DQM: Data Quality Metrics for AI Components in the Industry

Sabrina Chaouche[1*], Yoann Randon[1*], Faouzi Adjed[1*], Nadira Boudjani[1,2], Mohamed Ibn Khedher[1],

[1]IRT SystemX, 2 Boulevard Thomas Gobert 91120 PALAISEAU, France
[2]Valeo Brain Division, 6 Rue Daniel Costantini, 94000 CRETEIL, France
{sabrina.chaouche, yoann.randon, faouzi.adjed, mohamed.ibn-khedher}@irt-systemx.fr, nadira.boudjani@valeo.com

Abstract

In industrial settings, measuring the quality of data used to represent an intended domain of use and its operating conditions is crucial and challenging. Thus, this paper aims to present a set of metrics addressing this data quality issue in the form of a library, named DQM (Data Quality Metrics), for Machine Learning (ML) use. Additional metrics specific to industrial application are developed in the proposed library. This work aims also to assess various data and datasets types. Those metrics are used to characterize the training and evaluating datasets involved in the process of building ML models for industrial use cases. Two categories of metrics are implemented in DQM: *inherent data metrics*, are the ones evaluating the quality of a given dataset independently from the ML model such as statistical proprieties and attributes, and *model dependent metrics* which are those implemented to measure the quality of the dataset by considering the ML model outputs such the gap between two datasets in regards to a given ML model. DQM is used in the scope of the Confiance.ai program to evaluate datasets used for industrial purposes such as autonomous driving.

Introduction

The current paradigm in Machine Learning (ML) has been largely model-centric, where the research contributions are focused on enhancing models performances (Mazumder et al. 2024), whereas the learning of ML model is largely based on the information contained in the data. Recently, the Data Centric for Artificial Intelligence (DCAI) has emerged as a concept (Zha et al. 2023) to better master the ML model decisions. Furthermore, as reported by Mazumder et al., many ML industrialization difficulties and drops in performance often do not result from the model itself but from the data used to train it. Hence, taking into account the DCAI method would improve the AI life cycle process (Hutchinson et al. 2021; Polyzotis et al. 2018). The DCAI concept is also explored in the standardization and certification processes to be integrated in data management (Picard et al. 2020).

As stated before, the core of the data-centric approach is to select the best data from the broad pole of the available data. The evaluation of the quality of this selection have to be assessed and quantified following a set of specific metrics (Aroyo et al. 2022). In this work, we focused on the proposition and the implementation of a set of generic metrics to evaluate data quality for AI usage in the specific industrial domain. Our proposed work can be summarized into two main contributions:

- Proposition of industrial library assembling a set of relevant approaches for each selected data quality metric.
- Development of new metrics to integrate the industrial specification and requirements.

The rest of the paper is organized as follows: first, we review the state of the art related to data quality metrics. Then, give a description of the proposed approach, detailing the metrics for both categories, model-dependent and inherent data quality. Following this, we outline the implementation of these metrics, including the types of data already considered and the methods used. Next, we present the results of our experiments and provide a discussion on the findings. Finally, we conclude the paper and discuss the future perspectives, highlighting potential directions for further research and development.

Related Works

The multitude of data types (multivariate vectors, numerical, images, 3D clouds points, times series, etc) and the related requirements and objectives increase the number of data quality approaches. In addition, some specification needs may challenge a specific knowledge and expertise related to the application domain. Therefore, the dataset and its quality become central during the whole ML process, from the conception and specification until deployment going through training and test sets design. Indeed, as reported by Zha et al. (Zha et al. 2023), the scientific contributions considerably increased during the five past years in DCAI as a data-engineering strategy that improves the performance of a given AI system (Kumar et al. 2024; Polyzotis and Zaharia 2021). The strategy can be focused on data quality boosting, data augmentation, extrapolation, etc.

Several DCAI automated tools and libraries are proposed in the literature. Dcbench (Eyuboglu et al. 2022) was proposed to evaluate Data-centric AI development. Another benchmark named DataPerf was proposed by Mazumder

[*]These authors contributed equally.

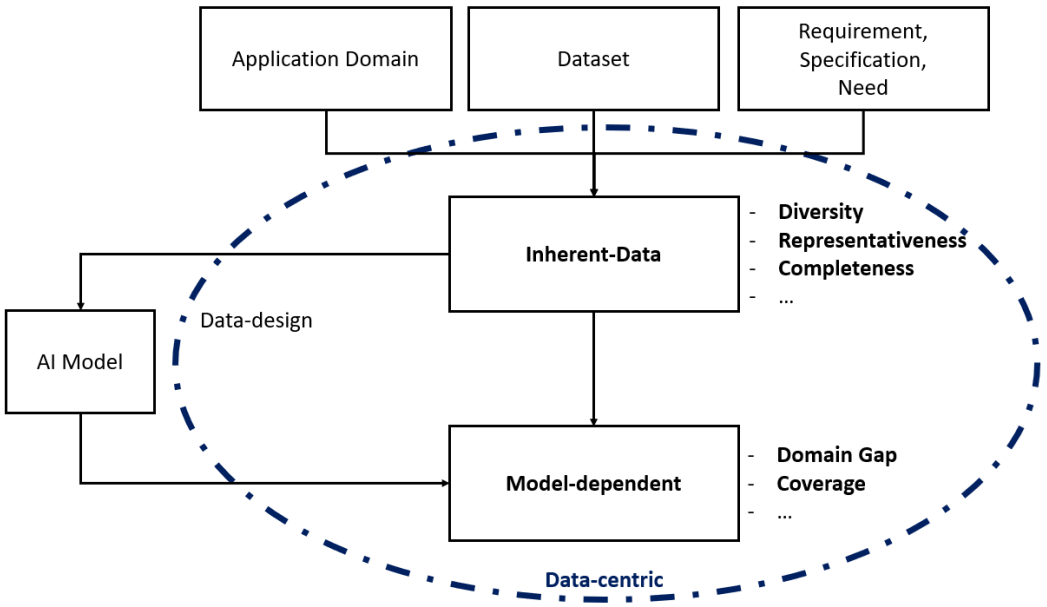

Figure 1: Data-centric approach view

et al. (Mazumder et al. 2024) to evaluate the impact of datasets. An other implementation of DCAI was proposed by Luley et al. (Luley et al. 2023). This implementation integrates the industrial constraints such as small datasets and specific context by adding expertise knowledge. By focusing on image data, Kastryulin et al. (Kastryulin et al. 2022) proposed a python library, under PIQ name, for image quality assessment.

Despite the approaches and libraries developed in the literature, the industrial requirements and specifications are rarely considered. In the current work, we propose a python library to qualify the quality of data used for ML purposes. Adjustments of these metrics is performed to link the data quality to the specifications and the requirements. An application of the developed methods is applied on industrial data use case.

Proposed Approach

In this section, we present our approach to assess the quality of datasets for ML use. This assessment methods are assembled in DQM (Data Quality Metrics), a Python library developed to quantify the quality of datasets used in ML systems in industrial settings. It aims to create a comprehensive framework able to assess the essential attributes of a dataset. The metrics implemented in our library are divided into two main categories: inherent data (data-dependent) metrics and model-dependent metrics as illustrated in Figure 1, which summarizes also the DCAI concept presented in the previous sections. In the following, we present each metric as defined by the Confiance.ai[1] program in its deliverable (Adjed et al. 2023), and describe the content of each category while detailing all its implemented methods.

Inherent-Data Metrics

Inherent-data or data-dependent metrics define the intrinsic properties and quality aspects of a dataset, from a statistical point of view, independent of any specific ML model or task; in other words, they assess the dataset's characteristics based solely on the data itself. Data-dependent metrics are measurements developed to quantify the quality of the dataset regarding the requirements and the specifications of the use case. Three metrics are developed in the proposed library which are 1) diversity, 2) representativeness and 3) completeness.

Diversity The diversity of a dataset is defined (in Confiance.ai program) as the assessment / verification of the presence of all required information to define the intended domain of use and its operating conditions. It is used to quantify to which extent the dataset fits the specifications described by the final user. The diversity is upheld once the presence of at least one occurrence of each requirement is verified. In other words, the existence of one data sample for each requirement is sufficient to qualify the dataset as diverse.

In the scope of this work, we implemented three methods to quantify the diversity of a given dataset. Two methods from the literature which are Simpson Index given by equation (1) and Gini-Simpson Index given by equation (2). These methods are mostly used in biology to quantify the diversity of a given environment. In addition to these two state-of-the-art

[1] A French community dedicated to the design and industrialisation of trustworthy critical systems based on artificial intelligence

www.confiance.ai

measures, we propose a new method, named Relative Diversity, in order to account for industrial requirements, given by equation (3).

$$\text{Simpson Index} = \frac{\sum_{i=1}^{n} n_i(n_i-1)}{N(N-1)} \tag{1}$$

where N is the total number of samples in the dataset and n_i the number of samples in each class. Simpson index ranges between 0 and 1; where a value of 1 indicates no diversity and 0, infinite diversity.

$$\text{Gini-Simpson Index} = 1 - \sum_{i=1}^{R} p_i^2 \tag{2}$$

where R is the number of types (classes) in the dataset, p_i is the proportion of each class in the dataset, with $p_i = n_i/N$ where n_i is the number of samples in each class and N is the total number of samples in the dataset.

$$\text{RD} = \sum_{i}^{n} \alpha_i d_i \tag{3}$$

where, d_i represents the diversity of the class i and α_i is a wight parameter with $\sum_i \alpha_i = 1$. The parameter α is set by default to $\alpha = \frac{1}{n}$, and can be adjusted to fit the requirements.

Representativeness Representativeness is crucial in determining how well the dataset reflects the population from which it is drawn. The Confiance.ai program defines the representativeness as the conformity of the distribution of the key characteristics of the dataset to a given specification (requirements, operating conditions, ...etc). The methods implemented to measure the representativeness are: the chi-squared (χ^2) and the Kolomogorov-Smirnov (KS) tests in addition to a new approach based on Entropy Information baptised Granular Relative Theoretical Entropy (GRTE).
Regarding the χ^2 test, it is a robust statistical method used for various purposes, for instance, to assess the goodness-of-fit of theoretical distributions, and to test the independence or homogeneity of variables. Additionally, several tests like Fisher and Student tests are derived from χ^2. In our work, we focused on the goodness-of-fit test, which is defined by equation (4)

$$\chi^2 = \sum_{j=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{4}$$

where n is the number of bins, O_i and E_i represent the observed and expected counts in the i-th bin, respectively.
The KS test is a non-parametric approach usually used to assess if the observed data corresponds to a specified theoretical distribution. This test is based on the comparison of observed and theoretical cumulative distributions. The measure of the KS test is defined by equation (5)

$$ks = \max\left(|F_t(x) - F_e(x)|\right) \tag{5}$$

where F_t and F_e are the theoretical and empirical cumulative distributions, respectively.
While entropy alone provides valuable insights into the randomness or the uncertainty within a dataset, it does not fully quantify the information required for representativeness. To address this, we propose a new method based on entropy, Granular Relative and Theoretical Entropy (GRTE). It compares the observed entropy with the expected entropy to provide a more comprehensive measure of data representativeness. GRTE is defined by equation (6)

$$GRTE = \exp\left(\frac{-2\,|H(\Pr(E)) - H(\Pr(O))|}{\alpha}\right) \tag{6}$$

where H represents the entropy, $\Pr(E)$ and $\Pr(O)$ denote the expected and observed values, respectively. The parameter $\alpha \geq 1$ adjusts the sensitivity of the measure; in our implementation, we used $\alpha = 1$.
GTRE values range from 0 (for non representative data) to 1 (for fully representative data). The granularity parameter is the number of bins which monitors the granularity of the required information in each bin.

Completeness In Confiance.ai program, the completeness metric is defined as the degree to which subject data associated with an entity has values for all expected attributes and related entity instances in a specific use case. It refers to the proportion of well-filled information in a given dataset (Juddoo and George 2020). Furthermore, completeness does not measure only the missing information, but it also encompasses the data to be excluded (Chehreghan and Ali Abbaspour 2018). The method implemented to asses the completeness is the Completeness Ratio given by equation (7)

$$\text{Completeness Ratio} = \frac{\#\text{ Filled items}}{\#\text{ Total items}} \tag{7}$$

where $\#$ defines the number of elements.

Model-Dependent Metrics

Model-depend metrics consider the couple dataset-model to characterize the overall quality. In the following, we discuss the implemented methods to measure the domain gap between two datasets and the coverage of the space of the intended purpose by the elements generated by the pair dataset-model.

Domain Gap Domain gap is defined as the distance between two distributions P and Q in a given space. In the context of a computer vision task, the domain gap between two images datasets refers to the difference in semantic, textures and shapes between the two. A significant domain gap would lead to an important drop of the model's performance which makes its outputs unreliable for industrial applications. In the following we define six methods to measure the domain gap between two images datasets: Central Moment Discrepancy (CMD), which quantifies the difference between distributions by comparing their central moments; Kullback-Leibler divergence for Multivariate Normal distributions (KLMVN), Maximum Mean Discrepancy (MMD), Wasserstein distance, Proxy A Distance (PAD), and Frechet Inception Distance (FID). Each method measures the distance between two distributions $P(\mu_P, \Sigma_P)$ and $Q(\mu_Q, \Sigma_Q)$ given specific conditions that will be indicated in its definition. For each method, the closer to zero the result is, the closer the distributions P and Q are.

CMD is a distance between two probability distributions P and Q on a compact interval $[a,b]^N$. For $X \sim P$ and $Y \sim Q$, CMD is defined by (Zellinger et al. 2019) as in equation (8)

$$\begin{aligned} \mathrm{CMD}(P,Q) = & \frac{1}{|b-a|}\|\mathbb{E}(X)-\mathbb{E}(Y)\|_2 \\ & + \sum_{k=2}^{\infty} \frac{1}{|b-a|^k}\left\|c_k(X)-c_k(Y)\right\|_2 \end{aligned} \tag{8}$$

where $\mathbb{E}(X)$ is the expectation of X, and

$$c_k(X) = \left(\mathbb{E}\left(\prod_{i=1}^{N}\left(X_i - \mathbb{E}\left(X_i\right)\right)^{r_i}\right)\right)_{\substack{r_1+\ldots+r_N=k \\ r_1,\ldots,r_n\geq 0}}$$

is the central moment vector of order k, where $k \in N^*$. For $k=2$, CMD corresponds to variance, $k=3$ to skewness and $k=4$ to kurtosis of probability distributions.

Wasserstein distance is used to compare finite probability distributions. It is defined by (Rüschendorf 1985) as in equation (9)

$$W_p(P,Q) = \left(\inf_{\gamma\in\Gamma(P,Q)}\int_{X\times Y} d(x,y)^p d\gamma(x,y)\right)^{1/p} \tag{9}$$

where $d(x,y)$ is a distance function and $\Gamma(P,Q)$ is the set of all joint distributions whose marginals are P and Q.

MMD is a distance between two probability distributions P and Q defined by (Gretton et al. 2012) as the distance between their mean embeddings in a given Reproducing Kernel Hilbert Space (RKHS) given by equation (10).

$$\begin{aligned} \mathrm{MMD}^2\left(P,Q\right) &= \|\mu_P - \mu_Q\|^2_{\mathcal{H}} \\ &= \mathbb{E}_{x,x}[k\left(x,y\right)] - 2\mathbb{E}_{x,y}[k\left(x,y\right)] \\ &+ \mathbb{E}_{y,y}[k\left(y,y\right)] \end{aligned} \tag{10}$$

where k is a RKHS.

PAD is an empirical approximation of H-divergence. It is defined by (Ganin et al. 2016) and given by equation (11)

$$\mathrm{PAD}(P,Q) = 2(1-2\epsilon) \tag{11}$$

where ϵ is the binary classifier model's error rate which has been trained to distinguish between samples from P and Q.

FID computes a distance between two normal multivariate distributions. It is mainly used in image generation task to compare the quality of the generated data with the real data. According to (Heusel et al. 2017) the formula is defined by equation (12)

$$\begin{aligned} \mathrm{FID}(P,Q) = & \|\mu_P - \mu_Q\|_2^2 \\ & + \mathrm{Tr}\left(\Sigma_P + \Sigma_Q - 2\sqrt{\Sigma_P\Sigma_Q}\right) \end{aligned} \tag{12}$$

where Tr defines the trace.

KLMVN is a distance that quantifies the gap between two normal distributions with positive defined covariance matrices μ_P and μ_Q. It is defined by (Contreras-Reyes and Arellano-Valle 2012) and given by equation (13)

$$\begin{aligned} D_{KL}\left(P\|Q\right) = & \frac{1}{2}\,\mathrm{tr}\left(\Sigma_Q^{-1}\Sigma_P\right) \\ & + \frac{1}{2}\left(\left(\mu_Q-\mu_P\right)^\top \Sigma_Q^{-1}\left(\mu_Q-\mu_P\right) - k\right) \\ & + \frac{1}{2}\ln\left(\frac{\det\Sigma_Q}{\det\Sigma_P}\right) \end{aligned} \tag{13}$$

where k is the vector dimension from the data.

Coverage According to Confiance.ai prgram definition, the coverage of a couple "Dataset + ML Model" is the ability of the execution of the ML Model on this dataset to generate elements that match the expected space. The approaches of coverage are integrated from neural coverage developed by (Yuan, Pang, and Wang 2023) provided in their repository [2].

Implementation

DQM is implemented as a Python library to promote its adoption and usage in the ML community. Table 1 gives a comprehensive overview of the metrics with the data types supported for each one in the current version. Figure 2 provides a representation of the structure of DQM; as shown, the metrics are classified as follow: 1) diversity, 2) representativness and 3) completeness for inherent data metrics; 4) domain gap and 5) coverage for model-dependent metrics.
As shown in Table 1, the current version of DQM is meant to handle only image datasets for domain gap methods. They rely on Pytorch for image transformations (e.g. reshape, rotation, normalization, etc) and pre-built models to extract features from data. However, the user can freely add his customized transformations and models via a configuration file. We provide the user with notebooks explaining how to use each method and guidelines to handle data and models dependencies.

Results and Discussion

In this section, we discuss some of the experiments conducted using DQM library on selected datasets for autonomous driving tasks. We start by presenting the datasets and the experimental protocol followed for each method then share and discuss the results.

Datasets

Valeo Deep Perception (VDP) is a proprietary dataset by Valeo of fish-eye cameras images of urban scenes captured in 4 different cities: Paris, Nuremburg, Stuttgart and California. It contains 132k images with some variableness in the driving environment (urban, highway and parking), weather, lightning level, etc. This dataset is shared and used in Confiance.ai program as one of the industrial use cases.

[2] https://github.com/Yuanyuan-Yuan/NeuraL-Coverage

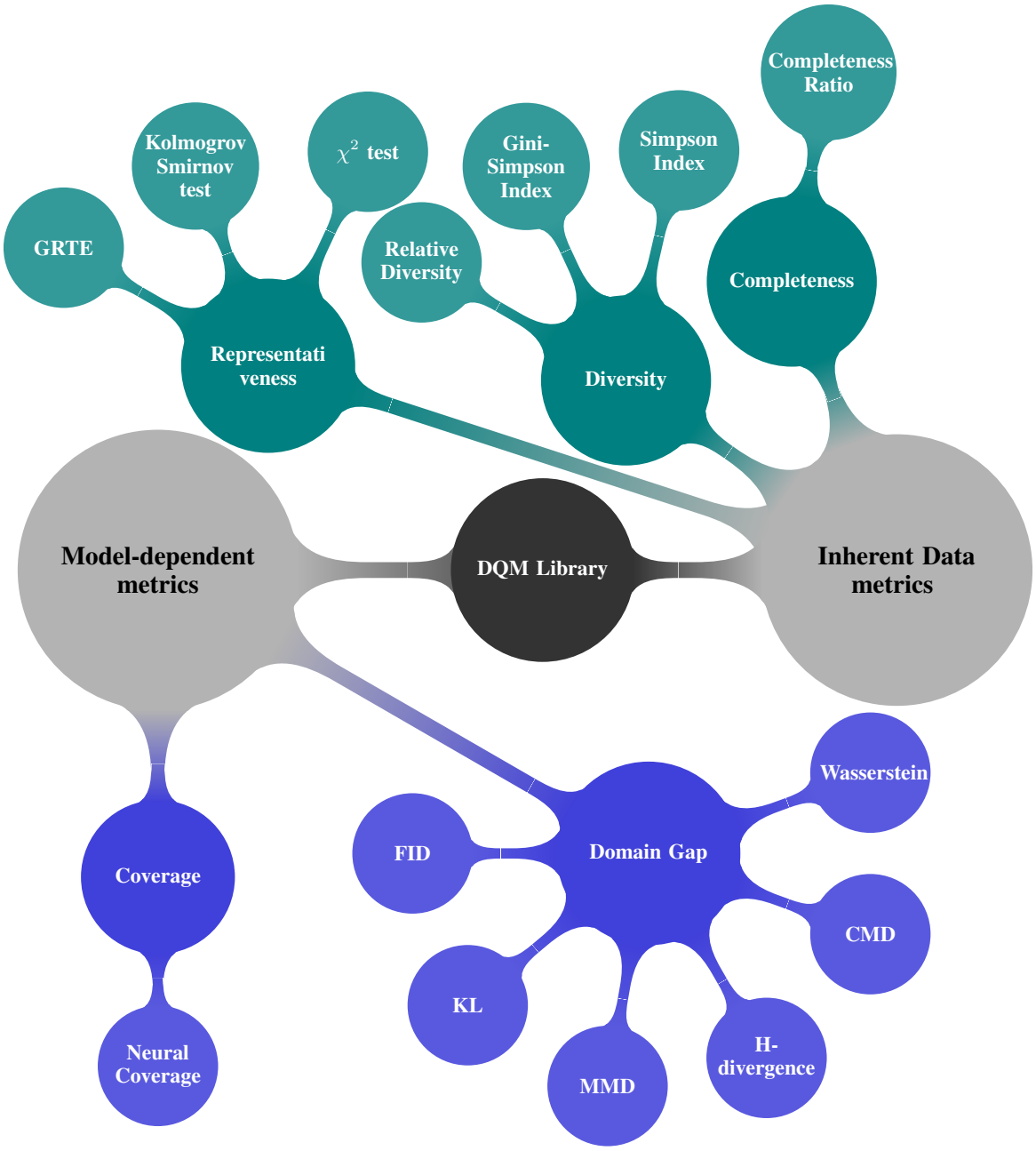

Figure 2: Mind Map of DQM Library

A preprocessing of the raw fish-eye images by a re-projection on a cylindrical viewpoint was made on the entire dataset to help increase the performance of convolutional neural networks based models. Valeo made a subset of the dataset available online under the name of Valeo WoodScape (Yogamani et al. 2019) for open source use.

BDD100K by (Yu et al. 2020) is a large-scale and diverse dataset that deals with the theme of autonomous driving tasks; containing 100k videos and more than 100k images with annotations such as weather condition, geographic location, bounding boxes for object detection, ...etc.

Synthetic Night Images is a small collection of 10k night images generated through a day-to-night style transfer using a JoliGEN model (JoliGEN 2024). The source images are selected from the daytime images in the BDD100K dataset. Figure 3 shows a sample of these images.

Experiments and Results

In the following, we share a part of our experiments on the above datasets to illustrate the potential of DQM for both categories, inherent and model-dependent metrics.

Metrics	Methods	Tabular	Images
Diversity	Simpson Index	✓	
	Gini-Simpson Index	✓	
	Relative Diversity	✓	
Representativeness	χ^2	✓	
	Kolmogorov-Smirnov Test	✓	
	GRTE	✓	
Completeness	Completeness Ratio	✓	
Domain Gap	Wasserstein		✓
	CMD		✓
	PAD		✓
	MMD		✓
	KLMVN		✓
	FID		✓

Table 1: Overview of data types supported for each metric in DQM.

Figure 3: Examples of night image generation using style transfer day to night.

Inherent-Data Metrics For inherent data metrics, we applied RD and GRTE methods to capture the quality of the VDP dataset from a statistical point of view.

Relative Diversity This experiment aims to evaluate how many of the expected modalities for the VDP dataset are present in the actual VDP dataset modalities. An example for sky cover modalities, is given in Table 2.

Required sky cover	VDP sky cover
sunny	cloudy
clear	obscured/invisible
cloudy	overcast
obscured/invisible	clear

Table 2: Example of required and present modalities for VDP's relative diversity assessment

The Relative Diversity method gives the following output:

- Categorical diversity = 0.6
- Same modalities = 3
- Missing modalities (requirements) = 1
- Additional modalities (in the dataset) = 1

GRTE In this experiment, we applied GRTE method to the VDP dataset on the *car* variable distribution to determine if it is closer to a normal or a uniform distribution. This evaluation is performed using discrete values; therefore, we used bins to have different precision levels for each distribution representation.
The results presented in Figure 4 show that the variable data distribution is closer to a normal distribution rather than a uniform one. In fact, normal distribution compatibility is more than 80% until 15 bins, while the uniform distribution matching does not exceed 60%. Starting 20 bins, both distributions matching decrease to fall towards zero. Thus, for the same dataset, the representativeness estimation is dependent on the granularity parameter (number of bins). For instance, if a granularity of 10 bins is sufficient for the given use case, then the dataset contains around 90% of the required information following a normal distribution, whereas, if a granularity of 50 bins is needed, then the dataset contains less than 20% of the required information following a normal distribution.

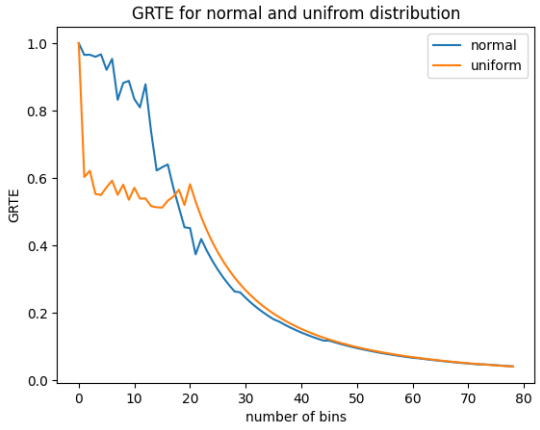

Figure 4: Results of GRTE metric depending on the granularity parameter (number of bins).

Model-Dependent Metrics In the following, we will discuss the experiments conducted to measure the domain gap between two datasets using the MMD method.

MMD The main goal of this experiment was to measure the domain gap between pairs of image datasets: VPD, BDD100K and a synthetic night images dataset. To better capture the domain gap induced by lightning conditions, we used subsets of each dataset representing daytime and nighttime scenes. In addition, we wanted to assess the gap between the images captured by similar cameras from different positions on the vehicle (front, left, right and rear view). Finally, identify the dataset size at which MMD stabilizes by varying the size of the datasets at each run.

Figure 5: Image Samples from the datasets and subsets used for assessing MMD metric.

Source Dataset	Target Dataset	dataset size = 10	dataset size = 100	dataset size = 400	dataset size = 1000
VDP	VDP	≈ 0	≈ 0	≈ 0	≈ 0
BDD100K	VDP	47.5	22.9	24.1	34.1
VDP night	VDP day front camera	42.5	55.6	56.3	55.7
BDD100K night	VDP night	38.6	44.3	45.6	47.2
VDP night	Synthetic night	38.6	45.1	43.4	42.6
BDD100K night	Synthetic night	4.6	2.3	2.8	2.5
VDP day front camera	Synthetic night	61.2	79.2	75.7	73.1
VDP day front camera	VDP day rear camera	12.5	6.0	7.7	6.9
VDP day front camera	VDP day left camera	11.1	7.7	8.2	7.8
VDP day right camera	VDP day left camera	5.0	9.3	7.9	8.5

Table 3: Summary of MMD scores on different datasets and subsets: BDD100K, VDP and a synthetic night images dataset.

Figure 5 gives a snap on the images contained in each dataset / subset and Table 3 gives a detailed view of the datasets and subsets used for each goal.

We used a Resnet18 (He et al. 2016) as a features extractor on both source and target datasets and selected a linear kernel for the MMD distance.

Table 3 gives the results of the experiments. We notice a large domain gap between datasets from different sources as in VDP vs BDD100K; and with different levels of lightning (daytime vs nighttime) as in VDP night vs VDP day front camera. The gap grows bigger when we have two datasets coming from different sources and with different lightning conditions as in VDP day front camera vs synthetic night experiment. The domain gap is relatively small in the subsets of the VDP dataset with different positions of the camera, given that the images are from similar cameras in the same lightning conditions. These observations remain consistent with human intuition. It is worth noting that the gap created by applying a style transfer on BDD100K images to create nighttime lightning conditions is small as shown by the experiment BDD100K night vs synthetic night; which can be a promising avenue to investigate for applying style transfer to simulate lightning and weather conditions.

Regarding the effect of the dataset size on the method stability, we notice that MMD converges starting dataset size of 100 samples. However, this observation depends on the diversity of the datasets involved.

Conclusion and Perspectives

In this work, we have addressed the critical challenge of evaluating data quality within the context of industrial ML processes, focusing on various metrics as identified in the Confiance.ai program. Our approach involved the implementation of a comprehensive set of metrics to assess data quality, categorized into inherent data metrics and model-dependent metrics.

To assess the inherent quality of data, we implemented methods from the state of the art, such as Simpson Index, Gini-Simpson Index for diversity, χ^2 and Kolmogorov-Smirnov tests for representativeness and a completeness measure. We developed a relative diversity measure to account for industrial requirements as well as an entropy representativeness method (GRTE). In the other hand, for model dependent metrics, we focused in this paper on the measures implemented for domain gap evaluation on image datasets using Central Moments Discrepancy, Wasserstein distance, Maximum Mean Discrepancy Kullback-Leibler divergence and H-divergence. We utilized our library DQM to assess the quality of different datasets both open source and proprietary used in actual industrial use cases provided by the Confiance.ai program. In addition, The library is integrated in a larger AI system called DebiAI ((Mansion et al. 2024)).

Future work includes expanding the existing methods to other data types for instance time series and scaling the library to allow its use on very large datasets intended for foundation models.

Acknowledgments

This work has been supported by the French Government under the "France 2030" program, as part of the SystemX Technological Research Institute. This work was conducted as part of the Confiance.ai program, which aims to develop innovative solutions for enhancing the reliability and trustworthiness of industrial AI-based systems.

References

Adjed, F.; Chaouche, S.; Randon, Y.; Le Coz, A.; Herbin, S.; Karaliolios, N.; Winckler, N.; and Feuilleaubois, E. 2023. Data Quality Assessment Metrics for Machine Learning Process. Technical report, Confiance.ai program - IRT SystemX.

Aroyo, L.; Lease, M.; Paritosh, P.; and Schaekermann, M. 2022. Data excellence for AI: why should you care? *Interactions*, 29(2): 66–69.

Chehreghan, A.; and Ali Abbaspour, R. 2018. An evaluation of data completeness of VGI through geometric similarity assessment. *International Journal of Image and Data Fusion*, 9(4): 319–337.

Contreras-Reyes, J. E.; and Arellano-Valle, R. B. 2012. Kullback–Leibler divergence measure for multivariate skew-normal distributions. *Entropy*, 14(9): 1606–1626.

Eyuboglu, S.; Karlaš, B.; Ré, C.; Zhang, C.; and Zou, J. 2022. dcbench: A benchmark for data-centric ai systems. In *Proceedings of the Sixth Workshop on Data Management for End-To-End Machine Learning*, 1–4.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; March, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *Journal of machine learning research*, 17(59): 1–35.

Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; and Smola, A. 2012. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13(25): 723–773.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Hutchinson, B.; Smart, A.; Hanna, A.; Denton, E.; Greer, C.; Kjartansson, O.; Barnes, P.; and Mitchell, M. 2021. Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 560–575.

JoliGEN. 2024. Style transfer on BDD100K.

Juddoo, S.; and George, C. 2020. A qualitative assessment of machine learning support for detecting data completeness and accuracy issues to improve data analytics in big data for the healthcare industry. In *2020 3rd International Conference on Emerging Trends in Electrical, Electronic and Communications Engineering (ELECOM)*, 58–66. IEEE.

Kastryulin, S.; Zakirov, J.; Prokopenko, D.; and Dylov, D. V. 2022. PyTorch Image Quality: Metrics for Image Quality Assessment.

Kumar, S.; Datta, S.; Singh, V.; Singh, S. K.; and Sharma, R. 2024. Opportunities and Challenges in Data-Centric AI. *IEEE Access*.

Luley, P.-P.; Deriu, J. M.; Yan, P.; Schatte, G. A.; and Stadelmann, T. 2023. From concept to implementation: the data-centric development process for AI in industry. In *2023 10th IEEE Swiss Conference on Data Science (SDS)*, 73–76. IEEE.

Mansion, T.; Braud, R.; Amrani, A.; Chaouche, S.; Adjed, F.; and Cantat, L. 2024. DebiAI: Open-Source Toolkit for Data Analysis, Visualisation and Evaluation in Machine Learning. In *ICAS 2024*.

Mazumder, M.; Banbury, C.; Yao, X.; Karlaš, B.; Gaviria Rojas, W.; Diamos, S.; Diamos, G.; He, L.; Parrish, A.; Kirk, H. R.; et al. 2024. Dataperf: Benchmarks for data-centric ai development. *Advances in Neural Information Processing Systems*, 36.

Picard, S.; Chapdelaine, C.; Cappi, C.; Gardes, L.; Jenn, E.; Lefèvre, B.; and Soumarmon, T. 2020. Ensuring dataset quality for machine learning certification. In *2020 IEEE international symposium on software reliability engineering workshops (ISSREW)*, 275–282. IEEE.

Polyzotis, N.; Roy, S.; Whang, S. E.; and Zinkevich, M. 2018. Data lifecycle challenges in production machine learning: a survey. *ACM SIGMOD Record*, 47(2): 17–28.

Polyzotis, N.; and Zaharia, M. 2021. What can data-centric AI learn from data and ML engineering? *arXiv preprint arXiv:2112.06439*.

Rüschendorf, L. 1985. The Wasserstein distance and approximation theorems. *Probability Theory and Related Fields*, 70(1): 117–129.

Yogamani, S.; Hughes, C.; Horgan, J.; Sistu, G.; Varley, P.; O'Dea, D.; Uricár, M.; Milz, S.; Simon, M.; Amende, K.; et al. 2019. Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9308–9318.

Yu, F.; Chen, H.; Wang, X.; Xian, W.; Chen, Y.; Liu, F.; Madhavan, V.; and Darrell, T. 2020. BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. arXiv:1805.04687.

Yuan, Y.; Pang, Q.; and Wang, S. 2023. Revisiting neuron coverage for dnn testing: A layer-wise and distribution-aware criterion. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, 1200–1212. IEEE.

Zellinger, W.; Grubinger, T.; Lughofer, E.; Natschläger, T.; and Saminger-Platz, S. 2019. Central Moment Discrepancy (CMD) for Domain-Invariant Representation Learning. arXiv:1702.08811.

Zha, D.; Bhat, Z. P.; Lai, K.-H.; Yang, F.; and Hu, X. 2023. Data-centric ai: Perspectives and challenges. In *Proceedings of the 2023 SIAM International Conference on Data Mining (SDM)*, 945–948. SIAM.