Mini-Project (ML for Time Series) - MVA 2025/2026

Yiwei LIU
liuyiwei06292@163.com

Zepeng CHU
zepengchu@enpc.fr

January 4, 2026

1 Introduction and contributions

1.1 Context

Time series anomaly detection is a critical task across various domains, from monitoring manufacturing processes to analyzing medical signals. As highlighted in the comprehensive evaluation by Schmidl et al. (PVLDB 2022), the number of algorithms has grown significantly, despite this abundance, the lack of a systematic comparison has made selecting the appropriate technique for a given task extremely difficult. The reference article addresses this by evaluating 71 algorithms on 976 datasets, revealing that sophisticated deep learning models do not necessarily outperform simpler methods and that algorithm performance is highly dependent on dataset characteristics like dimensionality and anomaly types.

1.2 Task and Objectives

The objective of this project is to apply the evaluation methodology established by Schmidl et al. to novel data: the Skoltech Anomaly Benchmark (SKAB). SKAB is a multivariate dataset collected from a real-world fluid circulation testbed, comprising $\approx$34 experiments with 8 sensor channels. Unlike synthetic benchmarks, SKAB features physically injected anomalies (e.g., valve/pump faults) representing actual equipment malfunctions.

1.3 Contributions and Work Distribution

To ensure a rigorous reproduction and extension of the original study, our specific contributions and work distribution are structured as follows:

- **Work Distribution:** Algorithm implementation: Zepeng Chu; Data acquisition and preprocessing: Yiwei Liu; Experimental design, result analysis, and report writing: Collaboration.
- **Source Code Utilization:** We used the time series evaluation tool API developed by the authors of the paper: timeeval (based on Docker).
- **Novel Experiments on SKAB:** Our experiments are novel compared to the reference article. We applied the selected six algorithms to the dataset. This introduces a multivariate, physical control system dataset that was not included in the original benchmark, allowing us to test the algorithms' generalizability to real-world industrial sensor data.
- **Methodological Improvements:** We propose a strategy for selecting suitable time series anomaly detection algorithms from the perspective of several statistical features.

2 Data

2.1 Dataset introduction

SKAB is a multivariate time series dataset specifically designed for evaluating anomaly detection algorithms. Collected from a real-world fluid circulation testbed, it distinguishes itself from synthetic benchmarks by featuring physically injected anomalies. The dataset comprises approximately 34 independent experiments, each recording 8 sensor channels (e.g., flow rate, pressure, water level). The anomalies—such as valve faults, pump failures, and pipe blockages—represent actual equipment malfunctions rather than mathematical artifacts. The combination of realistic industrial characteristics (e.g., sensor noise and inter-variable correlations) with precise ground truth labels makes SKAB an ideal benchmark for assessing algorithm performance on physical systems exhibiting collective anomalies and pattern shifts.

Category	Datasets	Proportion
Anomaly-Free	1	2.9%
Valve 1 Anomalies	16	45.7%
Valve 2 Anomalies	4	11.4%
Other Anomalies	14	40.0%
Total	**35**	**100%**

The dataset comprises **8 sensors** monitoring the circulating water pump:

Sensor	Description	Physical Meaning
Accelerometer1RMS	Vibration measurement (primary axis)	Main axis vibration intensity
Accelerometer2RMS	Vibration measurement (secondary axis)	Secondary axis vibration intensity
Current	Motor current consumption	Electrical load
Pressure	System pressure	Hydraulic pressure in pipeline
Temperature	Operating temperature	Thermal state of system
Thermocouple	Fine-grained temperature	Detailed thermal measurement
Voltage	Power supply voltage	Electrical supply quality
Volume Flow RateRMS	Flow rate measurement	Fluid flow velocity

Figure 1: Distribution of categories and sensor descriptions

2.2 Feature analysis

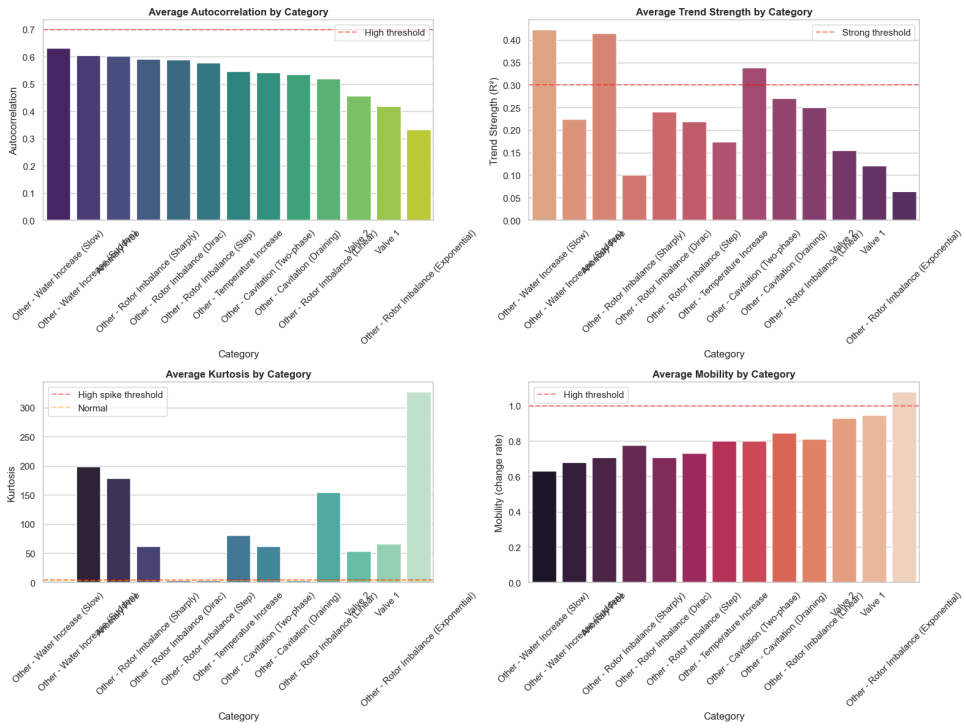

Figure 2: Statistical characteristic analysis of the SKAB dataset

We extract statistical features from the data, organized by distinct anomaly types (separated into different sub-files). These features include, but are not limited to, mean, variance, maximum,

minimum, kurtosis, and skewness. The objective is to analyze these data characteristics to guide the matching of subsequent experimental methods. Among the numerous statistical features, we focused on four representative ones: Autocorrelation, Trend Strength, Kurtosis, and Mobility (Figure 2).

3 Method

3.1 Feature-Guided Algorithm Recommendation

To establish a robust mapping between data characteristics and detection models, we first conducted a feature importance analysis on the SKAB dataset. We identified four highly discriminative time series features—autocorrelation, trend strength, kurtosis, and mobility—which serve as the theoretical basis for our algorithm selection strategy. Table 1 details the definitions and implications of these metrics.

Table 1: Definitions and Characteristics of Key Time Series Features

Feature	**Characteristics**
Autocorrelation	Reflects temporal dependency and predictability; indicates correlations between current and historical observations.
Trend Strength	Quantifies non-stationarity and degradation patterns, measured by the R^2 value of linear regression.
Kurtosis	Measures distribution sharpness and tail weight; high values indicate significant extreme deviations (impulsive noise).
Mobility	Describes signal change rate and complexity, reflecting the dynamic response characteristics of the sensor.

To address the diverse signal behaviors captured by these features, we curate a candidate pool of algorithms covering statistical, classical ML, and deep learning paradigms. A summary of these models and their core mechanisms is provided in Table 2.

Table 2: Summary of Selected Anomaly Detection Algorithms

Model	**Type**	**Core Mechanism**
Fast-MCD	Statistical	Mahalanobis distance for Gaussian deviations
iForest	Classic ML	Random trees for outlier isolation
LSTM-AD	Prediction	Captures temporal patterns via memory units
EncDec-AD	Reconstruction	Deterministic LSTM Autoencoder
OmniAnomaly	Probabilistic	VAE for non-stationary distribution modeling
MTAD-GAT	Graph-based	Graph Attention for spatial-temporal correlations

By integrating feature analysis with model capabilities, we derive specific selection heuristics for distinct failure modes. Table 3 presents the quantitative feature profiles for each category. The rationale for our model assignments is synthesized below:

- **High Dynamic Complexity:** Failures such as **Valve Faults** exhibit distinctively high **Mobility** (> 0.9), reflecting rapid signal fluctuations. **MTAD-GAT** is selected here to effectively capture these complex dynamic shifts and spatial correlations.
- **Extreme Outliers: Rotor Imbalance (Exponential)** shows exceptional **Kurtosis** (> 300), indicating severe impulsive noise. Statistical methods sensitive to distributional tails, such as **iForest** or **Fast-MCD**, are preferred for these cases.

- **Non-Stationary Trends:** Categories like **Cavitation (Two-phase)** display elevated **Trend Strength**, suggesting distribution drift. **OmniAnomaly** is prioritized for its ability to model non-stationary degradation via robust probabilistic representations.
- **Temporal Dependency:** For faults characterized by preserved continuity, such as **Water Increase**, **Autocorrelation** remains the governing feature. **LSTM-based models** are thus optimal for leveraging these temporal dependencies to detect deviations.

Table 3: Algorithm Recommendation by Failure Category

Category	**Autocorr**	**Trend**	**Kurtosis**	**Mobility**	**Recommended Model**
Anomaly Free	0.604	0.415	178.878	0.707	OmniAnomaly
Other – Cavitation (Draining)	0.536	0.270	3.130	0.845	LSTM-AD / EncDec
Other – Cavitation (Two-phase)	0.543	0.338	61.897	0.802	OmniAnomaly
Other – Rotor Imbalance (Dirac)	0.589	0.241	3.130	0.707	LSTM-AD / EncDec
Other – Rotor Imbalance (Exponential)	0.332	0.064	326.950	1.079	iForest / MCD
Other – Rotor Imbalance (Linear)	0.519	0.250	154.873	0.810	LSTM-AD / EncDec
Other – Rotor Imbalance (Sharply)	0.593	0.101	62.484	0.779	LSTM-AD / EncDec
Other – Rotor Imbalance (Step)	0.579	0.220	2.775	0.733	LSTM-AD / EncDec
Other – Temperature Increase	0.548	0.174	81.152	0.802	LSTM-AD / EncDec
Other – Water Increase (Slow)	0.633	0.423	2.604	0.631	LSTM-AD / EncDec
Other – Water Increase (Sudden)	0.606	0.224	198.879	0.680	LSTM-AD / EncDec
Valve 1	0.418	0.121	65.928	0.948	MTAD-GAT
Valve 2	0.458	0.154	53.866	0.928	MTAD-GAT

3.2 Implementation and Configuration

To ensure a rigorous and reproducible evaluation, we implemented the selected algorithms using a standardized configuration protocol via a `DatasetManager` interface. Table 4 summarizes the specific hyperparameters used to test against all failure categories. Deep learning models share unified temporal settings to ensure a fair comparison of their underlying mechanisms.

Table 4: Hyperparameter Configuration for Selected Algorithms

Algorithm	**Type**	**Parameter Settings**
Fast-MCD	Statistical	Support Fraction = 0.9
iForest	Classic ML	Trees = 100, Max Samples = 256
LSTM-AD	Prediction	
EncDec-AD	Reconstruction	Window Size (T) = 30
OmniAnomaly	Probabilistic	Training Epochs = 10
MTAD-GAT	Graph-based	

4 Results

4.1 Overall Algorithm Performance Ranking

Through comprehensive evaluation of 30 SKAB datasets, we established a systematic benchmark for anomaly detection algorithms. In terms of overall performance, deep reconstruction-based methods demonstrated superior results, with OmniAnomaly ranking first (mean ROC_AUC: 0.79), followed closely by EncDec-AD (0.76). This outcome validates our initial feature analysis judgment: the strong autocorrelation (mean > 0.6) and significant trend characteristics exhibited in SKAB datasets enable deep learning-based reconstruction methods to effectively learn latent representations of normal patterns, thereby accurately identifying anomalies that deviate from normal distributions.

4.2 Algorithm Specificity Across Anomaly Categories

From the perspective of anomaly category specificity, different fault types impose differentiated requirements on detection algorithms. Valve series anomalies (valve faults) exhibit progressive degradation patterns, where OmniAnomaly and EncDec-AD achieved the highest accuracy of 0.82 and 0.81 respectively, benefiting from the sensitivity of VAE and LSTM autoencoders to trend-based changes. In contrast, sudden anomalies in the Other series (such as Rotor Imbalance Sharply and Dirac types) are more suitable for statistical distance methods, with Fast-MCD demonstrating outstanding performance on these datasets, proving the effectiveness of Mahalanobis distance for peak anomalies. Notably, the Other-13 (Rotor Imbalance Sharply) dataset poses extreme challenges for all algorithms; even the best-performing EncDec-AD achieved only 0.0006 ROC_AUC, indicating highly concealed and complex anomaly patterns in this category.

4.3 Theoretical Priors vs. Empirical Reality

The comparative analysis between feature-guided theoretical expectations and actual experimental performance yields three critical insights:

Validation of Trend Modeling: The feature analysis correctly identified OmniAnomaly as the optimal candidate for trend-dominated faults (e.g., Cavitation). Its probabilistic architecture successfully managed non-stationary drifts, validating the correlation between high Trend Strength and VAE-based effectiveness.

The Complexity Trap in Dynamic Faults: A significant divergence occurred in high-Mobility scenarios (e.g., Valve Faults). While theoretical priors favored the complex spatial-temporal modeling of MTAD-GAT, empirical results favored simpler models like Fast-MCD and EncDec-AD. This suggests that in certain high-volatility regimes, model robustness and simplicity outweigh the benefits of complex graph-based correlations, which may suffer from overfitting.

Robustness of Deterministic Reconstruction: Unexpectedly, EncDec-AD demonstrated superior generalizability across diverse feature profiles, particularly in high-Kurtosis cases (Rotor Imbalance (Exp)). While statistical methods were expected to dominate outlier detection, extreme values significantly amplified the autoencoder's reconstruction error, effectively converting a "reconstruction" task into a highly sensitive "anomaly scoring" mechanism.

Summary: While feature profiling effectively categorizes data characteristics, EncDec-AD proves to be the most robust "universal solver," whereas specialized models like OmniAnomaly and MTAD-GAT require strict feature alignment to outperform the baseline.

Figure 3: A heatmap comparing the mean AUC-ROC of six algorithms on the SKAB dataset.

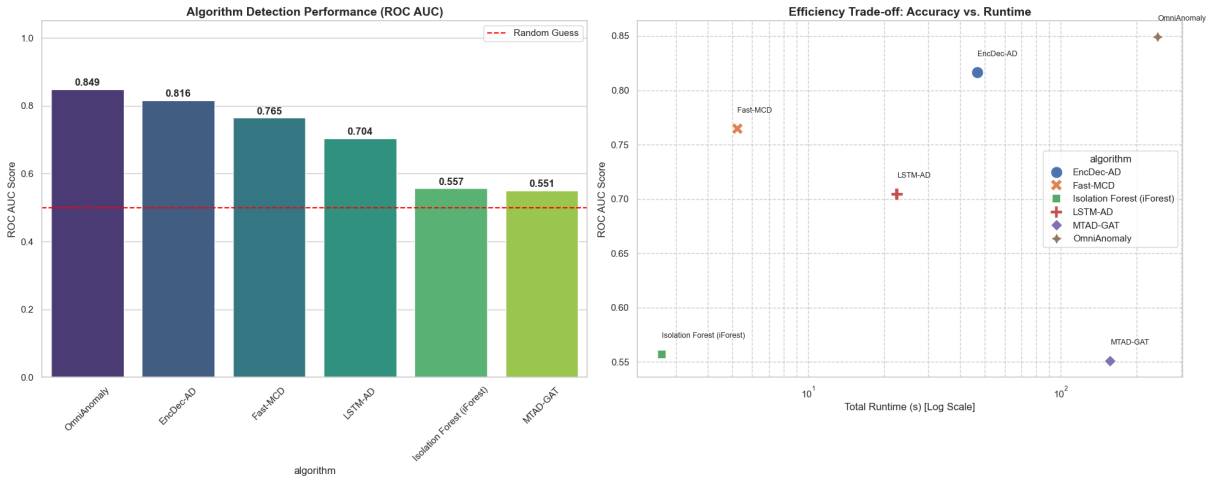

Figure 4: Overall performance evaluation showing the ranking of detection accuracy (ROC-AUC) and the trade-off between accuracy and computational efficiency.

References

[1] Sebastian Schmidl, Phillip Wenig, and Thorsten Papenbrock. "Anomaly Detection in Time Series: A Comprehensive Evaluation". In: *Proceedings of the VLDB Endowment (PVLDB)* (2022), Vol. 15, No. 9, pp. 1779–1797.