From Raw Measurements to Smart Data Products: Representation Learning, Augmentation, and Quality Qualification for Industrial Monitoring

***Fakhreddine, Ababsa*[1]**

[1]*PIMM Laboratory, ENSAM, CNRS, CNAM, Paris, France*
fakhreddine.ababsa@ensam.eu

Keywords: Data-Centric AI, Structural Health Monitoring (SHM), Data Augmentation, Multi-Modal Data Fusion, Spatio-Temporal Analysis

Abstract

Industrial monitoring increasingly depends on heterogeneous, multi-sensor, multi-scale and multi-temporal data, yet real-world performance is often constrained by two data-centric bottlenecks: (i) raw measurements are not engineered into learning-ready representations comparable across modalities and scales, and (ii) outputs are rarely quantified and qualified into consistent indicators and KPIs usable in industrial workflows. This paper proposes a data-centric framework for building smart, quantified industrial data products spanning micro-scale SEM microscopy (composite quality control), structural-scale SHM (Lamb-wave measurements), and macro-scale railway corridor monitoring (centimeter-level LiDAR and InSAR time series), with explicit attention to temporal consistency for evolution tracking. We introduce a representation-centric pipeline that maps each modality to standardized image-like or indicator-like objects: (i) Spatially Organized Images (SOI) from SHM damage indices and actuator–sensor path geometry, (ii) SEM segmentation maps coupled with KPI extraction (porosity, morphology, spatial distributions), and (iii) spatio-temporal change indicators and deformation hotspots from multi-epoch LiDAR, qualified with InSAR temporal trends. Data scarcity and labeling cost are addressed through two measurable levers: semi-automatic annotation and generative augmentation (controlled perturbations and VAE-based synthesis). We further propose data qualification metrics to assess representation quality and KPI consistency across samples and time.

1 Introduction

Industry 4.0 relies on cyber-physical systems and connected production lines, where quality control, availability, and maintenance increasingly depend on the use of data collected from multiple detection and inspection modalities. In such environments, data is both heterogeneous (signals, images, point clouds, time series) and multi-scale, ranging from the micrometric scale (microstructure and material defects) to the metric and kilometer scale (structures, networks, corridors) [1]. The main scientific challenge is not to deploy more sensors, but to build coherent and usable data representations that enable reproducible processing and meaningful comparisons in industrial environments. A second obstacle is multi-temporality: acquisitions are asynchronous, with different sampling frequencies and latencies, and they capture dynamics of a very different nature (transient events, slow changes, seasonality). As a result, the problem becomes intrinsically spatio-temporal: representations must remain consistent across time, trends must be estimated, and real changes must be separated from effects induced by acquisition or preprocessing. Heterogeneous temporal regimes and sampling patterns are widely recognized as major methodological obstacles to the implementation of robust large-scale monitoring pipelines [2]. Recent studies on predictive maintenance and industrial condition monitoring converge on a consistent observation: the robustness of real-world deployments depends critically on data preparation and structuring—including representation design, normalization, descriptor stability, and annotation quality—often more than on the systematic increase in model complexity [3]. In the same vein, the data-centric paradigm formalizes the idea that systematic data engineering (improving the quality, consistency, and relevance of datasets) is critical to the robustness and transferability of AI systems [4]. However, in practice, these operations are still often carried out on an ad hoc basis, adapted to a single modality and difficult to reproduce when moving from one scale or time frame to another.

In this paper, we propose a data-centric approach to generating intelligent, quantified industrial data products from heterogeneous, multi-scale, and multi-temporal measurements. Our main contribution is a representation-centric methodology that converts raw measurements into standardized analytical objects such as images or indicators designed for learning and quantification: (i) spatially organized images (SOI) derived from SHM damage indices and actuator-sensor path geometry, (ii) microscopic segmentation maps coupled with the extraction of key microstructural performance indicators (porosity, morphology, spatial distributions), and (iii) corridor-scale spatio-temporal change indicators/hot spots, constructed from multi-epoch observations and temporal trends extracted from InSAR time series. We address data scarcity and labeling costs using two measurable levers: semi-automatic annotation and generative/controlled augmentation. We also introduce a qualification step to ensure the comparability and reproducibility of KPIs and indicators across samples and time periods.

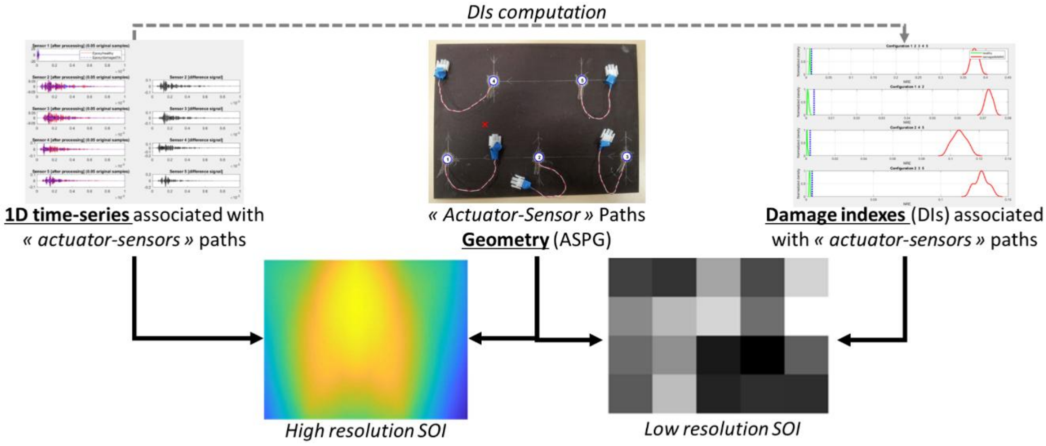


Fig. 1. Construction of Spatially Organized Images (SOI) from damage indices and actuator–sensor path geometry

The rest of the article is organized as follows: we first present the proposed methodology for representation, augmentation, and qualification; we then describe the datasets and three case studies using a common model; we report the results by isolating the impact of representation and qualification choices on robustness; finally, we discuss limitations and practical recommendations before concluding.

2 Data-Centric Methodology: Representation, Augmentation, and Qualification

In this work, we propose a data-centric approach aimed at building "smart" and qualified industrial data, designed to be directly usable in analysis pipelines. The approach is deliberately agnostic with regard to sensors and application domains: it specifies (i) the data artifacts to be produced, (ii) the transformation process used to obtain them from raw measurements, and (iii) the qualification mechanisms ensuring their comparability, traceability, and usability, upstream of any learning or decision layer. This explicit separation between data engineering and modeling gives the approach a generic and reproducible character, applicable to heterogeneous industrial contexts. The methodology is structured in three complementary steps—Representation, Augmentation, Qualification—which define a reproducible process for transforming raw measurements into quantified industrial data products

2.1 Step 1: Representation engineering

The first step is to define a formal representation of the data. The goal is not only to make the data compatible with learning algorithms, but also to ensure its comparability, semantic stability, and interpretability in an industrial context. The literature emphasizes that poorly defined or unstable representations are a major source of uncontrolled variability and performance degradation in real-world conditions [5]. Two categories of representations are considered. Image-type representations correspond to discretized structures (maps, grids, patches, volumes) defined on an explicit spatial, geometric, or temporal medium. They are designed to be directly usable by supervised learning models, while imposing a common framework for resolution, normalization, and analysis support. Indicator-type representations group together descriptors and KPIs calculated deterministically from data or image representations. These indicators are defined with explicit units, scales, and aggregation rules to ensure their reproducibility and interpretation by business experts. This step formalizes what can be described as the semantic contract of the data: it establishes the analysis medium, the transformation rules, the expected invariants, and the metadata necessary for interpretation. This framework is directly linked to the data quality dimensions identified in the literature (consistency, relevance of characteristics, interpretability), which are known to have a decisive influence on model behavior [6], [7].

2.2 Step 2: Managing scarcity: augmentation and efficient labelling

The second step explicitly addresses the scarcity of data and annotations, a structural constraint in most industrial applications. It relies on a combination of controlled augmentation and efficient annotation strategies, integrated as key components of the data product. Augmentation is designed as a constrained operator, not as a simple heuristic transformation. Each augmentation strategy is defined by: (i) a set of authorized operators (perturbations, synthetic generation, scenario injection), (ii) controlled parameter ranges, and (iii) validity rules to discard unrealistic samples. Recent work on data augmentation emphasizes that the effectiveness of these techniques strongly depends on respecting the semantics of the problem and controlling the induced distribution shift [8], [9]. At the same time, annotation production relies on assistance and correction mechanisms aimed at reducing human costs while maintaining label reliability. These strategies include pre-annotation tools, targeted correction protocols, and validation rules. The quality of annotations is recognized as a critical factor for model performance and stability, sometimes more decisive than the amount of data available [6], [10]. All these operations are documented in a generation and annotation log, ensuring data traceability (origin, type of transformation, parameters, corrections). This traceability is essential for

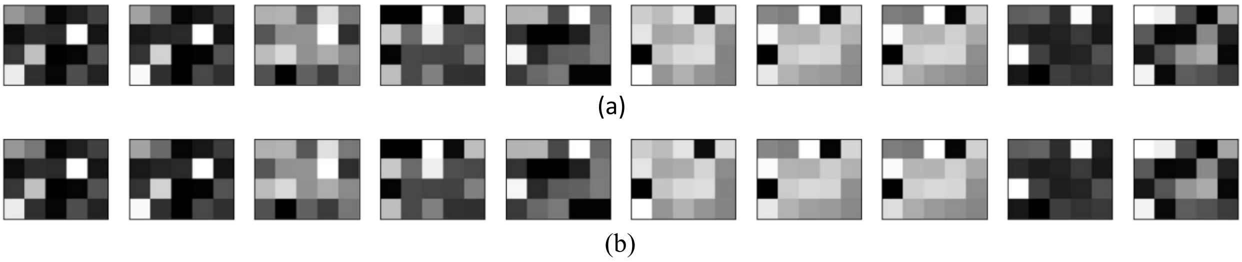


Fig. 2. VAE-based synthetic sample generation: (a) Test SOI Images. (b) Reconstructed SOI

scientific reproducibility and also meets industrial requirements for auditability.

2.3 Step 3: Data qualification and indicator reliability

The third step aims to associate explicit qualification mechanisms with the data produced, enabling their reliability and usability to be assessed. Unlike an assessment focused solely on predictive performance, this step introduces a clear distinction between the quality of model outputs and the reliability of derived indicators. At the model output level, qualification is based on metrics adapted to the type of task (classification, segmentation, detection), supplemented by operational criteria such as the stability of results in the face of controlled data variations, or deployment constraints (memory footprint, latency, reproducibility). Several studies in predictive maintenance and industrial monitoring show that these aspects strongly influence the operational adoption of AI systems [11]. With regard to indicators, qualification aims to ensure their comparability and robustness. It includes consistency analyses (stability with regard to discretization or aggregation choices), sensitivity analyses (impact of augmentation or generation parameters), and, when the data are temporal, inter-epoch consistency criteria. These dimensions are directly in line with the data quality and analytical system evaluation frameworks proposed in the literature [7], [12], [13], [14]. The outcome of this step is a qualification package associated with the data and indicators, providing measurable and auditable elements for interpretation and decision-making.

3 Application of the Data-Centric Methodology to Complementary Industrial Use Cases

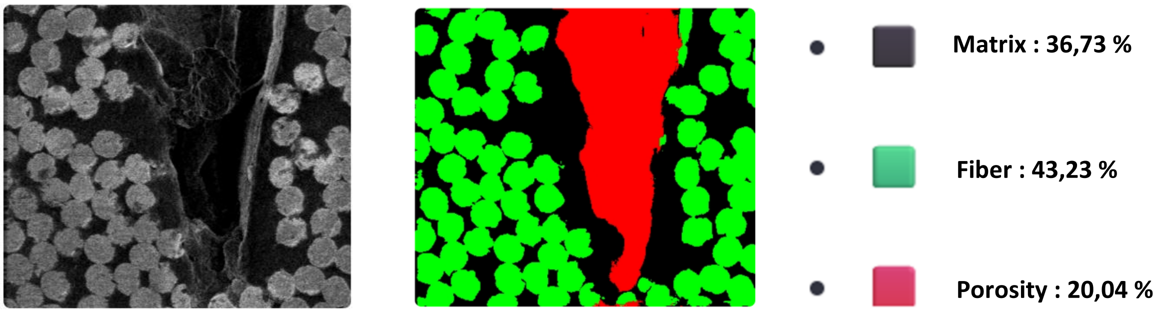
This section applies the data-centric methodology presented in Section 2 to three complementary industrial use cases, selected to highlight different dimensions of the proposed process: heterogeneity of modalities, changes of scale, and temporal regimes. The objective is to establish the feasibility and relevance of the methodology beyond a single application context, showing that the same model can be instantiated consistently in distinct contexts. Each use case is presented using a common report structure, focusing on (i) data representations and the resulting quantified indicators, (ii) the concrete scarcity strategy adopted, and (iii) the qualification signals used to support comparability and operational usability.

3.1 Use Case 1 : Structural health monitoring of composite structures

This first case study focuses on structural health monitoring (SHM) at the structural scale using guided Lamb waves, and serves as an initial validation of the data-centric methodology proposed in a context where raw measurements do not lend themselves directly to learning. The available data consist of one-dimensional time signals acquired along multiple actuator-sensor paths, with limited labeled samples and high sensitivity to experimental variability. This framework is representative of many SHM scenarios, where data scarcity and representation choices largely determine the effectiveness of data-driven approaches.

In accordance with the methodology presented in Section 2, the first step is to design a representation ready for learning. The raw Lamb-wave signals are first processed to calculate a set of damage indices, which are then reorganized into spatially organized images (SOI) by explicitly encoding the geometry of the actuator-sensor path (ASPG). Each SOI corresponds to a low-resolution structured 2D grid in which each cell represents the damage index associated with a specific actuator-sensor path. This transformation converts heterogeneous time series measurements into a standardized image-like representation with fixed semantic meaning and spatial support, suitable for convolutional learning while preserving the physical configuration of the sensor array (see Figure 1). The SOI is thus the central data product of the representation stage.

The second step addresses data scarcity through a controlled augmentation strategy. Rather than applying generic image-level augmentations that would break the physical semantics of the SOI layout, augmentation is performed upstream at the signal level by injecting controlled additive white Gaussian noise (AWGN) into the original waveforms. Damage indices are then recalculated and new SOIs are generated, ensuring that the augmented samples remain physically plausible and consistent with the underlying detection configuration. This signal-sensitive augmentation significantly increases the variability of the dataset while preserving the integrity of the representation. To further enrich the data distribution, a variational autoencoder (VAE) is trained on the SOI representations and used to generate additional synthetic samples in the latent space. This generative step provides a compact and controllable mechanism for extending the dataset beyond experimental acquisitions, while preserving the structural patterns necessary for classification (see Figure 2).



Original image Segmented Mask

Fig. 3. SEM-based microstructural analysis: example segmentation maps and corresponding quantitative KPIs (porosity, morphology, *spatial distributions)*

Finally, the data qualification phase allows the reliability of the data product obtained to be assessed. This qualification is reinforced by the controlled nature of the representation and augmentation process: ASPG mapping guarantees the reproducibility of the representation, while signal disturbance parameters (e.g., noise levels) allow explicit control of variability. The impact of data enrichment is evaluated through classification performance stability and dataset balance analysis, demonstrating that improvements are primarily due to representation consistency and data coverage rather than architecture adjustment. This case study therefore confirms that the SHM pipeline produces a traceable, reproducible, and quantified dataset that is consistent with the goals of the proposed data-centric methodology.

3.2 Use Case 2 : Microstructural Quality Control from SEM Imaging

The second case study focuses on microstructural quality control of composite materials based on images obtained by scanning electron microscopy (SEM) and illustrates the application of the proposed methodology at the microscopic scale, where data are inherently image-based, but where the cost of labeling and quantitative exploitation remain major challenges. Unlike in the case of structural health monitoring (SHM), the difficulty here lies not in converting signals into images, but in transforming high-resolution microscopic data into quantified and reliable indicators that can support industrial evaluation beyond visual inspection.

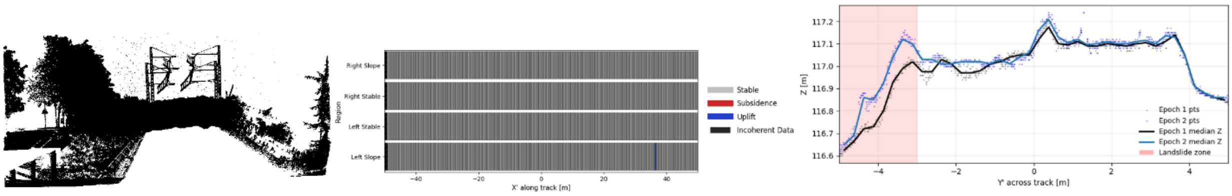
Following the representation stage defined in section 2, SEM acquisitions are first structured into a representation based on semantic segmentation. The images are processed to produce pixel-by-pixel maps corresponding to the main microstructural constituents (e.g., matrix, fibers, porosity). These segmentation maps are standardized image-type representations with explicit semantic meaning, enabling consistent analysis across samples and acquisitions. It is important to note that the segmentation result is not considered a final result, but rather an intermediate representation from which indicator-type data products are derived. Quantitative microstructural key performance indicators, such as phase fractions (e.g., porosity rate), morphological descriptors, and spatial distributions of defects, are calculated deterministically from the segmentation maps, providing interpretable measurements directly aligned with industrial quality criteria (see Figure 3).

The scarcity management phase primarily targets the cost and reliability of annotations, which are the main bottleneck in microscopy-based analysis. Manual pixel-level labeling is tedious and prone to variation, making it difficult to scale fully supervised pipelines. To overcome this limitation, a semi-automatic annotation strategy is adopted, combining model-assisted pre-segmentation and interactive correction. This approach significantly reduces the annotation effort while preserving label fidelity, and fits naturally into the data-centric framework by explicitly treating label production as part of the data product. Unlike synthetic augmentation, which can easily distort fine microstructural patterns, the enrichment strategy here focuses on label efficiency and consistency, ensuring that the available data accurately reflects the actual microstructure.

The qualification step evaluates both the segmentation results and the derived microstructural indicators. At the model output level, standard overlap measures such as the Dice coefficient and intersection over union are used to evaluate the accuracy and stability of segmentation across samples. These measures are complemented by practical considerations related to deployability, such as model complexity and robustness to variations in imaging conditions. Regarding indicators, qualification focuses on the consistency and reproducibility of KPIs, including stability between image tiles and sensitivity to segmentation uncertainty. This two-level qualification ensures that improvements in pixel-level accuracy translate into reliable quantitative indicators, rather than unstable or misleading measures.

This case study demonstrates that the proposed data-centric methodology naturally extends to high-resolution imaging scenarios, where the central challenge is not the availability of representation, but the qualification of data and the reliability of indicators. By explicitly separating representation, label effectiveness, and qualification, this approach enables the construction of traceable and quantified microstructural data products suitable for comparative analysis and integration into industrial quality control workflows.

3.3 Use Case 3 : Spatio-Temporal Monitoring of Railway Corridors



a) Raw LIDAR Data b) Patch level diagnostic map c) . Vertical enbankement profile

Fig. 4. Railway corridor monitoring: spatio-temporal change indicators derived from multi-epoch 3D data and complementary temporal trends used for indicator qualification

The third case study concerns the spatio-temporal monitoring of railway corridors and illustrates the application of the proposed methodology at the infrastructure level, where observations are spread over large spatial areas and acquired over several periods. In this context, the main challenge is not the lack of detection capabilities, but the transformation of heterogeneous, time-indexed observations into reliable and interpretable indicators suitable for long-term monitoring and early warning. This use case therefore emphasizes the methodology's ability to manage changes in scale and temporal consistency, which are central requirements in the monitoring of industrial infrastructure.

In accordance with the representation step defined in Section 2, raw observations are first structured into spatio-temporal representations based on indicators. Multi-temporal 3D measurements, such as LiDAR point clouds or derived digital elevation models, are processed to extract localized change descriptors and candidate hotspots to a coherent spatial medium. These descriptors form indicator-type representations that summarize geometric evolution while remaining interpretable at the segment or patch level along the corridor. In parallel, satellite measurements provide information on temporal trends at a larger spatial level, capturing long-term deformation dynamics. Rather than attempting to merge these heterogeneous sources at the model level, they are organized into complementary indicators that describe different aspects of the same phenomenon, thus forming a structured representation of spatiotemporal evolution (see Fig. 4).

Furthermore, reference data for subsidence or deformation events are generally scarce and incomplete, making fully supervised learning impractical. To mitigate this problem, physics-guided synthetic scenarios are used to emulate plausible geometric evolutions and to test the sensitivity of change detection. These scenarios are generated under explicit constraints in terms of shape, amplitude, and spatial extent, ensuring that the enriched data remains consistent with expected physical behavior. This strategy increases the coverage of representative situations while maintaining data traceability and interpretability.

Data qualification plays a central role in this case study, as infrastructure monitoring requires a high level of reliability before any operational intervention. This qualification is performed at the indicator level by combining geometric plausibility checks and temporal consistency analysis. Morphometric constraints eliminate geometrically inconsistent detections, thus reducing artifacts related to acquisition noise or recording errors. Temporal qualification relies on the persistence and consistency of indicators detected during different acquisition periods, supported by information on long-term trends captured from other modalities such as INSAR images. These criteria provide a structured method for distinguishing transient measurement effects from significant structural changes, without relying on a single predictive score.

This case study demonstrates that the proposed methodology naturally adapts to large-scale, multi-period monitoring scenarios, where the emphasis is on the reliability of indicators and their temporal consistency. By explicitly structuring representations, enrichment strategies, and qualification criteria, the approach makes it possible to produce auditable and quantifiable monitoring indicators suitable for integration into alert and decision-support systems.

4 Discussion and Analysis

This case study highlights three cross-cutting findings. First, the relevant design unit is not the model in isolation, but the data product in a data-centric sense, defined by an explicit representation specification. This specification formalizes the semantics of the analytical objects, the analysis framework (grid, patch, segment, time window), the normalization and aggregation rules, and the associated traceability elements. It directly conditions inter-campaign comparability and the auditability of the processing chain, prerequisites for the industrial exploitation of the indicators. Second, representation engineering should be understood as the explicit formalization of representation assumptions, particularly the assumptions of invariance and aggregation. It specifies what must remain stable (framework, scale, topology, aggregation) and what constitutes relevant variability. When these assumptions are not explicitly stated, acquisition variability and preprocessing choices affect the outputs, reducing robustness and making comparisons unreliable. Third, rarity must be treated as a constrained distribution coverage problem: enrichment is only relevant if it respects the representation assumptions and is accompanied by explicit traceability of the applied transformations. A distinction must be made between coverage-oriented enrichment (underrepresented plausible variability) and annotation efficiency strategies when label reliability is the limiting factor. Finally, qualification must be performed at two levels: (i) qualification of the outputs (stability under permissible disturbances, deployment constraints) and (ii) qualification of the KPIs/indicators (consistency, sensitivity to media/aggregation choices, temporal consistency across multiple epochs). This separation is crucial: operational value depends on the reliability of the quantified indicators, not solely on predictive performance.

5 Conclusion

In this article, we proposed a data-centric methodology for transforming heterogeneous industrial measurements into smart, qualified data products. This approach was instantiated across three complementary industrial scenarios: microscopy, structural-scale SHM, and spatiotemporal corridor monitoring. This demonstrates that a single framework can produce learnable, traceable, and quantifiable data across heterogeneous modalities and scales. The future directions of this work are structured around three main areas. First, we will continue formalizing representation specifications as reusable templates, including systematic recommendations for selecting the analysis medium and aggregation regimes in multi-scale and multi-temporal contexts. Second, we will enhance the qualification process with richer evidence, including acceptance criteria based on KPI sensitivity analysis and calibration analyses where relevant. Finally, we will explore scaling and automation strategies, including semi-automatic generation of representation metadata, improvement of annotation workflows, and integration of qualified data products into monitoring systems.

6 Acknowledgements

This research was funded by the Chair of Digital Twins for Construction and Infrastructure in their Environment, a research initiative supported by ESTP, ENSAM, SNCF Réseau, Schneider Electric, Egis, Bouygues Construction, and BRGM. We are grateful for their support and commitment to advancing the field of digital twins in the construction and infrastructure sectors.

7 References

[1] Morgan, J., Halton, M., Qiao, Y., Breslin, J. G.: ‘Industry 4.0 smart reconfigurable manufacturing machines’, Journal of Manufacturing Systems, 2021, 59, pp. 481-506

[2] von Brömssen, C., Fölster, J., Eklöf, K.: ‘Temporal trend evaluation in monitoring programs with high spatial resolution and low temporal resolution using geographically weighted regression models’, Environ Monit Assess, 2023, 195, article number 547

[3] Surucu, O., Gadsden, S.A., Yawney, J.: ‘Condition Monitoring using Machine Learning: A Review of Theory, Applications, and Recent Advances’, Expert Systems with Applications, 2023, 221, (119738), ISSN 0957-4174

[4] Malerba, D., Pasquadibisceglie, V.: ‘Data-Centric AI’, J Intell Inf Syst, 2024, 62, pp. 1493–1502

[5] Sculley, J. et al.: ‘Hidden Technical Debt in Machine Learning Systems’, NIPS/IEEE, 2015

[6] Mohammed, S., et al.:’The effects of data quality on machine learning’, Information Sciences, Elsevier, 2025

[7] Batini, C., Scannapieco, M.;’Data Quality: Concepts, Methodologies and Techniques’, Springer, 2006

[8] Shorten, C., Khoshgoftaar, T.: ‘A survey on Image Data Augmentation’, Journal of Big Data, Springer, 2019

[9] Mumuni, A., Mumuni, F.: ‘Data Augmentation: A Comprehensive Survey’, Array, Elsevier, 2022

[10] Northcutt, B., et al.:‘Pervasive Label Errors in Test Sets’, IEEE Data Engineering Bulletin, 2021

[11] Lee, J., et al.: ‘Predictive Maintenance of Industrial Systems’, Journal of Manufacturing Systems, Elsevier, 2014

[12] Redyuk, M., et al.: ‘Evaluation of monitoring indicators under uncertainty’, IEEE Transactions on Industrial Informatics, 2023

[13] Ababsa, F., Mallem, M.:’Robust camera pose tracking for augmented reality using particle filtering framework’, Machine Vision and Applications, 2011, 22, pp. 181–195

[14] Ababsa, F., Mallem, M.: ‘Robust camera pose estimation combining 2D/3D points and lines tracking’, IEEE Int. Symp. on Industrial Electronics, Cambridge, UK, 2008, pp. 774-779