Universidad Politécnica de Madrid

Escuela Técnica Superior de Ingenieros Informáticos

Grado en Máster Universitario en Innovación Digital

Data Mining and Time Series

Knowledge Discovery Project

Authors:

Emanuele Alberti
Leandro Duarte
Ottavia Biagi

Professor:

Aurora Perez

October 2025

Contents

1 Introduction

Anomaly detection in multivariate time series is a critical task across various domains, including aerospace systems monitoring, industrial equipment maintenance, and cybersecurity. This project explores the application of modern deep learning techniques for anomaly detection, specifically focusing on transformer-based approaches.

Our work is inspired by two papers in this field: **OmniAnomaly** [1], which uses stochastic recurrent neural networks with variational autoencoders, and **TranAD** [2], which leverages transformer networks with adversarial training. We aim to study these methodologies and apply transformer-based detection methods to well-established benchmark datasets in the anomaly detection domain.

Note: All project code, and documentation will be available at: `https://github.com/Leandr0Duar7e/DM-UPM-2526/tree/main`

---

2 Domain Understanding

Our focus is on multivariate time series anomaly detection, with particular emphasis on aerospace and industrial system monitoring. We work with NASA spacecraft telemetry data from SMAP (Soil Moisture Active Passive satellite) and MSL (Mars Science Laboratory rover) missions, as well as the Server Machine Dataset (SMD) containing 5 weeks of infrastructure metrics from 28 machines at a large internet company. In production server environments, entities log continuous Key Performance Indicators (KPIs) such as CPU utilization, memory load, disk I/O and network throughput, where anomalies typically arise from hardware degradation, resource contention, configuration errors or unexpected workload shifts manifesting across multiple correlated KPIs.

Anomaly detection in these systems presents challenges including multivariate dependencies where sensors interact and anomalies manifest across multiple dimensions, temporal patterns requiring both short and long-term dependency modeling, imbalanced data with rare anomaly events (4-13% of observations), and the need for unsupervised learning as training data lacks anomaly labels. Accurate monitoring contributes to operational reliability, cost reduction, and sustainability by avoiding wasted resources.

---

3 Data Understanding

We work with publicly available benchmark datasets commonly used in anomaly detection research. Table 1 summarizes the key characteristics of our target datasets.

Table 1: Dataset Summary Statistics

Dataset	Entities	Dimensions	Train Size	Anomaly %
SMAP	55	25	135,183	13.13%
MSL	27	55	58,317	10.72%
SMD	28	38	708,405	4.16%

The SMAP and MSL datasets contain time-series telemetry from NASA spacecraft operations with expert-labeled anomalies indicating system faults. The SMD dataset provides server resource utilization metrics (CPU, memory, disk I/O, network) from 28 machines across 3 groups. These datasets require normalization to the $[0, 1]$ range for stable model training, and a sliding window approach (typically size 10-50) is used to capture temporal context. Anomalies range from subtle deviations close to normal patterns to obvious significant outliers.

Exploratory Analysis. Our preliminary analysis of the SMD dataset focused on identifying representative subsets by ranking machines by mean KPI variance to capture dynamic entities. Correlation analysis revealed clusters of highly correlated KPIs (e.g., CPU-memory, network in-out) with correlation coefficients $\rho > 0.95$, allowing us to retain one representative KPI per cluster to reduce redundancy (Figure 1). The Augmented Dickey-Fuller test confirmed stationarity ($p \approx 3 \times 10^{-22}$) across most KPI sequences, validating the use of per-KPI normalization and fixed-length sliding windows. Temporal analysis (Figure 2) shows clear periodic cycles with local irregularities suggesting diurnal workload trends, while KPI distributions (Figure 3) exhibit right-skewed, multimodal behavior with heavy tails and rare extreme peaks. These characteristics motivate the use of sequence-based models with attention mechanisms and robust thresholding methods such as Peak-Over-Threshold (POT).

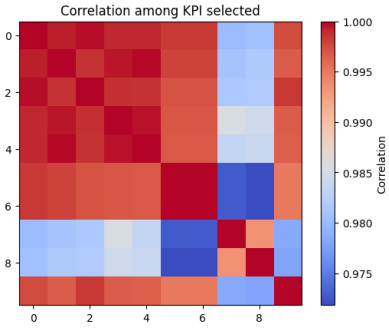

Figure 1: Correlation heatmap.

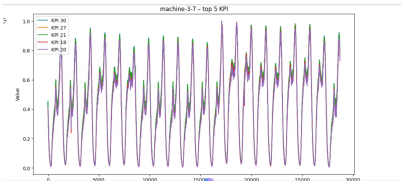

Figure 2: Temporal patterns.

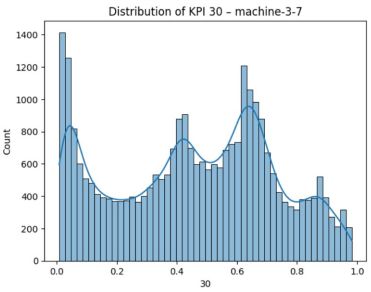

Figure 3: KPI distribution.

4 Project Goals

Our main goal is to study and apply transformer-based anomaly detection techniques to multivariate time series data while following best practices in the Knowledge Discovery process. We aim to understand state-of-the-art approaches including transformer architectures with attention mechanisms, reconstruction-based anomaly scoring, and unsupervised learning methodologies as presented in recent works by Su et al. [1] and Tuli et al. [2].

From a data mining perspective, this project focuses on two key tasks: anomaly detection as binary classification at the timestamp level, and anomaly diagnosis as multi-label classification to identify which specific dimensions exhibit anomalous behavior. These tasks provide excellent learning opportunities for understanding both time series analysis and multi-class classification problems in an unsupervised context.

Our approach emphasizes learning and incremental development rather than immediate complexity. We plan to start with a single dataset (likely SMD as it is readily available) and implement a simple baseline model such as an LSTM-based autoencoder or basic reconstruction model. From this foundation, we will progressively build toward more sophisticated transformer-based architectures as time and understanding permit. Throughout the process, we will maintain proper experimental validation using train/validation/test splits and appropriate metrics (Precision, Recall, F1-score, AUC-ROC) while comparing results against baseline methods.

This project plan remains flexible and open to refinement based on preliminary results and feedback from professor. We prioritize understanding the theoretical foundations and implementing a rigorous but manageable experimental pipeline over attempting to replicate overly complex state-of-the-art systems.

References

[1] Ya Su et al. "Robust Anomaly Detection for Multivariate Time Series through Stochastic Recurrent Neural Network". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (2019). URL: https://api.semanticscholar.org/CorpusID:196175745.

[2] Shreshth Tuli, Giuliano Casale, and Nicholas R. Jennings. *TranAD: Deep Transformer Networks for Anomaly Detection in Multivariate Time Series Data*. 2022. arXiv: 2201.07284 [cs.LG]. URL: https://arxiv.org/abs/2201.07284.