UNIVERSITÀ DEGLI STUDI
DI SALERNO

Data-Centric Approaches in Industrial Predictive Maintenance

A Systematic Literature Review

Student: Zakarya Boudraf (Mat. 0522501649)

Professors: Palmieri Francesco, Polese Giuseppe, De Lucia Andrea, Scarano Vittorio

Examination Date: 29/01/2026

Context: The Paradigm Shift

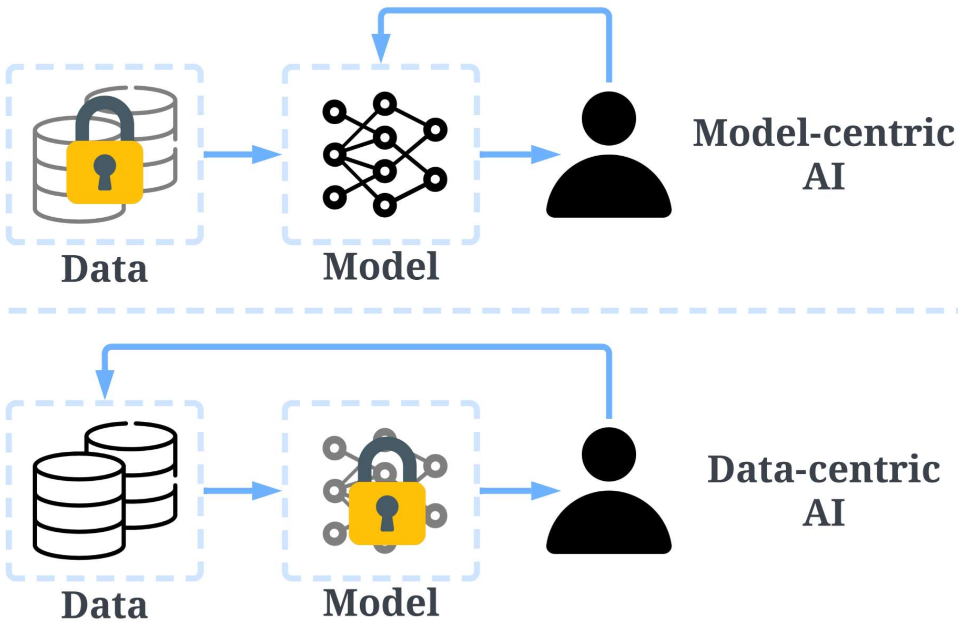
Predictive Maintenance (PdM)

Using data analysis to detect anomalies and predict equipment failure before they happen.

Data-Centric vs Model-Centric AI

- **Model-Centric:** Iterating on model / code while keeping data fixed.
- **Data-Centric:** Iterating on data (Augmentation, Labeling) while keeping code / model fixed.

In Industry 4.0, algorithms are mature, but data is often "dirty" or imbalanced.



The Industrial Data Problem

UNIVERSITÀ DEGLI STUDI DI SALERNO

Standard Deep Learning fails on industrial data due to three core issues:

Extreme Imbalance

99.8% Healthy data vs. 0.2% Fault data. Models become biased toward the "healthy" class.

High Noise

Sensor readings are corrupted by factory vibrations and electromagnetic interference.

Lack of Labels

Historical data exists, but "ground truth" (exact fault onset time) is often missing or inaccurate.

Review Objectives (Research Questions)

UNIVERSITÀ DEGLI STUDI DI SALERNO

RQ1: Augmentation

What data augmentation techniques are most effective for handling imbalanced datasets in industrial PdM?

RQ2: Benchmarks

Are public datasets realistic enough for industrial use?

RQ3: Preprocessing

What preprocessing and noise-reduction strategies are applied in harsh environments?

RQ4: Metrics

Which evaluation metrics are preferred over simple 'Accuracy'?

RQ5: Challenges

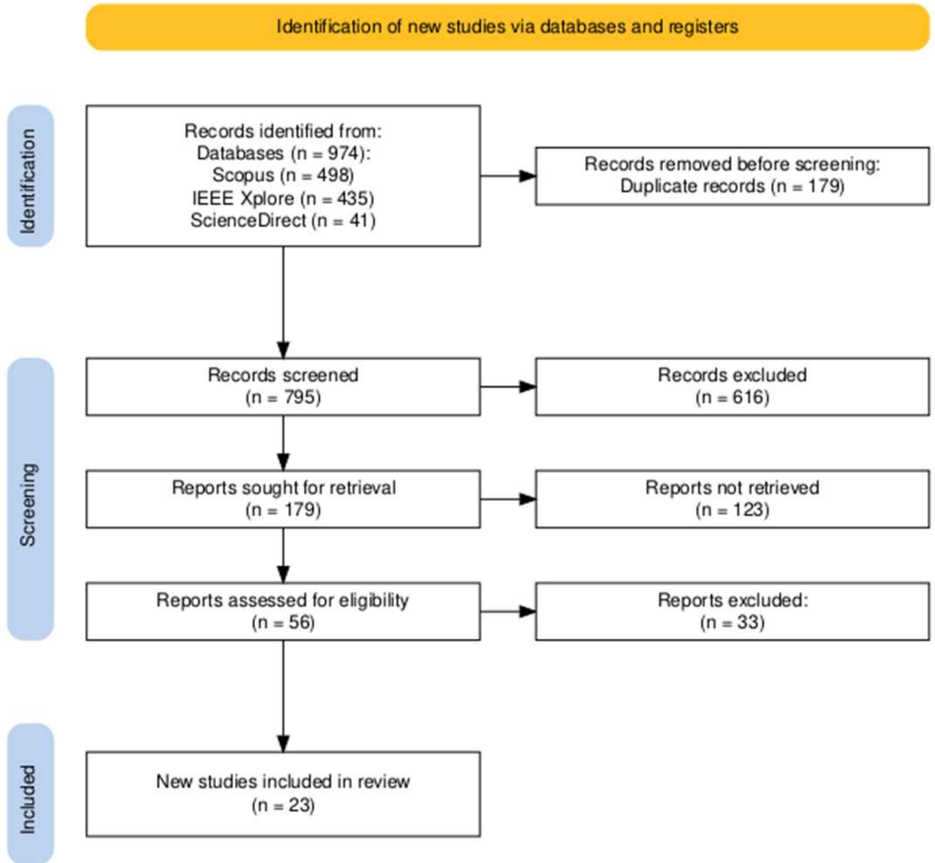
What challenges remain unresolved regarding data labeling, ground truth availability problem?

Review Protocol & Selection

UNIVERSITÀ DEGLI STUDI DI SALERNO

Search Strategy

- **Databases:** IEEE Xplore, Scopus, ScienceDirect.
- **Query:** ("Predictive Maintenance" OR "Fault Diagnosis" OR "RUL") AND ("Deep Learning" OR "Neural Networks") AND ("Data Augmentation" OR "Imbalance" OR "Noise" OR "Data Scarcity")
- **Query for ScienceDirect:** Title, abstract, keywords: "Deep Learning", "Anomaly Detection", "Industrial/IoT"
- **Filters:** English, Journal Articles, Conference papers, 2020-2025.

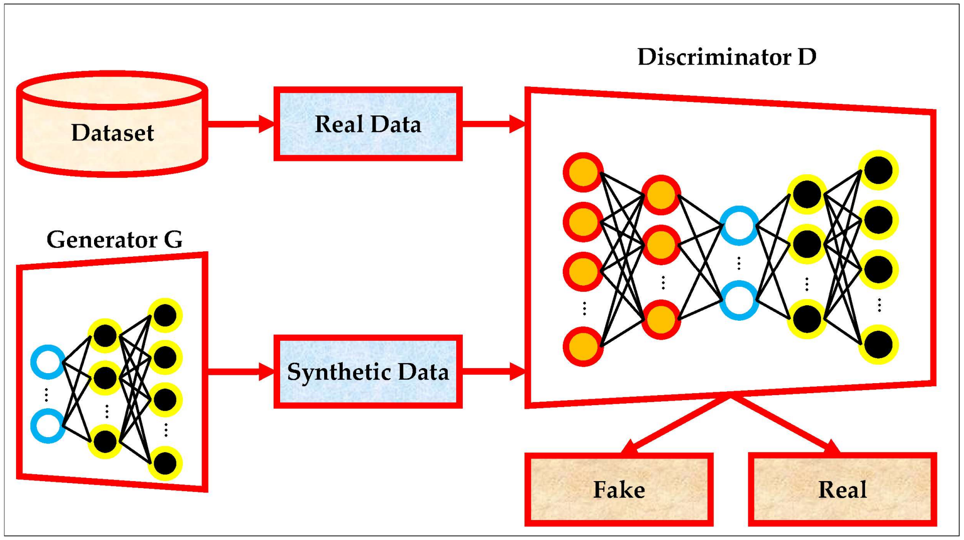


Findings: Data Augmentation (RQ1)

UNIVERSITÀ DEGLI STUDI DI SALERNO

Generative Methods dominate

- **GANs (Generative Adversarial Networks):** The dominant method. It uses a "Generator" to create fake fault data and a "Discriminator" to validate it.
- **VAEs (Variational Autoencoders):** Used to learn normal distributions for anomaly detection.
- **Impact:** These methods artificially balance the dataset, allowing models to learn features of rare faults.



Findings: Benchmarking (RQ2)

UNIVERSITÀ DEGLI STUDI DI SALERNO

Are public datasets realistic enough for industrial use?

Standard Public Benchmarks

- **NASA C-MAPSS:** Jet Engine Simulated Data sets that consist of multiple multivariate time series.
- **IMS Bearing:** A real-world vibration dataset from a run-to-failure test on rolling element bearings
- Other datasets like CWRU and SWaT
- These public datasets are widely used but considered "too clean" for modern evaluation, since they lack the complex noise of real factories.

The Shift to Real-world Proprietary Data

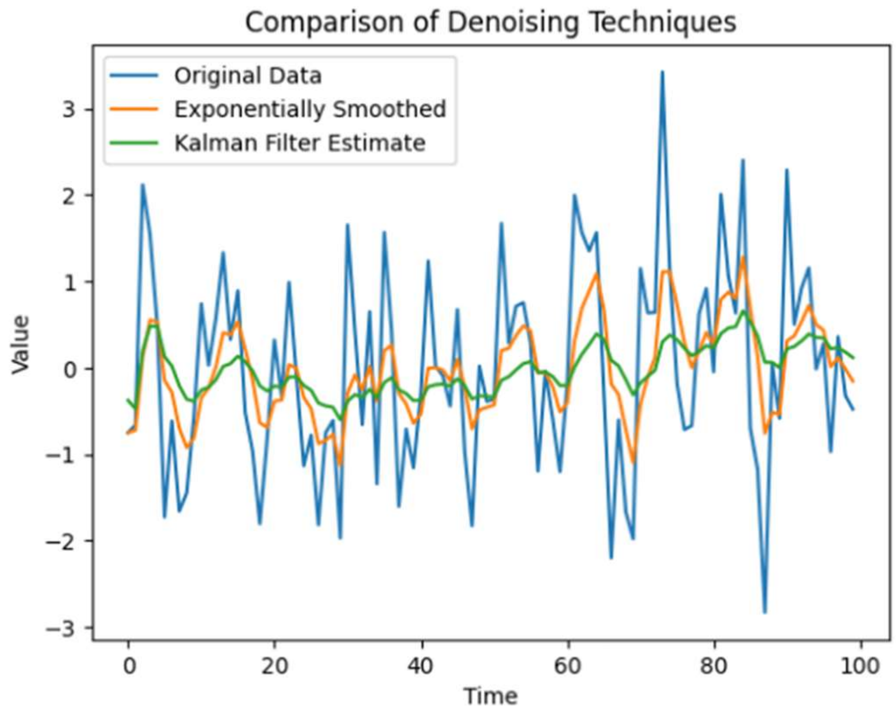
- **Trend:** Shift toward real-world proprietary data to more accurately predict failures and anomalies.
- **Why?** Real data contains noise, gaps, sensor drift, and variable operating conditions that simulations miss.

Findings: Noise Reduction (RQ3)

Handling Harsh Environments

Raw sensor data is rarely usable directly.

- **Signal Decomposition:** Techniques like *EEMD-ICA* separate the true mechanical signal from background factory noise.
- **Denoising Autoencoders (DAE):** Neural networks explicitly trained to reconstruct clean signals from corrupted inputs.



Findings: Evaluation Metrics (RQ4)

UNIVERSITÀ DEGLI STUDI DI SALERNO

Why 99% Accuracy is meaningless in imbalanced data:

AUC

Area Under Curve

Evaluates performance across all threshold settings, independent of class **balance**.

F1

F1-Score

Harmonic mean of Precision and Recall. Essential for minimizing false alarms.

RE

Reconstruction Error

Difference between the real data and the model's attempt to recreate it.

RQ5: Open Challenges

UNIVERSITÀ DEGLI STUDI DI SALERNO

1. Cost of Labeling

Manual annotation is expensive since it requires experts and prone to human error.

2. Domain Misalignment

Different machines have different data. Models trained on Machine A fail on Machine B.

3. Ground Truth Availability

In IIoT, exact fault start times are often unknown or inaccurate, complicating supervised training.

Conclusion

UNIVERSITÀ DEGLI STUDI DI SALERNO

Core Finding: Data imbalance is the primary bottleneck for PdM in IIoT.

- **Generative AI (GANs)** is effectively solving the data scarcity problem.
- **Robust Denoising** is essential for real-world harsh environments.
- **Using proper metrics** that are not affected by data imbalance provides more accurate representation.

→ **Future work** must focus on Unsupervised Learning and Transfer Learning to remove the need of manual labeling and to improve transferability.

Questions?

Thank You for Your Attention

Zakarya Boudraf

0522501649

29th January 2025