

Data Readiness for AI: A 360-Degree Survey

KAVEEN HINIDUMA, The Ohio State University, USA

SUREN BYNA, The Ohio State University, USA

JEAN LUCA BEZ, Lawrence Berkeley National Laboratory, USA

Artificial Intelligence (AI) applications critically depend on data. Poor quality data produces inaccurate and ineffective AI models that may lead to incorrect or unsafe use. Evaluation of data readiness is a crucial step in improving the quality and appropriateness of data usage for AI. R&D efforts have been spent on improving data quality. However, standardized metrics for evaluating data readiness for use in AI training are still evolving. In this study, we perform a comprehensive survey of metrics used to verify data readiness for AI training. This survey examines more than 140 papers published by ACM Digital Library, IEEE Xplore, journals such as Nature, Springer, and Science Direct, and online articles published by prominent AI experts. This survey aims to propose a taxonomy of data readiness for AI (DRAI) metrics for structured and unstructured datasets. We anticipate that this taxonomy will lead to new standards for DRAI metrics that would be used for enhancing the quality, accuracy, and fairness of AI training and inference.

Additional Key Words and Phrases: Data readiness, Data quality metrics, AI-ready data

ACM Reference Format:

Kaveen Hiniduma, Suren Byna, and Jean Luca Bez. 2022. Data Readiness for AI: A 360-Degree Survey. In . ACM, New York, NY, USA, 36 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 INTRODUCTION

Data readiness for artificial intelligence (AI) refers to the critical process of preparing and ensuring the quality, accessibility, and suitability of datasets before using them for AI applications. Readyng the data is a fundamental step, which involves collecting, cleaning, organizing, and validating the dataset not only to make them compatible with AI algorithms and models, but also to make certain that the datasets are appropriate and unbiased. By achieving data readiness, organizations can maximize the accuracy, efficiency, and effectiveness of their AI systems, ultimately leading to more informed decision-making and successful AI-driven outcomes.

Data readiness for AI (DRAI) is an important concern in AI applications, as evidenced by a survey conducted by Scale AI [63]. A significant number of participants encountered challenges related to data readiness within their machine learning (ML) projects. Similarly, a study [146] involving nearly 2,400 respondents from over 100 countries explains the time-intensive nature of data preparation for data scientists working with AI applications. It is crucial to recognize that the quality of outcomes generated by an AI system is heavily linked to the readiness of the input data. This connection highlights the significance of addressing the “garbage in, garbage out” saying, which emphasizes that flawed or insufficient input data will inevitably lead to inaccurate and unreliable results from AI algorithms [124]. Hence, ensuring the availability of well-prepared data for training machine learning (ML) models is critical, as it leads to more precise and dependable predictions.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2022 Association for Computing Machinery.

Manuscript submitted to ACM

With growing requirements of unbiased data for AI, the field of quantitative evaluation of data readiness with appropriate metrics is still evolving. Data Quality Toolkit (DQT) [39] provides a suite of tools and functionalities to streamline the preparation and cleaning of data. DQT's data quality report various dimensions of data quality metrics defined by Sidi et al. [128] such as completeness, consistency, accuracy, and timeliness, along with suggestions for improvement. Ravi et al. [45] focus on the critical process of making experimental datasets FAIR (Findable, Accessible, Interoperable, Reusable) for AI readiness. They emphasize the usage of current data infrastructure to establish a framework suitable for automatic AI-powered exploration. To achieve this objective, they publish FAIR and AI-ready datasets [20]. This study also illustrates usage of FAIR principle compliance [2] and AI-ready datasets for inference.

Although these separate efforts and tools are available to improve quality of data, there is a lack of a comprehensive study on effective metrics and standards for evaluating data readiness for AI. To address that gap, we perform a comprehensive examination of the existing metrics and tools that could be used for evaluating data readiness, covering both structured and unstructured data dimensions. We also describe the metrics designed to evaluate fairness and privacy related issues in data, which critically impact the process of decision-making in AI algorithms.

This survey refers to a comprehensive set of data readiness dimensions and data preparation techniques targeting data usage in AI. Metrics targeting data readiness for AI (DRAI) contain a subset of data quality dimensions including completeness, duplicates, correctness, and timeliness. This distinction between DRAI and data quality is critical for understanding our survey's scope. We identify the existing metrics and scoring mechanisms (§3) and existing frameworks or tools (§4) in the literature that could be used to measure DRAI. Based on the distillation of available literature, we propose a potential comprehensive definition of data readiness for AI using six dimensions (§5). We discuss gaps and challenges towards developing a DRAI assessment framework (§6). This survey is particularly aimed at data preparers for future AI use and data scientists who analyze the data to ensure their datasets are ready for AI applications.

2 SCOPE OF THE STUDY

Our literature review methodology was carefully structured to ensure a comprehensive and unbiased examination of existing research. The search queries used to obtain the sources for this study are presented in Table 1, categorized into general, structured data, and unstructured data search queries. As a result, we gathered nearly 30 papers from ACM Digital Library, over 20 papers from IEEE Xplore, 10 papers each from Springer and Science Direct, and more than 40 papers from journals including Nature, Springer and Science Direct, as well as several relevant books, to review. Additionally, we highlight discussions on six web articles and explore the metrics used in six commercially used tools.

Table 1. Data Readiness for AI Metrics: Summary of search terms used to identify literature to perform this study

General	Structured Data Related	Unstructured Data Related
"data readiness" AND "AI"	searched under each data readiness for AI dimension e.g., "discriminat" AND "metric" OR "measure" OR "evaluat"	"speech quality" AND "metric" OR "measure" OR "evaluat"
"data readiness"	"data quality dimension"	"audio quality" AND "metric" OR "measure" OR "evaluat"
"data readiness" AND "machine learning" OR "ML"	"data quality" AND "metric"	"video quality" AND "metric" OR "measure" OR "evaluat"
"AI ready"	"data prepare" AND "AI"	"image quality" AND "metric" OR "measure" OR "evaluat"
"data quality" AND "machine learning"	"data read" AND "metric"	"visual quality" AND "metric" OR "measure" OR "evaluat"
"data quality" AND "measure"	"data preprocess"	
"data quality" AND "evaluation"	"data clean"	
"data quality" and "AI"	"data quality" AND "survey"	

Papers and articles were included if they discussed data readiness metrics for AI or data quality metrics, covered structured or unstructured data dimensions, and addressed fairness and privacy issues related to DRAI. We included sources that provided metrics or tools for evaluating data readiness and those that were peer-reviewed or published in reputable journals. We also used web articles that provided valuable insights and were considered highly relevant to

our study. Additionally, if the metrics were related to AI data preprocessing, such as feature relevancy, class imbalance measures, and FAIR compliance, they were also included.

We excluded papers that did not focus on DRAI, those that were not peer-reviewed, and articles that lacked substantial contributions to understanding DRAI metrics. We also excluded duplicates and studies not available in English. Additionally, we excluded papers if the metrics were not quantifiable or unrelated to the pre-data training stage in AI.

Our survey aims to identify and analyze the existing metrics and scoring mechanisms for measuring DRAI, covering both conventional dimensions of data quality (e.g., completeness, outliers, timeliness, and correctness) and AI-specific dimensions (e.g., fairness, feature importance, class imbalance, and mislabeled data). We seek to provide insights into how these dimensions and metrics can be used to assess the preparedness of data for AI applications.

In forming this study, our focus included published literature across different timeframes – pre-2000, 2000–2010, and post-2010. In Fig. 1, we show the distribution of sources across different time frames. Paying attention to references prior to 2000 is particularly important because they include some well-established data quality metrics that remain relevant today. These early efforts provide the context and insights into an evolution of data readiness metrics over the years. Before 2000, work primarily focused on general data quality considerations without specific emphasis on applications in AI. From 2000 to 2010, AI research was focused on foundational computational methods and concepts. Post-2010, with the rise of big data and machine learning technologies, a notable shift toward creating metrics to assess data preparedness emerged, specifically for AI applications.

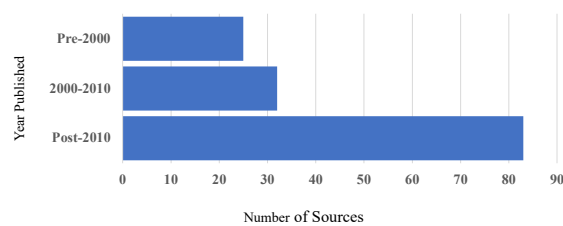


Fig. 1. Papers chosen for this survey from different time frames.

2.1 Existing Surveys and Gaps

Existing surveys on data quality ([17, 43, 86, 128, 136]) primarily focused on traditional dimensions like completeness, correctness, timeliness, and also the quality of textual and multimedia data, providing a strong foundation for understanding the general challenges in the field. However, with the rise of AI, there has been an increasing emphasis on AI-specific concerns like feature relevance, class imbalance, mislabeled data, privacy, and fairness ([83, 101, 108, 125, 139]). These emerging factors are critical alongside conventional quality benchmarks. Recognizing the importance of both aspects, our survey aims to address this shift towards a comprehensive view of data readiness, where traditional and AI-focused dimensions are equally essential.

Toward gaining an understanding of metrics for data readiness, a few studies surveyed data pre-processing stages in preparing data for AI. Priestley et al. [108] conducted a study highlighting the role of decision-makers and practitioners in improving data-focused practices. They highlighted the significance of data cleaning and pre-processing stages, including feature selection, duplicate elimination, outlier removal, consistency assurance, and handling of missing values. Documentation of these pre-processing steps was essential to ensure compatibility and identify potential dependencies and information leakages among features. Their insights, derived from an extensive literature survey, laid the foundation for recognizing the pivotal role of data readiness in AI applications. Yalaoui and Boukhedouma [148] presented a survey paper on data quality evaluation and enhancement. They compared existing frameworks, models, and methods for data quality evaluation and enhancement, and identified challenges.

Batini and Scannapieco's [17] book offers a comprehensive introduction to a broad range of data quality issues. This book provides a state of the art overview of data quality measurement practices by using probability theory, data mining, statistical data analysis, and machine learning.

A few studies have looked into metrics for unstructured data. Elmagarmid et al. [43] explored similarity metrics for duplicate record detection, specifically focusing on challenges posed by typographical variations in string data. Li et al. [83] concentrated on feature selection metrics, categorizing traditional approaches into wrapper, filter, embedded, and hybrid methods. Furthermore, Forman [52] explored feature selection metrics for text classification, examining their performance. The studies on image quality measures [136] and perceptual visual quality metrics [86] are important in understanding DRAI for unstructured data, particularly in visual data, where assessment of image quality and perceptual metrics are essential for effective AI applications.

In the context of bias and fairness, Shahbazi et al. [125] provided a survey of techniques focused on identifying and mitigating representation bias across diverse data types, such as structured data, image data, textual data, and graph data. They defined the problem and discussed causes and methods for measuring and quantifying representation bias in structured and unstructured datasets. Addressing the issue of discrimination in DRAI systems, Ntouts et al. [101] focused on the challenges and implications of biased data on the fairness and accuracy of AI-based decision-making. They emphasized the importance of understanding and mitigating biases in data to prevent the continuation of discriminatory practices. Similarly, in the context of privacy, Wagner and Eckhoff [139] reviewed over 80 privacy metrics across six domains, categorizing them based on the aspect of privacy measured, required inputs, and data types needing protection. They identified research gaps in areas such as metric combination and interdependent privacy, proposing a method for selecting appropriate metrics through nine key questions. Emphasizing the importance of employing multiple metrics to address diverse privacy aspects, the paper serves as a reference guide and toolbox for privacy researchers, aiding informed choices in metric selection for specific scenarios.

Our study aims to contribute to the evolving field of DRAI metrics, specifically focusing on structured and unstructured data metrics. While previous works have explored specific dimensions or concentrated on individual metrics, we address the lack of a comprehensive study including numerous DRAI dimensions. There are broader perspectives in addressing DRAI applications. We identify numerous important factors that define data readiness, such as data preparation, privacy leakage evaluation, data discrimination evaluation, compliance to FAIR principles, mislabeled data detection, feature relevancy analysis, bias-related issues, and quality evaluation of speech and multimedia data. To survey existing literature in these dimensions and to identify gaps, our effort aims to advance the field of data readiness for AI (DRAI) and inform best practices for evaluating the suitability of data for AI applications. Given the growing importance of DRAI, our work fills a critical gap in the published literature and gives insights to practitioners and decision-makers in



Fig. 2. 360° View of Mapping Data Readiness Dimensions for AI

the field. As illustrated in Fig. 2, the 360° plot view of data readiness incorporates a range of metrics that reflect the ongoing discussion in the literature for both structured and unstructured data. Additionally, we also define DRAI by introducing key pillars and the DRAI metrics for each pillar.

3 DRAI METRICS

This section provides an extensive summary of the existing metrics found in the literature that are used to measure data readiness for AI. We will explore these metrics for both structured and unstructured data, primarily focusing on structured data for which metrics have evolved more. Nevertheless, we describe many metrics related to assessing readiness of unstructured data, including textual, multimedia, image, speech, and video-related data. In Table 2, we provide a snapshot of all the dimensions and metrics discussed in this survey.

3.1 Structured data

Structured data is organized with a consistent format and follows a specific structure or schema. Data stored in spreadsheets, relational database tables, self-describing file formats, etc., are common forms of structured data. This

Table 2. Dimensions and Metrics for (Structured and Unstructured) DRAI

Structured Data		Unstructured Data		
Dimension	Metrics	Data Type	Dimension	Metrics
Completeness	Blake et al. [21], Bors et al. [24], Santos et al. [122], Pearson [105]	Textual Data	Lexical Diversity	Templin [135], McCarthy et al. [94], McCarthy et al. [95]
Outliers	Bors et al. [24], Li et al. [84], Breunig et al. [26], Pokraja et al. [107], Rosner et al. [118], Leys et al. [82], Rousseeuw et al. [119], Degirmenci et al. [38]		Term Importance	Luhn [91], Sparck Jones [133]
Mislabeled	Gupta et al.'s [44], Cohen's Kappa [32]		Readability Score	Rudolf Flesch [51], Coleman and Liau [33], Robert Gunning Associates [5],
Duplicate Values	Bors et al. [24], Levenshtein distance metric [80], Waterman et al. [144], Jaro's distance metric [68], Monge et al. [98], Russell et al. [120]		Topic Coherence	Röder et al. [117], Mimmo et al. [96], Newman et al. [99]
Feature Relevancy	Dai et al. [35], He et al. [57], Zhao et al. [150], Duda et al. [41], Nie et al. [100], Lewis [81], Robnik-Sikonja et al. [115], Davis et al. [37], Liu et al. [87], Gini [54], Hall et al. [55]		Bias Indicator	Papakyriakopoulos et al. [104],
Class Imbalance	Lu et al. [90], Francisco et al. [12], Ortigosa-Hernández et al. [103], Zhu et al. [151], Gupta et al. [44]	Multimedia Data	Image Quality	MSE and PSNR [8], Wang et al. [141], Wang et al. [142], Wang et al. [143], Sarnoff's JND-Metrix [6], Sheikh et al. [126], Chandler et al. [29], Lakhani's [78], Sabottke et al.'s. [121], Lin et al. [85], Marziliano et al. [93], PIQ [74]
Class Separability	Gupta et al.'s [44], Sejong [102], Borsos et al. [25]		Speech Quality	Mean Opinion Score [66], Rix et al. [114], Jayant et al. [69], Taal et al. [134], Beerends et al. [18], Itakura-Saito Spectral Distortion [67], Objective Difference Grade [4],
Discrimination Index	Azzalini et al. [15], Feldman et al. [50], Celis et al. [28], Simonetta et al. [130], Gupta et al. [44], Amazon SageMaker Developer Guide [76]		Video Quality	PSNR [8], Wang et al. [142], Sheikh et al. [126], Chandler et al. [29], Huynh-Thu et al. [62], Netflix [23], OPTICOM's PEVQ [1]
Data Split Ratio	Joseph [71], Affendras et al. [9]			
Data Point Impact	Ghorbani et al. [53], Wang et al. [140], Leave-One-Out [34], Koh et al. [77], Bachem et al. [16], Ribeiro et al. [113]			
Correctness	Kaiser et al. [73], Pipino et al.'s [106]			
Timeliness	Kaiser et al. [73], Heinrich et al. [58], Blake et al. [21]			
Privacy Leakage	Vatsalan et al. [137], Duddu et al. [42], Carlini et al. [27], Song et al. [132], Bezzi [19], Longpré et al. [89], Sevgi et al. [14], Aindo AI [3]			
Sample Size	Alwosheel et al. [13], Haykin [56]			
FAIR Compliance	Wilkinson et al. [145], Clarke et al. [30]			

section will discuss metrics related to various dimensions, such as completeness, outliers, labels, etc., as shown on the left half of Figure 2 and their sources listed in the left half of Table 2.

3.1.1 Completeness. Completeness refers to the presence or availability of required data and attribute values in a dataset. It indicates whether data points or entries are complete, with all relevant attribute values recorded and available.

Example: In a dataset containing information about credit card customer demographics, this metric verifies if an attribute (e.g., income) is available for all customers. Completeness ensures that a dataset is reliable and suitable for analysis, as there is no loss of information due to missing data.

Metrics in Literature: Blake and Mangiameli [21] propose a “completeness” metric to measure the presence of missing values in a dataset. This metric quantifies the proportion of null (missing) data records to the total number of data records.

Using this metric, researchers demonstrate how missing data can impact the results of classification tasks. For example, Jäger et al. [72] demonstrates that handling missing values enhances predictive model performance. They observe up to 20% improvement for classification tasks and 15% improvement for regression tasks, emphasizing the importance of addressing missing data in optimizing downstream ML outcomes. In addressing missing data, Santos et al. [122] use data imputation, which involves filling in or estimating missing values to maintain data integrity. In particular, the authors use the k-nearest neighbors (KNN) imputation technique in this process. They discuss the significance of choosing appropriate distance metrics, such as Heterogeneous Value Difference Metric (HVDM) and Heterogeneous Euclidean-Overlap Metric (HEOM), which effectively handle both nominal and continuous data while preserving data distribution during imputation.

Bors et al. [24] propose a different approach to quantify missing values in a dataset by using indicators to distinguish missing from non-missing values. Their method offers a practical tool for data preparation, allowing easy identification of missing data in AI applications.

Another type of completeness targets missing data “disguised” with default values. Pearson [105] discusses missing values are encoded or represented in ways that obscure their true nature, such as using “zero” values to indicate missing data. Such practices can severely distort analysis results. According to Vo et al. [138] these issues affect not only the model’s predictions but also its explainability, potentially skewing feature importance calculations crucial for interpreting complex AI systems.

Impact on AI: Complete and accurate data enhances AI systems’ accuracy and reliability. Identifying explicitly missing data is easy, while disguised missing values pose a greater challenge as they appear valid but are placeholders or incorrect entries. Disguised values can lead to biased outcomes, reduced accuracy, and misinterpretation of AI models.

Summary: *Metrics proposed by various researchers quantify the impact of missing values and suggest remedies such as KNN imputation with suitable evaluation metrics. Additionally, using indicators for distinction and incorporating completeness metrics that consider data types and relationships further improve the handling of missing data.*

3.1.2 Outliers. Outliers in a dataset refer to data points that significantly deviate from the typical or expected values within the dataset. They are points that are significantly distant from the majority of the data points and do not follow the general patterns present in the dataset.

Example: Consider a dataset of housing prices based on factors such as size, number of bedrooms and bathrooms, location, schools, etc. In this dataset, an outlier could be a property with extremely high or low prices that do not align with the average price range of similar properties. This outlier might be an exceptional case, such as a luxury mansion in an otherwise average neighborhood, or it could be a data entry error.

Metrics in Literature: In their research, Bors et al. [24] discuss the concept of plausibility as a metric to identify outliers in datasets for AI applications, which can disrupt statistical analyses and modeling. Data analysts employ two main approaches to quantify the number of outliers: robust statistics and non-robust statistics. Robust statistics use the median and the robust interquartile range estimator, which are more resistant to outliers. Li et al. [84] further explore the standard deviation and interquartile range-based outlier detection methods in their study. In contrast, non-robust statistics involve using the mean and standard deviation to identify entries that deviate significantly from the mean.

In contrast, Breunig et al. [26] introduce the Local Outlier Factor (LOF) as a metric for identifying outliers in a dataset. LOF quantifies the level of being an outlier for each data instance by considering the density of the dataset's distribution. Outliers are expected to have lower local densities compared to their surrounding instances. The LOF algorithm computes the LOF value for a specific instance by comparing the density of that instance's neighborhood with its neighboring instances. This neighborhood is a user-defined parameter (e.g., number of nearest neighbors). Higher LOF values indicate a higher probability of an instance being an outlier, which implies a notably lower density in its local neighborhood than in neighboring data points. Building on this foundation, Pokrajac et al. [107] introduced the ILOF (Incremental Local Outlier Factor) method, which uses the LOF metric to determine if a new data point is an outlier. By analyzing the computed LOF value, the ILOF method assigns a score to the incoming sample, indicating whether it is classified as an outlier or not. This approach allows for real-time outlier detection and updates the scores of existing points to measure the impact of the new data point.

Degirmenci and Karal [38] introduce RiLOF, which addresses limitations in existing statistical outlier detection techniques by introducing the MoNNAD (Median of Nearest Neighborhood Absolute Deviation) metric. This metric is calculated as the median of the absolute variances among the LOF values of the k -nearest neighboring data points and the LOF value of a given sample. This score indicates how much the sample deviates from its local neighborhood. In the RiLOF method, the MoNNAD score is used to label and score query samples. Samples are categorized as outliers when their MoNNAD scores are equal to or greater than a specific limit. The RiLOF method assigns more importance to the query sample, resulting in clearer differentiation between inliers and outliers. The study demonstrates that the MoNNAD metric, incorporated in the RiLOF method, successfully detects outliers, including outlier clusters, that other techniques fail to recognize.

The GESD (Generalized Extreme Studentized Deviation) technique, as introduced by Rosner [118], and the MAD (Median Absolute Deviation) technique, proposed by Leys et al. [82], are both outlier detection methods that share similarities in their underlying principles. Both approaches aim to identify outliers within datasets by using statistical measures to assess the deviation of data points from central tendencies. GESD identifies outliers by evaluating the maximum absolute difference between each sample and the dataset's mean and normalizing it by the standard deviation. MAD identifies outliers by considering the median of absolute differences between data records and the dataset's median. It incorporates a constant associated by assuming that data is normally distributed. Additionally, the Z-score method aligns with this principle by normalizing sample values using the mean and standard deviation or MAD. The robust Z-score version, introduced by Rousseeuw and Hubert [119], substitutes the median and MAD for more robust measures, demonstrating the shared concept of using statistical measures to detect outliers in datasets.

Impact on AI: Outliers can significantly impact AI systems by skewing data distributions and introducing biases that lead to inaccurate models and unreliable predictions. They can distort statistical measures such as mean and variance, affecting algorithms like linear regression and clustering. They also can reduce the generalization and robustness of classification and neural network models. Outliers can also introduce noise, complicating the ML process and potentially leading to false positives or negatives [64]. However, in specific contexts like fraud detection or rare disease diagnosis

(e.g., Markham [92]), outliers can be critical for identifying anomalies and should not be removed without a thorough analysis. Proper detection and management of outliers maintain the integrity and effectiveness of AI models.

Summary: *Outliers data points deviate significantly from most of the dataset and do not follow the general patterns. Different metrics and techniques proposed to identify and measure outliers in datasets include measures based on column heterogeneity, statistical measures like median, standard deviation, mean, and interquartile range, and techniques such as Local Outlier Factor (LOF), Generalized Extreme Studentized Deviation (GESD), Z-score.*

3.1.3 Misabeled Data. Misabeled data in the context of preparing a dataset for AI refers to instances or data points with inaccurate labels. It represents a form of labeling error or inconsistency within the dataset, where the assigned labels do not align with the true or expected labels.

Example: Consider a dataset for email spam classification, where each email is labeled as either “spam” or “not spam”. If some emails are mistakenly labeled as “not spam” when they should have been labeled as “spam”, or vice versa, it introduces mislabeled data. In this case, the mislabeled instances create discrepancies between the assigned labels and the actual content or characteristics of the emails.

Metrics in Literature: Gupta et al.’s [44] Data Quality Toolkit (DQT) introduces a label purity metric to measure the impact of adding random noise on the performance of a classifier trained on the dataset. In the example provided in the study, 10% random noise is introduced to 41 datasets from UCI (Dua and Graff [40]) and Kaggle ([7]) repositories, and the performance of an AutoAI classifier (Liu et al. [88]) is measured using 3-fold cross-validation. The results show a drop in classifier performance after inducing noise, with varying degrees of decrease observed across the datasets.

In evaluating the accuracy of labels assigned by multiple annotators, Cohen’s Kappa [32] is widely employed for assessing inter-rater reliability, especially with categorical or binary labels. Cohen’s Kappa calculates the agreement beyond chance, ranging from -1 to 1 , where values near 1 indicate substantial agreement. Lavitas et al. [79] contribute a credibility metric, assessing the likelihood of correct annotations based on multiple reviewers’ agreement. The metric ranges from $(N/2 + 1)/N$ to 1 , reflecting high credibility with close agreement and lower credibility with less agreement among reviewers, offering insights into the reliability of the annotation process involving multiple reviewers.

Impact on AI: Incorrect labels in training data lead to poor model performance because the AI learns from erroneous examples, resulting in skewed predictions and decisions [31]. This issue can cause substantial financial costs due to the need to retrain models and correct errors. Additionally, mislabeled data can undermine trust in AI systems, as stakeholders lose confidence in their outputs. Ethical implications are also significant, as mislabeled data can introduce or amplify biases that lead to discriminatory outcomes, such as flawed medical diagnoses or biased hiring algorithms. Addressing mislabeled data is crucial for maintaining the integrity and effectiveness of AI models.

Summary: *Mislabeled data refers to instances in a dataset that have inaccurate labels, creating discrepancies between assigned labels and the true or expected labels. Available approaches include a label purity metric for classifying performance under induced noise, label agreement among annotators, and credibility metric through reviewer consensus.*

3.1.4 Duplicate Values. This refers to the presence of duplicate or redundant records within a dataset. Duplicates appear when the same or similar data entries appear multiple times, potentially skewing the ML process.

Example: Consider a dataset of customer transactions where each entry represents a purchase made by a customer. Multiple entries of the same transaction refer to duplication, which may impact the analysis results.

Metrics in Literature: Bors et al. [24] propose a mechanism to identify duplicate entries in a dataset using a scoring system based on uniqueness. A user selects one or more columns in the dataset that are intended to have distinctive combinations of values. The system assigns a score of 1 (true) to values with a unique combination in the selected columns and a 0 (false) score to values found multiple times. By incorporating this score, the system effectively flags

duplicate entries in the dataset. Elmagarmid et al. [43] conducted a comprehensive survey exploring various similarity metrics for duplicate detection, addressing challenges in managing typographical variations in string data. One of the highlighted character-based similarity metrics in the survey is the Levenshtein distance metric [80]. This metric measures the number of operations needed to transform one string into another through edit operations (insertion, deletion, and character replacement). Waterman et al. [144] introduced the Affine Gap Distance metric to overcome the limitations of the standard edit distance metric in matching shortened or truncated strings. It introduces two edit operations: open and extend the gap. An open gap in sequence alignment indicates the start of missing or deleted characters in the sequence. In contrast, an extended gap accommodates consecutive missing characters by extending an existing one. This metric allows for smaller penalties for gap mismatches, resulting in more accurate measurements for truncated or shortened strings. Additionally, Jaro's distance metric [68] quantifies the similarity between two strings by identifying common characters that appear at the same positions in both strings and adding up the number of transpositions. The Jaro metric considers the number of shared characters, the lengths of the strings, and the number of transpositions. Monge and Elkan [98] introduced a token-based similarity metric designed to detect duplicates in text fields using atomic strings. Atomic strings are identified by punctuation characters, acting as delimiters, and consist of alphanumeric characters as individual units within the text fields. Two atomic strings are considered duplicates if they are either identical or if one is a prefix of the other. This approach helps identify duplicates in text fields by considering matching atomic strings and provides a similarity score to assess the degree of duplication between fields.

The Soundex algorithm, introduced by Russell [120], is a phonetic coding scheme used to detect duplicates by comparing the phonetic similarity of character strings, such as names. The algorithm transforms names into codes based on rules of phonetic similarity. It preserves the initial letter of the name as the prefix letter and assigns codes to each remaining letter according to specific phonetic groups. Vowels act as separators between consecutive consonants. Consecutive occurrences of the same code are merged, and if the resulting code has fewer than three characters, zeros are added as padding. By applying the Soundex algorithm, names are encoded into phonetic codes that capture their phonetic similarities. It enables the detection of similar-sounding names indicating potential duplicates.

Impact on AI: When training on duplicated data, models may overfit by learning redundant patterns that do not generalize well to unseen data. This over-representation can lead to skewed predictions and unreliable outcomes. Duplicates also increase the dataset size, storage costs, and computational resources required for training. Additionally, they can degrade data quality, making it harder to derive accurate insights and leading to inefficiencies in data processing.

Summary: *Duplicates refer to the presence of duplicate or redundant instances in a dataset, which can distort analysis and modeling. Various similarity metrics, based on Levenshtein distance, Affine Gap Distance, Jaro's distance, Monge et al.'s token-based algorithm, and the Soundex phonetic coding scheme, are available to measure duplicates.*

3.1.5 Feature Relevance. This refers to identifying and selecting the most informative features or variables that contribute to an AI model. In a dataset, various features or variables are typically collected for each instance, representing different aspects or characteristics of the data. However, not all features may be equally relevant or valuable for an AI model. Feature relevance metric aims to identify the subset of features that are most influential in making accurate predictions or capturing the underlying patterns in the data.

Example: Consider a dataset for predicting housing prices that includes number of bedrooms, square footage, location, schools, and proximity to amenities. In this case, feature relevance would involve analyzing the relationship between each feature and the target variable (house prices) to determine which ones have the strongest correlation or impact on the predictions. Features that have weak or negligible influence can be excluded.

Metrics in Literature: The column heterogeneity measure proposed by Dai et al. [35] uses soft clustering techniques and mutual information to quantify the relevance of features in a dataset. Soft clustering assigns fractional memberships to data points across multiple clusters. Mutual information measures the dependence between feature values and soft clustering results, capturing their association. The computed mutual information values are then used to derive column heterogeneity scores for each feature. Features with higher scores are considered more informative.

A survey by Li et al. [83] collectively explores similarity-based feature selection metrics, including Laplacian Score, SPEC, Fisher Score, Trace Ratio, and ReliefF. The Laplacian Score algorithm by He et al. [57] constructs affinity and Laplacian matrices to measure similarities and differences among data points, producing scores that prioritize features capturing underlying data structures. SPEC, an extension of the Laplacian Score, introduced by Zhao and Liu [150], emphasizes alignment with data structure through spectral analysis. Duda et al.'s [41] Fisher Score emphasizes comparability within classes and distinctiveness between classes, while the Trace Ratio criterion by Nie et al. [100] and ReliefF algorithm by Robnik-Šikonja and Kononenko [115] emphasize within-class similarity and between-class dissimilarity. Collectively, these algorithms highlight the significance of exploiting data relationships and class structures.

Li et al. [83] focus on information-theory-based feature selection methods, including Mutual Information Maximization (MIM) (Lewis [81]). MIM relies on the concept of entropy to evaluate the significance of features by measuring the reduction in uncertainty they bring to a classification task. MIM evaluates each feature's significance based on its correlation with class labels. Features with higher Mutual Information (MI) scores are considered more informative and are selected until the desired number of features is reached. Li et al. [83] also provides statistical methods, highlighting their roles and applications in various fields. Among these methods, the Low Variance method measures feature relevance by evaluating variances and removing features with variances below a specified threshold. In binary classification, the T-Score method, proposed by Davis and Sampson [37], quantifies a feature's capacity to differentiate classes by calculating T-scores based on class means and standard deviations, with higher scores indicating stronger discriminatory power. Conversely, the Chi-Square Score method, introduced by Liu and Setiono [87], assesses feature-class independence through an independence test derived from differences between observed and expected frequencies. Gini Index [54] evaluates a feature's partitioning potential across different classes, using class probabilities, considering how effectively its values divide the dataset. The Correlation-based Feature Selection (CFS) by Hall and Smith [55] evaluates feature subsets worth using a correlation-based heuristic. The CFS score balances predictive power with redundancy using symmetrical uncertainty.

Impact on AI: By identifying the most important features, AI models can concentrate on key data aspects, which reduces noise and computational demands and enhances performance. This process helps mitigate the “curse of dimensionality,” speeds up training, and can prevent over-fitting. It also enhances model interpretability by highlighting key factors driving predictions or decisions.

Summary: *Feature relevance helps in identifying and selecting the most informative and significant features that contribute to the predictive power of AI models. Existing feature relevance metrics use statistical techniques such as soft clustering, similarity, and information theory.*

3.1.6 Class Imbalance. In the context of data for AI, class imbalance refers to the highly skewed or uneven distribution of instances among different classes (categories) in a dataset. It means that one or more classes appear more than others, leading to an imbalanced representation.

Example: Class imbalance often appears in datasets with rare event detection, such as credit fraud detection, earthquake prediction, network intrusion detection, customer churn prediction, rare disease diagnosis, rare species of animal sightings, etc. In these cases, rare events or anomalies are the focus of detection.

Class imbalance can pose challenges during AI model training and evaluation. Models trained on imbalanced data tend to prioritize the majority class, resulting in poor prediction performance for the minority class. In an anomaly detection example, an imbalanced dataset may lead to a prediction model that performs well in predicting normal instances but performs poorly in identifying anomalies, which are often the class of interest.

Metrics in Literature: The Individual Bayes Imbalance Impact Index (IBI3), introduced by Lu et al. [90] assesses the impact of class imbalance on individual samples, providing insights into potential biases and dataset limitations. IBI3 quantifies the difference in posterior probabilities between balanced and imbalanced scenarios, revealing how class imbalance influences classification outcomes. It requires trained models and estimation of posterior probabilities to calculate, making it essential to have access to both for an accurate assessment. IBI3 measures the influence of class imbalance on classification outcomes for each minority class sample, with lower values indicating less impact.

Imbalance Ratio (IR), introduced by Alberto et al. [12], is a widely used metric to quantify the level of class imbalance in a dataset, especially in binary classification problems. It provides a numerical representation of the discrepancy between the majority and minority class instances. IR is calculated by dividing the count of instances in the majority class ($N_{majority}$) by the count of instances in the minority class ($N_{minority}$). A higher IR indicates a more imbalance.

Ortigosa-Hernández et al. [103] propose Imbalance Degree (ID) as a metric to measure class imbalance, considering specific characteristics of the class distribution. Despite its advantages, ID has drawbacks, such as sensitivity to the choice of distance function and potential unreliability in extreme cases. In contrast, Zhu et al. [151] introduce the Likelihood Ratio Imbalance Degree (LRID) to overcome ID's limitations. LRID uses the likelihood ratio (LR) test, providing a high-resolution measurement of imbalance by comparing empirical class distribution to a balanced distribution. Gupta et al. [44] propose the class parity metric, which considers various data properties, including the imbalance ratio, dataset size, and proportion of difficult samples in the extreme minority class.

Impact on AI: Class imbalance in datasets can significantly impact AI systems, particularly in classification tasks. Class imbalance can cause models to overlook important patterns in minority classes, potentially leading to misclassification.

Summary: *Class imbalance in AI-ready data is assessed using metrics like Imbalance Ratio (IR), with Imbalance Degree (ID) offering nuanced measurements. Likelihood Ratio Imbalance Degree (LRID) provides a high-resolution assessment through the likelihood ratio test. Gupta et al.'s class parity metric considers imbalance ratio, dataset size, and difficult samples.*

3.1.7 Class Separability. Class Separability refers to the degree of similarity or shared characteristics between different classes or categories within the dataset. It measures the overlap or sharing of common features among the data points from different classes.

Example: Consider a facial recognition dataset consisting of two classes: “smiling” and “not smiling.” The overlap in this dataset refers to the extent to which the facial features of individuals in these two classes. If the dataset contains many instances where individuals in both classes have similar facial expressions or features, it indicates a high overlap.

Metrics in Literature: Gupta et al.'s Data Quality Toolkit (DQT) [39] introduces a class overlap metric, which quantifies overlapping regions among different classes in a dataset by analyzing data points in overlapping regions of the data space. Additionally, the evaluation of class overlap in imbalanced classification settings is addressed through metrics such as the R -value and augmented R -value. Sejong's [102] R -value assesses the extent of overlap between classes by considering the proportion of instances in a specific class located in regions of the feature space shared with instances from other classes. Borsos et al. [25] enhance this approach with the augmented R -value, which considers the dataset's imbalance ratio (IR). It provides a weighted measure that combines class overlap and dataset imbalance for a more comprehensive understanding.

Impact on AI: Class separability impacts AI systems in classification tasks by influencing a model’s ability to accurately distinguish between different categories. The model can establish clear decision boundaries when classes are well-separated in the feature space. This leads to improved accuracy, faster training, better generalization to new data, and increased resilience against noise. This also enhances model interpretability, making the decision-making process more transparent. Conversely, low-class separability can lead to more misclassifications.

Summary: *Class separability is the level of similarity among diverse classes in a dataset. DQT introduces a metric to detect overlapping areas between classes, assessing data points that are close yet belong to different classes. The R-value is a measure proposed to quantify the degree of overlap in imbalanced classification problems.*

3.1.8 Discrimination Index. Discrimination in data refers to biases that may cause discriminatory outcomes in AI systems. It measures unfairness or unjust treatment towards individuals or groups that may be encoded in the data.

Example: Consider a company that uses an AI system to filter job applicants based on their resumes. The AI model is trained on historical data of successful candidates and their qualifications. If the historical data is biased and reflects discriminatory practices, such as favoring candidates from certain demographic groups, the AI model may unknowingly sustain those biases and lead to unfair outcomes in the hiring process.

Metrics in Literature: The Difference metric, introduced by Azzalini et al. [15], assesses the degree of bias within a dependency by comparing the confidence of that dependency with and without consideration of sensitive attributes. A higher Difference value indicates a stronger indication of unfair behavior. It is further supported by the Approximate Conditional Functional Dependency (ACFD). Additionally, the authors propose the P-Difference metric to measure the impact on dependency confidence by excluding one sensitive attribute at a time. This highlights the influential attributes contributing to unfairness.

Feldman et al. [50] introduce the “Likelihood Ratio” (LR_+) metric to measure disparate impact in a dataset, calculated based on sensitivity and specificity. It assesses the impact of the protected class on classification outcomes, but it requires a model trained on the dataset to generate results. Celis et al. [28] introduce two metrics for assessing discrimination based on sensitive attributes. The “representation rate” measures fairness by checking how well different attribute values are represented compared to a set threshold. The “statistical rate” evaluates fairness by analyzing the conditional probabilities of class labels given attribute values, helping to identify potential discrimination. These metrics provide quantitative fairness evaluation, offering flexibility based on specific application requirements.

Simonetta et al. [130] introduce two metrics that contribute to assessing fairness, bias, and completeness. The first metric, a “combinatorial metric,” evaluates dataset completeness by focusing on the distinct combinations of categories within specific columns. It quantifies completeness by comparing the total count of unique data points to the expected number of distinct combinations. In contrast, the second metric, based on “frame theory,” [36] offers a sophisticated approach to measuring fairness and bias. It treats the dataset as a matrix and applies operations to analyze the distribution of vectors within the matrix. Eigenvalues obtained from this matrix assessment measure the tightness of the frame, with uniform eigenvalue distribution indicating a balanced dataset. The Gini-Simpson index ([131]) is used to assess balance and homogeneity further. The combinatorial metric targets the representation of distinct combinations, while the frame theory-based metric considers the overall distribution and balance of the dataset’s vectors.

The Amazon SageMaker Developer Guide [76] uses various metrics for identifying bias in data. Class Imbalance (CI) measures sample distribution across the sensitive attributes, and Difference in Proportions of Labels (DPL) assesses outcome disparities. The guide also leverages various divergence metrics like Kullback-Leibler (KL), Jensen-Shannon (JS), and Lp-norm evaluate differences in the outcome distributions across demographic facets. Total Variation Distance (TVD) and Kolmogorov-Smirnov (KS) measure the degree of distribution divergence, and Conditional Demographic

Disparity (CDD) assesses outcome disparities within subgroups. In contrast, DQT [39] presents a disparate impact measure to quantify group discrimination, offering a score for assessing fairness. DQT also includes remediation strategies to mitigate bias in data.

Impact on AI: Unfair data can significantly impact AI systems by leading to biased and incorrect decisions. When AI is trained on biased or unrepresentative data, it can produce outcomes that systematically disadvantage certain groups based on race, gender, socioeconomic status, or other factors. This can result in discriminatory practices in critical areas such as hiring, lending, healthcare, and law enforcement.

Summary: *The discrimination index allows analysts to quantify and measure biases or discriminatory outcomes encoded in the data used for training and deploying AI models. The metrics, such as the Difference metric, P-Difference metric, Likelihood Ratio (LR+), representation rate, statistical rate, completeness metric, divergence metrics, and frame theory-based metrics provide quantitative measures to detect discriminatory behavior in the dataset.*

3.1.9 Data Split Ratio. Optimal data splitting in AI involves dividing a dataset into training, validation, and testing subsets to maximize the performance and generalization of the AI model. This metric aims to allocate the appropriate proportions of data for effective model training, hyperparameter tuning, and unbiased evaluation.

Example: Datasets are typically split with a ratio of 60/20/20 for training, validation, and testing. Split ratios of 70/15/15, 80/10/10 are also common. By splitting the dataset in this manner, we can ensure that the AI model is trained on diverse and representative data, fine-tuned for optimal performance, and tested on unseen instances, enabling a robust sentiment analysis system.

Metrics in Literature: Afendras and Markatou [9] suggests that irrespective of data distribution or analytic task, the optimal training sample size in cross-validation is identified as half of the total sample size. Similarly, Joseph [71] examines the ideal data splitting ratio for training and validation sets in linear regression models. The authors propose a ratio of $\sqrt{p} : 1$, where p is the number of parameters required to estimate a well-fitting linear regression model. The authors also present a strategy for determining p using variable selection methods. It suggests that this approach can be helpful in regression and classification tasks.

Impact on AI: An optimal split ensures that the model is trained on a sufficient amount of data to learn effectively while being validated and tested on separate, representative subsets to evaluate its generalization capabilities. An inappropriate split ratio can lead to overfitting or underfitting. Additionally, ensuring that the split maintains the statistical distribution of the data is crucial to avoid biases.

Summary: *The data split ratio, which involves dividing a dataset into training, validation, and testing subsets, is crucial for optimizing AI model performance. Affendras et al. suggest a guideline, proposing the training set to be half of the total dataset. With metrics like the $\sqrt{p} : 1$ ratio guide to ideal splits in linear regression models.*

3.1.10 Data Point Impact. It refers to the measure of the influence or significance of individual data points within a dataset. It quantifies the extent to which each data point contributes to an AI model or system's overall performance, accuracy, or behavior.

Example: Consider a patient medical record dataset with the patient's age, medical history, symptoms, and diagnostic outcomes. By analyzing the impact of data points, we can determine which specific patient records have a higher influence on the outcomes of an AI model built for disease diagnosis.

Metrics in Literature: Ghorbani and Zou [53] introduced the Data Shapley metric, based on the Shapley value from cooperative game theory. It assesses the impact of individual data points in supervised machine learning. The metric measures the contribution of each data point to the model's predictions, revealing its importance in model training. Various techniques, such as Monte Carlo and gradient-based methods, estimate a data point's impact by considering

its combinations with different subsets of the training data. Similarly, Wang and Jia [140] propose the Banzhaf value, a metric to assess data point value in the presence of noisy model performance scores. The authors investigate the robustness of data valuation in stochastic gradient descent, where randomness can lead to inconsistent value rankings.

In addition to these, several other methods have been developed to measure data importance, including the Leave-One-Out (LOO) [34] evaluation, which assesses model performance changes when individual data points are removed, influence functions [77] estimate a data point's effect based on loss function gradients, and k-nearest neighbors (KNN) approaches analyze proximity to decision boundaries. Core-set selection [16] identifies impactful subsets that perform similarly to the full dataset. Local Interpretable Model-agnostic Explanations (LIME) [113] provide insights by approximating complex models with interpretable ones around specific predictions.

Impact on AI: Evaluating the influence of specific data points allows for better understanding and optimization of the model, ensuring that the most relevant and informative data is used while minimizing noise and irrelevant information. This process can improve the accuracy, efficiency, and generalization capabilities of AI systems, leading to more reliable predictions and insights. Additionally, understanding the impact of data points helps identify and mitigate biases, ensuring fair and equitable AI outcomes.

Summary: *Data point impact refers to the measure of the influence or significance of individual data points within a dataset. In addition to feature importance metrics, such as Shapley and Banzhaf values and LIME, influence functions and ablation (removing data points) approaches like LOO are used to measure data point impact.*

3.1.11 Correctness. In terms of data values for AI, correctness refers to the degree of accuracy and fidelity in representing the information of the system being analyzed. It measures how closely the recorded data values align with the actual values they are supposed to represent. The goal of the metric is to minimize discrepancies between the recorded data and the ground truth.

Example: Consider a dataset containing temperature measurements from weather stations. The correctness of data values would involve ensuring that each recorded temperature value accurately reflects the actual temperature at the corresponding location and time. Inaccuracies in the recorded values compared to the actual temperature values would indicate a lack of correctness in the dataset.

Metrics in Literature: Pipino et al.'s [106] correctness metric quantifies data accuracy by calculating the complement of the error ratio. It focuses on clear criteria, like precision levels, and recognizes the contextual variations in error tolerance. This ensures a systematic evaluation of data correctness. Similarly, Kaiser et al. [73] involves comparing attribute values in the dataset (w_I) with their corresponding values in the real world (w_R). A domain-specific distance function, denoted as $d(w_I, w_R)$, quantifies the difference between these attribute values. The objective of the metric is to ensure normalization within the interval $[0, 1]$ without using a quotient.

Impact on AI: Correct and accurate data ensures that AI models can learn true patterns and make precise predictions to reduce the likelihood of errors and biases. Inaccurate data can lead to flawed outcomes, such as false positives or negatives, undermining the effectiveness of the AI application by spoiling user trust. It also enhances the generalization and robustness of models, enabling them to handle diverse inputs and perform well in real-world scenarios.

Summary: *In the context of data values for AI, correctness refers to how accurately the recorded data represents the ground truth. They involve calculations related to error ratios, precision criteria, contextual error tolerance, and comparisons between dataset attribute values and their real-world values.*

3.1.12 Timeliness. The timeliness of data refers to the time data collection and its relevance to the phenomenon or domain being studied in an AI application. It measures how closely the data captures the most relevant information available at the time of analysis or model training, ensuring that the data is up-to-date and reflects the present conditions.

The metric may vary depending on the question being solved by an AI application. In some applications, the latest data may be needed. In others, data relevant to the pattern an AI application is trying to predict. Existing metrics define timeliness based on the existence of the most recent data.

Example: Consider an AI system that predicts product demand for an e-commerce company. Timeliness of the data used for training the model would involve using the most recent sales data, customer preferences, and market trends. If the dataset contains sales data from several months ago, it may not accurately capture the current demand patterns and consumer behavior related to the current and near-future conditions. By ensuring timely data, such as incorporating daily or weekly sales updates, the AI model can better adapt to the changing market dynamics and provide more accurate demand predictions.

Metrics in Literature: Two studies, one by Kaiser et al. [73] and the other by Heinrich and Klier [58], introduce metrics for evaluating the timeliness of attribute values. Both use probability-based approaches to assess the freshness and relevance of data. Kaiser et al.'s metric uses an exponential distribution model to calculate attribute decline rates. It indicates the average proportion of outdated attribute values within a specified time frame. It quantifies attribute age based on data quality assessment time and data acquisition time to offer an automated and interpretable measure of timeliness. In contrast, Heinrich et al. propose a probability-based currency metric (PBCM) that assesses data item timeliness using a set of probabilities. These probabilities are derived from diverse data sources and methods, including expert assessments, historical data analysis, and machine learning algorithms. While both metrics share a foundation in probability theory, Kaiser et al.'s metric focuses on attribute-level timeliness. In contrast, Heinrich et al.'s PBCM assesses data item timeliness, offering flexibility for various data types and contexts.

Blake and Mangiameli [21] propose a method to assess data timeliness using a classification model. They introduce a metric called T , which measures the impact of introducing new and more current data into the dataset. To evaluate data volatility and timeliness, the authors replace a percentage of old instances with new ones in the training data while assuming a fixed currency. The T metric is computed based on the total quantity of records in the training and test data and the number of replacement records introduced for reclassification.

Impact on AI: Timely data is critical to AI applications trying to understand patterns, although the 'time' refers to the question an AI application is trying to answer. Applications trying to understand and predict patterns related to the latest trends require the most recent data.

Summary: *The timeliness metric assesses the recency and relevance of the data about the current state of the phenomenon or domain it represents. Kaiser et al. and Heinrich et al. introduced metrics for assessing data timeliness. Blake et al. evaluate data timeliness through a classification model by assessing data replacement's impact on model performance.*

3.1.13 Privacy Leakage. Data privacy in the context of AI refers to protecting and preserving sensitive information contained within datasets, particularly concerning the risk of unauthorized disclosure or inference of private details.

A notable technique used to assess privacy is Membership Inference Attacks (MIA). MIA determines whether a specific data record was included in the training dataset used to build an AI model. By exploiting patterns and characteristics of the model's outputs, one can infer whether or not a particular data point was part of the training set. This raises concerns about the privacy of individual data records and the potential for unauthorized access to such information.

Another dimension of AI-related privacy issues emerges with the use of synthetic data. Synthetic data is generated to mimic real data while preserving privacy by avoiding the use of actual sensitive information. However, suppose the synthetic data is too close to the real data. In that case, it can reveal private details, making it vulnerable to privacy attacks. The balance between creating useful synthetic data and ensuring privacy remains a significant challenge in AI.

Evaluating the closeness of the synthetic data to the real data can be useful in determining the potential privacy leaks that can emerge during AI applications.

Example: Consider a healthcare dataset used to train a machine-learning model for disease diagnosis. The dataset contains sensitive medical information about patients, including their symptoms, test results, and diagnoses. Without proper setting of privacy for data, one may determine whether a specific patient’s data was used during training, which poses a privacy risk as it could reveal a patient’s medical condition or other confidential information.

Metrics in Literature: Vatsalan et al. [137] focus on mitigating re-identification risks in released datasets within the education sector. In contrast to existing approaches that often assume prior knowledge, the proposed method employs a Markov Model to quantify re-identification risks by using all available information in the datasets, including event-level details associating multiple records with the same individual and exploring correlations between attributes.

In a broader context of privacy metrics in literature, works such as SHAPR introduced by Duddu et al. [42] and Song et al.’s [132] privacy risk metric are noteworthy. SHAPR quantifies the susceptibility of individual training data records to membership inference attacks by calculating Shapley values, emphasizing the influence of specific data points on model predictions. In contrast, Song et al.’s metric assesses the likelihood of a data record being present in a model’s training dataset, focusing on evaluating the privacy risk from an adversarial perspective. Both metrics aim to address privacy concerns by assessing the privacy risks associated with data records. However, they depend on the specific AI model used in the context.

In another study, Carlini et al. [27] introduced the Attack Success Rate (ASR) as a metric for privacy leakage. ASR measures the success of an attack in predicting if a specific example is part of the training dataset. It is calculated by training a model, performing an attack, and evaluating the attack’s success in correctly predicting membership. However, ASR can only be measured after training a model and conducting an attack. Alongside these contributions, Bezzi [19], Longpré et al. [89], and Arca and Hewett [14] proposed entropy-based metrics to offer valuable insights into dataset anonymity to measure unpredictability and disorder in released data.

Regarding synthetic data privacy leaks, Aindo AI’s [3] privacy score assesses the privacy risk of synthetic data by comparing proximity ratios between real and synthetic datasets. The process involves calculating the Train to Train Proximity Ratio (TTPR) for real data and the Train to Synthetic Proximity Ratio (TSPR) for synthetic data. The score is derived from the ratio of records below a specific threshold in both distributions. A score of 100 indicates minimal privacy risk, while lower scores reflect higher risk.

Impact on AI: Evaluating privacy in data impacts AI by ensuring that personal information is protected while enabling AI systems to function effectively. Privacy assessments help identify and mitigate risks associated with data breaches, intentional or unintentional or malicious accesses, and misuse, which are critical as AI technologies increasingly process personal data. Additionally, addressing privacy concerns helps avoid ethical and legal repercussions, such as biases in AI models and non-compliance with regulations like GDPR [48]. Privacy evaluations also encourage the adoption of privacy design principles in data, ensuring that AI systems are developed with privacy considerations. This will ultimately lead to more ethical and responsible AI deployment.

Summary: *The privacy leakage metric in AI aims to safeguard sensitive information from unauthorized disclosure. Privacy metrics range from a Markov Model to address re-identification risks to SHAPR to quantify membership privacy risk using Shapley values. Attack Success Rate (ASR) is a measure of privacy leakage in membership inference attacks. Entropy-based metrics are also explored to assess dataset anonymity.*

3.1.14 Sample Size. Sample size refers to the number of data points or instances selected from a population to be included in a dataset for training an AI model. It represents the subset of data used to make inferences or predictions.

Example: If researchers collected data from 500 patients to predict the likelihood of a disease based on patient characteristics, such as their age, gender, medical history, and test results, the sample size of the dataset would be 500.

Metrics in Literature: Alwosheel et al. [13] investigate the sample size requirements for accurate decision-making analysis using Artificial Neural Networks (ANNs). They introduce a new guideline, “factor 50,” recommending that the optimal dataset size for an ANN should be the number of adjustable parameters in the model multiplied by 50. This guideline is more conservative than the commonly used “factor 10” rule-of-thumb found in the literature [56]. However, determining the appropriate sample size for ANN-based decision-making is complex and depends on the model’s complexity, which is difficult to predict. To address this, the authors propose three approaches: evaluating ex-post if the training sample size was sufficient, using prior studies or literature to estimate the optimal number of neurons, or referring to existing literature to calculate the expected number of neurons required for the analysis.

Impact on AI: An adequate sample size ensures that AI models can learn effectively from the data, capturing the underlying patterns without overfitting or underfitting. Small sample sizes may lead to overfitting, where the model performs well on training data but poorly on new, unseen data due to learning noise or random variations rather than the true signal. Conversely, excessively large sample sizes can be resource-intensive and may not significantly improve model performance beyond a certain point. Rajput et al. [110] have shown varying impacts of sample size on model accuracy, with some models performing better with larger datasets, while others may not see significant gains past a certain threshold.

Summary: *In AI, sample size denotes the quantity of data points chosen from a population for analysis or model training. The “factor 50” determines sample size in decision-making with Artificial Neural Networks (ANNs), contrasting with the commonly used “factor 10”.*

3.1.15 FAIR Principle Compliance Score. FAIR compliance of a dataset for AI refers to the degree to which a dataset adheres to Findability, Accessibility, Interoperability, and Reusability (FAIR) principles [2]. In AI, a dataset should be well-documented, easily discoverable, accessible, formatted in a way that facilitates integration and analysis, and can be effectively reused for AI applications.

Example: Consider a dataset containing information about different types of cars for training an AI model tasked with predicting car prices. The FAIR principles applied to this dataset would be:

- **Findability:** The dataset should contain a unique identifier to be easily discoverable with standardized metadata containing details about car attributes, including make, model, year, mileage, engine type, and other existing features. Clear information on the dataset’s source and data collection methodologies is also essential.
- **Accessibility:** The dataset’s accessibility requires a platform with controlled access, considering privacy or licensing constraints. Secure access should be given through proper authentication and authorization methods.
- **Interoperability:** The dataset should be structured and formatted according to established storage standards like CSV or JSON to ensure easier integration with AI systems and tools. A well-defined schema must be included with the dataset, specifying the meaning and format of each attribute. This promotes consistency and compatibility across diverse AI models and applications.
- **Reusability:** The dataset should come with clear usage licenses or permissions that outline how the dataset can be used. Additionally, comprehensive documentation should be included with the dataset, providing details about data collection procedures, preprocessing steps (such as data cleaning or feature engineering), and any potential biases or limitations in the data.

By following these FAIR principles, a structured dataset of car information becomes a valuable resource for AI researchers and practitioners. It facilitates the development of accurate car price prediction models and promotes transparent and ethical AI practices.

Metrics in Literature: Wilkinson et al. [145] introduced a comprehensive FAIR compliance measurement framework, aligning with the four FAIR sub-principles. The framework includes 14 universal metrics corresponding to specific sub-principles, covering aspects such as identifier schemes, metadata accessibility, findability, access protocols, metadata longevity, knowledge representation languages, linking, adherence to standards, and provenance. This flexible approach facilitates the objective assessment and improvement of FAIR compliance in various digital resources applicable across scholarly domains. In a similar framework introduced by Clarke et al. [30], FAIR metrics and FAIR rubrics play an important role, allowing users to associate digital resources with existing metrics. The authors emphasized the manual or automated quantification of FAIR metrics and contextual assessments using FAIR rubrics. This empowers users to evaluate and enhance data correctness in diverse projects. DataONE (Data Observation Network for Earth) [70] is a community-driven initiative that has adopted metrics to measure FAIR principle compliance of research data. Based on the FAIR criteria, the DataONE FAIR suite generates comprehensive assessment scores based on the metadata.

Impact on AI: A dataset’s compliance with the FAIR principles can impact AI by enhancing data management and accessibility. By ensuring data is easily discoverable and accessible, AI systems can be trained on diverse and comprehensive datasets, leading to robust and accurate models. Interoperability allows different AI systems to work together seamlessly, while reusability enables researchers to build upon existing datasets and models, enabling reproducible innovation and efficiency. Moreover, FAIR compliance promotes collaboration and transparency that accelerates AI research and ensures the reproducibility of results. It is essential for building trust in AI technologies.

Summary: *The FAIR score measures the extent to which a dataset adheres to the principles of Findability, Accessibility, Interoperability, and Reusability. Both Wilkinson et al. and Clarke et al. created frameworks to assess the FAIR compliance of digital resources aligned with FAIR sub-principles. Several other FAIR compliance score evaluation websites use the same 14 principles that Wilkinson et al. defined.*

Data Readiness Metrics for Structured Data

In evaluating structured data readiness for AI applications, we considered metrics spanning completeness, outliers, mislabels, duplicates, feature relevancy, class imbalance, class separability, discrimination index, data split ratio, data point impact, correctness, timeliness, privacy leakage, sample size, and FAIR compliance score.

3.2 Unstructured Data

Unstructured data, including textual, image, and audio data, present unique challenges in evaluating and ensuring readiness for AI applications. In this section, we provide a brief overview of the metrics and scoring mechanisms (shown in the right half of Table 2) used to assess the suitability of unstructured data for AI. While the evaluation techniques for structured data are well-established, we will highlight the relevant measuring techniques from structured data that can be applied to unstructured data. By examining these evaluation methods, we gain insights into the readiness, relevance, and accuracy of unstructured data. It enables us to make informed decisions when using such data in AI models.

3.2.1 Textual Data.

3.2.1.1 Lexical Diversity. Lexical diversity measures the richness, variety, and complexity of the vocabulary used within the text. It offers insights into the level of linguistic expression, domain coverage, and potential challenges in

understanding and processing textual data. A higher lexical diversity score indicates a broader range of words and linguistic patterns, providing a solid foundation for AI to learn from the data.

Example: Consider two datasets: *A* with a low lexical diversity score and *B* with a high lexical diversity score. Dataset *A* has a repetitive and limited vocabulary, such as a chatbot dialogue focused on a specific topic. The low lexical diversity suggests a constrained range of language, potentially limiting the chatbot’s ability to respond effectively to diverse user inputs. In contrast, dataset *B* consists of a collection of news articles from various domains, exhibiting a wide range of vocabulary and language styles. The high lexical diversity in dataset *B* indicates a greater readiness for AI applications. It offers a more comprehensive representation of language usage, enabling models to generalize across different topics and understand a broader range of inputs.

Metrics in Literature: Type-Token Ratio (TTR) introduced by Templin [135] measures lexical diversity in textual data. It calculates the ratio of unique word types (vocabulary size) to the total number of tokens (words or other linguistic units) in the text. TTR provides an estimate of the text’s richness and variety of vocabulary. A higher TTR indicates greater lexical diversity, suggesting a more comprehensive range of word usage in the text.

McCarthy et al.’s [95] metrics, vocd-D and HD-D, focus on measuring lexical diversity in textual data. vocd-D uses type-token ratios (TTR) from randomly selected text samples to derive a D coefficient to represent lexical diversity. HD-D, on the other hand, employs the hypergeometric distribution to directly calculate the probabilities of word occurrence in randomly selected samples, resulting in the HD-D index. While both metrics assess lexical diversity, vocd-D uses random sampling and the D coefficient, whereas HD-D approximates results for all possible word arrangements. The correlation between HD-D and vocd-D is high, offering alternative methods of measuring lexical diversity. Additionally, McCarthy introduces the Measure of Textual Lexical Diversity (MTLD) [94], which quantifies lexical diversity by considering unique words and segment length. MTLD assesses the lexical diversity of longer texts, providing insights into vocabulary richness throughout the text.

Impact on AI: Measuring lexical diversity can significantly impact AI by influencing how language models are developed and evaluated. In AI, particularly in Natural Language Processing (NLP) and machine translation, assessing lexical diversity can help identify biases and limitations in language models. Reviriego et al. [112] state that AI-generated texts often exhibit lower lexical diversity compared to human-generated texts, leading to a feedback loop where AI models trained on such data become less effective over time. Ensuring high lexical diversity in AI outputs can improve the quality and accuracy of translations and other language-based AI applications. This ultimately contributes to more inclusive and representative AI systems.

Summary: *Lexical diversity, necessary for assessing linguistic richness in textual data, is measured by metrics like TTR and vocd-D, evaluating the ratio of unique word types to total tokens with standardized scores. HD-D offers a probability-based alternative for assessing diversity. MTLD divides the text into segments and calculates the average segment length where vocabulary richness falls below a threshold.*

3.2.1.2 Term Importance. Term importance evaluates the significance of individual terms in textual data. It measures the relevance and impact of terms in capturing the essence and meaning. Term importance considers various factors, such as the frequency of a term within the text, its rarity across the entire dataset or corpus, and its discriminative power in distinguishing the text from others. By assigning weights or scores to terms based on their importance, this metric enables AI models to focus on key terms that carry valuable semantic information and discard less informative or common terms, assisting in feature selection, document ranking, or topic extraction.

Example: In a dataset of news articles, the term “pandemic” might be considered highly important due to its relevance in conveying crucial information regarding COVID-19. On the other hand, common words like “the” or “and”

would be assigned lower importance scores as they provide little discriminative power or unique information. AI models can prioritize and focus on the most significant terms by analyzing term importance, enabling better understanding, classification, or summarization of textual data.

Metrics in Literature: TF-IDF (Term Frequency-Inverse Document Frequency) is a widely used quantitative metric to evaluate the importance of words in a document or collection of documents (Ramos et al. [111], Simha [129], Qaiser and Ali [109]). It combines term frequency (TF), measuring word occurrence in a document, and inverse document frequency (IDF), assessing the rarity of a word's appearance across the entire corpus. TF-IDF quantifies a word's significance by considering its frequency in a document and its discriminative power across the corpus. The TF-IDF score is computed by multiplying TF and IDF values for each word. High TF-IDF scores indicate that words that are both frequent in a document and rarely appear in the corpus, which makes them essential for the context. This metric helps identify essential features and characteristics, enabling information retrieval, document classification, and keyword extraction in natural language processing.

Impact on AI: Measuring term importance is needed for tasks such as information retrieval, text summarization, and sentiment analysis, where distinguishing between relevant and irrelevant terms can greatly impact the accuracy and relevance of the output. By measuring term importance accurately, AI systems can prioritize critical information that leads to precise and contextually relevant results. Furthermore, TF-IDF aids in reducing computational resource requirements in AI training by focusing on significant terms.

Summary: *Term importance assesses the significance of individual terms in textual data for AI applications. TF-IDF is a widely used metric that combines term frequency and inverse document frequency.*

3.2.1.3 Readability Score. Readability score is a quantitative metric used to assess textual data complexity and ease of understanding, enabling effective preparation for AI. It measures various linguistic factors, such as the length of sentences, choice of words, and syntactic structure, to determine the readability of a text. By considering these factors, readability scores provide valuable insights into the suitability of text for different target audiences and applications. AI applications can be optimized by selecting appropriate training data using readability scores, ensuring that the content aligns with the desired level of comprehension and avoids potential barriers to understanding. This metric is important in enabling the development of more accessible and contextually appropriate language models.

Example: Consider a scenario where an AI model is trained to generate educational content for an elementary school. In this case, readability scores can be used to assess the complexity of different texts and select appropriate training data. The readability scores can help identify texts that align with the target audience's reading abilities by analyzing factors such as sentence length, vocabulary difficulty, and grammatical complexity. This ensures that an AI model is trained on comprehensible and engaging content for young learners, promoting effective knowledge transfer and enhancing the overall learning experience.

Metrics in Literature: The Flesch-Kincaid Grade Level, introduced by Flesch [51] estimates the approximate education (grade) level needed to comprehend a given text using the average number of words in sentences and that of syllables in words. The resulting score represents the education level required to understand the text. Lower grade values indicate higher/easier readability. This metric provides a standardized measure for assessing text comprehension and enables content tailoring to suit specific audience reading abilities.

The Coleman-Liau Index, developed by Coleman and Liau [33], is another readability scoring method that assesses the reading level of a text based on factors like letter count and sentence length. Unlike the Flesch-Kincaid Grade Level, which considers syllable count, the Coleman-Liau Index calculates the grade level based on the average number of

letters and sentences per 100 words. A score of 5 on the index indicates that the text is at a reading level equivalent to that of a fifth grader in the US schooling system. It is widely used in schools and provides a quick measure of readability.

Furthermore, the Gunning Fog Index introduced by Robert Gunning Associates [5] offers an additional perspective on the readability of textual data for AI model training. Unlike the Coleman-Liau Index and the Flesch-Kincaid Grade Level, which focus on sentence and letter count, the Gunning Fog Index considers both the percentage of complex words and the sentence length. It generates a score between 0 and 20, with lower scores representing easier readability.

Impact on AI: By evaluating readability, developers can adjust the complexity of the training data to match the intended audience’s comprehension level, which is particularly important in fields like education and healthcare, where clear communication is essential. By ensuring that training data is readable, AI models can avoid perpetuating biases that arise from overly complex or inaccessible language. This will also reduce disparities in information access.

Summary: *The readability score is a quantitative metric used to assess the complexity of textual data and ease of understanding. Popular readability metrics calculate the grade level needed to comprehend a text based on average words per sentence and syllables per word and by considering the percentage of complex words and sentence length.*

3.2.1.4 Topic Coherence. This measure evaluates the readiness of textual data for AI by assessing the logical and semantic connectedness within a set of topics or a document. It quantifies the degree to which words within a topic exhibit meaningful relationships and contribute to a coherent theme. A higher coherence score shows stronger semantic coherence, indicating that the words are closely related and provide a clearer understanding of the topic. By evaluating topic coherence, AI practitioners can ensure that the textual data is well-structured, coherent, and ready for AI model training. This would promote accurate and meaningful text generation and facilitate better comprehension and usage of data by AI algorithms.

Example: Consider a collection of news articles about technology trends. Topic coherence can be measured to evaluate the readiness of this textual data for AI. Data can be segmented into topics like “Artificial Intelligence,” “Blockchain Technology,” and “Internet of Things” by applying topic modeling techniques. Topic coherence analysis assesses the semantic relationships between words within each topic. A high coherence score would indicate that words within a topic, such as “machine learning,” “algorithm,” and “predictive analytics” in the “Artificial Intelligence” topic, are closely related and contribute to a coherent theme. This demonstrates that the textual data is well-prepared for AI, as it exhibits clear and meaningful topic structures.

Metrics in Literature: Roder et al. propose the “CV coherence score” [117] to quantify topic coherence in textual data by assessing the semantic similarity between words within a topic. This is computed using Latent Dirichlet Allocation [22] to extract topics from the text and then measure the pairwise word similarity within each topic to evaluate how well the words contribute to a coherent theme. Higher scores indicate stronger semantic relatedness and better topic coherence. Despite its popularity, the CV coherence score has limitations, such as sensitivity to topic size, potential mismatch with human judgment, and inability to capture higher-level coherence aspects.

Mimno et al. introduce “UMass coherence score” [96] to assess the topic coherence of a set of topics extracted from a text corpus. It evaluates coherence by considering the probability of word co-occurrences within topics. The score is computed by aggregating the logarithm of the co-occurrence probabilities of all word pairs within and across topics. A higher UMass coherence score indicates stronger word co-occurrence patterns and better topic coherence, while a lower score suggests weaker word associations and less coherent topics. Compared to the CV coherence score, the UMass coherence score offers a more reliable evaluation of topic coherence, accounting for topic size, aligning better with human judgment, and directly measuring word co-occurrence probabilities.

Newman et al. proposed “UCI coherence score” [99] to assess the coherence of topics generated by a topic model. This measures the semantic relatedness and meaningful connections between words within a topic by calculating the semantic association between word pairs based on their co-occurrence in sliding windows. The score is computed using a specific equation considering the probabilities of observing individual words and word pairs in a sliding window. Higher UCI coherence scores indicate stronger associations between word pairs and better topic coherence.

Impact on AI: By evaluating and optimizing for topic coherence, developers can ensure that the AI models produce topics that align better with human understanding, which will improve the semantic interpretability of the model outputs. This evaluation can lead to accurate and reliable AI systems, as coherent topics are more likely to represent genuine patterns in the data rather than random associations.

Summary: *Topic coherence is a metric used to evaluate the logical and semantic connectedness within a set of topics or a document. Coherence scores, such as the CV, UMass, and UCI coherence scores, quantify the semantic similarity or word co-occurrence probabilities within topics to assess their coherence.*

3.2.1.5 Bias Indicator. A bias indicator is a measure used to prepare textual data for AI by quantifying and identifying potential biases in the text. It serves as a tool to assist in detecting and mitigating biased content, ensuring that AI systems can make more informed and fair decisions. The bias indicator analyzes various linguistic and semantic features within the text, such as word choice, sentence structure, and contextual references, to assess the potential presence of biases related to factors like gender, race, religion, or other sensitive attributes. By providing a quantitative assessment of bias, the indicator helps developers and users understand the underlying biases in the data. It enables them to address and mitigate these biases during AI model development, promoting more equitable and unbiased AI systems.

Example: Consider a dataset of job application statements used to train an AI system to evaluate and rank applicants. A bias indicator can be used to analyze the textual data, ensuring fairness and reducing bias. The indicator would examine language and semantic patterns to identify potential biases, such as gender, race, or age-related biases. For instance, if the indicator detects a bias where certain occupations or characteristics are consistently favored or discriminated against, it would flag it for further examination. This enables developers to address and mitigate biases in the dataset before training the AI model. This promotes equal opportunities and minimizes discriminatory outcomes during the evaluation process. The bias indicator plays a crucial role in preparing the textual data, allowing the creation of more objective and unbiased AI systems for job application assessments.

Metrics in Literature: Papakyriakopoulos et al. [104] propose a robust bias measurement technique using word embeddings and cosine similarity calculations. The method involves defining word pairs to represent different types of discrimination and creating a list of concepts for measuring bias. For variable concepts, which change based on social groups, the bias calculation equation considers the cosine similarity between word pair embeddings and concept embeddings. A modified bias calculation equation is used for non-variable concepts, which remain the same irrespective of social groups. By quantifying the differences in cosine distances, the proposed method comprehensively analyzes bias in different contexts, accounting for variable and non-variable concepts.

Impact on AI: Textual data often contains subtle biases related to gender, race, ethnicity, and other social factors, which can be encoded in language patterns, word associations, and context. If these biases are not identified and addressed before training, an AI model may learn to reinforce these biases, leading to unfair outputs in tasks such as language generation, sentiment analysis, or text classification. This preventive approach leads to ethical and unbiased AI systems.

Summary: *A bias indicator is used to quantify and identify potential biases in textual data for AI applications. A bias indicator that uses word embeddings and cosine similarity calculations is one of the most used metrics.*

3.2.2 Multimedia Data. In this subsection, we explore the key aspects of DRAI for image, speech, and video data. We concentrate on quality evaluation metrics for each data type. This exclusive emphasis on quality evaluation metrics is fundamentally crucial in determining the reliability and usability of these types of data for AI.

3.2.2.1 Image Quality. As a metric to measure the readiness of image data for AI, image quality refers to the degree to which an image accurately represents the visual content it intends to depict. It includes various aspects such as sharpness, color accuracy, clarity, and the absence of artifacts or distortions. Assessing image quality is crucial in AI applications as it directly impacts the reliability and effectiveness of algorithms that rely on visual data. High-quality images provide a solid foundation for AI to extract meaningful features, recognize patterns, and make accurate decisions.

Example: In an AI system designed for autonomous driving, image quality plays a vital role in ensuring the accuracy and reliability of object detection and recognition. High-quality images with clear details and accurate colors allow the AI system to accurately identify pedestrians, vehicles, and traffic signs, enabling it to make precise decisions in real time. Conversely, low-quality images with blurriness or artifacts may lead to misinterpretation of objects or false detection, compromising the safety and efficiency of autonomous driving systems.

Metrics in Literature: The field of image quality assessment is a heavily researched area, constantly evolving to meet the demands of various applications. In this discussion, we explore some of the favored and widely used measures developed to evaluate image quality accurately.

Mean Squared Error (MSE) and Peak Signal-to-Noise Ratio (PSNR) [8] are essential metrics for evaluating image quality and are widely used in AI applications. MSE calculates the average squared difference between the pixel values of a reference and processed image, providing a quantifiable representation of distortion. PSNR, derived from MSE, offers a perceptual quality metric expressed in decibels, comparing the maximum signal power with the average squared error. Higher PSNR values indicate high-quality images. While MSE alone may not align with human perception of quality, PSNR provides a standardized measure that accounts for human visual perception.

Wang and Bovik [141] present a universal image quality index, distinct from MSE and PSNR, defined mathematically for independence from specific images, viewing conditions, and observers. This index incorporates correlation coefficient, mean luminance difference, and contrast similarity, offering practical advantages in simplicity and universality. In related work, Wang et al. [142] introduce the Structural Similarity (SSIM) index as an alternative to traditional error-sensitive methods, emphasizing structural similarities in luminance, contrast, and structure to align more closely with human visual perception. Furthermore, their Multi-Scale Structural Similarity (MSSIM) approach, as proposed in [143], extends SSIM by considering variations in resolution and viewing conditions, providing increased flexibility and demonstrating improved performance.

Similarly, Sarnoff's JND-Metrix [6] takes into account Just Noticeable Differences (JND) based on the Human Visual System (HVS) principles. Unlike traditional metrics such as PSNR or MSE, the JND-Metrix considers the sensitivity and perception of the HVS to different types of image distortions. By incorporating knowledge of the HVS, including factors like contrast sensitivity, spatial masking, and visual attention, the perceptual impact of image distortions is quantified more accurately. JND-Metrix measures the visibility of distortions by estimating JND thresholds for various distortions, indicating the level of distortion perceptually distinguishable from the original image.

Two notable advancements in image quality assessment include Sheikh et al.'s [126] Visual Information Fidelity (VIF) criterion and Chandler et al.'s [29] Visual Signal-to-Noise Ratio (VSNR) metric. VIF, a full-reference method, evaluates the correlation between image information and visual quality, outperforming traditional metrics in various scenarios. On the other hand, VSNR uniquely considers human visual system properties, including near-threshold and

supra-threshold perception. It provides a more accurate representation of perceived quality by addressing contrast sensitivity and global precedence.

The PyTorch Image Quality (PIQ) [74] [75] provides a diverse array of image quality metrics designed for various assessment requirements. Among full-reference metrics, FSIM (Feature Similarity Index Measure) [149] is significant for its ability to evaluate image quality by analyzing structural and color features. GMSD (Gradient Magnitude Similarity Deviation) [147] is also noteworthy which focuses on the gradient magnitudes to assess image quality. In the no-reference category, BRISQUE (Blind/Referenceless Image Spatial Quality Evaluator) [97] stands out as it evaluates quality using natural scene statistics without needing a reference image. This makes it particularly useful in real-world applications. Within the distribution-based category, FID (Frechet Inception Distance) [59] is a key metric that is used for assessing generative models by measuring the similarity between distributions of real and generated images.

Lakhani [78] and Sabottke and Spieler [121] demonstrate that resolution is a critical image quality metric when developing deep learning models for medical imaging and radiology applications. Higher image resolutions can lead to improved AI model performance, especially in detecting specific medical conditions. However, it is essential to strike a balance with computational resources to avoid limitations in the training process. The appropriate image resolution for a given task can vary based on several factors, including the image data type, the specific AI model or algorithm used, and the application's requirements. Additionally, blurriness is another factor that can impact the performance of AI models, particularly those tailored for image recognition, object detection, and segmentation tasks. Additionally, blurriness, a significant factor affecting AI model performance, has been extensively studied. Lin et al.'s [85] method estimates contrast decrease on edges, while Marziliano et al. [93] gauge blurriness by analyzing edge spread, providing valuable insights into overall blurriness levels.

Impact on AI: Measuring image quality impacts AI training by ensuring that the data used to train models is clear, accurate, and free from distortions, which directly influences the model's ability to learn effectively. High-quality images allow AI models to extract meaningful features and recognize patterns more accurately, leading to better generalization and performance. Conversely, low-quality images with noise, blurriness, or artifacts can introduce errors, reduce the model's learning efficiency, and lead to inaccurate predictions.

Summary: *Image quality refers to how accurately an image represents its visual content and covers aspects such as sharpness, clarity, color accuracy, and the absence of artifacts. Commonly used metrics for image quality assessment include MSE, PSNR, SSIM, MSSIM, JND-Metrix, VSNR, and VIF. These metrics provide quantitative measurements of image distortion and quality. Additionally, considering factors like image resolution and blurriness is crucial to balance improved model performance and computational resources.*

3.2.2.2 Speech Quality. Speech quality refers to the overall perceived clarity, intelligibility, and fidelity of speech signals. It captures the extent to which speech data effectively conveys the intended information and is free from distortions, noise, or artifacts that could impact its understandability. Speech quality includes factors such as signal clarity, absence of background noise, the naturalness of speech, and the ability to capture and reproduce various linguistic and acoustic properties accurately. A high level of speech quality in the data ensures that AI models can effectively process and interpret speech inputs, leading to more accurate and reliable performance across speech-related tasks, such as speech recognition, synthesis, or understanding.

Example: Speech quality is important for the optimal functioning of voice-controlled virtual assistants. In a scenario with high speech quality, a user's command is delivered with clarity and minimal background noise. This ensures the virtual assistant accurately interprets and executes the request. With low speech-quality data with distortions and noise, an AI system may struggle to comprehend the command, leading to potential errors in execution.

Metrics in Literature: The ITU-T (International Telecommunication Union – Telecommunication Standardization Sector) introduced Mean Opinion Score (MOS) [66], which involves human listeners rating the perceived speech quality. MOS helps to understand how well the speech data aligns with human expectations, enabling improvements based on listener feedback. On the other hand, Perceptual Evaluation of Speech Quality (PESQ), introduced by Rix et al. [114], objectively quantifies the quality of processed or degraded speech signals. It uses a model that simulates human auditory perception to calculate prediction scores. Segmental Signal-to-Noise Ratio (SNRseg), introduced by Jayant and Noll [69], is another metric that gauges the ratio of the clean speech signal to the background noise within short segments. By providing a localized evaluation, SNRseg assists in identifying segments of speech that may be affected by noise, leading to targeted noise reduction or enhancement techniques.

Widely used objective metrics for evaluating speech quality and intelligibility are Short-Time Objective Intelligibility (STOI), Perceptual Speech Quality Measure (PSQM), Itakura-Saito Spectral Distortion (IS), and Objective Difference Grade (ODG). STOI, introduced by Taal et al. [134], primarily assesses the similarity between clean and degraded speech signals, focusing on how well the degraded speech retains intelligibility compared to the original signal. In contrast, PSQM, introduced by Beerends and Stemerdink [18], explores various factors affecting speech quality, including distortion, noise, and echo. IS [67] quantifies spectral distortion in the frequency domain, helping to understand the impact of different operations on speech quality. On the other hand, ODG, specified in [4], evaluates the perceived difference in quality between two speech signals, aiding in comparing speech processing algorithms or system configurations. These metrics offer distinct perspectives on speech quality, with STOI and PSQM emphasizing intelligibility and overall quality, IS focusing on spectral distortion, and ODG providing a comparative quality assessment.

Impact on AI: Evaluating speech quality can significantly impact AI systems, particularly in speech recognition, synthesis, and processing. By assessing audio quality, AI models can improve their accuracy in recognizing speech across various conditions, including noisy environments and diverse accents. For speech synthesis, quality evaluation leads to natural-sounding and intelligible artificial voices. In audio processing, it enables the development of better noise cancellation and audio enhancement algorithms. Moreover, speech quality assessment improves AI training by providing more accurate data labeling and helping to filter out low-quality samples.

Summary: *Speech quality refers to the perceived clarity, intelligibility, and fidelity of speech signals. Metrics such as Mean Opinion Score (MOS), Segmental Signal-to-Noise Ratio (SNRseg), Short-Time Objective Intelligibility (STOI), Perceptual Evaluation of Speech Quality (PESQ), Perceptual Speech Quality Measure (PSQM), Itakura-Saito Spectral Distortion (IS), and Objective Difference Grade (ODG) are used to assess speech quality objectively and subjectively.*

3.2.2.3 Video Quality. As a metric to evaluate the readiness of video data for AI, video quality refers to the overall visual fidelity and perceptual coherence of a video sequence. It assesses the accuracy with which the video represents the original content, ensuring that crucial details, structures, and visual cues are preserved without significant degradation or distortion. Evaluating video quality involves objective and subjective criteria, where objective assessments employ computational algorithms to analyze pixel-wise differences and feature similarities. In contrast, subjective evaluations incorporate human perception and user feedback. By considering video quality as a crucial factor, AI practitioners can determine the suitability of video data for their applications, ensuring that the data meets the necessary standards for achieving accurate and reliable AI-driven results.

Example: In preparation for developing an AI-powered autonomous driving system, a team of engineers collects a vast amount of video data from various in-car cameras and external sensors. To evaluate the readiness of this video data for AI training, they carefully assess its quality. In this context, video quality refers to the video sequences' overall visual fidelity and coherence, ensuring that critical details are preserved without significant degradation. The engineers

analyze the videos to detect potential artifacts or distortions that may impact the AI system’s perception algorithms. They also involve human evaluators to rate the video quality subjectively based on their perceived visual appeal.

Metrics in Literature: PSNR, SSIM, VIF, and VSNR are image quality metrics, discussed in Section 3.2.2.1, that can also be effectively applied to measure the quality of video data. When applied to videos, these metrics analyze individual frames’ spatial quality and temporal coherence to capture the visual information across consecutive frames. PSNR, SSIM, VIF, and VSNR provide objective insights into video quality and its perceptual fidelity by assessing the pixel-wise differences, structural similarity, information fidelity, and sensitivity to visual distortions. This adaptability makes them valuable tools for researchers and practitioners seeking to quantify and improve the visual experience in video applications. Furthermore, like speech quality assessments discussed in Section 3.2.2.2, video quality evaluations often employ the Mean Opinion Score (MOS) methodology. MOS in video quality involves presenting video samples to human viewers who rate their subjective perception of the video’s quality. The average MOS scores offer valuable insights into human preferences and perceptual experiences, complementing the objective metrics’ findings to make more informed decisions regarding video data suitability for various AI applications.

VQM (Video Quality Metric) introduced by Huynh-Thu and Ghanbari [62], VMAF (Video Multimethod Assessment Fusion) introduced by Netflix [23], and OPTICOM’s PEVQ (Perceptual Evaluation of Video Quality) [1] are advanced and specialized metrics specifically designed to assess the quality of video data. VQM focuses on replicating human visual perception to evaluate video quality accurately. By analyzing spatial and temporal characteristics of video frames, VQM provides a comprehensive measure of video fidelity, making it invaluable in video compression and transmission applications. VMAF takes a multifaceted approach by combining traditional metrics like PSNR and SSIM with machine learning techniques. VMAF predicts how viewers perceive video quality by leveraging a human-rated dataset, making it highly effective for video streaming, content delivery, and codec research. Meanwhile, PEVQ, standardized by ITU (International Telecommunication Union), offers both objective and subjective evaluations. Its computational model estimates video quality based on visual and temporal features, while human evaluators provide MOS-based subjective assessments. Widely used in video telephony and conferencing systems, PEVQ ensures that video communication meets acceptable quality standards.

Impact on AI: Evaluating video quality can significantly impact AI training, particularly for models focused on computer vision, video processing, and generation tasks. By assessing video quality, researchers can curate better training datasets, develop more effective data augmentation techniques, create more accurate labels, and improve the assessment of AI-generated content. This process enables the filtering out of low-quality or corrupted videos that could negatively impact model performance while also allowing for the development of better video generation and enhancement algorithms. Furthermore, incorporating human perception metrics into video quality evaluation helps train AI models to produce results that are visually appealing to human viewers.

Summary: *Video quality refers to overall visual fidelity and perceptual coherence, assessing its accuracy in representing the original content. Objective metrics such as PSNR, SSIM, VIF, and VSNR, commonly used for image quality assessment, can be effectively applied to measure video quality by analyzing spatial and temporal characteristics. Additionally, specialized metrics like VQM, VMAF, and PEVQ are specifically designed to address the challenges unique to video data, incorporating human perceptual aspects and machine learning techniques to predict viewer perception.*

3.2.2.4 Other Metrics for Unstructured Data. Many metrics mentioned in the structured data section (§3.1) are also applicable to measuring readiness of unstructured data. For instance, preparing unstructured data for AI applications involves adapting key dimensions typically applied to structured data (3.1). “Correctness” is fundamental, ensuring the accuracy and integrity of content within unstructured data like text, images, audio, and video to maintain AI model

reliability. “Feature Relevancy” is crucial for identifying informative elements within unstructured data, aiding pattern recognition and decision-making. “Privacy Leakage” safeguards sensitive information, requiring privacy-preserving techniques. Addressing “Class Imbalance” and “Class Separability” enhances classification tasks in unstructured data by ensuring balanced representation and category distinguishability. Lastly, “Timeliness” is vital, particularly in dynamic data domains, as it ensures AI models stay relevant and up-to-date with evolving data patterns. Compliance of unstructured data with FAIR principles is a major requirement to make certain the data is ready for AI applications. All these metrics collectively contribute to unstructured data readiness for AI usage.

DRAI Metrics for Unstructured Data

In examining unstructured DRAI, this study addresses both textual and multimedia domains. Textual data metrics, including lexical diversity, term importance, readability score, topic coherence, and bias indicators, are discussed with a focus on metrics highlighted in the literature. Simultaneously, multimedia DRAI metrics, such as image, speech, and video quality, are explored in the same context, featuring metrics highlighted in relevant literature.

4 EXISTING FRAMEWORKS

Although not specifically targeting data readiness for AI, numerous frameworks have been developed to evaluate various aspects of data quality. Targeting comprehensive AI readiness evaluation, we have recently developed AI Data Readiness Inspector (AIDRIN) [60]. We describe here various data quality evaluation frameworks along with AIDRIN.

General data quality toolkits include frameworks like Informatica’s data quality tool [65], an open-source solution for data profiling, cleansing, and monitoring that assesses metrics such as data completeness, accuracy, and reliability. DQLearn [127] is another tool in this category, focusing on systematically addressing data quality issues through detection, correction, and custom rule implementation. Additionally, Deequ [123] stands out as a library that allows unit tests for data, facilitating early data quality checks within data pipelines. Moreover, the PIQ [74] [75] library offers an extensive list of metrics, including both traditional and modern techniques, for evaluating image quality. This library is a valuable resource for researchers and developers working on AI applications, allowing them to select the most appropriate metric based on their specific requirements. PIQ includes metrics categorized as full-reference, no-reference, and distribution-based metrics.

A small number of toolkits target AI data analysis. Among them, the Data Nutrition Project (DNP) Label [61] provides a standardized format for presenting essential dataset information, including metadata, variable descriptions, and correlations, aiding in the preparation of quality training data for AI. The Data Readiness Report [10] focuses on generating detailed documentation to assist in data preprocessing for machine learning, offering insights into data quality and readiness across standardized dimensions. IBM’s Data Quality Toolkit (DQT)[44] emphasizes automated explanations of data quality across various dimensions, including completeness, feature relevance, label purity, and data fairness, simplifying data preparation and enhancing productivity for AI practitioners.

Domain-specific toolkits are frameworks designed for specific domains or dimensions of DRAI, such as fairness and FAIR compliance. IBM’s AI fairness 360 toolkit [46] is an open-source software aimed at detecting and mitigating biases in machine learning models, providing metrics like representation and statistical rates of sensitive attributes. For FAIR compliance, tools like FAIR Cookbook [116] and FAIRassist [49] guide users in implementing and measuring FAIR principles, while ESS-DIVE [47] evaluates FAIR compliance in earth sciences data repositories.

Toward a comprehensive evaluation of data readiness for AI, we have recently developed AIDRIN (AI Data Readiness Inspector) [60]. AIDRIN encompasses both traditional data quality assessments and metrics specifically designed for AI

readiness, such as data bias, privacy, feature relevance, correlation, and FAIR compliance. The framework allows users to select their assessment criteria and generate intuitive visualizations and reports to enhance the interpretation and usability of results in AI-related tasks.

Despite the increasing interest in evaluating data readiness, there is still a lack of comprehensive tools covering a broad range of metrics to evaluate the readiness of structured and unstructured data for a given AI task. This is a challenging endeavor, and this survey paper will serve as a reference for understanding the available metrics and developing strategies for incorporating them into tools to comprehensively assess data readiness.

5 PILLARS OF DATA READINESS FOR AI

Based on the knowledge gathered in this survey, we propose a high level taxonomy of metrics. We categorize the AI-ready data assessment metrics into six pillars. These are data quality, understandability and usability, data structure and organization, impact of data on AI, fairness and bias, and data governance (as shown in Fig. 3). These pillars contain a comprehensive set of aspects in data preparation, ensuring that data is prepared and readied to support AI systems effectively, ethically, and efficiently. Each pillar is supported by specific metrics that provide a structured framework for evaluating data readiness in AI contexts.

Of these categories of metrics, the first four, i.e., data quality, understandability and usability, structure and organization, and data governance, are agnostic to the AI method. They can be applied generically to a broad set of datasets and be used for a wide range of AI applications. In contrast, the last two categories, i.e., the impact of data and fairness, are more specific to the use case and the AI methodology, making them critical for specialized contexts but requiring tailored approaches.

Data quality ensures that the data used to train AI models is accurate, complete, and consistent. High-quality data minimizes the risk of errors in AI outputs, leading to more reliable and trustworthy models. When data quality is compromised, it can result in inaccurate and unstable models. Thus, maintaining rigorous data quality standards is essential for achieving effective and credible AI outcomes. Structured data quality can be evaluated using metrics described in Section 3, such as completeness, correctness, timeliness, mislabeling, multimedia quality.

Data understanding and usability are important for enabling AI systems to interpret and utilize data effectively. This category of metrics emphasize the importance of clear documentation, comprehensive metadata, reusability, and accessible data interfaces. When data is well-understood and easy to use, it accelerates AI model development. FAIR principle compliance metrics[3.1.15] can serve in evaluating data understanding and usability.

Data structure and data organization are important to integrate data into AI workflows seamlessly and efficiently. Adequate number of samples in data and proper data partitioning into training, testing, and validation sets allow for accurate model evaluation. In addition, the data model, i.e., the schema of the data, and data organization, i.e., how the data is stored, also play a role in the speed of AI training. also plays a role in the performance of AI applications.

Data Governance is essential for managing data in a way that is ethical, secure, and compliant with legal standards. This pillar covers key aspects such as data privacy, security, and the ethical use of data, which are necessary for building trust in AI systems. Without proper governance, AI systems risk violating privacy regulations, facing security breaches,

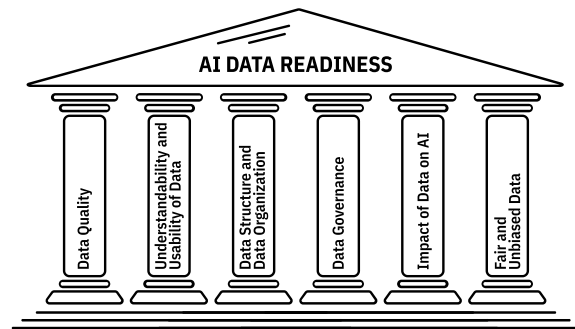


Fig. 3. A high-level view of data readiness metric categories for AI.

and engaging in unethical practices, which can harm public trust and lead to significant legal and reputation-related consequences. Metrics such as privacy leakage[3.1.13] are essential for understanding the extent of potential privacy risks within the data.

The **impact of data on AI** covers the importance of data content and its relevance to AI applications. Rich and high-impact data that provides meaningful insights is critical for deriving effective AI outcomes by enabling models to make accurate predictions and identify deep patterns. Feature relevance[3.1.5] and data point impact[3.1.10] serve as quantitative measures to assess the value of data for a given AI application.

Fair and unbiased data is a fundamental aspect of ensuring that AI systems produce equitable and unbiased outcomes. This pillar focuses on the representativeness of the data and the absence of biases that could lead to discriminatory practices. Fairness in AI models is not only an ethical concern but also crucial for maintaining public trust in AI technologies. When data used in AI models is biased or unrepresentative, it can result in skewed outcomes that may continue existing inequality and injustice. This undermines the societal benefits that AI promises to deliver. Metrics such as the discrimination (bias) index[3.1.8], class imbalance[3.1.6], and class separability[3.1.7] are crucial for assessing the fairness of data before its use in AI applications.

The six dimensions outlined above can be quantified using specific metrics discussed in this study to evaluate DRAI. While these metrics provide a foundational assessment, additional metrics are required to fully capture the breadth of each dimension. Aggregating the evaluations across the existing and future metrics will result in a comprehensive DRAI assessment that would lead to highly accurate and impactful AI solutions.

6 GAPS AND CHALLENGES

We discuss the challenges in defining the metrics for assessing data readiness for AI in structured and unstructured data. While structured data poses unique challenges regarding standardization, interpretability, and sensitivity, unstructured datasets present additional complexities due to their diverse formats, varying modalities, and contextual nuances.

Assessing data readiness for AI and data science applications, regardless of its structure, presents several challenges stemming from the absence of a unified framework that complicates the comparison and consistent application of metrics across diverse dimensions. This absence hampers the development of a cohesive and comprehensive data readiness assessment method explicitly tailored for different types of data. Although IBM DQT [44] and AIDRIN [60] have taken the initial steps in addressing this concern, their coverage is primarily focused on structured data, leaving a need for further advancements to include a broader range of dimensions related to structured data readiness and extend the toolkit's capabilities to address unstructured data challenges. Additionally, with the rapid growth of large datasets, there is a lack of parallel systems designed to evaluate data readiness effectively at scale. This highlights the need for further advancements to develop robust methods for handling diverse and extensive data environments.

A significant challenge in DRAI assessment is the evolving nature of data quality dimensions. As new data quality metrics emerge, they add complexity to the evaluation process. According to Batini and Scannapieco's [17], the development of metrics specific to various domains, such as the condition of archival documents, the integrity of statistical data, and the positional accuracy in geospatial data, further complicates comprehensive assessments of data quality and readiness. This ongoing evolution necessitates continual adaptation and refinement of evaluation frameworks to effectively address established and new metrics.

In the rapidly evolving fields of AI and data science, the emergence of new use cases and diverse data structures constantly challenges the evaluation of data readiness. To ensure these metrics remain effective in assessing data preparedness for the latest AI applications, they must adapt and evolve alongside the technology. Striking the right

balance between data readiness and quantity is another crucial challenge. While numerous metrics focus on data readiness, a comprehensive approach should consider the trade-off between data readiness and sufficiency. Sufficient data volumes are essential for meaningful analysis and effective AI model training, making finding the optimal equilibrium between data readiness and quantity a challenge. Additionally, clear interpretability of these readiness metrics is essential for stakeholders to grasp their implications on overall data readiness and their potential impacts on AI model performance. Enhanced visualization and explanation techniques can significantly improve the practical utility of these metrics, facilitating more informed decision-making processes.

Addressing the challenges in the dynamic field of AI and data readiness is important. Specific AI applications often require customized data readiness metrics due to varied data readiness standards, requiring domain-specific expertise for effective navigation. Simultaneously, addressing subjective judgments and human biases, particularly concerning fairness and privacy, adds another layer of complexity. Developing unbiased and ethical metrics that adapt to various data types and applications requires ongoing research and innovation. Furthermore, the ever-evolving nature of real-world datasets requires continuous data readiness assessment throughout an AI system's life cycle, extending beyond the initial data preparation phase.

Data access and ownership concerns can impede data readiness evaluation, particularly when datasets are restricted due to privacy and ownership issues. These limitations can delay a comprehensive data readiness assessment, requiring collaborative efforts and agreements between data providers and users to navigate these challenges effectively. Furthermore, the ongoing challenge lies in establishing meaningful thresholds that define acceptable data readiness levels. Given the context-dependent nature of data and the diversity of AI use cases, universal and context-independent threshold values are challenging to determine. Clear guidelines regarding data readiness thresholds are essential to ensure consistent and effective data preparation practices. Lastly, developing well-established benchmark datasets and evaluation protocols becomes crucial to compare and evaluate various data readiness metrics. Creating representative benchmarks spanning different industries and data structures can facilitate a more comprehensive comparison of diverse metrics and scoring mechanisms.

Data readiness assessment is critical for all AI applications, including large language models (LLMs)[11]. Ensuring data completeness, accuracy, and consistency, these metrics will build a robust foundation for LLMs. Correct, unbiased, and relevant data enhance the model's ability to generate coherent and reliable outputs. Assessing bias in datasets ensures fairness and mitigates skewed predictions. Therefore, applying comprehensive quality metrics is vital to preparing data that effectively supports the latest requirements of LLMs, leading to more accurate and contextually relevant model performance. These requirements highlight the need for more advanced metrics and tools to effectively prepare data to ensure LLMs are trained on high-quality inputs that meet the requirements of modern AI applications.

While having a comprehensive list of metrics in the toolbox is important, deciding the metrics to be used for a particular AI application is essential. When determining the suitable metrics for an AI application, it is crucial to start by defining the objective of the application and being aware of any constraints including data limits or legal considerations. Once the application is defined, one should explore the available metrics and understand each metric and its suitability to achieve the goals. That should consider the trade-offs involved and analyze which metrics best align with your objectives. Focusing on metrics that impact the AI application performance and efficiency achieves a balance between simplicity and depth. Ensuring that the AI model's performance is regularly monitored and that the metrics are flexibly adjusted as it evolves also plays a role in using the correct data. Additionally, reviewing relevant literature can provide valuable insights in selecting the most appropriate metrics. For example, Wagner and Eckhoff [139] proposed a method

for selecting among over 80 privacy metrics. The selected metrics assess factors such as data sensitivity, trade-offs, and performance expectations, providing meaningful insights and driving the application toward its intended outcomes.

Addressing these gaps and challenges in data readiness metrics urges collaboration among researchers, practitioners, and industry experts. Advancing the state-of-the-art in these metrics will contribute to more reliable utilization of data in AI applications, unlocking the maximum potential of this valuable resource.

7 CONCLUSION

This comprehensive survey underscores the critical role of data preparation toward the effective usage of data by AI applications. We explored the challenges, tools, and metrics associated with data readiness, emphasizing its significance in achieving accurate and dependable AI-driven outcomes. Our study highlights the need for a holistic approach, covering structured and unstructured data, and underscores the importance of incorporating fairness-related metrics to ensure unbiased AI decision-making. We have identified and showcased existing metrics and scoring mechanisms that effectively measure data readiness available in literature by studying more than 140 publications from reputable journals including ACM, IEEE, and other reputable journals, as well as web articles, spanning the past three decades. By thoroughly exploring these dimensions and metrics, we have summarized numerous data readiness metrics and proposed a taxonomy of them. This will develop a deeper understanding of data readiness for AI applications. As AI advances and data becomes an even more critical asset, staying up to date with the latest research and advancements in data readiness metrics is essential. This survey provides a foundational resource for researchers and practitioners, equipping them with the essential insights needed to navigate the complexities of data preparation for AI effectively and emphasizing that data readiness is not just a preliminary step but an ongoing commitment.

REFERENCES

- [1] 2008. PEVQ – the Standard for Perceptual Evaluation of Video Quality. <http://www.pevq.com/pevq.html> Accessed 22 July 2023.
- [2] 2024. FAIR Principles. <https://www.go-fair.org/fair-principles/>.
- [3] n.d. <https://docs.aindo.com/evaluation/privacy/>
- [4] n.d. BS.1387 : Method for Objective Measurements of Perceived Audio Quality. <https://www.itu.int/rec/R-REC-BS.1387/en>. Accessed 17 July 2023.
- [5] n.d. The Gunning Fog Index. <https://readable.com/readability/gunning-fog-index/>. Accessed 12 July 2023.
- [6] n.d. JNDmetrix Technology. http://www.sarnoff.com/products_services/video_vision/jndmetrix/. Accessed 12 July 2023.
- [7] n.d. Kaggle. <https://www.kaggle.com/> Accessed: Sept 2023.
- [8] n.d. PSNR. <https://www.mathworks.com/help/vision/ref/psnr.html> Data Accessed 7/26/2023.
- [9] Georgios Afendras and Marianthi Markatou. 2019. Optimality of training/test size and resampling effectiveness in cross-validation. *Journal of Statistical Planning and Inference* 199 (2019), 286–301. <https://doi.org/10.1016/j.jspi.2018.07.005>
- [10] S. Afzal, C. Rajmohan, M. Kesarwani, S. Mehta, and H. Patel. 2020. Data Readiness Report. In *IEEE Int. Conference on Smart Data Services (SMDS)*.
- [11] Telm AI. 2023. Demystifying Data Quality’s Impact on Large Language Models. <https://www.telm.ai/blog/demystifying-data-qualitys-impact-on-large-language-models/>. Accessed: [Your Access Date].
- [12] Francisco Alberto, Salvador García, Mikel Galar, Ronaldo Prati, Bartosz Krawczyk, and Francisco Herrera. 2018. *Learning from Imbalanced Data Sets*. Springer.
- [13] Ahmad Alwosheel, Sander van Cranenburgh, and Caspar G. Chorus. 2018. Is your dataset big enough? Sample size requirements when using artificial neural networks for discrete choice analysis. *Journal of Choice Modelling* 28 (2018), 167–182. <https://doi.org/10.1016/j.jocm.2018.07.002>
- [14] Sevgi Arca and Rattikorn Hewett. 2020. Is Entropy enough for measuring Privacy?. In *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*. 1335–1340. <https://doi.org/10.1109/CSCI51800.2020.00249>
- [15] Fabio Azzalini, Chiara Criscuolo, and Letizia Tanca. 2022. E-FAIR-DB: Functional Dependencies to Discover Data Bias and Enhance Data Equity. *J. Data and Information Quality* 14, 4, Article 29 (nov 2022), 26 pages. <https://doi.org/10.1145/3552433>
- [16] Olivier Bachem, Mario Lucic, and Andreas Krause. 2017. Practical Coreset Constructions for Machine Learning. In *Advances in Neural Information Processing Systems*.
- [17] Carlo Batini and Monica Scannapieco. 2006. *Data Quality: Concepts, Methodologies and Techniques* (1 ed.). Springer-Verlag Berlin Heidelberg, Berlin, Heidelberg. XIX, 262 pages. <https://doi.org/10.1007/3-540-33173-5>
- [18] J. Beerends and J. Stermerdink. 1994. A Perceptual Speech Quality Measure Based on a Psychoacoustic Sound Representation. *Journal of Audio Eng. Soc.* 42 (December 1994), 115–123.

- [19] Michele Bezzi. 2007. An entropy based method for measuring anonymity. In *2007 Third International Conference on Security and Privacy in Communications Networks and the Workshops - SecureComm 2007*. 28–32. <https://doi.org/10.1109/SECCOM.2007.4550303>
- [20] B. Blaiszik et al. 2016. The Materials Data Facility: Data Services to Advance Materials Science Research. *JOM* 68 (2016). <https://doi.org/10.1007/s11837-016-2001-3>
- [21] 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, Article 8 (feb 2011), 28 pages. <https://doi.org/10.1145/1891879.1891881>
- [22] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3 (January 2003), 993–1022. Submitted 2/02; Published 1/03.
- [23] Netflix Technology Blog. 2017. Toward a practical perceptual video quality metric. <https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652>
- [24] 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, Article 3 (may 2018), 26 pages. <https://doi.org/10.1145/3190578>
- [25] Z. Borsos, C. Lemnaru, and R. Potolea. 2018. Dealing with overlap and imbalance: a new metric and approach. *Pattern Anal Applic* 21, 2 (2018), 381–395. <https://doi.org/10.1007/s10044-016-0583-6>
- [26] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. 2000. LOF: Identifying density-based local outliers. In *Proc. ACM SIGMOD Int. Conf. Manage. Data*. 93–104.
- [27] Nicholas Carlini, Matthew Jagielski, Chiyuan Zhang, Nicolas Papernot, Andreas Terzis, and Florian Tramèr. 2022. The Privacy Onion Effect: Memorization is Relative. arXiv:2206.10469 [cs.LG]
- [28] L. Elisa Celis, Vijay Keswani, and Nisheeth K. Vishnoi. 2020. Data preprocessing to mitigate bias: A maximum entropy based approach. arXiv:1906.02164 [cs.LG]
- [29] Damon M. Chandler and Sheila S. Hemami. 2007. VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images. *IEEE Transactions on Image Processing* 16, 9 (2007), 2284–2298. <https://doi.org/10.1109/TIP.2007.901820>
- [30] Daniel JB Clarke et al. 2019. FAIRshake: toolkit to evaluate the FAIRness of research digital resources. *Cell systems* 9, 5 (2019), 417–421.
- [31] Cleanlab. 2024. Elevating Data Quality: The Crucial Role of Proper Data Annotation. <https://cleanlab.ai/blog/learn/data-annotation/>. Accessed: 2024-08-06.
- [32] J. Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46. <https://doi.org/10.1177/001316446002000104>
- [33] Meri Coleman and Ta Lin Liao. 1975. A computer readability formula designed for machine scoring. *J. of Applied Psychology* 60 (1975), 283–284.
- [34] R. Dennis Cook and Sanford Weisberg. 1982. *Residuals and Influence in Regression*. Chapman & Hall.
- [35] Bing Tian Dai, Nick Koudas, Beng Chin Ooi, Divesh Srivastava, and Suresh Venkatasubramanian. 2006. Rapid Identification of Column Heterogeneity. In *Sixth International Conference on Data Mining (ICDM'06)*. 159–170. <https://doi.org/10.1109/ICDM.2006.132>
- [36] Ingrid Daubechies. 1992. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics.
- [37] John C. Davis and Robert J. Sampson. 1986. *Statistics and Data Analysis in Geology*. Vol. 646. Wiley, New York.
- [38] Ali Degirmenci and Omer Karal. 2021. Robust Incremental Outlier Detection Approach Based on a New Metric in Data Streams. *IEEE Access* 9 (2021), 160347–160360. <https://doi.org/10.1109/ACCESS.2021.3131402>
- [39] IBM Developer. 2021. IBM Data Quality AI Toolkit. <https://developer.ibm.com/learningpaths/data-quality-ai-toolkit/overview/> Date accessed: June 12, 2023.
- [40] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. <http://archive.ics.uci.edu/ml>.
- [41] Richard O. Duda, Peter E. Hart, and David G. Stork. 2012. *Pattern Classification*. John Wiley & Sons.
- [42] Vasisht Duddu, Sebastian Szyller, and N. Asokan. 2022. SHAPr: An Efficient and Versatile Membership Privacy Risk Metric for Machine Learning. arXiv:2112.02230 [cs.CR]
- [43] Ahmed K. Elmagarmid, Panagiotis G. Ipeirotis, and Vassilios S. Verykios. 2007. Duplicate Record Detection: A Survey. *IEEE Transactions on Knowledge and Data Engineering* 19, 1 (2007), 1–16. <https://doi.org/10.1109/TKDE.2007.250581>
- [44] Nitin Gupta et al. 2021. Data Quality Toolkit: Automatic assessment of data quality and remediation for machine learning datasets. arXiv:2108.05935 [cs.LG]
- [45] Nikil Ravi et al. 2022. FAIR principles for AI models with a practical application for accelerated high energy diffraction microscopy. *Scientific Data* 9, 1 (nov 2022). <https://doi.org/10.1038/s41597-022-01712-9>
- [46] Rachel K. E. Bellamy et al. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. arXiv:1810.01943 [cs.AI]
- [47] S. Cholia et al. 2024. ESS-DIVE Overview: A Scalable, User-Focused Repository for Earth and Environmental Science Data. Scientific Data Division, Lawrence Berkeley National Laboratory, Berkeley, CA. <https://ess-dive.lbl.gov/>
- [48] European Parliament and Council of the European Union. 2016. *Regulation (EU) 2016/679 of the European Parliament and of the Council*. <https://data.europa.eu/eli/reg/2016/679/oj>
- [49] FAIRassist.org. n.d.. FAIRassist.Org. <https://fairassist.org>. Jan. 6, 2024.
- [50] Michael Feldman et al. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Sydney, NSW, Australia) (KDD '15)*. Association for Computing Machinery, New York, NY, USA, 259–268.

- [51] Rudolf Flesch. 1986. *The Art of Readable Writing* (19th print.-collier books ed ed.). MacMillan.
- [52] George Forman. 2003. An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *J. Mach. Learn. Res.* 3 (mar 2003), 17 pages.
- [53] Amirata Ghorbani and James Zou. 2019. Data Shapley: Equitable Valuation of Data for Machine Learning. arXiv:1904.02868 [stat.ML]
- [54] C. Gini. 1912. Variability and Mutability: Contribution to the Study of Statistical Distribution and Relations. *Studi Economico-Giuridici della R* (1912).
- [55] Mark A. Hall and Lloyd A. Smith. 1999. Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper. In *FLAIRS*. 235–239.
- [56] Simon S. Haykin. 2009. *Neural networks and learning machines* (third ed.). Pearson Education, Upper Saddle River, NJ.
- [57] Xiaofei He, Deng Cai, and Partha Niyogi. 2005. Laplacian Score for Feature Selection. In *NIPS*. 507–514.
- [58] Bernd Heinrich and Mathias Klier. 2015. Metric-based data quality assessment — Developing and evaluating a probability-based currency metric. *Decision Support Systems* 72 (2015), 82–96. <https://doi.org/10.1016/j.dss.2015.02.009>
- [59] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2018. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. arXiv:1706.08500 [cs.LG] <https://arxiv.org/abs/1706.08500>
- [60] Kaveen Hiniduma et al. 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 (Rennes, France) (SSDBM '24)*. Article 7, 12 pages.
- [61] 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]
- [62] Q. Huynh-Thu and M. Ghanbari. 2008. Scope of validity of PSNR in image/video quality assessment. *Electronics Letters* 44, 13 (Jun 19 2008), 1–2.
- [63] Helen Hwang. 2022. New AI readiness report reveals insights into ML lifecycle. <https://www.datacenterknowledge.com/machine-learning/new-ai-readiness-report-reveals-insights-ml-lifecycle>. Accessed on May 15, 2023.
- [64] Espire Infolabs. 2024. Outlier Detection Redefined: A Deep Dive into AI's Impact: Espire Blog. <https://www.espire.com/blog/posts/outlier-detection-redefined-a-deep-dive-into-ai-impact> Accessed: 2024-08-06.
- [65] Informatica. n.d.. *Data Quality Metrics & Measures - All You Need to Know*. Accessed Aug. 26, 2024.
- [66] International Telecommunication Union. 2018. *ITU-T Recommendation P.808: Subjective Evaluation of Speech Quality with a Crowdsourcing Approach*. Technical Report. International Telecommunication Union, Geneva.
- [67] F. Itakura and S. Saito. 1968. Analysis Synthesis Telephony Based on the Maximum Likelihood Method. In *Proc. 6th Int. Congr. Acoust.* Tokyo, Japan.
- [68] M.A. Jaro. 1976. *Unimatch: A Record Linkage System: User's Manual*. Technical Report. US Bureau of the Census, Washington, D.C.
- [69] N.C. Jayant and P. Noll. 1984. *Digital Coding of Waveforms: Principles and Applications to Speech and Video*. Prentice Hall, NJ, USA.
- [70] Matthew B Jones and Peter Slaughter. 2019. https://www.dataone.org/uploads/dataonewebinar_jonesslaughter_fairmetadata_190514.pdf
- [71] V. Roshan Joseph. 2022. Optimal Ratio for Data Splitting. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 15, 4 (August 2022), 531–538. <https://doi.org/10.1002/sam.11583>
- [72] Sven Jäger, Anders Allhorn, and Felix Bießmann. 2021. A Benchmark for Data Imputation Methods. *Frontiers in Big Data* 4 (2021), 693674. <https://doi.org/10.3389/fdata.2021.693674>
- [73] M. Kaiser, Mathias Klier, and Bernd Heinrich. 1970. [PDF] how to measure data quality? - A metric-based approach. <https://www.semanticscholar.org/paper/How-to-Measure-Data-Quality-A-Metric-Based-Approach-Kaiser-Klier/afcd53c5a88f3320c861ad3f09f28237b6744cb>
- [74] Sergey Kastrulyin, Dzhamil Zakirov, and Denis Prokopenko. 2019. PyTorch Image Quality: Metrics and Measure for Image Quality Assessment. <https://github.com/photosynthesis-team/piq> Open-source software available at <https://github.com/photosynthesis-team/piq>.
- [75] Sergey Kastrulyin, Jamil Zakirov, Denis Prokopenko, and Dmitry V. Dyllov. 2022. PyTorch Image Quality: Metrics for Image Quality Assessment. <https://doi.org/10.48550/ARXIV.2208.14818>
- [76] Martin Kemka. 2019. Learning Amazon Sagemaker. <https://docs.aws.amazon.com/sagemaker/latest/dg/clarify-data-bias-metric-cddl.html>
- [77] Pang Wei Koh and Percy Liang. 2017. Understanding Black-box Predictions via Influence Functions. In *International Conference on Machine Learning*.
- [78] Paras Lakhani. 2020. The Importance of Image Resolution in Building Deep Learning Models for Medical Imaging. *Radiology: Artificial Intelligence* 2, 1 (2020), e190177. <https://doi.org/10.1148/ryai.2019190177>
- [79] Liliya Lavitas, Olivia Redfield, Allen Lee, Daniel Fletcher, Matthias Eck, and Sunil Janardhanan. 2021. Annotation quality framework-accuracy, credibility, and consistency. In *NEURIPS 2021 Workshop for Data Centric AI*.
- [80] V.I. Levenshtein. 1965. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Doklady Akademii Nauk SSSR* 163, 4 (1965), 845–848. Original in Russian—translation in *Soviet Physics Doklady*, vol. 10, no. 8, pp. 707–710, 1966.
- [81] David D. Lewis. 1992. Feature Selection and Feature Extraction for Text Categorization. In *Workshop on Speech and Natural Language*. 212–217.
- [82] Christophe Leys, Christophe Ley, Olivier Klein, Pierre Bernard, and Laurent Licata. 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *J. Exp. Social Psychol.* 49, 4 (2013), 764–766.
- [83] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P. Trevino, Jiliang Tang, and Huan Liu. 2017. Feature Selection: A Data Perspective. *ACM Comput. Surv.* 50, 6, Article 94 (dec 2017), 45 pages. <https://doi.org/10.1145/3136625>
- [84] Peng Li, Xi Rao, Jennifer Blase, Yue Zhang, Xu Chu, and Ce Zhang. 2021. CleanML: A Study for Evaluating the Impact of Data Cleaning on ML Classification Tasks. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*. 13–24. <https://doi.org/10.1109/ICDE51399.2021.00009>

- [85] Weisi Lin, Li Dong, and Ping Xue. 2005. Visual distortion gauge based on discrimination of noticeable contrast changes. *IEEE transactions on circuits and systems for video technology* 15, 7 (2005), 900–909.
- [86] Weisi Lin and C.-C. Jay Kuo. 2011. Perceptual visual quality metrics: A survey. *Journal of Visual Communication and Image Representation* 22, 4 (2011), 297–312. <https://doi.org/10.1016/j.jvcir.2011.01.005>
- [87] Huan Liu and Rudy Setiono. 1995. Chi2: Feature Selection and Discretization of Numeric Attributes. In *ICTAI*. 388–391.
- [88] Sijia Liu et al. 2020. An ADMM-based Framework for AutoML Pipeline Configuration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 4892–4899.
- [89] Luc Longpré, Vladik Kreinovich, and Thongchai Dumrongpokaphan. 2017. Entropy as a Measure of Average Loss of Privacy. *Thai Journal of Mathematics* (2017), 7–15. <https://api.semanticscholar.org/CorpusID:6672504>
- [90] Yang Lu, Yiu ming Cheung, and Yuan Yan Tang. 2019. Bayes Imbalance Impact Index: A Measure of Class Imbalanced Dataset for Classification Problem. arXiv:1901.10173 [cs.LG]
- [91] H. P. Luhn. 1957. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development* 1, 4 (1957), 309–317. <https://doi.org/10.1147/rd.14.0309>
- [92] Chris Markham. 2024. How AI Can Uncover Data Outliers and Patterns in Patient Behavior. <https://www.saama.com/how-ai-can-uncover-data-outliers-and-patterns-in-patient-behavior/>. <https://www.saama.com/how-ai-can-uncover-data-outliers-and-patterns-in-patient-behavior/>
- [93] Pina Marziliano, Frederic Dufaux, Stefan Winkler, and Touradj Ebrahimi. 2002. A no-reference perceptual blur metric. In *Proceedings. International conference on image processing*, Vol. 3. IEEE, III–III.
- [94] Philip M McCarthy. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.D. Dissertation. The University of Memphis.
- [95] Peter M. McCarthy and Scott Jarvis. 2010. MTLD, VOCB-D, and HD-D: A Validation Study of Sophisticated Approaches to Lexical Diversity Assessment. *Behavior Research Methods* 42, 2 (2010), 381–392. <https://doi.org/10.3758/BRM.42.2.381>
- [96] David Mimno et al. 2011. Optimizing Semantic Coherence in Topic Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Edinburgh, United Kingdom) (*EMNLP '11*). Association for Computational Linguistics, USA, 262–272.
- [97] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. 2012. No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Transactions on Image Processing* 21, 12 (2012), 4695–4708. <https://doi.org/10.1109/TIP.2012.2214050>
- [98] A.E. Monge and C.P. Elkan. 1996. The Field Matching Problem: Algorithms and Applications. In *Proc. Second Int'l Conf. Knowledge Discovery and Data Mining (KDD '96)*. 267–270.
- [99] David Newman et al. 2010. Evaluating Topic Models for Digital Libraries. In *Proceedings of the 10th Annual Joint Conference on Digital Libraries* (Gold Coast, Queensland, Australia) (*JCDL '10*). 215–224. <https://doi.org/10.1145/1816123.1816156>
- [100] Feiping Nie, Shiming Xiang, Yangqing Jia, Changshui Zhang, and Shuicheng Yan. 2008. Trace Ratio Criterion for Feature Selection. In *AAAI*.
- [101] E. Ntoutsi et al. 2020. Bias in data-driven artificial intelligence systems—An introductory survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 10, 3 (2020), e1356. <https://doi.org/10.1002/widm.1356>
- [102] Sejong Oh. 2011. A new dataset evaluation method based on category overlap. *Computers in Biology and Medicine* 41, 2 (2011), 115–122. <https://doi.org/10.1016/j.combiomed.2010.12.006>
- [103] Jonathan Ortigosa-Hernández, Iñaki Inza, and Jose 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>
- [104] Orestis Papakyriakopoulos et al. 2020. Bias in Word Embeddings. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) (*FAT* '20*). 446–457. <https://doi.org/10.1145/3351095.3372843>
- [105] Ronald K. Pearson. 2006. The problem of disguised missing data. *SIGKDD Explor. Newsl.* 8, 1 (jun 2006). <https://doi.org/10.1145/1147234.1147247>
- [106] Leo L. Pipino, Yang W. Lee, and Richard Y. Wang. 2002. Data Quality Assessment. *Commun. ACM* 45, 4 (apr 2002), 211–218. <https://doi.org/10.1145/505248.506010>
- [107] Dragoljub Pokrajac, Aleksandar Lazarevic, and Longin Jan Latecki. 2007. Incremental local outlier detection for data streams. In *Proc. IEEE Symp. Comput. Intell. Data Mining*. 504–515.
- [108] Maria Priestley, Fionntán O'Donnell, and Elena Simperl. 2023. A Survey of Data Quality Requirements That Matter in ML Development Pipelines. *J. Data and Information Quality* (apr 2023). <https://doi.org/10.1145/3592616> Just Accepted.
- [109] Shahzad Qaiser and Ramsha Ali. 2018. Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. *International Journal of Computer Applications* 181 (07 2018). <https://doi.org/10.5120/ijca2018917395>
- [110] Deepak Rajput, Wanjun Wang, and Cheng-Chung Chen. 2023. Evaluation of a decided sample size in machine learning applications. *BMC Bioinformatics* 24 (2023), 48. <https://doi.org/10.1186/s12859-023-05156-9>
- [111] Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, Vol. 242. Citeseer, 29–48.
- [112] Pedro Reviriego, Javier Conde, Elena Merino-Gómez, Gonzalo Martínez, and José Alberto Hernández. 2023. Playing with Words: Comparing the Vocabulary and Lexical Richness of ChatGPT and Humans. arXiv:2308.07462 [cs.CL] <https://arxiv.org/abs/2308.07462>
- [113] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. “Why Should I Trust You?” Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

- [114] A.W. Rix et al. 2001. Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*.
- [115] Marko Robnik-Šikonja and Igor Kononenko. 2003. Theoretical and Empirical Analysis of ReliefF and RReliefF. *Machine Learning* 53, 1-2 (2003).
- [116] 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).
- [117] Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (Shanghai, China) (WSDM '15)*. 399–408. <https://doi.org/10.1145/2684822.2685324>
- [118] Bernard Rosner. 1983. Percentage points for a generalized ESD many-outlier procedure. *Technometrics* 25, 2 (1983), 165–172.
- [119] Peter J. Rousseeuw and Mia Hubert. 2018. Anomaly detection by robust statistics. *WIREs Data Mining Knowl. Discovery* 8, 2 (Mar. 2018), e1236.
- [120] R.C. Russell. 1922. Index. <http://patft.uspto.gov/netahtml/srchnum.htm>
- [121] Carl F. Sabottke and Bradley M. Spieler. 2020. The Effect of Image Resolution on Deep Learning in Radiography. *Radiology: Artificial Intelligence* 2, 1 (2020), e190015. <https://doi.org/10.1148/ryai.2019190015>
- [122] Miriam Seane Santos, Pedro Henriques Abreu, Szymon Wilk, and João Santos. 2020. How distance metrics influence missing data imputation with k-nearest neighbours. *Pattern Recognition Letters* 136 (2020), 111–119. <https://doi.org/10.1016/j.patrec.2020.05.032>
- [123] 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.
- [124] Ron Schmelzer. 2019. The Achilles' Heel of AI. <https://www.forbes.com/sites/cognitiveworld/2019/03/07/the-achilles-heel-of-ai/?sh=20e53e4d7be7>
- [125] Nima Shahbazi, Yin Lin, Abolfazl Asudeh, and H. V. Jagadish. 2023. Representation Bias in Data: A Survey on Identification and Resolution Techniques. *ACM Comput. Surv.* (mar 2023). <https://doi.org/10.1145/3588433> Just Accepted.
- [126] H.R. Sheikh and A.C. Bovik. 2006. Image information and visual quality. *IEEE Transactions on Image Processing* 15, 2 (2006), 430–444. <https://doi.org/10.1109/TIP.2005.859378>
- [127] S. Shrivastava, D. Patel, N. Zhou, A. Iyengar, and A. Bhamidipaty. 2020. DQLearn: A Toolkit for Structured Data Quality Learning. In *2020 IEEE International Conference on Big Data (Big Data)*. Atlanta, GA, USA, 1644–1653. <https://doi.org/10.1109/BigData50022.2020.9378296>
- [128] Fatimah Sidi et al. 2012. Data quality: A survey of data quality dimensions. In *2012 International Conference on Information Retrieval & Knowledge Management*. 300–304. <https://doi.org/10.1109/InfRKM.2012.6204995>
- [129] Simha. 2021. Understanding TF-IDF for Machine Learning. <https://www.capitalone.com/tech/machine-learning/understanding-tf-idf/>
- [130] A. Simonetta, A. Trenta, M. C. Paoletti, and A. Vetrò. 2021. Metrics for Identifying Bias in Datasets. *SYSTEM* (2021).
- [131] E. Simpson. 1949. Measurement of Diversity. *Nature* 163, 688 (1949), 688. <https://doi.org/10.1038/163688a0>
- [132] Liwei Song and Prateek Mittal. 2021. Systematic Evaluation of Privacy Risks of Machine Learning Models. In *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, 2615–2632. <https://www.usenix.org/conference/usenixsecurity21/presentation/song>
- [133] Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *J. of Documentation* 28, 1 (1972), 11–21.
- [134] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. 2010. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *2010 IEEE Int. Conf. on Acoustics, Speech and Signal Processing*. 4214–4217. <https://doi.org/10.1109/ICASSP.2010.5495701>
- [135] Maxine Templin. 1957. *Certain Language Skills in Children*. University of Minnesota Press, Minneapolis.
- [136] Kim-Han Thung and Paramesran Raveendran. 2009. A survey of image quality measures. In *2009 International Conference for Technical Postgraduates (TECHPOS)*. 1–4. <https://doi.org/10.1109/TECHPOS.2009.5412098>
- [137] Dinusha Vatsalan et al. 2022. Privacy risk quantification in education data using Markov model. *British Journal of Educational Technology* 53, 4 (2022), 804–821. <https://doi.org/10.1111/bjet.13223> arXiv:<https://bera-journals.onlinelibrary.wiley.com/doi/pdf/10.1111/bjet.13223>
- [138] Tuan L. Vo, Thu Nguyen, Hugo L. Hammer, Michael A. Riegler, and Pal Halvorsen. 2024. Explainability of Machine Learning Models under Missing Data. arXiv:2407.00411 [cs.LG] <https://arxiv.org/abs/2407.00411>
- [139] Isabel Wagner and David Eckhoff. 2018. Technical Privacy Metrics: A Systematic Survey. *ACM Comput. Surv.* 51, 3, Article 57 (jun 2018), 38 pages. <https://doi.org/10.1145/3168389>
- [140] Jiachen T. Wang and Ruoxi Jia. 2023. Data Banzhaf: A Robust Data Valuation Framework for Machine Learning. arXiv:2205.15466 [cs.LG]
- [141] Zhou Wang and A.C. Bovik. 2002. A universal image quality index. *IEEE Signal Processing Letters* 9, 3 (2002), 81–84. <https://doi.org/10.1109/97.995823>
- [142] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612. <https://doi.org/10.1109/TIP.2003.819861>
- [143] Z. Wang, E.P. Simoncelli, and A.C. Bovik. 2003. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, Vol. 2. 1398–1402 Vol.2. <https://doi.org/10.1109/ACSSC.2003.1292216>
- [144] M.S. Waterman, T.F. Smith, and W.A. Beyer. 1976. Some Biological Sequence Metrics. *Advances in Math.* 20, 4 (1976), 367–387.
- [145] Mark D. Wilkinson, Susanna-Assunta Sansone, Erik Schultes, Peter Doorn, Luiz Olavo Bonino da Silva Santos, and Michel Dumontier. 2018. A design framework and exemplar metrics for fairness. <https://www.nature.com/articles/sdata2018118>
- [146] Alex Woodie. 2020. Data Prep Still Dominates Data Scientists' Time, Survey Finds. <https://www.datanami.com/2020/07/06/data-prep-still-dominates-data-scientists-time-survey-finds/>. Accessed on May 15, 2023.
- [147] Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C. Bovik. 2014. Gradient Magnitude Similarity Deviation: A Highly Efficient Perceptual Image Quality Index. *IEEE Transactions on Image Processing* 23, 2 (Feb. 2014), 684–695. <https://doi.org/10.1109/tip.2013.2293423>
- [148] Mehdi Yalaoui and Saida Boukhedouma. 2021. A survey on data quality: principles, taxonomies and comparison of approaches. In *2021 International Conference on Information Systems and Advanced Technologies (ICISAT)*. 1–9. <https://doi.org/10.1109/ICISAT54145.2021.9678209>

- [149] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. 2011. FSIM: A Feature Similarity Index for Image Quality Assessment. *IEEE Transactions on Image Processing* 20, 8 (2011), 2378–2386. <https://doi.org/10.1109/TIP.2011.2109730>
- [150] Zheng Zhao and Huan Liu. 2007. Spectral Feature Selection for Supervised and Unsupervised Learning. In *ICML*. 1151–1157.
- [151] Rui Zhu, Ziyu Wang, Zhanyu Ma, Guijin Wang, and Jing-Hao 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>