International Journal of Gender, Science and Technology
ISSN: 2040-0748

UGC Care Group I Journal.
Vol-13 Issue-02 Dec 2024

A Survey of Performance Enhancement on MLModel through Data Quality

Jagatjeeta Mohanty[#], Sushama Rani Dutta[*]

AIML Department
St.Peter's Engg College,Hyderabad
#jagatjeeta.mohanty@gmail.com
*sushamadutta@gmail.com

Abstract— The fitness of the systems which are using Machine Learning (ML) model depends on the quality of the data. Specifications of a good-quality data set have traditionally been defined by the needs of the data users—typically analysts and engineers. Using a review of recent literature, at the intersection of ML, data management and human-computer intervention , data management plays a major role. The data quality should be according to the need of the stakeholder, software developer or organization. We therefore propose a new treatment of data quality by structuring them into two dimensions (1) the stage of ML life cycle where use cases occur (2)main category of data quality that can be pursued(intrinsic, contextual, representational and accessibility). To illustrate this works, we contribute a temporal mapping of the various data quality requirements that are important at different stages of the ML data pipeline. We also share some implications for data practitioners and organizations that wish to enhance their data management routines in preparation for ML model.

Keywords— **Data Quality, Machine Learning,Dimension of data,ML pipe line**

I. INTRODUCTION

In today's world, voluminous amount of data is being produced that gets used by Artificial Intelligence(AI) systems. In the field of ML, which relies on the use of data to classify or detect patterns in existing information as well as using past data to train algorithms to solve new task. Our article focuses on the latter subset of ML, which is growing in popularity in systems to predict probable outcomes based on certain inputs, or to make recommendations about which decisions would be optimal in a given scenario. These systems work with structured as well as unstructured data (e.g., text, images,audio) to address practical use cases in fields such as clinical diagnosis, criminal justice, financial lending, manufacture and autonomous vehicles, among others. In the remainder of this article, we will refer to ML systems as software systems in which ML models or algorithms are deployed, typically for the purposes of solving a problem in the real world. Poor- quality datasets and data science pipelines can compromise ML systems in several ways. Messy or inaccurate data can also disturb the operational efficiency of businesses, with estimates of 10% to 30% of revenue being spent on resolving data quality issues [1]. The importance of data quality is therefore increasingly being recognised by private and public stakeholders who want to mitigate social risks, reduce costs and support the effective assimilation of ML technologies in society. Our article aims to help data practitioners to navigate these challenges by distilling some of the key concepts from recent literature in the fields of ML, data management and Human-Computer Interaction (HCI). Our contributions include the following:

An overview of some of the key data quality requirements that matter in ML systems.

An illustration of how these requirements map onto traditional data quality criteria.

A structure for identifying the most salient data quality requirements depending on the stage of the ML lifecycle where the data use case occurs.

II. RELATED WORK

Training data for ML algorithms can be collected in a variety of ways. In their comprehensive survey of data collection methods for ML, Roh et al. [2] group these into three categories: (1) data acquisition (including discovery, augmentation and generation), (2) data labelling (using manual or semi-supervised approaches) and (3) improvement (cleaning the data itself or improving the model built upon it). The extent to which these data collection methods are used varies depending on the use case and the type of data upon which an ML system relies. In larger organisations and complex innovation ecosystems, the data may pass through multiple stakeholders and can be transformed in various ways before it reaches an ML practitioner or their resultant product. Because of this, the topic of data quality is beginning to transcend beyond the field of data management, which accommodates holistic considerations such as how people search for relevant datasets [3], how developers perceive data work [4] and the best ways of using crowdsourcing to generate, evaluate or label data [5]. Although the role of these dynamic processes and multi-stakeholder configurations is increasingly being recognised by data practitioners, it is less clear how traditional data quality frameworks and notions of data accountability are adapting to ML development pipelines [6].

2.1 Data Quality Means Meeting the Needs of Different Users

Traditionally, Data quality compliance has meant meeting the needs of immediate data user (analyst or engineers) who value

clean machine readable data. This singular focus can flatten the variety of uses and data quality requirements that are facedat various stages of ML development pipeline. For example, data quality aspects that are important to ML developers are likely to be different from software developers and organisation. Software developers may have their own preferences for specific data quality aspects like security, provenance, legal compliance and capacity to meet business goal in real world. Data quality specification include list of 60 dimensions created by Black and Van Nederpelt [7] includes qualities related to data accuracy, coverage, legal compliance and usability.

2.2 Four dimensions of Data Quality

2.2.1 Intrinsic data quality has traditionally been understood to reflect the extent to which data values conform to the actual or true values [8]; this includes specific requirements such as accuracy, provenance and cleanliness, the latter of which covers practices such as the addressing missing values and redundant cases. Besides the usual data qualities needed for statistical analysis (e.g., addressing missing data, anomalies), an intrinsic quality that is increasingly valued by ML practitioners and regulators relates to data lineage and traceability. For data that require multiple pre-processing steps or transactions between organisations, the origins of their features becomes important. Traceability makes it possible to interpret and audit the history that precedes the output of ML algorithms [9], but despite recent regulations on Explainable AI (XAI), traceability is not yet shortlisted in the data quality framework used by the UKgovernment suggesting that this data quality characteristic may need to be promoted in the context of ML.

2.2.2 Contextual data quality

It relates to the extent to which data are pertinent to the task of data user. It includes dimensions such as relevance, timeliness, completeness and appropriateness. An essential questions that is considered here is the extent to which the sample of cases contained in the dataset diverges from true distribution of cases. When ML model is deployed. Possible sources divergence includes historical time or geographical representation. For example temporality has been flagged as a potential source of difficulty in textual data, where models trained on historical text corpora such as Google News have produced past social stereotypes like for example, the word 'man' being associated with computer programmer and 'woman' with homemaker. If this biases are further allowed then it will amplify the bias and continue in society. Other contextual biases has been detected in image data with publicly available image corpora such as ImageNet and Open Images that comes from Eurocentric contexts. Insufficient representation of some geographic reasons such as Asia & Africa has resulted in less information learned by ML algorithms. This results in solutions that performs poorly for under-represented groups. Example electronic soap dispenser does not respond to darker skin. These cases urge practitioners to think critically about the contexts captured by their dataset and extent to which it reflects the use cases and experiences of end users.

2.2.3 Representational Data Quality

It refers to the degree to which data are representedshould be clear which should include aspects like being interpretable and easy to comprehend. These aspects can be implemented by practices like standardisation and documentation. Standardisation refers to method for capturinginformation in a consistent manner including data structures and formats for capturing specific attributes (eg. Date, location, measurement error). This helps engineers to understand how dataset is correlated to physical world so that the training data or model output can be transformedaccordingly. When the limitation of dataset is explicitly mentioned in the document, the users take measures to improve the quality of dataset for their specific use cases. Some solution even allows for dataset to remain same while ML algorithm are tuned to produce more robust or socially equitable results.

2.2.4 Accessibility

It refers to the degree to which data are available and secure. The advancement of big data & ML application has allowed publishing of dataset in an open manner. But there are some restricted dataset which can be accessed by some secure access mechanism, so that their value can be realised . For ML stakeholder who work with commercially sensitive data face issues related to security and take legal precautions

2.3 Why Knowledge of Desirable Data Quality Practices is important

In ML, finding out required data quality is a complex task. Organisation and practitioners have to specify which data quality attributes are required for the use cases and howto define them. In a study of organisation that was applied to ISO/IEC 25012 data quality standard, Gualo et. al[10] found that researchers find it difficult to identify and describe the data quality rules that are applied to their use cases. Thepractitioners found that providing examples of what the requirement can look like helps to guide them in clarifying their own rule. Another challenge is that the list of requirement is very lengthy, so there is information overload. Therefore, practitioners cannot apply traditional standards. The above two challenges occur at the beginning stage of planning and it is a complex task to define which data qualities to evaluate. Without planning, it becomes difficult to develop the right quality rules and right tools to enforce them.

2.4 Data Quality Planning precedes Implementation

Our aim is to support ML practioners and data managers at the planning stage of their data quality speculation. By focusing on the requirements that thrives, practitioners are in advantageous position to select the appropriate andmeaningful data quality control, assurance and improvement steps for their use cases. Our aim in this article is to inform thepractioners of data quality need and practices that exist that

are meaningful in the field of ML. This will be done by using recent academic literature and grasping the suggestion according to the aspect of data quality that are present in the field of data management. Secondly to help readers in selecting a smaller set of data quality practices that may be applied to uses cases. In doing so, it becomes easier for organisation and individuals to prepare data management routine for ML.

3 METHODOLOGY

Our literature review was conducted using a systematic mapping protocol [54] to select a small set of relevant articles from the much larger collection of literature emerging at the intersection of data quality and ML. In the following, wepresent the research questions, inclusion criteria and search strategy that were used to select articles for review.

3.1 Research Questions

Our review aimed to identify and discuss the data quality requirements that are important to ML development, and how they differ from more established data management practices. For this purpose, we defined the following research questions:

- Where do the data quality requirements of ML sit in relation to traditional data quality frame- works from data and information management?

•Does ML present any new challenges that are not yet accommodated by traditional data quality frameworks?

The preceding questions deal with data quality management planning as opposed to implemen- tation. This is a distinction that has previously been recognised in industry standards such as ISO 8000-61, as depicted in Figure 1.

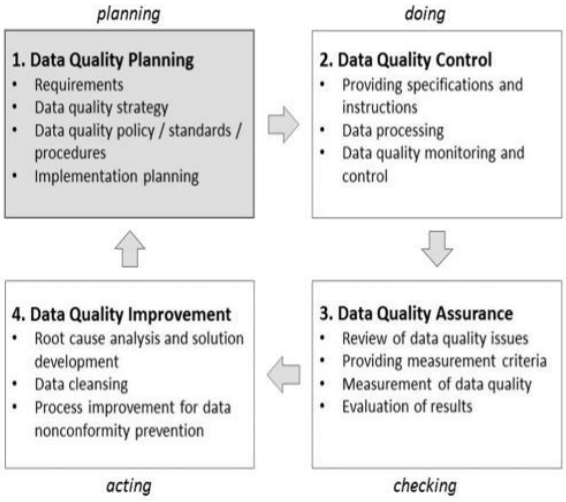

Fig. 1. Data quality management process

3.2 Selection Criteria

Our interest in data quality planning (as distinct from implementation) helped to limit the scope of our literature review and make the topic small enough to be discussed in a single paper. Specifically, our targeted articles dealt with philosophical or experiential perspectives on data quality frameworks, as opposed to articles that evaluated specific data management techniques or pro- posed new solutions for managing data quality. Our inclusion criteria were as follows:

•The abstract of the paper must discuss conceptual frameworks for defining data quality re- quirements in relationto ML, or experiences of how these requirements have been defined in practice.

•The paper was published between 2015 and 2022, to provide a contemporary overview.

•The paper is peer reviewed and published in a journal, conference or workshop.

•The paper may come in the form of a full-length article, extended abstract or workshop description.

Our exclusion criteria were as follows:

•The abstract of the paper focuses only on techniques for data quality processing, assurance or improvement rather than conceptual frameworks for defining the data quality requirements.

•The abstract of the paper only considers the data quality requirements of a specific industry that uses ML (e.g., healthcare, finance, materials science).

•The paper does not contain information about the publisher.

•The paper is an early iteration of a later work (e.g., if a similar workshop was delivered by the same authors multiple times, we selected only the latest version).

3.3 Search Strategy

Our literature search strategy consisted of three stages: (1) pre-selected articles that were already known to us, (2) automatic search on Google Scholar and selected conference proceedings, and

(3) forward and backward snowballing to identify further articles.

4.RESULTS

[22] proposed a sequence of nine stages that constitute the task of knowledge discovery in datasets.8 The authors suggested that the process typically begins with developing an understanding of the application domain and use case, followed by data collection, pre-processing and reduction, before moving on to identifying and applying relevant data mining methods, as well as interpreting and acting on their insights. Although the authors recognised that knowledge discovery workflows also include challenges related to data accessibility, HCI and model scaling, their pipeline focused onthe granular steps contained within data mining. A similar focus on data is adopted by the upcoming industry standard ISO/IEC 5259, whose provisional data processing framework is illustrated in the upper part of Figure 3 [11].Recent academic discussions of the ML pipeline have been moredetailed in separating out the different stages undergone by ML data. Specifically, they explore model development, verification, deployment and monitoring, which pose different requirements in terms of organisational and operational considerations [12, 13].

For the purposes of this article, we organise our findings into a series of stages listed in the first column of Table 3.

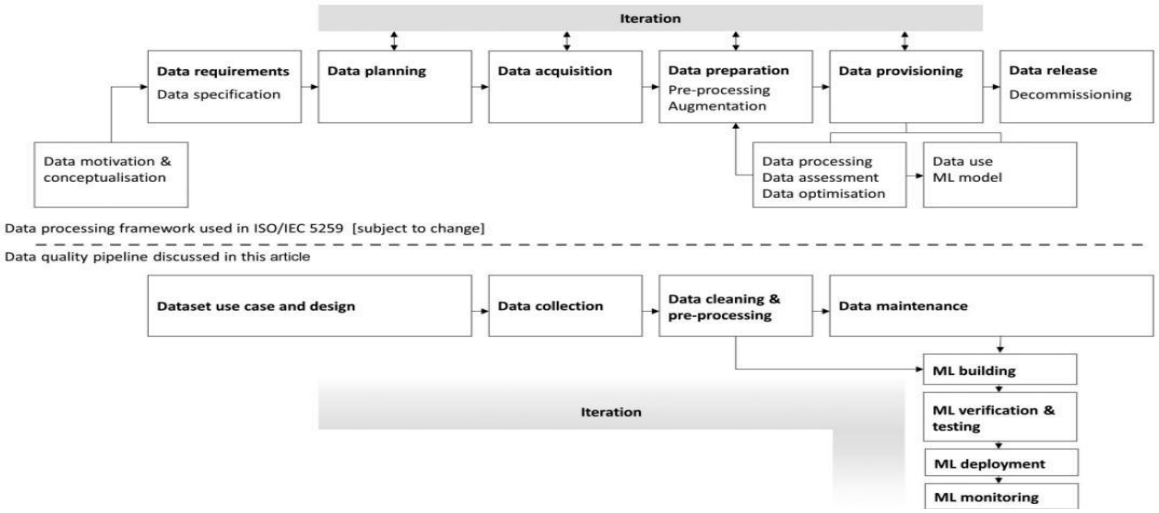

Fig. 3. An illustration of how our data quality pipeline (lower) maps to the data processing framework used in ISO/IEC 5259 (upper part). Diagram adapted from Chang [18]

Table 3. ML Data Quality Considerations Classified According to Different Categories of Quality (Horizontal) and Stages of the ML Development Pipeline (Vertical)

Development Stage	Data Quality Category			
	Intrinsic	Contextual	Representational	Accessibility
Dataset use case and design	Accuracy of data can be supported by hiring human annotators and field experts in advance [49, 52, 53].	Relevance of data can be ensured by determining what features are required in advance [9, 36, 53].	Clarity and credibility of the metadata can be improved by including documentation on user requirements and dataset design [35].	Availability of data can be supported by infrastructure for data collection and management (particularly in large organisations) [25, 52, 57]. Validity of data for online learning can be assured by putting in place runtime verification tools [21, 50].
Data collection	Accuracy can be improved by: · Human-in-the-loop approaches for data labelling and augmentation [49, 73]. · Data collection tools that raise actionable alerts to warn users of unexpected values in advance [38, 57]. · Screening and training of data workers [49, 70, 73].	Context coverage can be supported by institutional guidelines on potential power imbalances, ethics and inclusivity [9, 36, 62, 70].	Clarity of the metadata can be supported by documenting the data collection process (e.g., using datasheets, checklists) [9, 23, 35, 48, 58]. Consistency of data can be improved using standardisation [25, 33].	Regulatory compliance can be supported by institutional frameworks and procedures for consent, transparency, ethics and privacy [9, 36, 70].

	Intrinsic	Contextual	Representational	Accessibility
Data cleaning and pre-processing	Uniqueness of data entries and features can be improved by removing redundant cases and reducing the complexity of the features [5, 38, 57]. Completeness can be supported by automated pre-processing and ML aids for augmentation and annotation [5, 27, 38, 70].	Contextual bias can be detected using ground-truth correlations [32, 52, 53]. Contextual validity can be improved by balancing the classes and measuring how well the dataset fits the real-world problem [3, 5, 8, 9, 25, 38, 70].	Clarity of the data pre-processing sequence can be improved using documentation and publication of code [52, 72]. Consistency of data sourced from heterogeneous sources can be supported by reformatting standards, normalising and aggregation [38, 70]. Precision can be improved by using representational standards that allow for uncertainty [3, 5].	Security of sensitive data supported by anonymisation [33, 70].
Data maintenance		Contextually biased data can be improved using curation, including infrastructure, tools and practices for maintaining nonstatic datasets that grow over time [3].	Maintainability at scale is supported by standards [3]. Clarity of the dataset can be supported by user interfaces for dataset exploration [25, 32, 33, 52, 57]. Clarity of the metadata can be supported by documentation on: · Data content (e.g., nutrition labels) [25, 32]. · Maintenance plan [36]. · Mission statement [36].	Availability of data can be facilitated by infrastructure for differential access and sharing (e.g., via data trusts) [32, 35]. Identifiability of the correct dataset (out of multiple versions) can be guided by version control and DOIs [25, 32, 35, 57].

6 CONCLUSION

Shifting data practices from current priorities driven by availability or convenience towards high- quality data will require the effort of decision makers and practitioners at every level of organi- sations and policy. It is our hope to have contributed a useful vocabulary for perceiving and ar- ticulating some of the nuanced data quality requirements that can be resolved by practitioners in different parts of the ML pipeline

REFERENCES

[1] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on explainable artificial intelli- gence (XAI). IEEE Access 6 (2018), 52138–52160.

[2] Ariful Islam Anik and Andrea Bunt. 2021. Data-centric explanations: Explaining training data of machine learning systems to promote transparency. In Proceedings of the 2021 CHI Conference on HumanFactors in Computing Systems. 1–13.

[3] Lora Aroyo, Matthew Lease, Praveen Paritosh, and Mike Schaekermann. 2022. Data excellence for AI: Why should you care? Interactions 29, 2 (2022), 66–69.

[4] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, et al. 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and chal- lenges toward responsible AI. Information Fusion 58 (2020), 82– 115.

[5] Rob Ashmore, Radu Calinescu, and Colin Paterson. 2021. Assuring the machine learning lifecycle: Desiderata, meth- ods, and challenges. ACM Computing Surveys 54, 5 (2021), 1–39.

[6] Jacqui Ayling and Adriane Chapman. 2021. Putting AI ethics to work: Are the tools fit for purpose? AI and Ethics
2 (2021), 405–429.

[7] Yang Baolong, Wu Hong, and Zhang Haodong. 2018. Research and application of data management based on Data Management Maturity Model (DMM). In Proceedings of the 2018 10th International Conference on Machine Learning and Computing. 157–160.

[8] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. IBM Journal of Research and Development 63, 4-5 (2019), Article 4, 15 pages.

[9]Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigat- ing system bias and enabling better science. Transactions of the Association for Computational Linguistics 6 (2018), 587–604.

[10] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM Conference on Fairness, Account- ability, and Transparency. 610–623.

[11] Laure Berti-Equille. 2019. Learn2Clean: Optimizing the sequence of tasks for web data preparation. In Proceedings of the World Wide Web Conference. 2580–2586.

[12] Leopoldo Bertossi and Floris Geerts. 2020. Data quality and explainable AI. Journal of Data and Information Quality
12, 2 (2020), 1–9.

[13] A. Black and P. van Nederpelt. 2020. Dimensions of Data Quality (DDQ) Research Paper. Retrieved June 10, 2021 from http://www.dama-nl.org/wp-content/uploads/2020/09/DDQ-Dimensions-of-Data-Quality-Research-Paper- version-1.2-d.d.-3-Sept-2020.pdf.