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



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.

UI Enhancements

Shown below are some user interfaces of AIDRIN 2.0 Structured Data Readiness • • ed structured dataset to evaluate it on the metrics available below Data Quality framework that covers a broad range of readiness CSV (.csv) Data Quality Metrics dimensions that aids in quantitatively and qualitatively evaluating the readiness of data. Impact of Data on Al Correlation Analysis Fairness and Bias Assess Data Readiness Fairness Metrics Class Imbalance Data Governance File Uploaded: adult.csv **Privacy Preservation** Understandability and Usability **FAIR Compliance Evaluation** Data Structure and Organization **Clear Uploaded File**

APPFL Integration

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



APPFL Advanced Privacy-Preserving Federated Learning Framework

APPFL Server 資 * Model Zoo FL-Alg Zoo AIDRIN integrates to CNN, Resnet, LSTM, FedAvg, FedCompass, evaluate data Transformer... IIADMM... characteristics at 1/2 Configuration each edge client and Loss & Metric M Epoch, privacy budget, CrossEntropy, MSE, assess their impact MAE... optimizer, lr... on overall model performance. Data Readiness AIDRIN Completeness, Outliers, Fairness, FAIR compliance . . . APPFL Communicator 7 MPI GRPC 😂 SZ Documentation and code APPFL Client Trainer Zoo Private Data Common Optim, PEFT, MNIST, CIFAR, Flamby, LEAF, Power Usage... Personalized FL...

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

		Quality	Fair and Unbiased	
		Completeness, outliers, duplicates	Statistical rates, class imbalance etc.	
	Select Readiness Criteria	Understandability and Usability	Governance	
		FAIR principle compliancek-anonymity, l-divK-anonymity, l-divk-anonymity, l-div		
		Structure and Organization	Impact on Al	
	Upload Data and Metadata	Data types, structuring of data	Feature importance, correlation analysis	
		Number	of Attributes	
	Conorato Summary		of records	





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





References

- 1. 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).
- 2. 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.
- 3. 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.
- 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

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.