UNIVERSITÀ DEGLI STUDI DI SALERNO

DIPARTIMENTO DI INFORMATICA

Systematic Literature Review

Data-Centric Approaches in Industrial Predictive Maintenance

Professors:

DE LUCIA ANDREA
PALMIERI FRANCESCO
POLESE GIUSEPPE
SCARANO VITTORIO

Student:

Zakarya BOUDRAF
Mat. 0522501649

EXAMINATION DATE: 29/01/2025

COURSE YEAR: 2024/2025

CONTENTS

ABSTRACT

The paradigm of Industry 4.0 has established Predictive Maintenance (PdM) as a cornerstone of operational efficiency. However, the efficacy of data-driven models is frequently compromised by data quality issues, such as class imbalance, noise, and scarcity. This Systematic Literature Review (SLR) investigates data-centric challenges in industrial PdM by analyzing 23 primary studies selected from an initial pool of 974. The review follows a strict protocol divided into planning, conducting, and reporting phases. The results indicate that while the C-MAPSS dataset remains a dominant benchmark, there is a critical need for real-world noise handling and unsupervised domain adaptation to address the limitations of current supervised approaches.

CHAPTER 1

INTRODUCTION

1.1 Context and Motivation

Industrial Internet of Things (IIoT) systems generate massive volumes of sensor data, enabling the transition from reactive to Predictive Maintenance (PdM). While deep learning algorithms (DL) have shown remarkable potential in diagnosing equipment faults, their performance is heavily dependent on the quality of the training data. In real-world industrial settings, data is often "dirty" characterized by high noise levels, missing values, and severe class imbalance due to the rarity of failure events. Existing reviews predominantly focus on model architectures (e.g., CNN vs. LSTM), often overlooking the critical data-centric challenges that determine successful deployment.

1.2 Research Objectives

The primary objective of this thesis is to perform a Systematic Literature Review (SLR) to identify and evaluate data-centric methodologies in PdM. Specifically, this study aims to answer the following Research Questions (RQs):

- **RQ1:** What data augmentation techniques are most effective for

handling imbalanced datasets in industrial predictive maintenance?

- **RQ2:** Which public datasets are used to benchmark these models, and are they representative of real-world "open-set" anomaly scenarios?
- **RQ3:** What preprocessing and noise-reduction strategies are applied to raw sensor data to ensure model robustness in harsh industrial environments?
- **RQ4:** Which evaluation metrics are preferred over traditional point-based metrics to accurately assess time-series anomaly detection?
- **RQ5:** What challenges remain unresolved regarding data labeling, ground truth availability, and the "misalignment" problem?

CHAPTER 2

RESEARCH METHODOLOGY

This review adheres to the Systematic Literature Review (SLR) guidelines proposed by Kitchenham and Charters. The methodology is structured into three distinct phases: **Planning**, **Conducting**, and **Reporting**.

2.1 Phase 1: Planning

The planning phase established the protocol for the review, defining the search strategy and the criteria for study selection to ensure reproducibility and minimize bias.

2.1.1 Search Strategy

The search was conducted across three major academic databases: **Scopus**, **IEEE Xplore**, and **ScienceDirect**.

For **Scopus** and **IEEE Xplore**, the primary search string was designed to capture the intersection of Deep Learning, Predictive Maintenance, and Data Quality issues:

("Predictive Maintenance" OR "Fault Diagnosis" OR "RUL")

AND ("Deep Learning" OR "Neural Networks") AND ("Data Augmentation" OR "Imbalance" OR "Noise" OR "Data Scarcity")

For **ScienceDirect**, due to the limitations of its advanced search interface regarding boolean complexity, a simplified query was employed focusing on Title, Abstract, and Keywords:

Title, abstract, keywords: "Deep Learning", "Anomaly Detection", "Industrial/IoT"

2.1.2 Inclusion and Exclusion Criteria

To ensure the relevance and quality of the selected studies, specific criteria were established.

Inclusion Criteria (IC)

Papers were included if they met the following conditions:

- **Timeframe:** Published after 2020 (2020–2025).
- **Language:** English only.
- **Type:** Journal articles and conference papers only.
- **Topic:** Explicitly discussed data challenges (augmentation, noise, datasets) or metrics in an Industrial/IoT context.
- **Method:** Utilized Deep Learning techniques (CNN, LSTM, GAN, VAE).

Exclusion Criteria (EC)

Papers were excluded based on the following reasons:

- **EC1 (Review/Survey):** The paper is a review, survey, or roadmap rather than a primary study.

- **EC2 (Non-Industrial):** The application domain is non-industrial (e.g., Medical, Agriculture, Finance).
- **EC3 (Traditional ML):** The study relies solely on traditional Machine Learning (e.g., SVM, Random Forest) without Deep Learning components.
- **EC4 (Network Security):** The focus is on pure network security (DDoS, IDS, Blockchain) unrelated to machinery health.
- **EC5 (Infrastructure):** The focus is on Cloud/Edge infrastructure issues (scheduling, routing) rather than equipment health monitoring.

2.2 Phase 2: Conducting

After defining the protocol, the conducting phase involved executing the search and rigorously selecting primary studies. This was achieved through a multi-stage filtration process organized into four main steps.

2.2.1 Step 0: Initial Query and Filters

The initial search was executed across the three databases, yielding a total of 974 results before deduplication.

- **Scopus:** 498 papers (using Title/Abstract/Keywords query).
- **IEEE Xplore:** 435 papers (using All Metadata query).
- **ScienceDirect:** 41 papers (using simplified query).

Immediately following the query execution, automated filters were applied directly within the database engines. Results were restricted to the timeframe **2020–2025**, **English language** only, and document types limited to **Journal Articles and Conference Papers**.

2.2.2 Step 1: Deduplication

The filtered results were exported into Zotero. Using the Zoplicate plugin, we performed bulk deduplication to identify and remove identical records existing across multiple databases. This process consolidated the initial 974 results into a unique set of 795 candidate papers.

2.2.3 Step 2: Title Screening and Abstract Scanning

This step involved a two-tier manual filtration of the 795 unique records.

1. **Title Screening:** We first screened titles to remove obviously irrelevant topics (e.g., Network Intrusion Detection, Medical Imaging), reducing the list to 179 papers.

2. **Abstract Scanning:** We then read the abstracts of the remaining candidates to ensure specific coverage of data-centric topics (e.g., imbalance, noise) rather than generic Deep Learning applications. This qualitative check reduced the list to 56 papers.

2.2.4 Step 3: Full-Text Evaluation

The final step combined accessibility checks with detailed quality assessment.

1. **Availability Check:** We checked for full-text availability. Papers behind strict paywalls or inaccessible via institutional repositories were excluded. This procedural limitation reduced the pool from 56 to 39 papers.

2. **Final Reading:** We read the full text (specifically Introduction and Conclusion) of the remaining 39 papers. We assessed them against our Quality Assessment (QA) criteria to ensure robust metrics and industrial applicability. The final selection resulted in 23 primary studies.

Table 2.1: Summary of the Selection Process

Step	Activity	Input	Output
Step 0	Initial Query & Filters	-	974
Step 1	Deduplication (Zotero)	974	795
Step 2	Title & Abstract Screening	795	56
Step 3	Full-Text Evaluation (Access & QA)	56	**23**

2.2.5 Quality Assessment (QA)

To ensure the scientific rigor of the selected studies, a Quality Assessment was performed during Step 3. Each paper was evaluated against the following questions:

- **QA1:** Is the dataset clearly defined and accessible (or described in detail if proprietary)?
- **QA2:** Are the preprocessing and augmentation steps reproducible?
- **QA3:** Does the study use appropriate performance metrics (e.g., F1-score, AUC) suitable for imbalanced data?

Only papers satisfying these criteria were included in the final set.

2.3 Phase 3: Reporting

The reporting phase documents the findings of the systematic review. The data extracted from the 23 primary studies is synthesized and presented in the following chapters (3 and 4), structured according to the defined Research Questions.

CHAPTER 3

LITERATURE REVIEW

To provide context for the results, this chapter briefly defines the core data-centric concepts encountered in the primary studies.

3.1 Data Imbalance in PdM

In industrial environments, machines operate in a healthy state for the vast majority of their lifecycle. Fault data is rare, leading to severe class imbalance (e.g., 99.8% normal vs 0.2% anomalous). This imbalance causes standard Deep Learning models to bias towards the majority class, necessitating augmentation strategies.

3.2 Synthetic Data Generation

Generative models such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) are increasingly used to synthesize artificial fault data. These models learn the underlying distribution of the minority class to generate realistic samples, thereby balancing the dataset.

3.3 Common Datasets

Two public datasets appear frequently as benchmarks:

- **C-MAPSS:** A simulated turbofan engine degradation dataset provided by NASA.
- **IMS Bearing:** A real-world vibration dataset from a run-to-failure test on rolling element bearings.

CHAPTER 4

RESULTS

This chapter presents the synthesis of the 23 selected studies, organized by the defined Research Questions.

4.1 Data Augmentation Techniques (RQ1)

Generative and adversarial techniques are emerging as the most effective methods for addressing data scarcity and class imbalance.

- **Adversarial Autoencoders (AAE):** The Confidence Adversarial Autoencoder (CAAE) addresses the "absence of anomaly labels" in multivariate time series. It employs a confidence network and a reconstruction network trained via adversarial learning to expand the decision boundary between normal and abnormal data [1].
- **WAE-GAN:** In metro train maintenance (MetroPT dataset), a Wasserstein Autoencoder (WAE) regularized with a GAN allows the model to learn stable representations of normal behavior from unbalanced data, enabling failure detection up to two hours in advance [2].

- **Reconstruction GANs:** For visual anomaly detection (e.g., traffic signs), GANs are employed for reconstruction. Anomalies are detected when the reconstruction error between the noisy input and the GAN-generated clean output is high [3].
- **Other Approaches:** Studies also explored few-shot learning (FS-GAN) [4] and synthetic out-of-distribution data generation (MGS) [5] to handle extreme scarcity.

4.2 Benchmarking Datasets (RQ2)

While public datasets are standard for benchmarking, recent literature highlights a shift toward real-world proprietary datasets to capture heterogeneous scenarios.

- **Public Benchmarks:** The **CWRU** dataset remains widely used for bearing fault diagnosis, often serving as a "source domain" for transfer learning [6, 7]. The **SWaT** dataset is standard for Cyber-Physical Systems [8].
- **Real-World Representativeness:** Many studies reject public datasets in favor of real-world data to capture true complexity. Examples include semiconductor manufacturing data handling heterogeneous sensors [9] and industrial screw tightening data characterized by high imbalance (0.2% failures) [10, 11].

4.3 Preprocessing and Noise Reduction (RQ3)

Robustness against "harsh environments" is achieved through advanced signal decomposition and dedicated denoising architectures.

- **Signal Decomposition:** A method combining Ensemble Empirical Mode Decomposition (EEMD) and Independent Component Analysis

(ICA) decomposes vibration signals, using Fuzzy Entropy to threshold and remove noise [6].

- **Denoising Architectures:** Denoising Autoencoders (DAE) are explicitly trained to remove noise from sensor data before passing it to classifiers [7].
- **Variational Mode Decomposition (VMD):** Used in elevator guideway diagnosis to separate multi-scale signals and suppress noise [12].

4.4 Evaluation Metrics (RQ4)

There is a clear preference for metrics that handle imbalance and threshold independence over traditional accuracy.

- **AUC (Area Under the Curve):** In highly imbalanced scenarios, AUC is preferred as it assesses performance across all thresholds, avoiding the bias of fixed-threshold metrics [10].
- **Reconstruction Error:** Studies often analyze the distribution of Reconstruction Error (MSE) rather than simple binary metrics, using dynamic statistical thresholds for detection.
- **Standard Metrics:** Precision, Recall, and F1-Score remain dominant for direct comparison of supervised models [13].

4.5 Open Challenges (RQ5)

The primary unresolved challenges center on the cost of supervision and model transferability.

- **Cost of Labeling:** The high cost of manual labeling drives the adoption of unsupervised or one-class learning approaches.

- **Domain Misalignment:** Industrial data is heterogeneous. Unresolved challenges include developing scalable models that adapt to new machines via Transfer Learning without extensive retraining [9].

CHAPTER 5

THREATS TO VALIDITY

5.1 Search Bias

The search was limited to English-language papers in three specific databases. This may have excluded relevant studies published in other languages or indexed in other repositories.

5.2 Selection Bias

The initial screening of papers was based on titles and abstracts. While strict criteria were applied, the potential for false negatives during the manual screening process exists.

5.3 Accessibility Bias

A significant reduction in the primary study pool occurred due to accessibility constraints. Of the 56 potentially relevant papers identified after abstract screening, 17 were excluded because the full text was not accessible via institutional subscriptions or open access. This limitation may have excluded

high-quality proprietary research.

CHAPTER 6

CONCLUSION

This review of 23 primary studies confirms that data quality is the linchpin of successful Predictive Maintenance in Industry 4.0. While Data Augmentation techniques like GANs and VAEs are maturing, the field relies too heavily on clean, simulated benchmarks like C-MAPSS. The results indicate a critical need for robust, real-world noise handling strategies and unsupervised domain adaptation to address the persistent lack of labeled failure data. Future work must focus on developing standardized evaluation frameworks that accurately reflect the costs and constraints of industrial deployment.

REFERENCES

[1] Jiahao Shan et al. “Unsupervised Multivariate Time Series Data Anomaly Detection in Industrial IoT: A Confidence Adversarial Autoencoder Network”. In: *IEEE Open Journal of the Communications Society* (2024).

[2] Miguel E. P. Silva et al. “Predictive Maintenance, Adversarial Autoencoders and Explainability”. In: *Machine Learning and Principles and Practice of Knowledge Discovery in Databases* (2023).

[3] Barbara Villarini et al. “Detection of Physical Adversarial Attacks on Traffic Signs for Autonomous Vehicles”. In: *2023 IEEE International Conference on Industry 4.0*. 2023.

[4] J. Kim et al. “FS-GAN: Few-Shot Anomaly Detection for Surface Defects”. In: *Journal of Manufacturing Systems* (2025).

[5] S. Kafunah et al. “MGS: Synthetic Out-of-Distribution Data Generation for Fault Diagnosis”. In: *IEEE Transactions on Instrumentation and Measurement* (2023).

[6] Hanting Zhou et al. “Intelligent machine fault diagnosis with effective denoising using EEMD-ICA-FuzzyEn and CNN”. In: *International Journal of Production Research* (2023).

[7] Seokju Oh et al. “Multi-Scale Convolutional Recurrent Neural Network for Bearing Fault Detection in Noisy Manufacturing Environments”. In: *Applied Sciences* (2021).

[8] A. Ramachandran et al. “Aquila Optimization with Machine Learning-Based Anomaly Detection Technique in Cyber-Physical Systems”. In: *Computer Systems Science & Engineering* (2023).

[9] Simone Tedesco et al. “A Scalable Deep Learning-Based Approach for Anomaly Detection in Semiconductor Manufacturing”. In: *2021 Winter Simulation Conference (WSC)*. 2021.

[10] Diogo Ribeiro et al. “A Comparison of Anomaly Detection Methods for Industrial Screw Tightening”. In: *Springer* (2021).

[11] Diogo Ribeiro et al. “Isolation Forests and Deep Autoencoders for Industrial Screw Tightening Anomaly Detection”. In: *Computers* (2022).

[12] Zhiwei Zhou et al. “TL-MC-ShuffleNetV2: A Lightweight and Transferable Framework for Elevator Guideway Fault Diagnosis”. In: *International Journal of Advanced Computer Science and Applications* (2025).

[13] Artur D. Surowka and Ruomu Tan. “Performance of Machine-Learning-Based Algorithms for Anomaly Detection in Variable Frequency Drives”. In: *2023 IEEE SDEMPED*. 2023.

LIST OF FIGURES

LIST OF TABLES