AI Data Readiness Inspector (AIDRIN) for Quantitative Assessment of Data Readiness for AI

Kaveen Hiniduma
Suren Byna
hiniduma.1@osu.edu
byna.1@osu.edu
The Ohio State University
Columbus, Ohio, USA

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

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

ABSTRACT

Garbage In Garbage Out is a universally agreed quote by computer scientists from various domains, including Artificial Intelligence (AI). As data is the fuel for AI, models trained on low-quality, biased data are often ineffective. Computer scientists who use AI invest a considerable amount of time and effort in preparing the data for AI. However, there are no standard methods or frameworks for assessing the "readiness" of data for AI. To provide a quantifiable assessment of the readiness of data for AI processes, we define parameters of AI data readiness and introduce AIDRIN (AI Data Readiness INspector). AIDRIN is a framework covering a broad range of readiness dimensions available in the literature that aid in evaluating the readiness of data quantitatively and qualitatively. AIDRIN uses metrics in traditional data quality assessment such as completeness, outliers, and duplicates for data evaluation. Furthermore, AIDRIN uses metrics specific to assess data for AI, such as feature importance, feature correlations, class imbalance, fairness, privacy, and FAIR (Findability, Accessibility, Interoperability, and Reusability) principle compliance. AIDRIN provides visualizations and reports to assist data scientists in further investigating the readiness of data. The AIDRIN framework enhances the efficiency of the machine learning pipeline to make informed decisions on data readiness for AI applications.

KEYWORDS

Data readiness metrics, data quality assessment, Data readiness for AI, FAIR principles

ACM Reference Format:
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 (SSDBM).* ACM, New York, NY, USA, 12 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

In the age of decision-making guided by data, the quality and readiness of datasets are critical to achieving success with artificial intelligence (AI) applications. As the reliance on AI continues to grow across various domains, the need for a complete tool to assess and ensure the suitability of datasets is important.

The current state of data evaluation tools presents a lack of comprehensive solutions. Many tools concentrate on specific aspects of data quality ([17, 30, 46, 47]), while others emphasize some factors of data readiness ([5, 25, 29]). These tools have improved our capabilities in assessing datasets. However, their disjoint nature creates challenges. Often, users should rely on multiple tools to evaluate different dimensions of data readiness, leading to inefficiencies and inconsistencies in the evaluation process. Additionally, the lack of standardized metrics and methodologies in these tools restricts the establishment of benchmarks for data assessment. As a result, there remains a need for a unified and comprehensive framework to address these drawbacks and to provide a quantitative evaluation of data readiness for AI applications.

To address this gap, we introduce AIDRIN (AI Data Readiness INspector), a system designed to comprehensively evaluate datasets through a diverse range of metrics, giving an overall perspective on their readiness for AI applications. While existing data quality tools [17, 30, 46, 47], data readiness frameworks [5, 25, 29], and data readiness dimension-specific (e.g., fairness, FAIR (Findability, Accessibility, Interoperability and Reusability) compliance) evaluation frameworks [8, 19, 43] have made significant contributions, AIDRIN distinguishes itself by addressing the limitations in these tools. Many current systems focus on specific dimensions, leaving gaps in the overall assessment. AIDRIN bridges these gaps by incorporating a set of metrics covering traditional data quality parameters, such as completeness, duplicates, outlier evaluations, AI-related metrics such as fairness considerations, privacy measures, feature relevance metrics, class imbalance evaluations, and broader FAIR principle compliance metrics. This comprehensive approach sets AIDRIN apart, providing users with a unified system for assessing the readiness of datasets for AI applications.

AIDRIN aims to simplify dataset evaluation by offering user-friendly data and metadata upload capabilities and simplifying the process for data practitioners and researchers. Its standout feature is in its scoring mechanism, derived from an extensive list of metrics covering a broad range of data readiness considerations. By filling the gap in comprehensive dataset evaluation shown by existing systems, AIDRIN allows users across diverse domains to measure dataset suitability for AI applications easily and efficiently.

In this paper, we describe the design and implementation of AIDRIN, a comprehensive data evaluation toolkit for measuring AI readiness of data. In the remainder of the paper, we describe the

SSDBM, July 10–12, 2024, Rennes, France

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

related work (§2), definition of AI readiness (§3), the metrics and the AIDRIN framework (§4), and its evaluation (§5).

2 RELATED WORK

In this section, we review existing toolkits and frameworks, specifically focusing on those designed for assessing the quality and readiness of data, with a particular focus on applications of AI. While the literature lacks research explicitly dedicated to data readiness for AI, we categorize the existing works into three main groups: toolkits designed for evaluating general data quality, toolkits explicitly targeting to assess data readiness, and toolkits designed to address specific domains or dimensions of readiness, such as FAIR compliance and bias.

Data Quality Toolkits. Within the area of data quality toolkits, Informatica's Data Quality Tool [30] is an open-source solution designed for data profiling, cleansing, and monitoring capabilities. It provides metrics for data completeness, accuracy, consistency, and reliability. Similarly, DQLearn [47] stands out as a framework designed for organized quality operations. It systematically addresses four main tasks: detecting data quality issues with the aid of metrics, correcting identified problems, implementing rules for custom assessment using metrics, and explaining data quality aspects. In designing AIDRIN, we incorporate a comprehensive range of metrics covering not only traditional data quality parameters but also offering assessments of AI readiness using metrics such as data bias, privacy, feature relevance, correlation, and FAIR compliance.

Deequ [46] is another data quality assessing library designed for defining "unit tests for data," enabling the measurement of data quality in large datasets. It allows users to set explicit assumptions about their data, such as attribute types, absence of null values, and more, in the form of data quality checks. These checks can then be verified on a sample of data to identify errors by defining checks on various properties of data early in the data pipeline, before feeding it to consuming systems or machine learning algorithms. In contrast, AIDRIN stands out by offering a comprehensive approach to ensure the readiness of data for machine learning tasks, encompassing a wider range of considerations beyond data quality checks alone. Additionally, in AIDRIN we offer a user-centric approach in which the users can select their assessment criterion and easily visualize the assessment results for improved interpretation.

Data Readiness for AI Toolkits. The Data Nutrition Label [29] framework offers a standardized format for presenting essential information about datasets, including metadata, provenance, variable descriptions, statistics, pair plots, synthetic data generated from probabilistic models, and ground truth correlations. Gupta et al. [25] designed a toolkit to provide automated explanations across various dimensions of data quality, such as metadata data information, data provenance, simple statistics of data features, and linear correlations between features to simplify data preparation and improve the quality of training data for AI. To accommodate varying technical skills among practitioners, the design prioritizes user-friendly features and visualizations. The toolkit increases overall productivity by allowing data specialists to efficiently examine and choose datasets to accelerate AI model building and deployment. Data Readiness Report [5] focuses on addressing challenges in data preprocessing within machine learning pipelines. The proposed solution generates helpful documentation for a data quality and readiness assessment. This report offers detailed insights into data quality across standardized dimensions, documenting properties, quality issues, and data operations by various individuals.

Table 1: Data Quality Assessment Tools

Tool	Completeness	Outliers	Duplicates	Privacy	Fairness	FAIR Compliance	Feature Correlations	Feature Relevancy	Class Imbalance
Informatica [30]	✓								
DQLearn [47]	✓	✓	✓						
Gupta et al [25]	✓	✓	✓		✓		✓	✓	✓
Data Readiness Report [5]	✓	✓				✓	✓		
AI360 [8]					✓				
FAIR Cookbook [43]						✓			
FAIRassist [19]						✓			
ESS-DIVE FAIR [18]						✓			
AIDRIN	✓	✓	✓	✓	✓	✓	✓	✓	✓

We designed AIDRIN to quantify readiness factors and produce relevant visualizations, significantly simplifying the analysis process by reducing the time and effort required by data specialists. Notably, AIDRIN's focus on addressing emerging concerns like privacy, fairness, and FAIR compliance assessments emphasizes its commitment to overcoming evolving challenges within the field. This approach not only enhances the efficiency and effectiveness of data analysis but also highlights AIDRIN's goal of promoting responsible and ethical practices in data-driven challenges.

Domain Specific Toolkits. In the recent commitments towards obtaining fairness in AI models, considerable attention has been directed towards the evaluation of the fairness of data before its application in AI systems. One of the tools addressing this aspect is IBM's AI Fairness 360 toolkit [8] (AIF360), an open-source software designed to overcome biases in ML models. The toolkit incorporates bias detection metrics such as representation and statistical rates of sensitive attributes, allowing users to assess and address biases related to their specific circumstances.

Recent research has focused on providing FAIR (Findable, Accessible, Interoperable, and Reusable) data, with a focus on developing toolkits that evaluate the level of FAIR compliance in data. FAIR Cookbook [43] addresses challenges in implementing FAIR principles in scientific data management, recognizing the nature of FAIR principles and the absence of a formalized standard. The cookbook evaluates the costs and benefits associated with FAIR data, highlighting the value of operational changes in delivering FAIR data and services. In parallel, FAIRassist [19], an educational component of the FAIR sharing resource, guides users on how to measure and enhance the FAIR compliance of digital objects. ESS-DIVE [18] is a data repository for earth sciences data, which evaluates the FAIR principle compliance for the data being published. ESS-DIVE uses quantitative assessments based on metrics such as search performance, metadata completeness, and user feedback.

Despite the availability of various tools and frameworks that tackle different pieces of the AI readiness evaluation, no single

system evaluates all the aspects using quantitative metrics and visualizations. Towards addressing this gap, we present AIDRIN, a comprehensive framework that addresses various dimensions of data readiness, including FAIR principles compliance. In Table 1, we compare various features of existing tools with AIDRIN to highlight the scope and features offered by our framework.

3 DEFINING DATA READINESS FOR AI

The concept of "Data Readiness for AI" lacks a standard definition and is still evolving. To find a comprehensive definition of data readiness for AI, we have recently conducted a survey of existing metrics and partial definitions [27]. In our effort to design a comprehensive framework and metrics for AI readiness, we propose seven categories of evaluation and several metrics in each category. As a standard for defining AI readiness of data is still evolving, we anticipate a few changes to this classification. Given below is our proposal of categories of data readiness for AI.

Categories of Data Readiness for AI

- **Quality**
  - **Completeness**: Measures the presence of all required data.
  - **Outliers**: Identifies anomalous data points that deviate from the norm.
  - **Duplication**: Evaluates the presence of duplicate records.
  - **Data preparation practices**: Assesses the robustness of the methods used to prepare the data.
  - **Timeliness**: Ensures data is up-to-date and relevant.
- **Understandability**
  - **Metadata availability and quality**: Ensures comprehensive metadata is present to describe the dataset.
  - **Provenance**: Tracks the origin and lineage of the data, ensuring accuracy and completeness.
  - **User interfaces for data access**: Evaluates the ease of accessing and interacting with the data.
- **Structural quality**
  - **Used data types**: Assesses the appropriateness and consistency of data types.
  - **Quality of data schema**: Evaluates the design and structure of the data schema that supports normalized forms and fast data storage and access.
  - **File format and used data storage system**: Reviews the efficiency and suitability of file formats and storage systems.
  - **Data access performance**: Measures the speed and reliability of data retrieval.
- **Value**
  - **Feature importance**: Assesses the significance of different features within the dataset.
  - **Labels**: Examines the availability, quality, and correctness of labels for supervised learning.
  - **Data point impact**: Evaluates the influence of individual data points on the overall dataset.
  - **Uncertainty in data**: Measures uncertainty or confidence in the data using uncertainty quantification methods.
- **Fairness and bias**
  - **Class imbalance**: Assesses the distribution of classes within the dataset.
  - **Class separability**: Measures how well different classes can be distinguished.
  - **Discrimination index**: Identifies potential biases in the data.
  - **Population representation**: Ensures diverse and representative sampling of the population.
- **Governance**
  - **Collection**: Reviews consent, sampling methods, ethical considerations, regulatory compliance, and funding sources.
  - **Processing and curation**: Assesses anonymization, curation, and de-identification methods used.
  - **Application**: Evaluates usage restrictions and potential biases in data analysis.
  - **Security**: Reviews data sensitivity, access control mechanisms, and sharing protocols.
  - **Privacy**: Assesses privacy requirements, budgets, and scores.
- **AI application-specific metrics**
  - **Model-specific metrics**: Evaluates metrics specific to the AI models and their intended applications, ensuring the data meets the requirements for successful model training and deployment. Users can define data metrics that are specific to the AI model they are developing.

4 AI DATA READINESS INSPECTOR (AIDRIN)

Based on the definition of AI readiness of data described above, we designed AIDRIN as a comprehensive framework to provide data assessment metrics. In this section, we provide the details of metrics and visualizations provided in AIDRIN. We highlight the specific metrics integrated to assess data readiness across various dimensions. Among the categories described in Section 3, we provide Quality, Understandability (using FAIR principle compliance), Value, Fairness, and Bias metrics in AIDRIN. While the privacy aspect of Governance is already integrated into AIDRIN, other components addressing Governance and Structural Quality are currently in development.

4.1 Analysis Capabilities of AIDRIN

In Figure 1, we highlight the current design of AIDRIN. For the data and metadata users enter as input, AIDRIN analyzes and provides three types of inspection results: summary statistics of data, data readiness metrics, and visualizations. The summary statistics provide general information regarding the data, such as the number of attributes and data records, statistics such as percentiles, min, max, average, standard deviation, and distribution. Users can then dive deep with selected attributes (features) to inspect a plethora of AI readiness metrics, as shown in Table 1. We demonstrate metrics with corresponding plots and visualizations. Each metric also accompanies the definition and descriptions for ease of understanding. We describe the various metrics and visualizations below.

4.1.1 Completeness. Completeness signifies the existence of necessary data and attribute values within a dataset, indicating the extent to which data points or entries fully exist with relevant attribute values. In this study, we use the completeness metric proposed by Blake et al. [9] to assess this dimension of readiness. As defined by Blake et al., completeness refers to the presence of missing values in a dataset. AIDRIN quantifies this by measuring the proportion of missing values within each feature of the dataset.

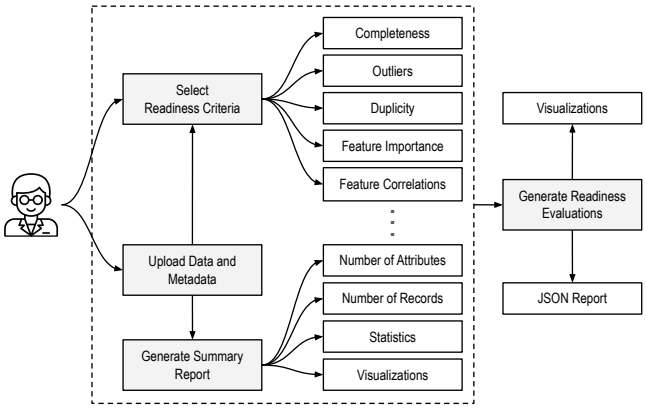
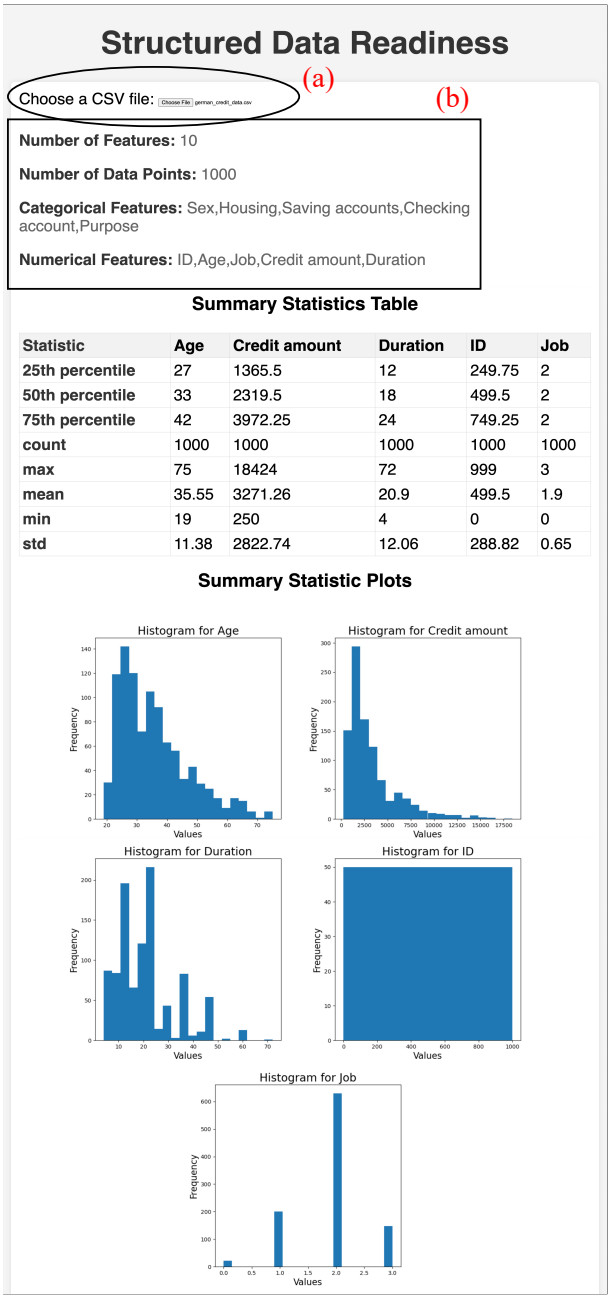


Figure 1: An overview of the AIDRIN workflow. The AI readiness assessment starts by processing data and metadata to generate summary statistics. Users can then select metrics across different data readiness dimensions for evaluation. AIDRIN generates corresponding visualizations and a comprehensive report of the selected metrics, aiding in data readiness analysis and decision-making.

4.1.2 Outliers. An outlier in a dataset denotes a data point or instance that deviates significantly from the anticipated values within the dataset. There are other metrics discussed in the literature to evaluate outliers like Local Outlier Factor (LOF) by Breunig et al. [11] and Incremental Local Outlier Factor (ILOF) by Pokraja et al. [41]. Although LOF and ILOF methods have been effective in identifying outliers by leveraging local density information, the Interquartile Range (IQR) method appears to be a better choice under various circumstances. The IQR method is a statistical tool used to assess the dispersion of data in a dataset. It centers on the middle 50% of the data, calculated as the range between the first quartile and the third quartile. By computing the IQR as the difference between Q1 and Q3, the method establishes bounds for potential outliers. Data points lying beyond these bounds, determined by a user-defined constant k (usually 1.5), are identified as outliers. This approach is especially valuable when dealing with datasets open to extreme values that could distort traditional measures of central tendency. The resulting outlier score, ranging from 0 to 1, quantifies the proportion of outliers in each feature of the dataset, providing a robust assessment of data variability.

he IQR method is more reliable for detecting outliers because it is less affected by extreme values than the LOF method. The simplicity and interpretability of the IQR method, based on quartiles, make it more accessible to both experts and non-experts in statistics. Additionally, the IQR method's applicability to non-normally distributed data and its consistent performance across diverse datasets highlight its flexibility. Unlike LOF, which may require careful parameter tuning, the IQR method is parameter-free, reducing sensitivity to parameter choices. These advantages position the IQR method as a reliable and straightforward option for outlier detection, particularly when dealing with datasets with varied distributions and potential outliers. Hence, AIDRIN uses IQR to assess the proportion of outliers in each feature of the dataset.



Statistic	Age	Credit amount	Duration	ID	Job
25th percentile	27	1365.5	12	249.75	2
50th percentile	33	2319.5	18	499.5	2
75th percentile	42	3972.25	24	749.25	2
count	1000	1000	1000	1000	1000
max	75	18424	72	999	3
mean	35.55	3271.26	20.9	499.5	1.9
min	19	250	4	0	0
std	11.38	2822.74	12.06	288.82	0.65

Figure 2: Interface for uploading files and summarizing dataset statistics in AIDRIN. Users can import datasets and view key statistical summaries, the dimensions of the dataset, and the numerical and categorical features.

4.1.3 Duplicates. Duplicates in a dataset refer to the existence of duplicate or redundant instances, potentially distorting the AI modeling process by introducing repetitive data entries. Bors et al. [10] introduce a scoring system for detecting duplicate entries that rely on uniqueness. A score of 1 (true) is assigned to values with a unique combination in the chosen columns, while a score of 0 (false) is assigned to values appearing multiple times. AIDRIN

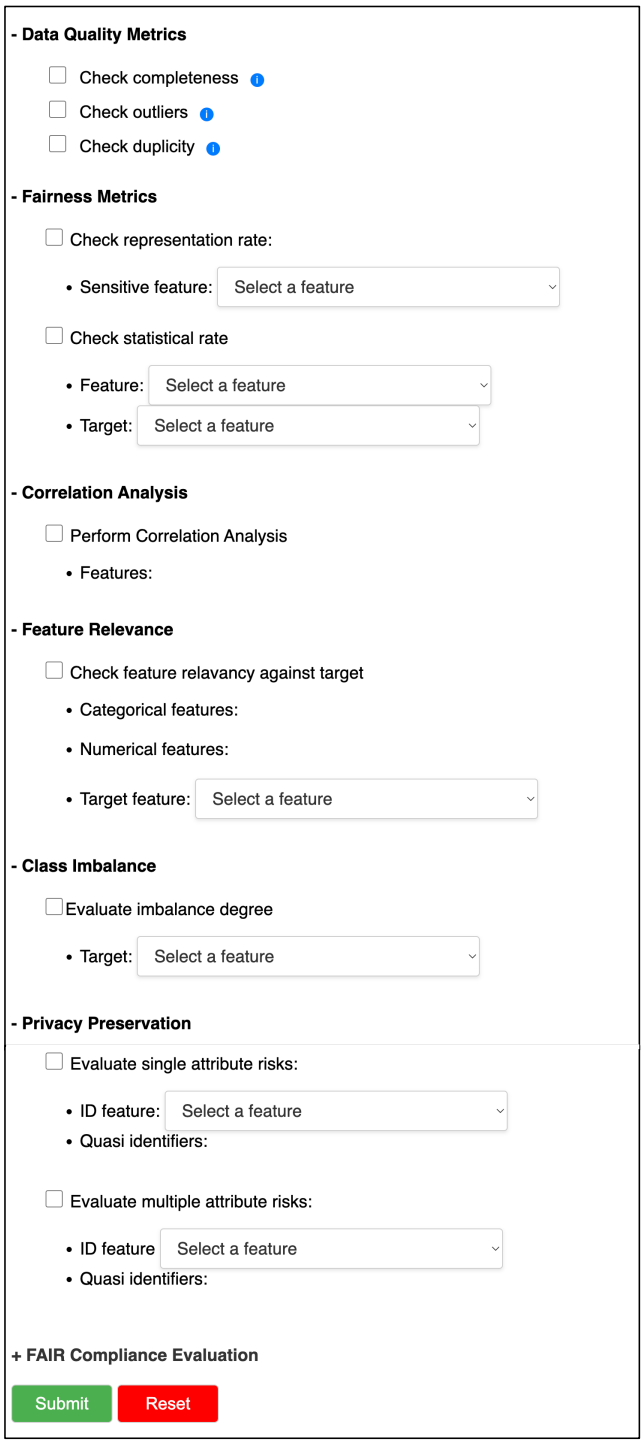


Figure 3: AIDRIN enables a tailored evaluation of the data and models through customized analysis.

uses this method to examine the number of identical items present in datasets. By comparing the number of unique items to the total number of items, it generates a single score indicating the level of duplication throughout the dataset.

4.1.4 Privacy. A privacy measure is a metric or method used to quantify the level of privacy protection in a given context, especially in the handling and release of sensitive data. It aims to assess the risk of re-identification of personally identifiable information in datasets. In the context of Vatsalan et al.'s work [50], privacy is focused on reducing the re-identification risks in datasets in the education sector. The authors propose the "MM risk score", which uses the Markov model to quantify re-identification risks. Unlike existing approaches that often assume prior knowledge, this method considers all available information in the dataset, incorporating event-level details that link multiple records to the same individual, as well as investigating correlations among attributes. Markov model assesses the relationships among various features present in the dataset. By analyzing the transitions between the features, it identifies the correlations across the features that could lead to potential re-identifications of individual data records in the dataset. The event-level details capture the sequential occurrences between different records of the same individual. For example, three separate features analyzed collectively could disclose sensitive information about an individual compared to considering only a single attribute. AIDRIN uses this approach for assessing and managing re-identification risks, emphasizing a more detailed analysis of dataset information to improve privacy protection.

Comparing this to other privacy metrics in the literature, Vatsalan et al.'s method offers distinct advantages. Unlike metrics such as SHAPr introduced by Duddu et al. [16] and privacy risk metric by Song et al.'s [48], which is tied to assessing privacy risks in the context of a specific AI model, Vatsalan's approach takes a more collective view. Using the Markov model, it considers complex relationships within the data, going beyond model-specific considerations. This broader perspective allows for a comprehensive evaluation of re-identification risks, making it potentially applicable to various datasets and scenarios.

Furthermore, Vatsalan et al.'s approach offers a preventive strategy compared to metrics like Attack Success Rate (ASR) introduced by Carlini et al. [12], which can only be measured after model training and an attack. The dynamic nature of their method allows for potential risk mitigation strategies to be implemented during the dataset release process, improving its practical usage.

4.1.5 Feature Correlations. Correlation analysis is a fundamental statistical technique used to measure the direction and strength of the correlation between two quantitative variables. It quantifies changes in one variable correspond to changes in another, with the correlation coefficient being a crucial metric to measure the degree of their association. The coefficient ranges from $[-1, 1]$, with 1 indicating a strong positive correlation, -1 indicating a strong negative correlation, and 0 suggesting no linear relationship.

Three well-known methods for measuring correlations include the Pearson correlation coefficient [21], Theil's U [49], and Cramer's V [14]. The Pearson coefficient is well-suited for continuous variables, capturing linear relationships. Theil's U, designed for categorical variables, incorporates information theory concepts, while Cramer's V extends this idea to measure association strength.

While both Theil's U and Cramer's V are suitable for categorical data, Theil's U often has an advantage. Theil's U is particularly robust due to its foundation in information theory, providing a complete measure of association. It has a scoring system, ranging from

0 (no association) to 1 (perfect association). Theil's U is preferred over Cramer's V for its ability to handle unequal distributions and its consideration of conditional probabilities. Theil's U is also asymmetric, meaning the association between the features X and Y may differ from that between Y and X. This is because it considers the direction of information flow and the conditional probabilities between the features. Therefore, the correlation matrices it generates will observe asymmetry as Theil's U's characteristic of considering directional relationships. In instances where categorical variables exhibit different levels of entropy, Theil's U provides more reliable and informative results.

AIDRIN utilizes the Pearson correlation coefficient to gauge the strength and direction of numerical feature correlations. On the other hand, for the assessment of categorical feature correlations, AIDRIN uses Theil's U. This dual approach reflects the flexibility of AIDRIN's correlation analysis, enabling a more meaningful examination of the complex inter-dependencies present in the data.

4.1.6 Fairness and Bias. Fairness in a dataset refers to the equal treatment of different groups or individuals, particularly concerning sensitive attributes such as race, gender, or socioeconomic status. According to Mehrabi et al. [36] unfair biases in datasets can lead to discriminatory outcomes when machine learning models are trained on them, potentially extending and increasing existing societal inequalities. To ensure ethical AI applications, it is necessary to assess and address these biases in the data.

Several metrics have been proposed to quantify and evaluate fairness in datasets. The Difference and P-Difference metrics, introduced by Azzalini et al. [7], compare the confidence of a dependency with and without the consideration of sensitive attributes to evaluate bias, identifying important attributes contributing to unfairness.

Feldman et al. [20] introduced the metric Likelihood Ratio (LR_+), which assesses the disparate impact in a dataset based on specificity and sensitivity. Celis et al. [13] contributed two metrics: the representation rate, which quantifies fairness in representing different attribute values, and the statistical rate, where the fairness evaluation is conducted by analyzing the conditional probabilities of class labels based on attribute values.

In the context of AIDRIN, the representation rate and statistical rate emerge as effective metrics for measuring group fairness. The representation rate assesses the distributions of different sensitive attributes in the dataset. The statistical rate, on the other hand, evaluates fairness through conditional probabilities, ensuring that the target representations do not discriminate certain groups based on sensitive attributes. In addition to their effectiveness in measuring group fairness, the representation rate and statistical rate metrics stand out for their ease of understanding and visualization. These qualities contribute to their practical advantage in guiding decision-makers and stakeholders toward a clearer understanding of the fairness issues within the dataset.

However, many traditional unified statistical rates are limited to binary-sensitive attributes, leaving non-binary attributes unaddressed. To tackle this limitation, we introduce the Target Standard Deviation (TSD) metric. This metric goes beyond binary groups by considering the average differences across all possible subgroups for a given target. Capturing the average differences across all sensitive groups for a given target considers variations that may exist within non-binary attributes. Moreover, its unified scalar value facilitates straightforward and interpretable evaluations, enabling stakeholders to compare and prioritize decisions effectively. However, the TSD metric may not capture contextual factors and proxies of the sensitive attribute within the dataset. It also cannot consider multiple sensitive attributes simultaneously, limiting its applicability.

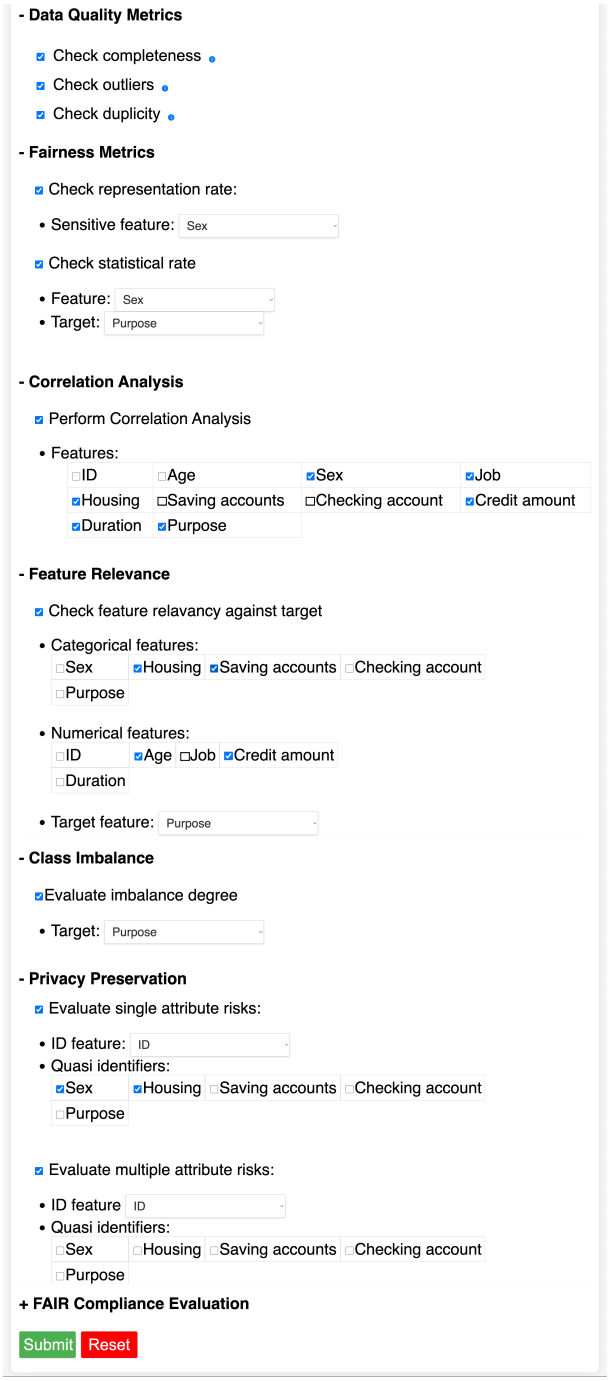
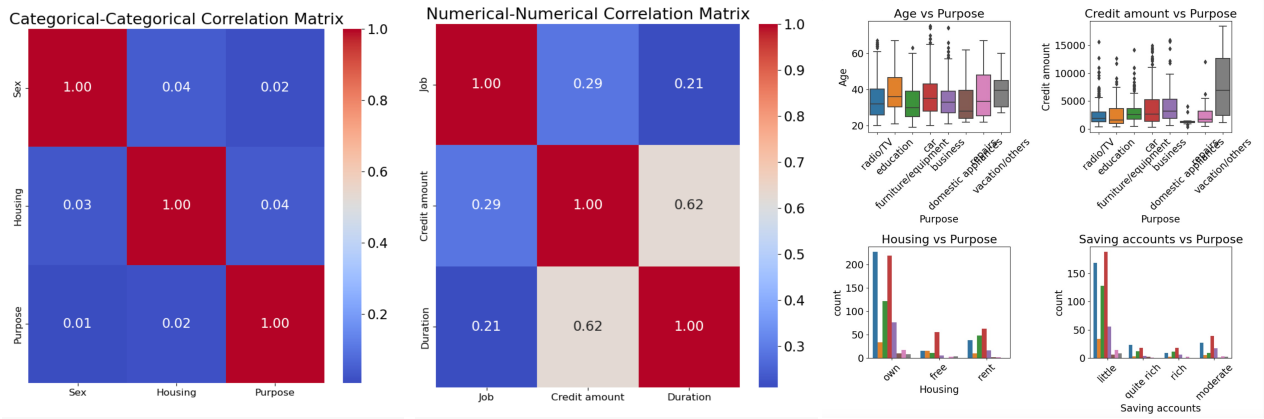


Figure 4: AIDRIN user interface for selecting metrics for the German Credit Datasets [28] dataset.



Categorical correlations are calculated using Theil's U, with values ranging from 0 to 1. A value of 1 indicates a perfect correlation, while a value of 0 indicates no correlation

Numerical correlations are calculated using Pearson's correlation coefficient, with values ranging from -1 to 1. A value of 1 indicates a perfect positive correlation, -1 indicates a perfect negative correlation, and 0 indicates no correlation

For numerical target column, scatter plot is generated between each feature and the target variable. For categorical target column, box plot is generated between each feature and the target variable.

Figure 5: A sample of the visualizations generated by AIDRIN based on user-selected metrics during SGC dataset evaluation.

Consider the categorical sensitive attribute provided by the user (that exists in the dataset) as $A \in [1, N]$ with N possible values (e.g., ethnicity), and let $Y \in [Y_1, Y_2, Y_3...]$ represent the targets, respectively. To formulate the TSD of Y_1:

$$\text{TSD} = \sqrt{\frac{1}{N}\sum_{n=1}^{N}(Pr\,(Y = Y_1|A = n) - \mu)^2} \tag{1}$$

Here, μ denotes the average target probability across all groups. A lower TSD for a particular target suggests that the groups have similar treatment within that target category. Poulain et al. [42] introduced this concept, where their metric evaluates algorithmic fairness across different sensitive groups by computing the standard deviation of the groups' true positive rates.

4.1.7 FAIR Principle Compliance. Metadata plays a crucial role in managing data. FAIR principles [23] have emerged to guide data publishers to improve the quality of data. AIDRIN assesses the FAIR (Findable, Accessible, Interoperable, Reusable) compliance of a given set of metadata.

AIDRIN currently supports the assessment of FAIR compliance of metadata using two metadata standards: DCAT [3] and Datacite [2]. In the DCAT format, the evaluation of the "Findable" category involves verifying the existence of essential keys associated with identification, description, and title. Special attention is given to descriptive keywords to improve metadata discoverability. Within the "Accessible" domain, the function checks keys related to access levels and publisher information. "Interoperability" is examined through keys that enhance compatibility with established standards. Under the "Reusable" principle, it examines keys associated with licensing and other keys that make the data trustworthy for reuse.

In Table 2, we show the DCAT elements divided into each subcategory under the FAIR principles [23]. It categorizes metadata keys into specific aspects of FAIR, counts the checks for each category, calculates a FAIR compliance score, and generates visualizations illustrating the distribution of checks across the four aspects. For Findability, AIDRIN checks for six metadata keys, while for accessibility, it examines five metadata keys. Interoperability involves three checks and the data is reusable if the four involving metadata keys under reusability are present. The overall FAIR compliance score is presented as a percentage, reflecting the proportion of fulfilled checks out of the total possible checks. The higher the score, the more FAIR the metadata is considered. This concept was inspired by the work conducted by Cholia et al.[18] in FAIR assessment of ESS-DIVE (Environmental System Science) data. The "Other" category captures keys not falling into predefined FAIR categories, providing a comprehensive metadata analysis beyond the core principles. The function calculates a FAIR compliance score by assessing the fulfillment of checks against the total possible checks, presenting the results in a pie chart as illustrated in the pie chart in Figure 6. This approach offers an overview of metadata quality in adherence to FAIR principles within the DCAT framework. Similarly, AIDRIN supports corresponding FAIR elements in the DataCite standard. The FAIR principle assessment module in AIDRIN is extendable to other metadata standards.

4.1.8 Class Imbalance. Class imbalance in a dataset occurs when the number of instances in one class significantly differs from the number of instances in another class, particularly common in binary classification problems. Hassanin et al. [26] highlighted that class imbalance can lead machine learning models to discriminate their predictions against the minority class. To address this, various metrics have been proposed in the literature to measure and understand the extent of class imbalance.

One such metric is the Imbalance Ratio (IR), introduced by Francisco et al. [6]. This metric provides a numerical representation of the disparity between the majority class instances and minority class instances in a dataset. A higher IR indicates a more imbalanced dataset, while a lower IR suggests a relatively balanced distribution. Another metric, the Imbalance Degree (ID), proposed

Table 2: DCAT FAIR Compliance Categorization

Principle	Subcategory	Elements	Reasoning
Findable	F1	identifier	Uniquely identifies the metadata, ensuring global uniqueness and persistence.
	F2	title, description, keyword, theme	Provide rich metadata, making it easier for users and computers to understand and find relevant datasets.
	F3	identifier	Reiterating the importance of explicitly linking metadata with the data they describe.
	F4	landingPage	Registering and indexing metadata in a searchable resource, enhancing discoverability.
Accessible	A1	distribution, downloadURL	Enable retrieval of data using a standardized protocol, providing a consistent and interoperable access mechanism.
	A1.1	format	Ensures openness, freedom, and universality in implementing the communication protocol.
	A1.2	accessLevel	Include information about authentication and authorization procedures, ensuring secure access when necessary.
	A2	publisher	The publisher information serves as a reference point for users seeking information about the dataset, even if the data is no longer accessible.
Interoperable	I1	format, conformsTo	Contribute to formal, accessible, and broadly applicable knowledge representation.
	I2	format, conformsTo	Support the use of vocabularies aligned with FAIR principles, ensuring compatibility and adherence to standards
	I3	references	Allows metadata to include qualified references, fostering interoperability by linking related datasets.
Reusable	R1	format, description, license	Contribute to rich metadata descriptions, supporting optimal reuse.
	R1.1	license	Ensures that data are released with clear and accessible licensing information, promoting understanding and adhering to usage terms.
	R1.2	programCode, bureauCode	Provides detailed information about the origin and history of the data, enhancing transparency and trust in data reuse.
	R1.3	conformsTo	Ensures that metadata and data follow domain-relevant community standards, facilitating integration and reuse within specific communities.

by Ortigosa-Hernández et al. [40], extends the measurement of class imbalance by considering specific characteristics of the class distribution. However, ID has limitations, including sensitivity to the choice of distance function and potential unreliability in extreme cases. In response to these challenges, the Likelihood Ratio Imbalance Degree (LRID), introduced by Zhu et al. [51], emerges as a novel metric. It uses the likelihood ratio (LR) test, offering a fine measurement of imbalance by comparing the existing class distribution to a balanced distribution. LRID aids researchers in making informed decisions on data preprocessing to mitigate the impact of class imbalance on AI models.

AIDRIN uses the ID as the chosen metric to measure class imbalance. Several factors contribute to the selection of ID over other metrics. ID considers specific characteristics of the class distribution, providing a more accurate representation of class imbalance. This advanced representation allows for a deeper understanding of the dataset's imbalance. The ID is flexible, offering a comprehensive measure of imbalance without relying on complex statistical tests. This simplicity makes it easier to understand and apply, enhancing its practical utility in various machine-learning scenarios. The code for ID is openly available, facilitating easier implementation for researchers and practitioners. This transparency encourages widespread adoption and promotes collaboration in the community. Researchers can readily access and integrate the ID metric into their workflows, easing the process of addressing class imbalance.

4.1.9 Feature Importance. In training AI models on datasets with numerous features, identifying the important features that contribute to the prediction task is necessary. AIDRIN uses Shapley values [33] as a method to measure feature importance, a technique widely used in various existing works (Aas et al. [4] Frye et al. [22] Merrick and Taly [37]). Shapley values provide a means to quantify the impact of each feature on a model's predictions.

AIDRIN uses a systematic method to calculate Shapley values, starting with data preprocessing and cleaning to optimize it for model training. Before the preprocessing and cleaning stage, the user uploads the data and specifies the feature set and target feature. This preprocessing step is automated but not extensively rigorous. It handles tasks like managing missing values, duplicates, and outliers, as well as performing one-hot encoding for specified categorical variables. We emphasize that ensuring clean data during the initial upload stage contributes to better results. Since AIDRIN does not conduct an extensive data cleaning process, the quality of the input data greatly impacts subsequent analyses and model performance. AIDRIN currently supports classification tasks in generating the Shapely values. After the data is transformed to meet the requirements of model training, a Random Forest Regressor is trained to perform the classification task. Our focus was not entirely on maximizing utility but rather on observing the important features most likely to contribute. Therefore, since Shapley values are computed based on the contributions of the features for the prediction task, they play an important role in assessing the influence of each feature on model predictions.

To visualize the impact of selected features on the target variable, AIDRIN offers a range of plots based on the nature of the target variable. For numerical target variables, scatter plots are generated to showcase relationships between numerical features and the target. This visualization aids in identifying patterns and trends within the data. On the other hand, for a categorical target variable, box plots are generated to illustrate the distribution of numerical features concerning different categories of the target (Figure 5 right). These box plots provide insights into the variability of features across different target categories, facilitating a deeper understanding of the feature dynamics. The framework can also generate categorical-categorical bar charts (Figure 5 right) to provide insights into the distribution of categories and their impact on the target variable. These bar charts offer a visual representation of the counts or proportions of each category within the selected features concerning the target variable.

We highlight that AIDRIN's simplified user interaction requires only uploading the data and selecting the readiness criteria. However, the more information provided to AIDRIN, the better its results can be. For instance, evaluating feature importance via Shapely values directly correlates with the thoroughness of data cleaning and

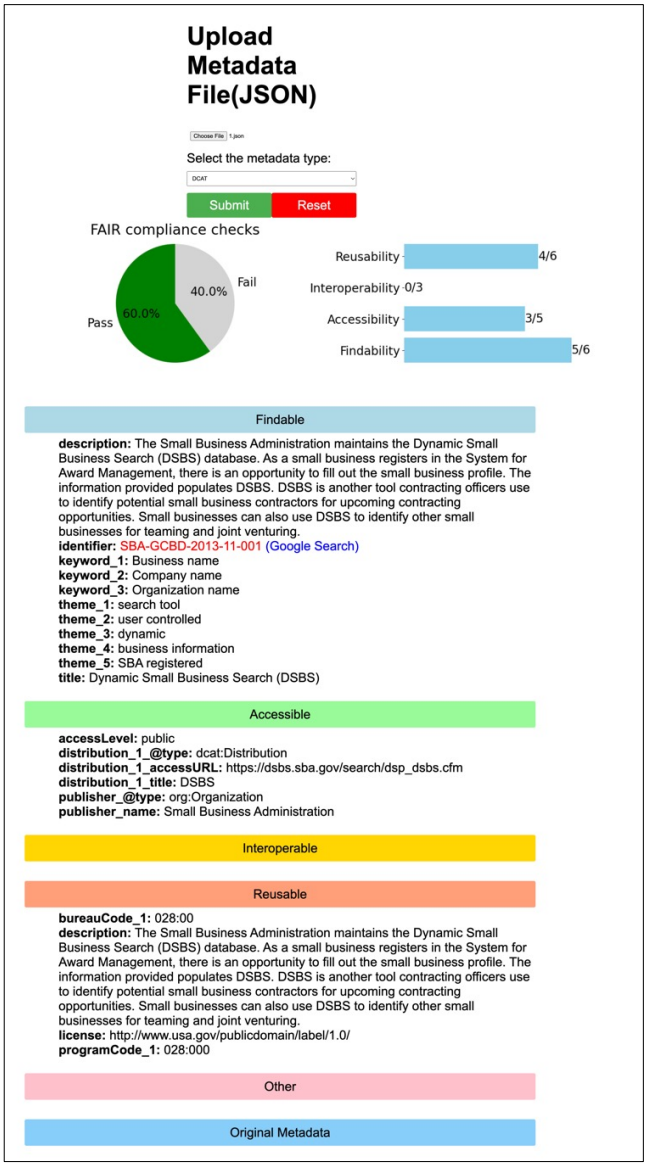


Figure 6: Illustration demonstrating the compliance check for FAIR principles applied to a DCAT data catalog.

preprocessing. Currently, that stage in AIDRIN is limited to removing missing records, managing duplicates and outliers, and one-hot encoding. However, users can expect to achieve a better feature importance analysis by providing thoroughly processed data.

4.2 Implementation of AIDRIN Framework

We have implemented the AIDRIN framework using Python. The essential Python packages required to use AIDRIN are Flask, Pandas, Matplotlib, and Scikit-learn. Flask is the web application framework that powers the AIDRIN user interface component. Pandas handles and processes the datasets efficiently, while Matplotlib enhances AIDRIN with visualization capabilities.

AIDRIN offers multiple ways for users to analyze data. The web interface, developed using HTML and JavaScript, allows users to upload datasets and start the analysis process. Web requests from the user will then launch the Flask server and generate the analysis results and visualizations.

Additionally, AIDRIN offers a PyPI (Python Package Index) package [1] to users who are proficient in Python and comfortable in using environments like Jupyter Notebooks. Users can install the AIDRIN PyPI package via the command line and use it for data readiness assessment. This package can also be used to improve the modularity of AIDRIN by allowing contributors to expand the accessibility of AIDRIN to different domains.

AIDRIN offers flexibility for users to select specific metrics applicable to their evaluation objectives. Subsequently, the system dynamically produces visualizations and scores aligned with the selected metrics. This comprehensive report serves as a valuable insight into the implementation of AI algorithms, providing users with insightful perspectives on their data.

AIDRIN allows users to download the generated report in JSON format, enabling convenient reference and sharing. Figure 1 provides a visual representation of the essential stages within AIDRIN, including data upload (Figure 2 (a)), dimensions of the data including the numerical and categorical attributes (Figure 2 (b)), summary statistics generation (Figure 2), and metric selection (Figure 3). Figure 4 illustrates the metric selection phase, after the data upload is completed, where the attribute names in the dataset are automatically extracted for the user to define the evaluation criteria. In this report, AIDRIN uses the German Credit dataset [28].

4.3 Towards an Aggregate AI Readiness Score

AIDRIN currently has a set of metrics to measure the readiness of data. Our goal is to aggregate a list of these metrics to develop an "AIDRIN Score" that summarizes the readiness of a given dataset for AI. This rating will provide a broad view of the dataset's quality, usability, and suitability for AI applications. It will enable researchers and practitioners to make informed data preparation and usage decisions. By establishing a robust and comprehensive framework, AIDRIN aims to standardize the assessment process. However, because the standards of data readiness for AI are still evolving, we are in the process of defining an "AIDRIN Score".

5 EVALUATION OF AIDRIN

We evaluate AIDRIN to test its performance, usability, and features. To evaluate performance, which is needed for all data analysis tools, we present completion time in generating all metrics using diverse

Table 3: Performance Evaluation of AIDRIN across diverse datasets of varying domains and sizes.

Dataset	Features	Records	Completeness	Duplicates	Outliers	Fairness	Feature Relevance	Correlation Analysis	Class Imbalance	Privacy	Time(s)
RPA [35]	58	782	✓	✓	✓	✓	✓	✓	✓	✓	0.8
SGC [28]	10	1K	✓	✓	✓	✓	✓	✓	✓	✓	0.98
MIDRC [38]	21	60K	✓	✓	✓	✓	✓	✓	✓	✗	1.5
SEHM [34]	10	416K	✓	✓	✓	N/A	✓	✓	N/A	N/A	4.8
MetroPT-3 [15]	17	1.5M	✓	✓	✓	N/A	✓	✓	N/A	N/A	5.3

structured datasets. We have examined the functionality and performance of AIDRIN in varying scenarios, examining its outputs and visualizations. The resulting analysis provides a detailed account of AIDRIN's efficacy, interpreting its flexibility in handling five datasets (shown in Table 3 and showcasing the results that contribute to a better understanding of its capabilities.

We have conducted a user study with the help of various groups of researchers to evaluate the usability of the tool. We evaluate the features of AIDRIN using two case studies, i,e., German Credit (SGC) dataset [28] and Cancer Genome Atlas Lung Adenocarcinoma (TCGA-LUAD) clinical data of patients accessed from the National Cancer Institute NCI Data Portal [24]. For SGC, we used the web interface to produce visualizations, and for the Cancer dataset, we used the Jupyter Notebook interface. We conducted all our experiments on a laptop with a 12-core Apple M2 Max chip with 32GB RAM.

5.1 Performance Evaluation

To evaluate the responsiveness of data analysis tasks in AIDRIN, we used execution time as a metric. The execution time represents both UI-based and Python Notebook interfaces. In Table 3, we show the datasets used in our study and an overview of AIDRIN's performance in assessing a range of data evaluation metrics. This assessment involves four distinct datasets acquired from the UCI Machine Learning Repository [31]: the Single Elder Home Monitoring(SEHM) dataset [34], the MetroPT-3 Dataset [15], the Regensburg Pediatric Appendicitis (RPA) dataset [35], and Statlog - German Credit (SGC) dataset [28]. We also evaluated the MIDRC medical cases dataset [38] provided by NIH. Analysis times of these datasets exhibit a diverse range, with collections of below 1000 records to datasets exceeding 1.5 million records, with feature dimensions ranging from 10 to over 50. This selection of varying-sized datasets comprehensively evaluates AIDRIN's performance, assessing its efficacy across small to large-scale usage scenarios and domains, including healthcare, transportation, and finance.

In our performance evaluation, the most time-intensive aspect of AIDRIN was identified as the computation of Shapley values to evaluate crucial features within a dataset. This is attributed to the necessity of training a random forest regressor for 100 estimators based on user-specified features and target attributes. For instance, when compared to statistical feature relevance methods, the computation time for calculating the Shapley values alone for a single numerical feature ('age') and one categorical feature ('sex') on the SGC dataset took 1.62 seconds, which is almost 2× more. This value increases significantly as the number of features is increased. For two categorical features ('sex' and 'housing') and two numerical features ('age' and 'credit amount'), the Shapley value computation took 5× more time compared to the statistical feature relevance computations. Conversely, from this analysis, the statistical measurements of feature relevance exhibited more efficient processing in comparison. Moreover, the risk scores for privacy assessment exhibited increased time demands with growing dataset sizes.

As shown in the performance evaluations detailed in Table 3, it becomes evident that the time required to generate results gradually increases from 0.8s to 5s as the dataset size increases from 782 records to 1.5 million records. Notably, for the evaluation of privacy alone, the MIDRC datasets required around 106.36 seconds for result generation on one attribute ('sex'). For two attributes ('sex' and 'race'), the privacy evaluation computations required 195.53 seconds to generate the results. As the data dimensions increase, the Markov model requires processing a greater number of possible states and combinations, resulting in longer processing times.

Overall, AIDRIN performs well, even on a laptop. AIDRIN's analysis time depends on the number of tasks, number of data records, and computation resources. Improving the performance of data loading and using advanced computing resources is ongoing.

5.2 Usability of AIDRIN: A User Study

We have conducted a user study involving a diverse group of participants to enhance the presentation of evaluations in AIDRIN. This included three experts, three PhD students, one postdoctoral scholar, and three computer scientists, all with experience in developing data management tools and AI algorithms. This diverse participant group provided valuable insights that allowed us to refine AIDRIN and expand its potential applications.

Our evaluation included an initial phase focused on gathering feedback from experts and PhD students who tested the web interface of AIDRIN. Their assessments primarily centered on functionality, usability, and areas for improvement. Simultaneously, the computer scientists applied AIDRIN in their ongoing projects focused on privacy-preserving federated learning [44], providing practical insights into its effectiveness for assessing data readiness. This phase was important in identifying usability challenges and including enhancements. Based on the feedback, we improved the function of AIDRIN. These included enhancing the user interface and developing a PyPI package to integrate with Python environments, particularly Jupyter Notebooks. These enhancements were designed to address specific user needs identified during the evaluation, thereby enhancing the tool's usability and functionality.

In subsequent evaluations, AIDRIN's improved version was tested again, with a focus on validating these enhancements. The PyPI package, specifically designed based on user feedback, was evaluated during the NCI CRDC Artificial Intelligence Data-Readiness (AIDR) Challenge in 2024 [39], where its integration with Jupyter Notebooks facilitated effective data assessment. This phase confirmed the robustness and applicability of AIDRIN.

Key insights from the user study highlighted interest among participants in integrating AIDRIN into frameworks such as APPFL [32] for federated learning. This reflects AIDRIN's potential to enhance data readiness assessments within federated learning environments. We have successfully integrated data readiness assessment into the APPFL framework, which is available publicly [44].

5.3 Data Readiness Evaluation

5.3.1 SGC Case Study dataset analysis using the web interface. In this case study, we present the findings obtained from the readiness evaluation conducted on the SGC dataset. AIDRIN summarized the data initially by displaying the dimensions of the data along with a detailed description of the summary statistics (Figure 2) of the existing features in the data.

After selecting the readiness metrics for the SGC dataset according to Table 3 and selection in Figure 4, in terms of completeness (Figure 7), we observed that two features were around 80% and 60% complete respectively. In contrast, the remaining features were fully complete. The number of duplicates in the dataset was 0, meaning

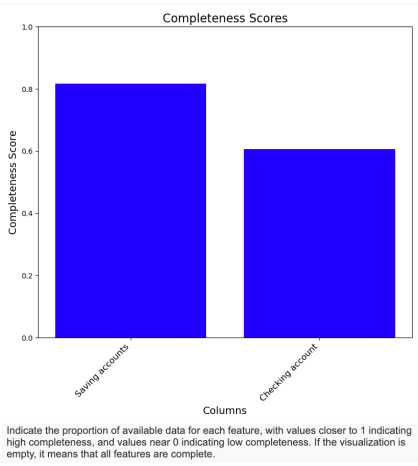


Figure 7: Visualization generated by AIDRIN of the completeness metric for the SGC dataset, illustrating the percentage of completeness for each feature.

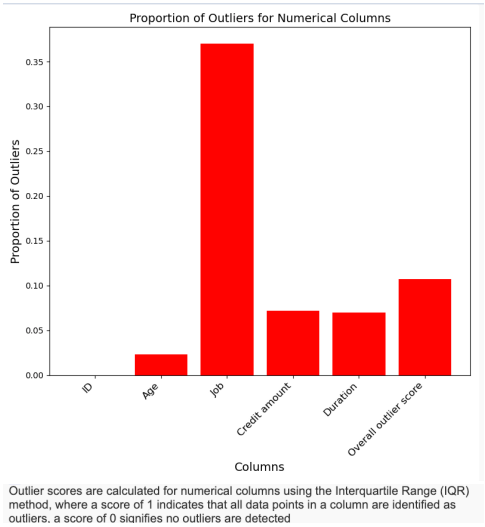


Figure 8: Visualization of outliers metric for the SGC dataset, showing the percentage of outliers for each numerical feature and the overall outlier score.

there were no duplicate records in the data. Regarding outliers (Figure 8), apart from one feature (35%), the numerical features exhibited less than 15% outliers.

Fairness analysis revealed biased representation rates (Figure 9), with over 75% of the representation being males and the remainder being females. The class imbalance was evident when selecting the "Purpose" attribute, with the most popular class comprising 33.7% and the minority class only 1.2%. The imbalance degree score of 4.49 indicates significant class distribution disparity. When considering the privacy risk scores associated with the 'Housing' feature, AIDRIN calculated a mean value of the re-identification risk scores to be 0.45. This finding suggests the presence of a potential re-identification risk associated with the 'Housing' feature.

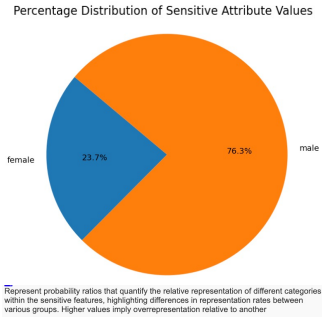


Figure 9: Visualization of representation rate metric for the SGC dataset, showing the percentage distribution of values across the selected sensitive attribute

Additionally, the visual outputs generated by AIDRIN helped to easily interpret these evaluations including other readiness evaluations such as feature relevance, and correlation analysis. Figure 5 illustrates a sample of the specific visual outputs generated based on the selections made in Figure 4, enhancing the user's ability to interpret and utilize the evaluation results effectively. Additionally, users have the option to download a JSON file containing detailed evaluation results. These findings display the important role of AIDRIN in assessing the readiness of the dataset for AI applications.

5.3.2 TCGA-LUAD Case Study dataset analysis using Jupyter Notebook. Another case study was conducted using the PyPI package of AIDRIN within a Jupyter Notebook. The dataset used for this case study was the Cancer Genome Atlas Lung Adenocarcinoma (TCGA-LUAD) clinical data of patients accessed from the National Cancer Institute NCI Data Portal [24]. This readiness analysis was conducted as a part of NCI CRDC Artificial Intelligence Data-Readiness (AIDR) Challenge in 2024 [39].

We observed that the overall completeness of the data was 56% with attributes having varying degrees of completeness scores, the lowest being 22%. The number of duplicates in the data was 0%, ensuring that each patient's information is uniquely represented in the dataset. AIDRIN's evaluation of representation rates revealed female counterparts accounting for 65% of the data records. In terms of race, 87% of the individuals were identified as whites, being the majority group. This imbalance raises concerns as healthcare decisions can exhibit a disparity among different racial groups.

Furthermore, AIDRIN observed the dataset contained a class imbalance with an ID score of 0.27. We considered vital status as the class attribute categorized into two classes: 'dead' and 'alive'. This score suggests that there is a skew from a balanced class distribution that could potentially result in a skewed decision-making process. The TSD score across the gender attribute for both classes was minimal (0.01). Notably, the TSD score across racial groups for both classes was 0.15. The scores reflect potential discrimination in class attributes between racial groups compared to gender groups.

The two case studies above illustrate AIDRIN's capabilities, with one using the web interface and the other using the Jupyter Notebooks. These evaluations empower scientists by providing them with valuable insights about their data to make informed decisions before proceeding to the next steps of ML pipelines.

6 CONCLUSION

The importance of high-quality and AI-ready data emphasizes the success of automated decision-making. However, the data assessment process of existing tools is inefficient and inconsistent. To address these challenges, we provide a definition of data readiness parameters focused for AI and introduce AIDRIN, a comprehensive toolset designed to assess the AI readiness of data. By integrating a diverse range of metrics, AIDRIN provides users with a complete framework to assess the readiness of their data both quantitatively and qualitatively. Moreover, AIDRIN's relevance extends into emerging areas like Federated Learning. It is important to evaluate the readiness of the edge client's data in such settings. We have integrated AIDRIN into the Advanced Privacy-Preserving Federated Learning ([32, 44, 45]) framework to evaluate data readiness.

AIDRIN's current functionality is focused on tabular data, presenting a limitation in its applicability to other data modalities. Another limitation is its performance when handling huge datasets, which may impact efficiency and scalability. Our future work will cover a broader range of data modalities (e.g., images, text, and audio) as modern AI applications rely on diverse data formats. We are also designing a comprehensive scoring method that is meaningful to users in improving the readiness of data for AI.

Acknowledgements

This research is supported by the U.S. Department of Energy, Office of Science, under contract numbers DE-AC02-06CH11357 (ANL), GR134836 (OSU), and DE-AC02-05CH11231 (LBNL).

REFERENCES

[1] [n. d.]. *AIDRIn: AI Data Readiness Inspector.* test.pypi.org/project/aidrin/0.5.4
[2] [n. d.]. DataCite Metadata Schema. DataCite Schema. https://schema.datacite.org/ Accessed 18 Feb. 2024.
[3] [n. d.]. DCAT-US Schema v1.1 (Project Open Data Metadata Schema). Project Open Data Metadata Schema. resources.data.gov/resources/dcat-us/ Feb. 2024.
[4] K. Aas, M. Jullum, and A. Løland. 2019. Explaining Individual Predictions When Features Are Dependent: More Accurate Approximations to Shapley Values. *arXiv preprint arXiv:1903.10464 [cs, stat]* (March 2019). http://arxiv.org/abs/1903.10464
[5] S. Afzal, C. Rajmohan, M. Kesarwani, S. Mehta, and H. Patel. 2020. Data Readiness Report. In *IEEE Int. Conference on Smart Data Services (SMDS)*. 42–51.
[6] F. Alberto, S. García, M. Galar, R. Prati, B. Krawczyk, and F. Herrera. 2018. *Learning from Imbalanced Data Sets.* Springer.
[7] F. Azzalini, C. Criscuolo, and L. Tanca. 2022. E-FAIR-DB: Functional Dependencies to Discover Data Bias and Enhance Data Equity. *J. Data and Information Quality* 14, 4 (November 2022), Article 29. https://doi.org/10.1145/3552433
[8] 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).
[9] Roger Blake and Paul Mangiameli. 2011. The Effects and Interactions of Data Quality and Problem Complexity on Classification. *J. Data and Information Quality* 2, 2 (February 2011), Article 8. https://doi.org/10.1145/1891879.1891881
[10] Christian Bors, Theresia Gschwandtner, Simone Kriglstein, Silvia Miksch, and Margit Pohl. 2018. Visual Interactive Creation, Customization, and Analysis of Data Quality Metrics. *J. Data and Information Quality* 10, 1 (2018), Article 3.
[11] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. 2000. LOF: Identifying Density-Based Local Outliers. In *ACM SIGMOD Int. Conf. Manage. Data.*
[12] N. Carlini, M. Jagielski, C. Zhang, N. Papernot, et al. 2022. The Privacy Onion Effect: Memorization is Relative. *arXiv:2206.10469 [cs.LG]* (2022).
[13] L. E. Celis, V. Keswani, and N. K. Vishnoi. 2020. Data Preprocessing to Mitigate Bias: A Maximum Entropy Based Approach. *arXiv:1906.02164 [cs.LG]* (2020).
[14] Harald Cramér. 1946. *Mathematical Methods of Statistics.* Princeton University Press, Princeton. 282 pages.
[15] N. Davari, B. Veloso, R. Ribeiro, and J. Gama. 2023. MetroPT-3 Dataset. UCI Machine Learning Repository. https://doi.org/10.24432/C5VW3R
[16] V. Duddu, S. Szyller, and N. Asokan. 2022. SHAPr: An Efficient and Versatile Membership Privacy Risk Metric for Machine Learning. *arXiv:2112.02230* (2022).
[17] Lehmberg et al. 2020. datafold: data-driven models for point clouds and time series on manifolds. *Journal of Open Source Software* 5, 51 (2020), 2283.
[18] S. Cholia et al. 2024. ESS-DIVE Overview: A Scalable, User-Focused Repository for Earth and Environmental Science Data. https://ess-dive.lbl.gov/
[19] FAIRassist.org. [n. d.]. FAIRassist.Org. https://fairassist.org. Jan. 6, 2024.
[20] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *21st ACM SIGKDD.*
[21] David Freedman, Robert Pisani, and Roger Purves. 2007. *Statistics (International Student Edition)* (4th ed.). WW Norton & Company, New York.
[22] C. Frye, I. Feige, and C. Rowat. 2019. Asymmetric Shapley Values: Incorporating Causal Knowledge into Model-Agnostic Explainability.
[23] GO FAIR. 2022. GO FAIR Principles. https://www.go-fair.org/fair-principles/.
[24] R. L. Grossman, A. P. Heath, V. Ferretti, H. E. Varmus, D. R. Lowy, W. A. Kibbe, and L. M. Staudt. 2016. Toward a Shared Vision for Cancer Genomic Data. *New England Journal of Medicine* 375, 12 (2016), 1109–1112.
[25] Nitin Gupta, Hima Patel, et al. 2021. Data Quality Toolkit: Automatic Assessment of Data Quality and Remediation for Machine Learning Datasets. *arXiv preprint arXiv:2108.05935* (2021).
[26] Tarek Hasanin, Taghi M Khoshgoftaar, Jaimie L Leevy, and et al. 2019. Severely imbalanced Big Data challenges: investigating data sampling approaches. *Journal of Big Data* 6, 1 (2019), 107. https://doi.org/10.1186/s40537-019-0274-4
[27] Kaveen Hiniduma, Suren Byna, and Jean Luca Bez. 2024. Data Readiness for AI: A 360-Degree Survey. arXiv:2404.05779
[28] Hans Hofmann. 1994. Statlog (German Credit Data). *UCI Machine Learning Repository.* https://doi.org/10.24432/C5NC77
[29] 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]
[30] Informatica. [n. d.]. *Data Quality Metrics & Measures - All You Need to Know.* Accessed Jan. 10, 2024.
[31] M. Kelly, R. Longjohn, and K. Nottingham. [n. d.]. The UCI Machine Learning Repository. https://archive.ics.uci.edu.
[32] Zilinghan 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.
[33] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4765–4774.
[34] D. Marín López, D. Marín, J. Fonollosa, J. Llano, A. Perera, and Z. Haddi. 2023. Single Elder Home Monitoring: Gas and Position. UCI ML Repository.
[35] R. Marcinkevičs et al. 2023. Regensburg Pediatric Appendicitis Dataset (1.01) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7669442
[36] Ninareh Mehrabi et al. 2022. A Survey on Bias and Fairness in Machine Learning. arXiv:1908.09635 [cs.LG]
[37] L. Merrick and A. Taly. 2019. The Explanation Game: Explaining Machine Learning Models with Cooperative Game Theory. *arXiv:1909.08128* (2019).
[38] MIDRC. [n. d.]. The Medical Imaging and Data Resource Center (MIDRC). https://www.midrc.org/
[39] National Cancer Institute Center for Biomedical Informatics and Information Technology. n.d.. CRDC insights. https://datacommons.cancer.gov/news/nci-crdc-artificial-intelligence-data-readiness-aidr-challenge.
[40] J. Ortigosa-Hernández, I. Inza, and J. A. Lozano. 2017. Measuring the Class-Imbalance Extent of Multi-class Problems. *Pattern Recognition Letters* 98 (2017), 32–38. https://doi.org/10.1016/j.patrec.2017.08.002
[41] D. Pokrajac, A. Lazarevic, and L. J. Latecki. 2007. Incremental Local Outlier Detection for Data Streams. In *IEEE Symp. Comput. Intell. Data Mining.* 504–515.
[42] Raphael Poulain et al. 2023. Improving Fairness in AI Models on Electronic Health Records: The Case for Federated Learning Methods. *arXiv preprint arXiv:2305.11386* (2023).
[43] P. Rocca-Serra, W. Gu, V. Ioannidis, et al. 2023. The FAIR Cookbook - The Essential Resource for and by FAIR Doers. *Sci Data* 10 (2023).
[44] M. Ryu et al. [n. d.]. *APPFL: Advanced Privacy-Preserving Federated Learning.*
[45] Minseok Ryu et al. 2022. APPFL: Open-Source Software Framework for Privacy-Preserving Federated Learning. In *IPDPS Workshops.* IEEE, 1074–1083.
[46] Sebastian Schelter, Dustin Lange, Philipp Schmidt, Meltem Celikel, Felix Biessmann, and Andreas Grafberger. 2018. Automating Large-Scale Data Quality Verification. *Proc. VLDB Endow.* 11, 12 (August 2018), 1781–1794.
[47] S. Shrivastava et al. 2020. DQLearn: A Toolkit for Structured Data Quality Learning. In *International Conference on Big Data (Big Data).* 1644–1653.
[48] L. Song and P. Mittal. 2021. Systematic Evaluation of Privacy Risks of Machine Learning Models. In *30th USENIX Security Symposium (USENIX Security 21).*
[49] H. Theil. 1992. Some Reflections on Static Programming under Uncertainty. In *Henri Theil's Contributions to Economics and Econometrics*, B. Raj and J. Koerts (Eds.). Advanced Studies in Theoretical and Applied Econometrics, Vol. 24.
[50] Dinusha Vatsalan et al. 2022. Privacy Risk Quantification in Education Data Using Markov Model. *British Journal of Educational Technology* (2022), 804–821.
[51] R. Zhu, Z. Wang, Z. Ma, G. Wang, and J.-H. Xue. 2018. LRID: A New Metric of Multi-class Imbalance Degree Based on Likelihood-Ratio Test. *Pattern Recognition Letters* 116 (2018), 36–42. https://doi.org/10.1016/j.patrec.2018.09.012