure, privacy-aware evaluations across edge and sensitive environments.
- Saves Time, Compute & Rework**
Detects data issues early in the ML pipeline, optimizing efficiency and accelerating trustworthy AI development.

PILLARS OF DATA READINESS FOR AI

We comprehensively explored key metrics and dimensions from the literature to define data readiness for AI^[1,2]

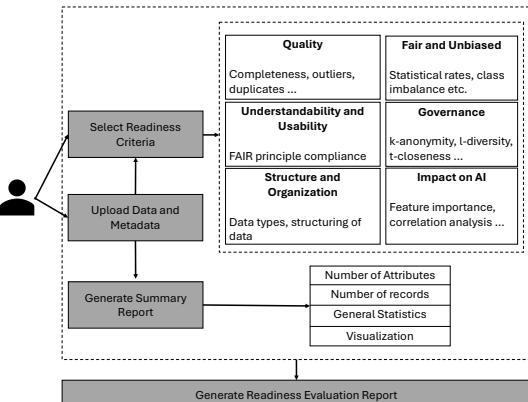


AIDRIN assesses data readiness across these pillars to provide a comprehensive evaluation of data suitability for AI

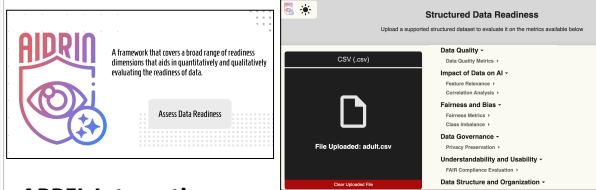
This work enhances AIDRIN by:

- Implementing an intuitive, interactive user interface with clear visual insights for centralized AI environments
- Enabling integration with APPFL^[3] framework to support secure, federated and decentralized evaluations

AIDRIN OVERVIEW

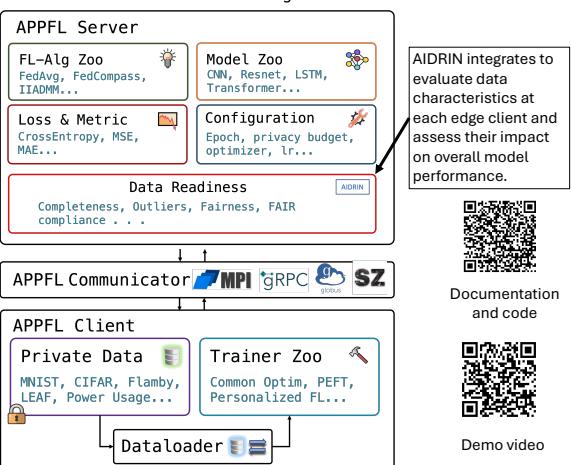


UI Enhancements
Shown below are some user interfaces of AIDRIN 2.0



APPFL Integration

 APPFL Advanced Privacy-Preserving Federated Learning Framework



AIDRIN integrates to evaluate data characteristics at each edge client and assess their impact on overall model performance.

Documentation and code: 

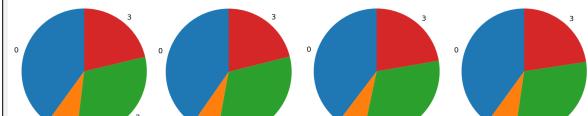
Demo video: 

• Shown below are two actual AIDRIN reports generated on the server during a federated learning task using the AI-READI dataset.

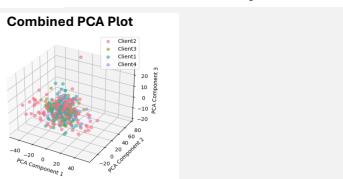
Data Readiness Report

Client IDs	Class Imbalance	Data Range	Data Shape	Data Mean
Client1	0.27	-2.12 to 2.62	(220, 3, 32, 32)	1.4
Client2	0.28	-2.10 to 2.62	(162, 3, 32, 32)	1.47
Client3	0.25	-2.12 to 2.29	(113, 3, 32, 32)	1.4
Client4	0.25	-2.12 to 2.64	(202, 3, 32, 32)	2.22

Distribution of Classes - Client 1 Distribution of Classes - Client 2 Distribution of Classes - Client 3 Distribution of Classes - Client 4



Combined PCA Plot



• Based on the client's selected metrics before training, the data is evaluated locally to preserve privacy. The resulting evaluations and visualizations are then sent to the server to generate the final aggregated data readiness report

References

- Kaveen Hiniduma, Suren Byna, Jean Luca Bez, and Ravi Madduri. 2024. 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 (SSDBM '24).
- Kaveen Hiniduma, Suren Byna, and Jean Luca Bez. 2025. Data Readiness for AI: A 360-Degree Survey. ACM Comput. Surv. 57, 9, Article 219 (September 2025), 39 pages.
- Z. Li, S. He, P. Chaturvedi, T.-H. Hoang, M. Ryu, E. Huerta, V. Kindratenko, J. Fuhrman, M. Giger, R. Chard et al., "APPFLX: Providing privacy-preserving cross-silo federated learning as a service," in 2023 IEEE 19th International Conference on e-Science (e-Science), IEEE, 2023, pp. 1-4.

Acknowledgements

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.

Figure 3: Accepted poster for SSDBM '25