Information and Software Technology 162 (2023) 107268

Contents lists available at ScienceDirect

Information and Software Technology

journal homepage: www.elsevier.com/locate/infsof

A survey on dataset quality in machine learning

Youdi Gong [a,b], Guangzhen Liu [a], Yunzhi Xue [a], Rui Li [a], Lingzhong Meng [a,*]

[a] Institute of Software Chinese Academy of Sciences, Beijing, 100190, China
[b] Beihang University, Beijing, 100191, China

ARTICLE INFO

Keywords:
Dataset
Dataset quality
Machine Learning

ABSTRACT

With the rise of big data, the quality of datasets has become a crucial factor affecting the performance of machine learning models. High-quality datasets are essential for the realization of data value. This survey article summarizes the research direction of dataset quality in machine learning, including the definition of related concepts, analysis of quality issues and risks, and a review of dataset quality dimensions and metrics throughout the dataset lifecycle and a review of dataset quality metrics analyzed from a dataset lifecycle perspective and summarized in literatures. Furthermore, this article introduces a comprehensive quality evaluation process, which includes a framework for dataset quality evaluation with dimensions and metrics, computation methods for quality metrics, and assessment models. These studies provide valuable guidance for evaluating dataset quality in the field of machine learning, which can help improve the accuracy, efficiency, and generalization ability of machine learning models, and promote the development and application of artificial intelligence technology.

1. Introduction

1.1. Background

Machine Learning (ML) has seen significant progress in recent years due to advances in technology and the availability of large datasets. However, the success of ML models heavily relies on the quality of the dataset used to train and evaluate the models. The importance of high-quality datasets for training and evaluating models cannot be overstated. A high-quality dataset is one that accurately represents the real-world phenomena, is comprehensive, and is free from biases. The quality of the dataset can have a significant impact on the accuracy and effectiveness of the ML model. Therefore, understanding the different aspects of dataset quality in machine learning is essential to ensure the success of ML projects.

One crucial aspect of data quality is the identification and clarification of quality problems within the dataset, such as missing data, data corruption, or data errors. By locating these issues, data cleaning can be performed to improve the quality of the dataset. For machine learning models to learn useful knowledge, a high-quality dataset is a prerequisite. The saying “garbage in, garbage out” emphasizes the importance of using a high-quality dataset in machine learning. Therefore, research on dataset quality is critical for improving the quality of datasets, leading to improved efficiency in machine learning tasks.

In this paper, we present a survey on dataset quality in machine learning. We explore the different factors that contribute to dataset quality, such as data collection, cleaning, preprocessing, bias and fairness, data augmentation, and evaluation. We examine the importance of each of these factors and how they impact the quality of the dataset. Furthermore, we provide a comprehensive overview of the entire process of dataset quality evaluation, including evaluation dimensions, evaluation metrics, metric computation, and evaluation models.

Our work aims to provide a survey of the various aspects of dataset quality in machine learning, including the evaluation of dataset quality. We hope that our survey will be useful for researchers, practitioners, and anyone interested in developing datasets. The survey can serve as a guide for developing high-quality datasets and for selecting appropriate metrics to evaluate datasets accurately. Ultimately, our goal is to provide researchers and practitioners in machine learning with a valuable resource for improving their understanding of dataset quality and its impact on the performance of machine learning models.

1.2. Outline

This paper is structured to provide a comprehensive overview of the various aspects of dataset quality assessment in machine learning.

In Section 1, we highlight the importance of dataset quality assessment and outline the structure of the paper. In Section 2, we provide some background knowledge, including definitions of key terms used in this survey and a summary of the dataset under consideration. In

[*] Corresponding author.
E-mail address: lingzhong@iscas.ac.cn (L. Meng).

https://doi.org/10.1016/j.infsof.2023.107268
Received 12 September 2022; Received in revised form 22 May 2023; Accepted 24 May 2023
Available online 1 June 2023

Section 3, we review quality issues related to dataset content and their potential impact on machine learning models. Based on the dataset lifecycle in Section 4, we propose a dataset quality assessment framework, which includes eight quality dimensions and 32 evaluation metrics. Based on the existing research, we analyze the measurable evaluation metrics in Section 5 and introduce the measurement methods of the evaluation metrics. Then in Section 6, the evaluation method of dataset quality is elaborated. In Section 7, we describe the literature review methodology used in this study. Finally, in Section 8, we summarize the contributions of our work and analyze the value of dataset quality research to stakeholders.

By following this structure, our paper aims to provide readers with a comprehensive understanding of dataset quality evaluation in machine learning, ranging from the definition of quality issues, dimensions, and evaluation metrics, to the development of a quality assessment framework, and the evaluation of dataset quality using different methods. We hope that this paper will serve as a valuable resource for researchers, practitioners, and stakeholders interested in the development and evaluation of high-quality datasets.

2. Preliminary knowledge

This section aims to provide essential background knowledge related to dataset quality evaluation. By clarifying key concepts and definitions, the section makes the entire paper self-contained and allows readers to better understand subsequent discussions on dataset quality. This section serves as a foundation for readers to grasp the key aspects of dataset quality evaluation.

2.1. Definitions

Dataset: A collection of instances used for constructing or evaluating the performance of a machine learning model.

At the top level, the data could be categorized as:

• Structured data: The data also known as row data, is logically expressed and realized by a two-dimensional table structure, strictly follows the data format and length specifications, and is mainly stored and managed by a relational database.

• Unstructured data: The data whose data structure is irregular or incomplete, there is no predefined data model, and it is inconvenient to use the two-dimensional logic table of the database to represent the data. Including all formats of office documents, text, pictures, XML, HTML, various reports, images and audio/video information, etc. [1]

According to the data type, datasets can be further classified into:

• Text datasets: A set of text for building or evaluating a machine learning model.

• Image datasets:A set of image for building or evaluating a machine learning model.

• Voice datasets: A set of voice for building or evaluating a machine learning model.

Dataset life cycle: The various stages of dataset from generation to destruction, including data collection, data annotation, data storage, data testing and data destruction.

Dataset Quality: The planning, implementation and control of activities that apply quality management techniques to dataset to ensure that dataset is fit for consumption and meets the needs of dataset consumers.

Quality Issues: The quality problem that may exist in the dataset itself or in the process of use.

Quality Dimension: Refers to different perspectives of quality, not a fixed number.

Metrics: The carrier used to describe and express the basic situation determined after evaluation problem and evaluation objects are determined. In this paper, the evaluation problem is the quality issues of the dataset, and the evaluation object is the dataset.

Assessment: Comparing the measurement results of dataset quality evaluation metrics with the needs of dataset consumers.

2.2. Datasets

This section provides a brief overview of different types of datasets used in various fields. Three main types of datasets are discussed: text datasets, image datasets, and speech datasets. Text datasets are primarily used in the field of natural language processing, while image datasets are commonly used in computer vision. Speech datasets, on the other hand, are used in the domain of speech signal processing. Understanding the different types of datasets and their applications is crucial for evaluating the quality of the datasets in their respective fields.

Text datasets play a crucial role in Natural Language Processing (NLP), where they are widely used in various tasks such as sentiment analysis, text classification, and semantic analysis. The data is typically collected from websites or online forums and stored in the form of text or XML files. Text datasets contain various attributes such as field descriptions, URLs, article titles, and content. In the case of sentiment analysis, the dataset includes annotation information for sentiment categories. Some commonly used public text datasets in NLP tasks are 20Newsgroups [2] and AG-news Datasets [3]. For sentiment analysis, widely used text datasets include Amazon Review Datasets [4], NLPCC2013 [5], SST-1 [6], and Yelp Datasets [7]. These datasets serve as a benchmark for evaluating the performance of different NLP algorithms and models.

Image datasets are commonly used in tasks such as image classification, image recognition, and object detection in the field of Computer Vision. These datasets can be obtained from two sources: online and manual collection based on specific task requirements. Image datasets are usually stored in image files such as jpg and png, and contain information such as images, annotation files, and semantic images. The most commonly used image datasets for different tasks include Labelme datasets [8], Pascal VOC datasets [9], ImageNet [10], Stanford canine datasets [11], Places datasets [12], and CIFAR datasets [13] for image classification tasks. For image recognition tasks, commonly used datasets include ImageNet, COCO datasets [14], COIL100 datasets [15], Places datasets, and CelebFaces datasets [16].

Speech datasets are crucial for developing and evaluating speech recognition models in the field of Speech Signal Processing. The most commonly used speech datasets include Mozilla Common Voice datasets which cover 104 different languages as of 2023, LibriSpeech datasets [17], 2000 HUB5 English datasets [18], VoxForge datasets [19], TIMIT which is the English speech recognition dataset [20], CHIME datasets [21], MUSAN [22], TAU Urban Acoustic Scenes 2022 Mobile [23], and Acoustic Event Dataset [24]. These datasets contain speech samples in various acoustic environments, speakers, and accents, providing a diverse range of data for training and testing speech recognition models. The speech datasets are usually stored in audio file formats such as WAV or MP3, and some datasets may also include transcriptions or annotations for specific tasks. These datasets serve as valuable resources for advancing the development of speech recognition technologies.

According to the classification and summary of the above datasets, we analyzed the main application task types, data volume and data content of the datasets, as shown in Table 1. Most of the public datasets are currently open-sourced from project-based datasets developed by scientific research institutions or made available through challenge projects to facilitate technical exchanges. These datasets provide a valuable resource for researchers to benchmark their models and for practitioners to develop and test their applications.

3. Dataset quality issues and risks

In this section, we will discuss several scenarios related to dataset quality and the associated risks they pose. Specifically, we will review quality issues related to dataset content and the potential implications

Table 1
Datasets summary.

Data	Datasets name	Application task	Data amount	Data content
Text	Amazon Review	Sentiment analysis	34 GB	Amazon customer reviews and star ratings
	IMDB[25]	Word Segmentation, Sentiment analysis	80 MB	Movie reviews
	NLPCC2013	Sentiment analysis	15 MB	7 moods marked, a total of 14,000 Weibo, 45,431 sentences
	20Newsgroups [2]	Word Segmentation	63 MB	20 groups of English news data on different topics
	SST-1[6]	Sentiment analysis	946 KB	Movie reviews
	Yelp [7]	Sentiment analysis	6.4 GB	Yelp merchant, review and user data
	AG-news [3]	Word Segmentation	11 MB	News articles
Image	Labelme [8]	Image Classification, Object Detection	0.9M	187,240 images, 62,197 annotated images, and 658,992 labeled objects
	Pascal VOC [9]	Image Classification, Object Detection, Semantic Segmentation	4 GB	Image of 20 categories of objects
	ImageNet [10]	Image Classification, Object Detection	146.4 GB	20,000+ categories
	COCO [14]	Semantic Segmentation, Object Detection	25.2 GB	Tagged images
	COIL100[15]	Object Detection	129.84 MB	100 different objects, imaged in full 360°
	Stanford Canine [11]	Image Classification	2 GB	120 categories
	Places	Image Classification, Object Detection	27 GB	205 scene categories and 2.5 million images with category labels
	CelebFaces [16]	Object Detection	21.7 GB	Celebrity images
	CIFAR [13]	Image Classification	324 MB	CIFAR-10 and CIFAR-100
Voice	Mozilla Common Voice	Speech Recognition	1TB	Have 104 different languages as of 2023
	LibriSpeech [17]	Speech Recognition	100 GB	Audiobook dataset including text and speech
	VoxForge [19]	Speech Recognition	10 GB	The sample data of English, French, German, Spanish, Italian, and Russian are relatively rich
	TIMIT [20]	Speech Recognition	1 GB	Eight major dialects of American English
	CHIME [21]	Speech Recognition	4 GB	Contains real simulated and clean voice recordings
	MUSAN [22]	Speech Recognition	11 GB	A corpus of music, speech, and noise recordings
	TAU Urban Acoustic Scenes 2022 Mobile [23]	Speech Recognition	27.5 GB	1-seconds audio segments from 10 acoustic scenes, total 64 h of audio
	Acoustic Event Dataset [24]	Speech Recognition	1.2 GB	28 class acoustic event sounds

for machine learning models. This discussion lays the foundation for subsequent evaluation and assessment of dataset quality.

The assessment of dataset quality is a crucial step in evaluating the overall quality of a dataset. There are two main aspects to consider: inherent quality and used quality. Inherent quality is evaluated based on data source and characteristics, while used quality is evaluated based on the specific machine learning task. As big data technology continues to advance and become more prevalent in fields such as artificial intelligence, medicine, and business, it is important to take into account factors such as data management, processing, and user needs in order to fully measure the quality of big data [26]. Additionally, it is important to consider the definition of data quality, the classification and sources of data quality problems, the data quality assessment framework, and data cleaning methods [27].

3.1. Dataset quality issues

This subsection discusses the risks associated with insufficient metadata leading to data misinterpretation.

According to a recent MIT study most of the well-known AI datasets are full of labeling errors, so we have to pay attention to dataset quality issues [28]. A quality assessment process, developed in [29], defines data quality issues and metadata quality issues during the identification stage of data quality problems. Data quality issues include duplicate data, inconsistent data, missing data, and incorrect data, while metadata quality issues include incomplete metadata for a metric or entity and imprecise metadata for a metric or entity. The assessment scale is constructed based on the problem definition to conduct the quality assessment work.

During the preprocessing stage, analysts spend a lot of time and energy dealing with large amounts of data. To address this problem, Christian et al. proposed MetricDoc [30], an interactive environment for evaluating data quality. The tool raises metrics for missing values, invalid data, incorrect data generation, and data duplication, and uses detection rules to identify data quality issues.

Similarly, Song et al. [27] identified data quality issues such as spelling errors, duplicate records, conflicting fields, and data inconsistencies, proposing data cleaning methods such as N-gram-based duplicate record detection to improve preprocessing efficiency. In response to the possible problems in the dataset, Simone et al. [31] proposed three problems: selection bias, framing bias and label bias.

In face recognition datasets, images with incorrect ID labels are common due to collection from the internet using search engines. Guo et al. [32] propose an ID tag cleaning method to retain more low-quality face images, provide diverse data, and train better face models.

In supervised computer vision tasks such as target detection, large-scale labeled data is obtained using artificial annotation crowdsourcing platforms, leading to a decrease in the quality of annotations. This will seriously affect the model's training, as proposed by Xie et al. [33].

In conclusion, the literature reviewed above highlights quality issues such as data duplication, data inconsistency, data missing, data incorrect, data invalid, data inconsistent, data incomplete, data inaccurate, and ID label error. These problems can adversely affect feature extraction, ultimately impacting the model's accuracy.

3.2. Risks associated with dataset quality issues

Dataset quality issues pose risks to the performance and practical applications of machine learning models, potentially resulting in inaccurate or unreliable predictions. The following are some common risks associated with dataset quality issues:

(1). Decline in model performance: Dataset quality issues may lead to a decline in model performance, lowering metrics such as accuracy, precision, or recall.

(2). Unreliable model: Dataset quality issues may lead to unreliable or unstable predictions, which can have a serious impact on practical applications.

(3). Misleading conclusions: Dataset quality issues may lead to erroneous or misleading conclusions, bringing risks and losses to business decisions.

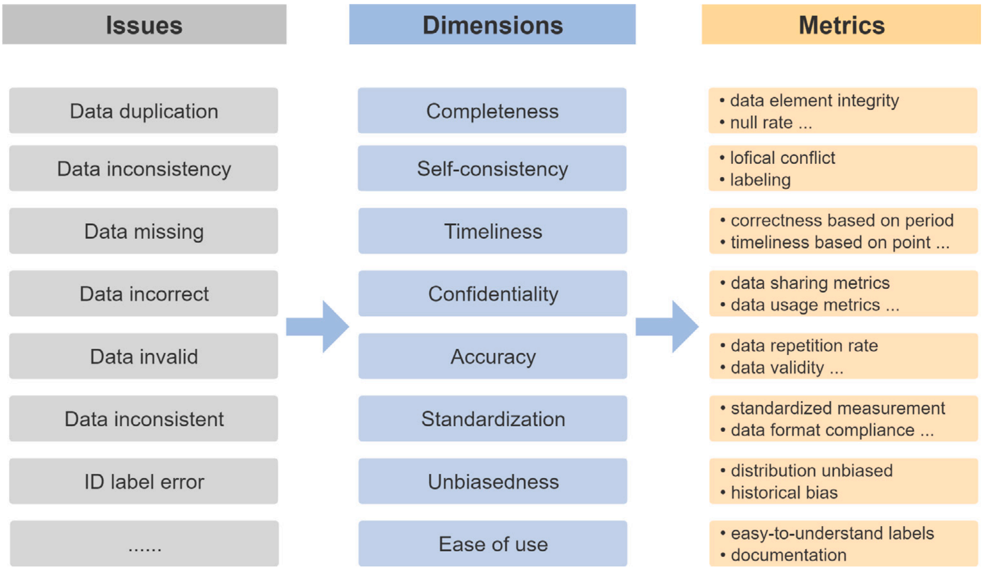

Fig. 1. Process of dataset quality analysis.

(4). Security risks: Dataset quality issues may pose security risks, such as privacy breaches and malicious attacks.

To reduce these risks, it is necessary to conduct data quality analysis, data cleaning, and data preprocessing, among other things, to ensure dataset quality. Many data quality assessment methods and metrics have been proposed in relevant works, such as data completeness, data consistency, data accuracy, and data reliability, which can help identify and correct issues in the dataset and improve model performance and practical application reliability.

4. Dataset quality evaluation framework

This section provides a comprehensive analysis of dataset quality, covering various dimensions and evaluation metrics. Starting with an overview of the dataset issues, we then introduce the dataset quality dimension and analyze evaluation metrics from this perspective, as shown in Fig. 1. By summarizing the quality of the dataset, we review each stage of the dataset's life cycle and present eight quality dimensions and 32 evaluation metrics. These dimensions and metrics are based on survey results and complement the ones proposed in this paper, providing a comprehensive framework for evaluating dataset quality.

4.1. Lifecycle-based dataset quality evaluation

The quality of datasets used in machine learning tasks is influenced by various factors throughout their lifecycle, from the initial data collection stage to the use of data to complete the task. It is important to analyze evaluation metrics at each stage of the lifecycle to ensure dataset quality.

The big data lifecycle consists of collecting data from different sources, storing data, computing or analyzing data, and visualizing results [26]. The lifecycle of datasets can be divided into several stages, including data generation/initial stage, data acquisition, data storage, data processing and analysis, and data visualization [34]. Taleb et al. [35] proposed the data life cycle, including data generation, data acquisition, data storage, and data analysis. Dataset quality is divided into three categories: original quality, process quality, and result quality, according to the lifecycle of datasets. Additionally, the quality of datasets is closely related to the task execution process, and can be further divided into the inherent quality of data, the quality of data expression, the quality of the situation related to the system, the quality of data utility, and the quality of user experience data [36].

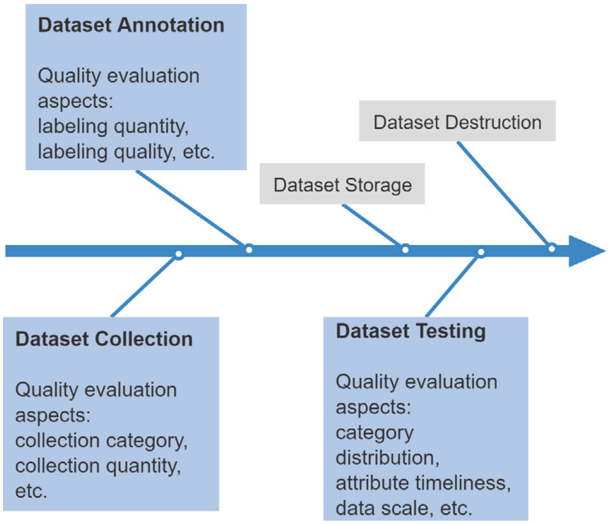

Fig. 2. Quality analysis in dataset life cycle.

Based on the above literature, the lifecycle of a dataset can be divided into several stages: dataset collection, dataset annotation, dataset storage, dataset testing, and dataset destruction, as shown in Fig. 2. During the lifecycle, the data collection and annotation phases have a significant impact on the inherent quality of the dataset, while the data testing phase evaluates the quality of the dataset from the perspective of applicable tasks. Dataset quality assessment metrics at each stage of the dataset's lifecycle are summarized in the Table 2.

4.1.1. Data collection

During the data collection process of datasets, it is essential to conduct research on the quality of the original data. The quality constraints of the data sources in datasets can be considered from two aspects: the standardization of data sources and the security and stability of data sources. In the process of data acquisition, it is important to ensure the real-time, error-free, and integrity of data acquisition, and to guarantee the quality of the data acquisition process. Additionally, there are quality evaluation metrics such as data definition consistency and data arrival rate [37]. The assessment of raw data quality includes three quality characteristics: accuracy, authenticity, and integrity. The

Table 2
Dataset lifecycle and quality evaluation.

Lifecycle	Metrics	Description
Data collection	Real-time data collection [37]	Ensure the timeliness and value of big data.
	Error-free data collection [37]	Ensure the accuracy of the data collected. There cannot be data descriptions that are inconsistent with objective facts.
	Integrity of data collection [37]	Ensure the integrity of the required data as much as possible during data collection, which can be represented by metrics such as the frequency of null values.
	Data content correctness [36]	The data content meets the collection requirements.
	Data format conformity [36]	Whether the format of the collected data meets the format requirements.
	Data repetition rate [36]	The proportion of duplicate data in the collected data.
	Data uniqueness [36]	Whether the data in the collected data is unique.
	Dirty data occurrence [36]	Proportion of dirty data in collected data.
	Datasets equivalence [38]	Some features of the datasets were compared with those used by previous researchers to demonstrate the adequacy of the selected datasets.
	Selection bias [39]	When the datasets is not representative of the expected population or does not solve the problem.
	Information bias [40]	Accuracy of automatically generated datasets, misclassification and labeling of the data that will be used.
	Negative set bias [41]	Select rich and unbiased negative cases in the training data.
Data annotation	Labeling accuracy [42]	The proportion of data labeling errors in the datasets.
	Word error rate [43]	Indicates some words that need to be replaced, deleted or inserted in order to keep the recognized word sequence consistent with the standard word sequence.
	Sentence error rate [44]	Identify if a word recognition error occurs in a sentence.
	Character error rate [45]	Chinese labels generally use CER to represent WER.
	Participle unambiguous	The participle is consistent with the word in the dictionary and has no ambiguity.
	Sentiment rating categories accuracy	Correct sentiment rating categories for labeled sentences.
	Semantically accuracy	Semantically correct words and sentences.
	Fluency of text data [46]	Compare the degree of overlap between the n-grams in the candidate passage and the reference passage. The higher the degree of overlap, the higher the quality of the text.
Data testing	Data imbalance [47]	Inter-class imbalance is the difference in the number distribution of each category in the dataset, and intra-class imbalance is within the same category.
	Data distribution consistency [48]	When dividing the training set, validation set, and test set, ensure that the distribution of each dataset is consistent.

accuracy metric includes data content correctness, data format compliance, data repetition rate, data uniqueness, and dirty data rate. The authenticity metric is related to the objective truth of data, while the integrity metric includes data element integrity and data record integrity [36].

In the data acquisition stage, four evaluation metrics are proposed: selection bias, datasets equivalence, information bias, and negative set bias [40]. Selection bias means that datasets cannot represent the expected population or cannot solve the problem [39]. The equivalence of datasets is used to compare some characteristics that threaten effectiveness with data used by previous personnel to prove the adequacy of datasets [38]. Information bias is related to the accuracy of automatically generated datasets, and negative set bias is used to select abundant and unbiased negative cases in training data [41].

Therefore, in the data collection stage, it is crucial to evaluate the quality of datasets using metrics such as real-time data collection, error-free data collection, the integrity of data collection, data content correctness, data format compliance, data repetition rate, data uniqueness, dirty data occurrence, selection bias, datasets equivalence, information bias, and negative set bias. These metrics ensure that datasets are of high quality and can be used effectively for machine learning tasks.

4.1.2. Data annotation

Data annotation is a crucial process in machine learning tasks that involves converting unprocessed primary data such as images, voice, text, and video into machine-recognizable information. The quality of the annotation is a significant factor that affects the performance of machine learning models.

In image annotation, the quality of the annotation depends on the accuracy of the bounding box marked around the object, and the smaller the pixel distance between the box and the object, the higher the quality of the annotation. Another important aspect of annotation quality is the correctness of the annotated objects. To measure the impact of mislabeled training data on model learning, a labeling accuracy metric has been proposed [42].

In voice annotation, it is essential to ensure that the time axis of the pronunciation of voice data is synchronized with the phonetic symbols in the marked area [49]. For English speech annotation, the word error rate [43] and sentence error rate [44] are commonly used evaluation metrics. In Chinese speech, the character error rate is generally used [45].

Text annotation has different quality standards for different tasks. For example, in Chinese word segmentation tasks, the evaluation metric is the consistency of the segmentation with the words in the dictionary without ambiguity. In the sentiment labeling task, the evaluation metric is the correctness of the sentiment rating category of the labeled sentence. In semantic tasks, the evaluation metric is the semantic correctness of words and sentences. Lavie proposed an evaluation metric for the fluency of text data and designed an evaluation algorithm to evaluate the quality of manual text annotation based on this metric [46].

In the data annotation stage, evaluation metrics are organized from three aspects: image, audio, and text, including labeling accuracy, word error rate, sentence error rate, character error rate, word segmentation without ambiguity, sentiment rating category accuracy, semantic accuracy, and text data fluency.

4.1.3. Data testing

In machine learning tasks, the characteristics of datasets may also affect the accuracy of task execution. In this stage, evaluation metrics of quality used are sorted out. The imbalance of datasets is proposed, and the imbalance is divided into inter-class imbalance and intra-class imbalance [47]. Inter-class imbalance is that the number distribution of different types of datasets varies greatly. Intra-class imbalance is when the characteristics of elements in the same category are not significantly different, resulting in that elements in the same category cannot cover all characteristics. In machine learning tasks, the use of datasets is divided into training sets, validation sets, and test sets. The data distribution in different sets requires metrics to measure quality. An evaluation metric of data distribution consistency is proposed for the generation process of random sample division of big data [48].

In the data testing phase, dataset quality evaluation metrics include data imbalance and data distribution consistency.

4.2. Quality dimensions

The quality of data is a critical aspect of data analysis and evaluation, and various evaluation metrics have been proposed in the literature. While there are some cognitive differences across different fields and individuals, standardized data quality assessment is becoming more common.

In one widely accepted standard, “GB/T 36344-2018 Information Technology Data Quality Evaluation Indicators”[50], six evaluation metrics are proposed, including normativity, integrity, accuracy, consistency, timeliness, and accessibility. In the context of machine learning systems, Picard et al. [51] proposed seven quality dimensions, which include accuracy, accessibility, consistency, relevance, timeliness, traceability, and usability. The dataset quality dimensions given by Chug et al. [52] include Provenance, Dataset Characteristics, Uniformity, Metadata coupling, Statistics, and Correlations.

Song et al. [27] suggest that the most suitable data quality dimension should be selected to evaluate the quality of data based on business needs. They provide a set of basic dimensions for data quality projects, which include data specification, data integrity fundamentals, duplication, accuracy, consistency and synchronization, timeliness and availability, ease of use and maintainability, data coverage, presentation quality, perception, relevance and trust, data decay, and transactability. Hongxun et al. [53] proposed six quality dimensions of redundancy, integrity, accuracy, consistency, timeliness, and intelligence and gave check rules for dataset and single data.

Different scholars have constructed metrics systems with different evaluation dimensions based on their evaluation objectives. Table 3 summarizes the common quality evaluation dimensions and their definitions from each literature. Combining these dimensions with those proposed in the life cycle, we identify eight quality dimensions, including completeness, self-consistency, timeliness, confidentiality, accuracy, standardization, unbiasedness, and ease of use.

(1). Completeness

Definition 1. Completeness is a critical quality dimension that reflects the degree to which subject data associated with an entity has values for all expected attributes and associated strength values in a given environment.

Completeness is a commonly used metric in data analysis and sorting applications, which mainly refers to missing values [30]. In the context of big data, where the analysis and mining of overall and full data are emphasized, data collection integrity metrics are also essential [37]. Rosli et al. [29] propose metadata completeness, which is evaluated based on two aspects: identifying metric labels without metric metadata and entity labels without entity metadata. Completeness includes six evaluation metrics, including data element integrity, data record integrity, metadata integrity, null rate, integrity of data collection categories, and the integrity of data collection quantity. These metrics can be used to assess the quality issues that may arise during dataset construction.

(2). Self-consistency

Definition 2. Self-consistency is defined as the degree to which there is no contradiction in the semantics of the data in a given context.

Self-consistency includes two evaluation metrics, logical conflict, and labeling, to evaluate the quality of dataset content involving semantics.

(3). Timeliness

Definition 3. Timeliness is defined as the degree to which data has properties that characterize its correct longevity in a given context.

Timeliness is to ensure that data keeps pace with the times and is not outdated. Data timeliness can be divided into timestamp-based data timeliness and rule-based data timeliness [66]. The timeliness of datasets can be evaluated in two aspects: correctness based on time period and timeliness based on the time point [36]. The timeliness includes three evaluation metrics: correctness based on period, timeliness based on point and time series, and evaluates the timeliness quality of data elements in the dataset.

(4). Confidentiality

Definition 4. Confidentiality is defined as the degree to which the data itself can only be accessed and interpreted by authorized users in a given environment.

Confidentiality includes six evaluation metrics: original data metrics, labeled data metrics, data sharing metrics, data analysis metrics, data usage metrics, and data discard metrics. Confidentiality is aimed at the entire life cycle of the dataset, from the original data collected to the final destruction of the dataset, to assess the quality of data confidentiality at different stages.

(5). Accuracy

Definition 5. Accuracy is defined as the degree to which data has attributes that correctly represent the true value of the relevant attributes of a concept or event in a given environment.

A record in datasets is accurate and valid if they are marked as records and correctly marked. Inaccurate or invalid data leads to data noise, and mislabeled data leads to label noise. For data validity, validating the data is a critical part of the analysis, because invalid input may hinder measurement or distort statistical evaluation. The definition of effectiveness is whether the preprocessed data are valid and whether big data analysis can be performed. Specific measurement metrics are given, including whether the data are stable and user-available within its validity period [37]. Datasets should contain a set of (input, output) pairs that fully conform to the specification. All (input, output) pairs must be validated against the specification before being used in the training process [51]. For data accuracy, it is defined as the degree of correct value [64]. The checking rule on the accuracy of datasets is after removing redundant data, the difference of the total amount of datasets in the link of adjacent data flow should be within a reasonable range [53]. Accuracy includes five evaluation metrics: data content correctness, data repetition rate, data uniqueness, data validity, and label accuracy. Accuracy is evaluated from the content and form of the dataset, including valid data, correct labels and correct data in content, and data duplication and uniqueness in form.

(6). Standardization

Definition 6. Standardization is defined as the degree to which data, in a given environment, conforms to data standards, data models, business rules, metadata, or authoritative reference data.

standardization includes six evaluation metrics: data standard measurement, authoritative reference data measurement, business rule measurement, security specification measurement, standardized measurement, and data format compliance. Standardization is different from the data content, data usage, and data source in the dataset. Aspects evaluate how well a dataset matches the relevant rules.

Table 3
Summary of dataset quality dimension in literatures.

Time	Literature	Evaluation dimension
2023	A survey of data quality requirements that matter in ML development pipelines [54]	Intrinsic, Contextual, Accessibility, and Representational
2023	Textured Mesh Quality Assessment: Large-Scale Dataset and Deep Learning-based Quality Metric [55]	GRAPHICS-LPIPS(Learned Perceptual Image Patch Similarity)
2022	ISO/IEC JTC 1/SC 42(AI)/WG 2(Data) Data Quality for Analytics and Machine Learning (ML)[56]	Accuracy, Precision, Completeness, Representativeness, Consistency, Relevance, Data scalability, Context coverage, Portability, Timeliness, Currentness, Identifiability, Auditability, Credibility, Understandability, Balance, Effectiveness, and Similarity
2021	Exploring big data traits and data quality dimensions for big data analytics application using partial least squares structural equation modeling [57]	Big data traits, Accuracy, Believability, Completeness, Timeliness, and Ease of operation
2021	Statistical Learning to Operationalize a Domain Agnostic Data Quality Scoring [52]	Provenance, Dataset Characteristics, Uniformity, Metadata coupling, Statistics, Correlations
2021	Non-empirical problems in fair machine learning [58]	Algorithmic fairness in the dataset
2021	Moving beyond algorithmic bias is a data problem [59]	Algorithmic bias in the dataset
2020	Studies on Data Quality Evaluation Index System for Internet Plus Government Services in Big Data Era [36]	Accuracy, Authenticity, Integrity, Standardization, Understandability, Consistency, Traceability, Accessibility, Timeliness, Security
2020	Redundancy and Complexity Metrics for Big Data [37]	Complexity, Redundancy, Density
2020	Ensuring Dataset Quality for Machine Learning Certification [50]	Accuracy, Accessibility, Consistency, Timeliness, Traceability, Usability, Relevance
2019	An Association-Based Intrinsic Quality Index for Healthcare Dataset Ranking [60]	Association-based intrinsic Quality Index(AQI)
2019	longSil: an Evaluation Metric to Assess Quality of Clustering Longitudinal Clinical Data [61]	Tightness, Degree of separation, LongSil-longitudinal silhouette coefficient
2018	How textual quality of online reviews affect classification performance: a case of deep learning sentiment analysis [62]	The word count, Review readability
2018	Assessing data quality — A probability-based metric for semantic consistency [63]	Semantic consistency
2018	Context-aware data quality assessment for big data [64]	Accuracy, Completeness, Consistency, Distinctness, Precision, Timeliness, Volume
2018	Construction of Big Data Quality Measurement Model [37]	Normative, Security and stability, Real-time Metrics, Error-free, Completeness, Consistency, Accuracy, Applicability, Rationality, Effectiveness, Ease of understanding, Value
2018	Visual Interactive Creation, Customization, and Analysis of Data Quality Metrics [30]	Completeness, Validity, Plausibility, Time Interval Metrics, Uniqueness
2018	Data Quality Assessment for On-line Monitoring and Measuring System of Power Quality Based on Big Data and Data Provenance Theory [53]	Intelligence, Redundancy, Integrity, Accuracy, Consistency, Timeliness
2018	Big Data Quality: A Survey [34]	Intrinsic, Contextual, Representational, Accessibility
2017	Datasets from Fifteen Years of Automated Requirements Traceability Research Current State, Characteristics, and Quality [65]	Accessibility, Intrinsic, Contextual, Representational

(7). Unbiasedness

Definition 7. Unbiasedness is defined as the degree to which the distribution of data categories or features in a dataset differs in a given environment.

Unbiasedness includes two evaluation metrics, distribution unbiased and historical bias, and is mainly used to evaluate the quality of the class imbalance and content subjective bias in the dataset.

(8). Ease of use

Definition 8. Ease of use is defined as the degree to which a dataset can be used quickly when applied in a given environment.

Ease of use includes two evaluation metrics: easy-to-understand labels and documentation, which measure how easy it is for researchers to use the dataset.

The dataset quality dimensions proposed above apply to all types of datasets, covering quality assessment at all stages of the dataset life cycle. Moreover, there is no requirement for the format of data elements, and each dimension is applicable to pictures and text.

At present, the quality dimension of evaluation datasets is mainly from the eight major aspects of completeness, self-consistency, timeliness, confidentiality, accuracy, standardization, unbiasedness, and ease of use. Starting from these eight quality dimensions, 32 evaluation metrics were obtained through in-depth analysis. Table 4 summarizes the evaluation dimensions and evaluation metrics of dataset quality, defines quality characteristics, and explains the evaluation metrics in the framework to provide guidance for subsequent model applications. However, according to different applications and different scholars, new evaluation metrics are also proposed, such as understandability, value, traceability, security, redundancy, and other metrics. In the field of machine learning, some scholars also proposed a single metric to evaluate the quality of datasets, such as complexity, redundancy, density, association-based intrinsic quality index, and intelligence metrics. The current evaluation dimension is still based on data quality, and there are few studies on the quality dimension of machine learning features.

5. Dataset quality evaluation metrics

The previous section deeply discussed the research status of dataset quality assessment from three aspects: dataset, dataset quality dimension, and dataset quality evaluation metric. Evaluation metrics for datasets are generally evaluated from both qualitative and quantitative aspects, and some metrics need to be evaluated manually, but some metrics can be calculated using algorithms to obtain results. This section mainly analyzes measurable evaluation metrics, introduces some measurement methods of evaluation metrics, and classifies evaluation

Table 4
Dataset quality metrics.

Characteristics	Metrics	Definition
Completeness	data element integrity	According to the requirements of business rules, the assignment degree of the data elements that should be assigned values in the dataset.
	data record integrity	According to the requirements of business rules, the degree of assignment of data records that should be assigned values in the dataset.
	metadata integrity	The integrity of the data information in the dataset.
	null rate	Null values represent missing information in the dataset, counting the number of fields in the dataset that are empty.
	integrity of data collection categories	When obtaining raw data, it is necessary to collect data of various categories.
	integrity of data collection quantity	When obtaining raw data, the amount of data required for each category also requires certain requirements.
Self-consistency	logical conflict	Identify conflicts in data semantics.
	labeling	The data can be labeled correctly.
Timeliness	correctness based on period	Based on the degree to which the number or frequency distribution of records within a date range meets business requirements.
	timeliness based on point	The degree to which the number of records, frequency distribution, or latency based on timestamps meet business needs.
	time series	The relative timing relationship between data elements of the same entity in a dataset.
Confidentiality	original data metrics	Cases where raw data is encrypted.
	labeled data metrics	When labeling data, control access and inspection of data.
	data sharing metrics	When sharing data, perform operations such as blanking sensitive data.
	data analysis metrics	During data analysis, authority control is performed on data statistics and other operations.
	data usage metrics	When the data is in use, perform migration encryption and integrity verification on the data.
	data discard metrics	Separation of permissions when data is discarded.
Accuracy	data content correctness	Whether the data content is the expected data.
	data repetition rate	Metrics that are unexpectedly repeated for a specific field, record, file, or dataset.
	data uniqueness	A measure of the uniqueness of a particular field, record, file, or dataset.
	data validity	The elements in the dataset conform to the requirements of the dataset design.
	label accuracy	The accuracy of labeling the elements in the dataset.
Standardization	data standard measurement	Judging whether the data meets the data standard according to the requirements of the business data.
	authoritative reference data measurement	Whether it meets the requirements of authoritative reference data.
	business rule measurement	Whether the data in the dataset complies with the business rules.
	security specification measurement	Security specifications are rules on security and privacy, including data rights management, data desensitization, etc.
	standardized measurement	The researchers preprocessed the data to convert it into the same format, so users do not need background knowledge to use it.
	data format compliance	Whether the data format includes (data type, data range, data length, precision, etc.) meets the expected requirements.
Unbiasedness	distribution unbiased	The overall distribution of the dataset is unbiased.
	historical bias	At the time of data collection, there may be bias based on collection equipment and collector preferences.
Ease of use	easy-to-understand labels	Datasets are labeled to make it easy for users to understand.
	documentation	The description of the data set is detailed, and the data set can be quickly familiarized with.

metrics from two directions: statistical information calculation and machine learning model calculation. Fig. 3 shows the main research ideas in this section, from the dataset quality evaluation framework to the dataset quality metric evaluation framework.

5.1. Dataset quality metric evaluation framework

By analyzing the characteristics of the dataset and quantifying the evaluation metrics, the above-mentioned dataset quality evaluation framework is tailored and applied to the quality measurement and evaluation tasks of the dataset. The dataset has the characteristics of large-scale data, annotation information, and relatively standardized datasets. Combined with the characteristics of the collection and the data, the quality evaluation system of the dataset is cut and supplemented to obtain the quality of the dataset. The model contains dataset quality with four quality dimensions and five evaluation metrics. The following are definitions of the five evaluation metrics:

(1). File Completeness: The extent to which files are missing in the dataset.

(2). Attribute timeliness: The degree of correctness of the features of the annotated attributes of the dataset over time.

(3). Data validity: The degree to which various types of data in the dataset are useful for extracting attribute features.

(4). Label accuracy: The degree to which labels corresponding to element in an dataset are real entities.

(5). Imbalance: The degree of difference in the distribution of the amount of data in each category in the dataset.

5.2. Evaluation metrics based on statistical information

For the five evaluation metrics proposed, three of the evaluation metrics are measured based on statistical information and two are based on machine learning. This section describes in detail the metric evaluation metrics with statistical information, including the meaning

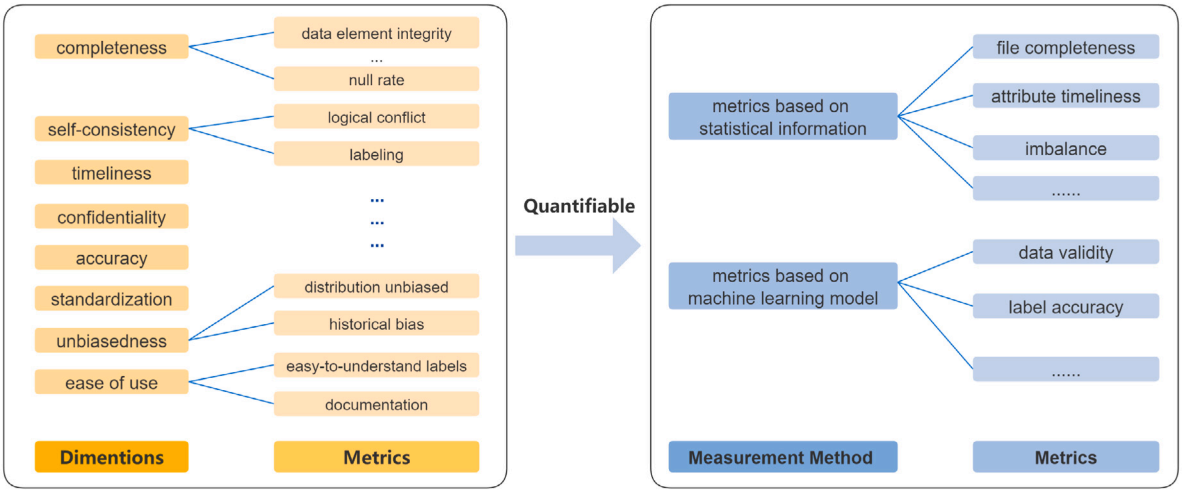

Fig. 3. Dataset quality evaluation framework and measurable metrics.

of the metric, the calculation formula, and the description of the algorithm.

5.2.1. Dataset completeness metrics

Quality issues with completeness measures include mismatches between elements in the dataset and requirements and in-dataset annotations that do not correspond to element files. Let us use an image dataset as an example to illustrate how this metric is measured [29].

There is a corresponding relationship between the image file and the configuration file of the image dataset. If the corresponding relationship is missing, the file is incomplete, and then continue to judge whether each file has a corresponding file. If it does not exist, the file is missing. There is a corresponding relationship between the image file and the configuration file of the image dataset. If the corresponding relationship is missing, the file is incomplete, and then continue to judge whether each file has a corresponding file. If it does not exist, the file is missing. The measurement result of this metric is calculated by Eq. (1).

$$\mathrm{X} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{N1}} \mathrm{a_i} + \sum_{\mathrm{i}=1}^{\mathrm{N2}} \mathrm{b_i}}{2 * \max\{\mathrm{N1},\ \mathrm{N2}\}} \tag{1}$$

$N1$: Number of image files.

$N2$: Number of annotation files.

a_i: Traverse the image folder, and for the ith image, check whether the annotation file corresponding to the image exists in the annotation folder. If it exists, $a_i = 0$, if it does not exist, $a_i = 1$.

b_i: Traverse the annotation folder, and for the ith annotation file, check whether the image file corresponding to the annotation file exists in the image folder. If it exists, $b_i = 0$, and if it does not exist, $b_i = 1$.

5.2.2. Dataset timeliness metrics

When using datasets, different rules may cause data invalidation and unavailability in different periods, so you need to consider the timeliness of the dataset. Let us take the image dataset as an example to illustrate the timeliness measurement method, the overall content of the image dataset may become old over time, and the features contained in the dataset account for a smaller proportion of the overall task requirements. Set a library of objects that change significantly over time, such as the COVID-19 dataset [67], over time, there will be new CT images with new data features.

The input of attribute timeliness is the image dataset, and the output is the timeliness level. This attribute is the annotation content in the image dataset. First, obtain a list of unique labels in the image datasets, and compare the attribute list with the timeliness rules to determine the timeliness quality of the dataset to be tested. The higher the timeliness, the greater the impact of temporal changes on the use of image datasets.

Let image datasets be the relation $R = (A_i, \dots, A_n)$, where A_i is the ith attribute. If attribute A satisfies the timeliness rule, the timeliness of the attribute is high. There are features with high timeliness. We judge that the dataset has high timeliness. Generally, experts set the timeliness rules for data in various fields. Objects with high timeliness have the characteristics of instability and fast iteration. The formulation of aging rules mainly depends on the speed of data changes, such as e-commerce data and COVID-19 virus image datasets used in analyzing sales tasks.

5.2.3. Dataset imbalance metrics

The imbalance metric measures the degree to which the data of each category is too different. Count the number of label categories, calculated according to Eq. (2), and gain the quantitative value of the category difference.

$$X = \frac{\sum_{i=1}^{N} |x_i - \bar{x}|}{N} \tag{2}$$

x_i: the count of each category.

N: the total number of objects.

By traversing all the label files, all categories of the dataset to be tested are obtained and calculated. First, count the number of categories in each tag file, analyze the tags in each tag file, get the label information of the objects tag file, and count all the tagged objects in the tag file. Then traverse the label files in the folder, and update the total number of categories once every cycle. Finally, obtain the number statistics of each category in the dataset to be tested and the measurement result of imbalance by calculating the formula.

5.3. Evaluation metrics based on machine learning model

5.3.1. Dataset validity metrics

In the process of using the dataset, it is defined that the samples that do not provide help for the attribute feature extraction of the dataset are invalid data [68]. Let us take the image dataset as an example to give the measurement method. The calculation formula for the validity of the image dataset is Eq. (3).

$$\mathrm{X} = \sum_{i}^{c} A_i / N \tag{3}$$

c: The total number of categories in dataset to be tested.

A_i: Number of invalid data in class i.

N: The total number of data in dataset to be tested.

The measurement algorithm of data validity, the CNN model encodes and decodes all data based on the Autoencoder principle. The training goal is to load the input value and output value into the encoder using the parameters trained by Autoencoder, and the feature space output by the encoder module can be used as the feature space extracted from the dataset. First, calculate the center of the feature space, then calculate the distance d between all vectors and the center point vector, the calculation method is shown in Eq. (4), and obtain the distance vector $D = \left(d_1, \ldots, d_j, \ldots, d_n\right)$.

$$d\left(x_i, x_j\right) = \sqrt{\sum_{t=1}^{m} \left(x_{it} - x_{jt}\right)^2} \tag{4}$$

The distance vector is normalized and then classified and the output and the data in the $(0.9, 1]$ interval are recorded as invalid data.

5.3.2. Dataset accuracy metrics

Labeling errors can significantly impact the accuracy of a measurement dataset. Inaccuracies may arise due to different types of labeling errors, and current methods for locating errors include manual review and machine learning model review. To illustrate a model-based metric measurement algorithm, we will use an image dataset as an example.

To address labeling errors in the image dataset, we propose the following algorithm to filter out errors and obtain an accurate measurement result [28]:

Step 1: Count — Estimate the joint distribution of noise labels and true labels.

Step 2: Clean — Identify and filter out incorrect samples.

Step 3: Re-training — After filtering out the incorrect samples, re-adjust the weights of each sample category and re-train the model.

We estimate the joint distribution by using quadruple cross-validation with Resnet architecture. The algorithm takes two input variables — the original sample label (referred to as the noise label due to potential errors) and the probability of each sample under different label categories predicted by cross-validation on the training set. This probability is an A probability matrix of size $n\ x\ m$ (n being the dataset size and m being the total number of label classes).

The count matrix is derived from the probability matrix, where we calculate the four metrics of True Positives, False Positives, False Negatives, and True Negatives based on the average probability. The sum of these counts equals the total number of manually labeled samples. We then weigh the original total number of samples and normalize the probability to obtain a joint distribution. Finally, we screen the samples based on the distribution obtained from the previous steps.

6. Dataset quality assessment methods

Firstly, the quality assessment methods used are introduced, including Analytic Hierarchy Process(AHP), Alternative Metrology, and Weight Grade Method.

AHP [69] is a qualitative and quantitative decision-making method, which can help to solve problems with a large number of subjective and objective factors, especially when the final expectations are relatively vague, using AHP to analyze is extremely effective. Through the analytic hierarchy process, the index weight of the evaluation system is calculated according to the contribution degree of the evaluation index, which makes the evaluation system more reasonable.

Alternative metrology [70] is a measure of the overall impact of diverse academic outcomes. Alternative metrology can be divided into a narrow sense and a broad sense. The narrow sense of alternative metrology is devoted to the study of new online metrics relative to traditional metrics based on citations, with particular emphasis on metrics based on social network data. The overall impact evaluation index system of achievements aims to replace the traditional quantitative scientific research evaluation system that relies only on citation metrics, and at the same time promote the comprehensive development of open science and online communication.

The Weight Grade Method [71], also known as the comprehensive weighted average method, uses several observations of the same variable in the past arranged in time sequence and uses the time sequence number as the weight to calculate the weighted arithmetic mean of the observed values.

The results of Dataset Completeness Metrics, Dataset Timeliness Metrics, Dataset Imbalance Metrics, Dataset Validity Metrics, and Dataset Accuracy Metrics are obtained according to the metric algorithm. The evaluation threshold of each evaluation metric is set by the quality evaluator, and the threshold of each evaluation metric is set.

According to the threshold and measurement results of Dataset Completeness Metrics, Dataset Timeliness Metrics, Dataset Imbalance Metrics, Dataset Validity Metrics, and Dataset Accuracy Metrics, the result data of the accuracy of the intermediate variables are obtained. Then, the measurement data and its threshold of the evaluation index are weighted and summed to obtain the quality evaluation results of the dataset, and then the quality is layered according to the obtained data.

7. Literature analysis

This section outlines the criteria and methods employed in the selection of review articles and provides an analysis of the chosen articles.

The selection criteria for the literature were established considering three primary factors: content, time, and authority.

For content: We initially defined the topic's keywords. These keywords were then used to search for relevant literature. To ensure a strong correlation with the topic, each piece of literature's abstract was meticulously reviewed.

Regarding time: No restrictions were imposed on the literature's age for basic research. Instead, the selection was based on the number of citations, prioritizing classic literature. For literature discussing recent research progress, we narrowed our selection to articles published within the last five years.

Concerning authority: We took steps to guarantee the credibility of the literature sources. Our search commenced with top-tier journals and conferences, such as ACM, AI, and JMLR, among others. Subsequently, we delved into academic literature databases with significant impact factors, like the Web of Science. Finally, we performed a secondary search on the citations of the literature initially selected.

According to the above criteria to complete the relevant literature search.

We first searched well-known literature databases at home and abroad using keywords such as "dataset", "dataset quality", and "image/text/voice dataset quality", and screened out articles related to this review. Literature databases include HowNet, Association for Computing Machinery, ScienceDirect, IEEE Xplore, etc.

Then, according to the references of the articles selected by the search and related authors published articles, we further supplemented the relevant literature.

Finally, we selected 71 relevant papers, of which 24 were related to datasets, 7 were related to dataset quality issues, 30 were related to dataset quality indicators, and 10 were other related. Most of the papers were published in 2018 and beyond. See Fig. 4 for the content quantity distribution of papers and the quantity distribution in the past five years.

8. Conclusion

This paper provides a comprehensive review of the research on dataset quality. Firstly, we provide definitions of relevant terms and discuss various datasets. Next, we summarize the different types of problems related to dataset quality analysis. Then, we review evaluation metrics related to dataset quality across the various stages of a dataset's lifecycle. We present an evaluation framework consisting of eight quality dimensions and 32 evaluation metrics. Furthermore, we discuss the computation of these evaluation metrics. Finally, we discuss dataset quality evaluation and identify some limitations of current

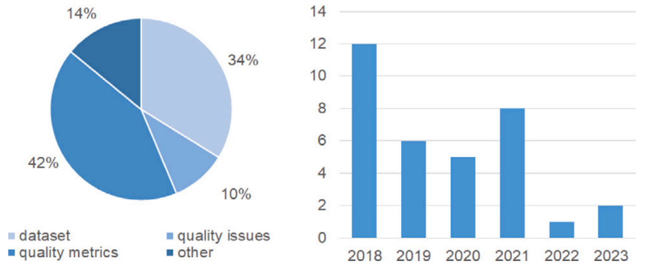

Fig. 4. Number of papers distributed.

research in this area. Overall, this review highlights the progress made in the field of dataset quality evaluation and suggests avenues for future research.

While current research on dataset quality has achieved significant results, particularly in the dimension of dataset quality, there are still some limitations that need to be addressed. These include:

(1). Lack of uniformity in the evaluation metrics used during the construction of datasets. Current dataset construction is based on the respective standards of each agency, making it challenging to obtain a unified description of the quality assurance of dataset construction.

(2). The dataset quality dimension is not yet comprehensive. At present, some of the quality dimensions of datasets are still analyzed based on the quality of traditional databases, and new quality dimensions need to be evaluated according to the characteristics of the machine learning field.

(3). Further research is needed to deepen the measurement of dataset quality evaluation metrics.

To address these limitations, future research can focus on clarifying dataset quality standards during dataset construction to achieve greater uniformity in the construction and use of datasets. Additionally, new quality dimensions can be added by focusing on the characteristics of tasks in the field of machine learning. Finally, deep learning, natural language processing, and other technologies can be utilized to vectorize the evaluation metrics of datasets to enhance the measurement of dataset quality evaluation metrics.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] I. Taleb, M.A. Serhani, R. Dssouli, Big data quality assessment model for unstructured data, in: IIT 2018 : 13th International Conference on Innovations in Information Technology, 2018.

[2] K. Lang, NewsWeeder: Learning to filter netnews, Mach. Learn. Proc. 1995 (1995) 331–339.

[3] G.D. Corso, A. Gullí, F. Romani, Ranking a stream of news, in: International Conference on World Wide Web, DBLP, 2005, p. 97.

[4] J. Ni, J. Li, J. Mcauley, Justifying recommendations using distantly-labeled reviews and fine-grained aspects, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP, 2019.

[5] S. Li, S.Y.M. Lee, W. Gao, C.R. Huang, Semi-supervised text categorization by considering sufficiency and diversity, in: G. Zhou, J. Li, D. Zhao, Y. Feng (Eds.), Natural Language Processing and Chinese Computing, NLPCC 2013, in: Communications in Computer and Information Science, vol. 400, Springer, Berlin, Heidelberg, 2013.

[6] R. Socher, A. Perelygin, J.Y. Wu, et al., Recursive deep models for semantic compositionality over a sentiment treebank, in: Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2013.

[7] https://www.yelp.com/dataset.

[8] http://labelme.csail.mit.edu/Release3.0/.

[9] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, Int. J. Comput. Vis. 88 (2) (2010) 303–338.

[10] J. Deng, W. Dong, R. Socher, et al., ImageNet : A large-scale hierarchical image database, in: Proc. CVPR, Vol. 2009, 2009.

[11] http://vision.stanford.edu/aditya86/ImageNetDogs/main.html.

[12] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba, Places: A 10 million image database for scene recognition, IEEE Trans. Pattern Anal. Mach. Intell. 40 (6) (2018) 1452–1464, http://dx.doi.org/10.1109/TPAMI.2017.2723009.

[13] A. Krizhevsky, G. Hinton, Learning multiple layers of features from tiny images, in: Handbook of Systemic Autoimmune Diseases, Vol. 1, no. 4, 2009.

[14] T.Y. Lin, M. Maire, S. Belongie, et al., Microsoft COCO: Common Objects in Context, in: European Conference on Computer Vision, 2014.

[15] S.A. Nene, S.K. Nayar, H. Murase, Columbia Object Image Library (COIL-100), Columbia University, 1996.

[16] Z. Liu, P. Luo, X. Wang, et al., Deep learning face attributes in the wild, 2014, arXiv e-prints.

[17] V. Panayotov, G. Chen, D. Povey, et al., Librispeech: An ASR corpus based on public domain audio books, in: ICASSP 2015-2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2015.

[18] Linguistic data consortium, 2002, 2000 HUB5 English Evaluation Transcripts LDC2002T43. Web Download. Philadelphia: Linguistic Data Consortium.

[19] http://www.voxforge.org/.

[20] John S. Garofolo, et al., TIMIT acoustic-phonetic continuous speech corpus LDC93s1, 1993, Web Download. Philadelphia: Linguistic Data Consortium.

[21] https://chimechallenge.github.io/chime6/download.html.

[22] D. Snyder, G. Chen, D. Povey, MUSAN: A music, speech, and noise corpus, Comput. Sci. (2015).

[23] https://zenodo.org/record/6337421#.ZGQ6dk9ByuJ.

[24] N. Takahashi, M. Gygli, B. Pfister, et al., Deep convolutional neural networks and data augmentation for acoustic event recognition, Interspeech (2016).

[25] A.L. Maas, R.E. Daly, P.T. Pham, et al., Learning Word Vectors for Sentiment Analysis, Association for Computational Linguistics, 2011.

[26] M. Abdallah, Big data quality challenges, in: 2019 International Conference on Big Data and Computational Intelligence, ICBDCI, 2019.

[27] J. Song, C. Shuang, D. Guo, et al., Data quality and data cleaning methods, 2013, Command Information System and Technology.

[28] C.G. Northcutt, L. Jiang, I.L. Chuang, Confident learning: Estimating uncertainty in dataset labels, 2021.

[29] M.M. Rosli, E. Tempero, A. Luxton-Reilly, Evaluating the quality of datasets in software engineering, J. Comput. Theor. Nanosci. 24 (10) (2018) 7232–7239.

[30] B. Christian, G. Theresia, K. Simone, et al., Visual interactive creation, customization, and analysis of data quality metrics, J. Data Inf. Qual. 10 (1) (2018) 1–26.

[31] S. Fabbrizzi, S. Papadopoulos, E. Ntoutsi, et al., A survey on bias in visual datasets, 2021.

[32] G. Guo, M. Jazaery, Automated cleaning of identity label noise in a large face dataset with quality control, IET Biometrics 9 (1) (2019).

[33] G. Xie, L. Guo, J. Gao, et al., Conceptual cognitive modeling for fine-grained annotation quality assessment of object detection datasets, Discrete Dyn. Nat. Soc. 2020 (2020).

[34] I. Taleb, M.A. Serhani, R. Dssouli, Big data quality: A survey, in: Big Data Congress 2018, 2018.

[35] I. Taleb, M.A. Serhani, C. Bouhaddioui, R. Dssouli, Big data quality framework: A holistic approach to continuous quality management, J. Big Data 8 (1) (2021) 1–41, 2021.

[36] Y. Li, H. Song, Y. Xu, Studies on data quality evaluation index system for internet plus government services in big data era, 2020, 012014.

[37] Construction of big data quality measurement model, in: Information Studies:Theory and Application, 2018.

[38] D. Diaz, G. Bavota, A. Marcus, et al., Using code ownership to improve IR-based traceability link recovery, program comprehension (ICPC), in: 2013 IEEE 21st International Conference on, IEEE, 2013.

[39] V. Gervasi, D. Zowghi, Supporting traceability through affinity mining, in: Requirements Engineering Conference, IEEE, 2014.

[40] W. Zogaan, P. Sharma, M. Mirahkorli, et al., Datasets from fifteen years of automated requirements traceability research: Current state, characteristics, and quality, in: Requirements Engineering Conference, IEEE, 2017.

[41] M. Mirakhorli, J. Cleland-Huang, Detecting, tracing, and monitoring architectural tactics in code, IEEE Trans Softw Eng 42 (3) (2016) 1.

[42] X. Zhang, X. Zhu, S.J. Wright, Training set debugging using trusted items, 2018.

[43] E. Ruckhaus, M. Vidal, S. Castillo, et al., Analyzing linked data quality with LiQuate, in: Proc. of the European Semantic Web Conf., 2014, pp. 488–493.

[44] N. Ruiz, M. Federico, Phonetically-oriented word error alignment for speech recognition error analysis in speech translation, in: Proc. of the Automatic Speech Recognition and Understanding, 2016, pp. 296–302.

[45] J.P. Escudero, J. Novoa, R. Mahu, et al., An improved DNN-based spectral feature mapping that removes noise and reverberation for robust automatic speech recognition, 2018, arXiv:1803.09016.

[46] C. Lin, ROUGE:A package for automatic evaluation of summaries, in: Proc. of the Meeting of the Association for Computational Linguistics, 2004, pp. 74–81.

[47] N. Japkowicz, Concept-learning in the presence of between-class and within-class imbalances, in: Proceedings of the 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence, 2001, pp. 67–77.

[48] H. Yulin, J. Yi, D. Dexin, H. Baihao, H. Jiajie, A new method for measuring the distribution consistency of mixed-attribute datasets, J. Shenzhen Univ. (Sci. Technol. Ed.) 38 (02) (2021) 170–179.

[49] L. Cai, S.T. Wang, J.H. Liu, Y.Y. Zhu, Survey of data annotation, J. Softw. 31 (2) (2020) 302–320.

[50] GB/T 36344-2018 Information technology—Evaluation indicators for data quality.

[51] S. Picard, C. Chapdelaine, C. Cappi, et al., Ensuring Dataset Quality for Machine Learning Certification, IEEE, 2020.

[52] S. Chug, P. Kaushal, P. Kumaraguru, et al., Statistical learning to operationalize a domain agnostic data quality scoring, 2021.

[53] T. Hongxun, W. Honggang, Z. Kun, et al., Data quality assessment for on-line monitoring and measuring system of power quality based on big data and data provenance theory, 2018, pp. 248–252.

[54] M. Priestley, F. O'Donnell, E. Simperl, A survey of data quality requirements that matter in ML development pipelines, J. Data Inf. Qual. (2023) http://dx.doi.org/10.1145/3592616.

[55] Y. Nehmé, J. Delanoy, F. Dupont, J.P. Farrugia, P.L. Callet, G. Lavoué, Textured mesh quality assessment: Large-scale dataset and deep learning-based quality metric, ACM Trans. Graph. (2023) http://dx.doi.org/10.1145/3592786.

[56] W. Chang, ISO/IEC JTC 1/SC 42(AI)/WG 2(data) data quality for analytics and machine learning (ML), 2022, https://jtc1info.org/wp-content/uploads/2022/06/01_06_Wo_2022_05_24_ISO-IEC-JTC1-SC42-WG2-Data-Quality-for-Analytics-and-Machine-Learning-Wo-Chang-NIST-final.pdf (Online; Accessed 01/12/2022).

[57] M. Wook, N.A. Hasbullah, N.M. Zainudin, et al., Exploring big data traits and data quality dimensions for big data analytics application using partial least squares structural equation modelling, J. Big Data 8 (1) (2021) 1–15.

[58] T. Scantamburlo, Non-empirical problems in fair machine learning, Ethics Inf. Technol. 23 (4) (2021) 703–712, http://dx.doi.org/10.1007/s10676-021-09608-9.

[59] S. Hooker, Moving beyond algorithmic bias is a data problem, Patterns 2 (4) (2021).

[60] J. Shi, J. Zhang, Y. Ge, An association-based intrinsic quality index for health-care dataset ranking, in: 2019 IEEE International Conference on Healthcare Informatics, ICHI, 2019, pp. 1–8, http://dx.doi.org/10.1109/ICHI.2019.8904553.

[61] D.T.A. Luong, P. Singh, M. Ramezani, et al., longSil: An evaluation metric to assess quality of clustering longitudinal clinical data, J. Healthc. Inf. Res. 3 (1) (2019) 1–19.

[62] L. Li, T.T. Goh, D. Jin, How textual quality of online reviews affect classification performance:A case of deep learning sentiment analysis, Neural Comput. Appl. (2018).

[63] B. Heinrich, M. Klier, A. Schiller, Wagner G., Assessing data quality–A probability-based metric for semantic consistency, Decis. Support Syst. 110 (2018) 95–106, (2018).

[64] D. Ardagna, C. Cappiello, W. Samá, et al., Context-aware data quality assessment for big data, Future Gener. Comput. Syst. 89 (DEC.) (2018) 548–562.

[65] N.W. Zog Aa, P. Sharma, M. Mirahkorli, et al., Datasets used in fifteen years of automated requirements traceability research, 2017.

[66] L.I. Mohan, L.I. Jianzhong, Data currency determination: Key theories and technologies, Intell. Comput. Appl. (2016).

[67] X. He, X. Yang, S. Zhang, et al., Sample-efficient deep learning for COVID-19 diagnosis based on CT scans, 2020.

[68] V. Birodkar, H. Mobahi, S. Bengio, Semantic redundancies in image-classification datasets: The 10% you don't need, 2019.

[69] J.L. Jin, Y.M. Wei, J Ding, Fuzzy comprehensive evaluation model based on improved analytic hierarchy process, J. Hydraul. Eng. (2) (2004) 144–147.

[70] J. Priem, D. Taraborelli, P. Groth, et al., Altmetrics: A manifesto. [2010-10-26]. http://altmetrics.org/manifesto.

[71] T. Ju, chun, W.U. Jian, et al., New study on determining the weight of index in synthetic weighted mark method, Syst. Eng.-Theory Pract. 21 (8) (2001) 43–48.