AIDRIN: A Comprehensive Toolset for Automating Data Preparation for AI

Kaveen Hiniduma*
Ohio State University
Columbus, OH, USA
hiniduma.1@osu.edu

Jean Luca Bez*
Lawrence Berkeley Laboratory
Berkeley, CA, USA
jlbez@lbl.gov

Ravi Madduri
Argonne National Laboratory
Argonne, IL, USA
madduri@anl.gov

Suren Byna
Ohio State University
Columbus, OH, USA
byna.1@lbl.gov

ABSTRACT

High-quality, ethically-governed, and efficiently structured data is important for effective AI. However, organizations often lack a unified method to assess whether datasets are ready for AI modeling. AIDRIN (AI Data Readiness Inspector) provides a comprehensive, multi-pillar framework that quantifies AI data readiness across six dimensions: Quality, Impact on AI, Understandability and Usability, Fairness and Bias, Structure and Organization, and Governance. The tool enables data teams to identify issues early, prioritize remediation, and make informed modeling decisions. AIDRIN is accessible as a web application, a Python package on PyPI, and openly developed on GitHub for community use and contribution, making it flexible for various workflows. Its interactive visualizations and interpretable reports help both technical and non-technical users understand dataset strengths and weaknesses. We extend AIDRIN by adding a customizability module, allowing users to define their own metrics and remedies to evaluate and prepare data for AI.

ACM Reference Format:
Kaveen Hiniduma, Jean Luca Bez, Ravi Madduri, and Suren Byna. 2025. AIDRIN: A Comprehensive Toolset for Automating Data Preparation for AI. In . ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

AI systems have both the strengths and weaknesses of their training data. Despite the widely recognized principle of "garbage in, garbage out," there is no standardized, end-to-end process for assessing whether a dataset is AI-ready [6]. Practitioners often rely on multiple independent tools: one for quality checks, another for fairness, another for metadata compliance, lacking a unified assessment.

AIDRIN provides an integrated solution that evaluates datasets across six pillars, including *Quality*, *Impact on AI*, *Understandability and Usability*, *Fairness and Bias*, *Structure and Organization*, and *Governance* by converting diverse dimensions of readiness into quantifiable metrics and visual reports. Through its customizability module, it further supports application-specific aspects by allowing users to define their own metrics and remedies to help data teams assess readiness for AI, reduce downstream risks, and communicate issues effectively.

2 RELATED WORK

Existing tools for assessing data readiness typically focus on specific dimensions. General-purpose data quality tools like Deequ [10] focus on profiling and constraint validation, while AI-specific solutions, such as the Data Nutrition Label [8], emphasize metadata but lack comprehensiveness. Specialized frameworks, like AIF360 [3] for fairness or FAIR evaluation tools [4, 5], focus on individual dimensions but fail to integrate multiple aspects. AIDRIN stands out by synthesizing these dimensions into a cohesive, quantifiable framework to offer interpretable metrics and visualizations for a complete assessment of AI data readiness.

3 AIDRIN'S SIX PILLARS OF DATA READINESS

AIDRIN evaluates datasets through six interconnected pillars, ensuring a multi-dimensional readiness profile. Furthermore, a customizability module enables users to define application-specific metrics, which can be categorized into any of the six pillars:

(1) **Quality:** Assesses completeness, accuracy, consistency, validity, and the presence of outliers or duplicates to ensure reliable inputs for AI models.
(2) **Impact on AI:** Evaluates feature relevance, correlations, and data distributions to identify variables likely to enhance predictive performance.
(3) **Understandability and Usability:** Measures compliance to FAIR principles (Findable, Accessible, Interoperable, Reusable) to ensure datasets are user-friendly and interoperable.
(4) **Fairness and Bias:** Analyzes class imbalance, representation rates, and statistical biases to mitigate fairness risks.
(5) **Structure and Organization:** Examines data types, schema structure, and access performance to support seamless integration into AI pipelines.

*These authors contributed equally to this work

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

(6) **Governance:** Quantifies privacy measures, authentication protocols, and adherence to ethical data collection standards to meet regulatory requirements.

This multi-pillar approach provides a comprehensive readiness report that allows data teams to address weaknesses systematically. Apart from the standard metrics, AIDRIN also offers customizability, allowing users to incorporate their metrics for data evaluation, as most metrics are AI-agnostic and can be adapted to specific datasets or organizational needs. For example, in medical imaging applications, a metric that assesses specific sensor calibration is not AI-agnostic. By using the customizability module, users can define such domain-specific metrics and ensure that readiness assessments accurately reflect application-specific requirements.

4 SYSTEM DESIGN AND ACCESSIBILITY

AIDRIN allows users to upload datasets and metadata, select relevant metrics, and receive interactive reports featuring visualizations like correlation heatmaps, feature importance charts, class balance dashboards, and FAIR compliance summaries. Each metric is accompanied by clear definitions and documentation to ensure traceability and usability for both technical and non-technical users. AIDRIN's open-source nature and customizable framework allow users to tailor metrics to specific AI use cases. This feature has proven effective not only in centralized AI settings but also in federated learning contexts, which was demonstrated in our prior research [7].

5 EVALUATION AND PERFORMANCE

AIDRIN was evaluated on diverse public datasets and their corresponding metadata (e.g., [1, 2, 9, 11]), demonstrating efficient metric computation and high usability. The tool's unified dashboard enabled users to identify data issues early, streamline cleaning and transformation tasks, and communicate findings effectively to stakeholders.

As an example, Figure 1 shows a visualization from the AIDRIN web application illustrating the FAIR compliance evaluation for the [11] metadata. A full demonstration of this functionality, including comprehensive results, is available in [12]. The FAIR assessment is used to evaluate the *Understandability and Usability* dimension of data readiness as it ensures that data is findable, accessible, interoperable, and reusable. This component of AIDRIN was inspired by best practices from the ESS-DIVE repository [5].

6 FUTURE DIRECTIONS

AIDRIN's key strengths include its multi-dimensional approach, which balances data readiness and its interpretable outputs, which cater to diverse audiences. Its customizability allows users to define task-specific metrics to enhance flexibility. However, AIDRIN currently focuses on structured data formats such as CSV, NPZ, JSON, HDF5, and XLSX, with ongoing efforts to support multimodal datasets (e.g., text, images, audio) and also very large datasets. Some governance and structural metrics are still in development. A composite AIDRIN scoring model is in the works to summarize readiness while preserving detailed insights across pillars. Future work will expand compatibility and introduce metrics to establish AIDRIN as a standard for responsible AI data management.

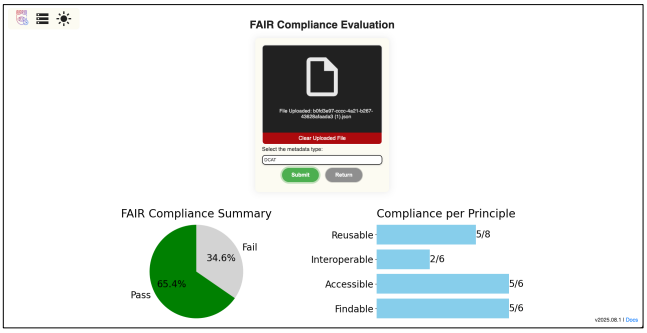

Figure 1: Screenshot of the AIDRIN web application showing FAIR compliance evaluation for the[11] metadata. This assessment reflects the *Understandability and Usability* dimension in the multi-pillar framework.

7 CONCLUSION

AIDRIN offers a unified, multi-pillar framework to assess AI data readiness by integrating quality, fairness, usability, structure, governance, and impact on AI into interpretable metrics and visual reports. Its customizability module allows users to define domain-specific metrics and remedies, making the framework adaptable across diverse applications. AIDRIN lays the foundation for a flexible and practical standard in responsible AI data management.

REFERENCES

[1] Mohammad Aljubran and Roland N. Horne. 2025. Stanford Thermal Earth Model for the Conterminous United States. Dataset, Stanford University, Department of Energy; accessed via Data.gov. https://catalog.data.gov/dataset/stanford-thermal-earth-model-for-the-conterminous-united-states-b1c34 Metadata updated on January 20, 2025; licensed under Creative Commons Attribution.

[2] Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

[3] Rachel K. E. Bellamy et al. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *arXiv preprint arXiv:1810.01943* (2018).

[4] A. Devaraju and R. Huber. 2023. F-UJI: An Automated FAIR Data Assessment Tool. https://doi.org/10.5281/zenodo.6361400 Accessed: 2023-08-17.

[5] S. Cholia et al. 2024. ESS-DIVE Overview: A Scalable, User-Focused Repository for Earth and Environmental Science Data. https://ess-dive.lbl.gov/

[6] Kaveen Hiniduma, Suren Byna, and Jean Luca Bez. 2025. Data Readiness for AI: A 360-Degree Survey. *ACM Comput. Surv.* 57, 9, Article 219 (April 2025), 39 pages. https://doi.org/10.1145/3722214

[7] Kaveen Hiniduma, Zilinghan Li, Aditya Sinha, Ravi Madduri, and Suren Byna. 2025. CADRE: Customizable Assurance of Data Readiness in Privacy-Preserving Federated Learning. In *Proceedings of the 2025 IEEE 21st International Conference on e-Science (eScience).* to appear.

[8] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. (2018). arXiv:arXiv:1805.03677 [cs.DB]

[9] Pacific Northwest National Laboratory. 2025. EGS Collab Experiment 1 Stimulation Data. Dataset, Pacific Northwest National Laboratory, Department of Energy. https://catalog.data.gov/dataset/egs-collab-experiment-1-stimulation-data-0593b Metadata updated January 20, 2025; licensed under Creative Commons Attribution 4.0. DOI: 10.15121/1651116.

[10] Sebastian Schelter, Dustin Lange, Philipp Schmidt, Meltem Celikel, Felix Biessmann, and Andreas Grafberger. 2018. Automating Large-Scale Data Quality Verification. *Proceedings of the VLDB Endowment* 11, 12 (Aug. 2018), 1781–1794.

[11] Paul Siratovich. 2025. GOOML Big Kahuna Forecast Modeling and Genetic Optimization Files. Dataset, Upflow, Department of Energy, reproduced via Data.gov. https://catalog.data.gov/dataset/gooml-big-kahuna-forecast-modeling-and-genetic-optimization-files-01985 Metadata updated January 20, 2025. DOI: https://doi.org/10.15121/1812319.

[12] YouTube. 2025. AIDRIN FAIR compliance evaluation. https://youtu.be/GfnaBPU6R64. YouTube video, published.