Explaining Temporal Data in Manufacturing using LLMs and Markov Chains

Jan Šturm
jan.sturm@ijs.si
Jožef Stefan Institute
Jožef Stefan International
Postgraduate School
Ljubljana, Slovenia

Maja Škrjanc
maja.skrjanc@ijs.si
Jožef Stefan Institute
Jožef Stefan International
Postgraduate School
Ljubljana, Slovenia

Oleksandra Topal
oleksandra.topal@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Inna Novalija
inna.koval@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Dunja Mladenić
dunja.mladenic@ijs.si
Jožef Stefan Institute
Jožef Stefan International
Postgraduate School
Ljubljana, Slovenia

Marko Grobelnik
marko.grobelnik@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

Monitoring and understanding complex industrial processes from high-dimensional IoT sensor data remains a significant challenge. While advanced modeling techniques like Hierarchical Markov Chains can abstract raw data, their outputs are often difficult for domain experts to interpret, creating a gap between data-driven insights and operational management. Existing explainability methods often focus on feature importance rather than providing holistic, semantic descriptions of system states. This paper introduces a framework that bridges this gap by transforming the abstract states of a process model into intuitive, human-readable concepts. The methodology leverages the StreamStory (Hierarchical Markov Chain) tool approach to generate behavioral profiles based on log-likelihood calculations within sliding temporal windows. StreamStory states are summarized using an LLM to assign semantic labels and descriptions. This approach reduces the initial reliance on domain experts for analysis, aids the understanding of complex system dynamics, and provides a transparent foundation for identifying both normal and anomalous operational patterns. The result is a more interpretable representation of industrial processes, facilitating improved predictive maintenance and operational efficiency.

Keywords

Multivariate Timeseries, Explainable AI, LLMs, Markov Chains

1 Introduction

The proliferation of Internet of Things (IoT) sensors in industrial environments has generated vast streams of multivariate time-series data. While this data holds immense potential for process optimization and predictive maintenance, its complexity often surpasses human cognitive capacity. Tools like StreamStory [6] have emerged to model these complex systems using Hierarchical Markov Chains, abstracting raw data into a more manageable set of states and transitions. However, a fundamental challenge persists: a disconnect between the model's statistical outputs and the experiential knowledge of domain experts.

The motivation for this work stems from this challenge. Domain experts, who possess invaluable implicit knowledge of a system, often struggle to interpret the statistical outputs of process models. Conversely, data scientists may identify patterns that lack the necessary operational context for effective action. Presenting experts with a graphical representation of states and transitions is a step forward, but it does not fully bridge the semantic gap. They may not understand what a specific state represents in the physical world or why a particular transition is significant. This leads to a bottleneck where valuable data-driven insights are not fully utilized, hindering efforts to improve system management and efficiency.

To address this, the paper proposes a methodology that enhances the interpretability of hierarchical process models. This approach creates a new layer of understanding that is accessible to operational personnel without requiring deep data science expertise. By translating abstract model states into meaningful, semantically rich descriptions, it provides a tool that allows the system's behavior to be understood, validated, and ultimately, better managed. This work introduces a methodology to automatically generate these descriptions, moving from complex data to clear, actionable insights.

2 Related Work

The field of time-series anomaly detection has evolved from interpretable statistical models like ARIMA and classical machine learning such as Isolation Forest to high-performance deep learning architectures including LSTMs, Transformers, and Autoencoders [5, 4, 7]. While these advanced models excel at pattern recognition, their complexity necessitates post-hoc XAI tools like LIME and SHAP to explain their decisions, which are limited to providing low-level feature attributions [1].

Recent work also demonstrates the utility of Hidden Markov Models (HMMs) for anomaly detection, for instance, by designing active search strategies to locate an evolving anomaly among multiple processes [2], or by learning normal temporal dynamics from remote sensing data to detect, localize, and classify crop-related deviations [3]. However, while effective for detection, the abstract nature of HMM states can be difficult for domain experts to interpret. The present work addresses this by transforming the state sequence into a multi-scale behavioral profile, which

Information Society 2025, 7–11 October 2025, Ljubljana, Slovenia

enables a Large Language Model (LLM) to generate rich, semantic explanations of system behavior.

This approach innovates by first classifying each multivariate data point into a state within a pre-built Markov Chain model and then calculating log-likelihoods from the state sequence to form a multi-scale representation. Crucially, this representation allows for the recognition of regular system behavior and various anomalies. By analyzing the statistical distribution of these profiles—identifying dense regions of regular behavior and sparse outliers corresponding to anomalous states—an LLM can then assign rich, human-readable descriptions, connecting abstract data to operational knowledge.

3 Methodology

The framework is designed to post-process models generated by the StreamStory system. Figure 1 outlines this multi-stage process, which begins with the statistical features from the Markov model and culminates in semantically enriched explanations of system behavior. The core of this methodology is the transformation of abstract machine states into meaningful concepts using a combination of statistical feature engineering and LLM interpretation. The process focuses on creating robust representations of system behavior and leveraging an LLM to translate these representations into human-understandable language.

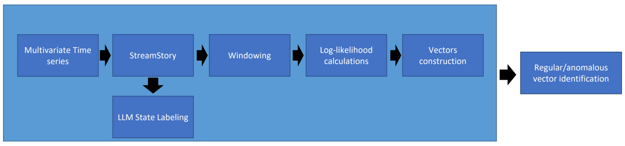

Figure 1: Proposed methodology for identifying and explaining normal and anomalous operational profiles.

3.1 Log-Likelihood Score Calculation

The input to the pipeline is a pre-existing Hierarchical Markov Chain model of an industrial process, which includes a history of state transitions over time. The first step is to create a rich feature representation that captures the system's dynamics. A sliding window (Figure 2) approach moves across the sequence of historical state transitions. For each window of a given size, a single feature is calculated: the log-likelihood of that specific sequence of transitions occurring. This score is calculated by summing the log-transformed transition probabilities for each step in the sequence, as defined by the underlying Markov model. The score effectively quantifies how "normal" or "expected" a particular sequence of behavior is according to the learned model. Highly probable sequences yield higher log-likelihood scores (closer to zero), while rare sequences result in large negative scores.

3.2 Behavior Profile Construction

To capture dynamics over multiple time scales, several sliding windows of different sizes are used simultaneously. The log-likelihood score calculated from each window is concatenated to form a single feature vector for each time step. This multi-scale vector, termed a *behavior profile*, serves as a rich representation of the system's dynamics at that moment, encapsulating both short-term and longer-term patterns. This profile is a crucial output, as it provides a quantitative basis for distinguishing between different modes of operation.

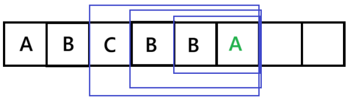

Figure 2: Sliding windows of different lengths are applied to a sequence of system states over time. A log-likelihood score is then calculated for the sequence within each window.

3.3 Ranking System Behavior via Anomaly Scoring

Following the construction of the behavior profiles, their distribution is analyzed to identify distinct operational patterns. An unsupervised density-based approach is employed to score each profile's typicality. The Isolation Forest algorithm is used for this purpose because it does not assume a specific data distribution and excels at identifying outliers in a high-dimensional space. Profiles that are common and lie in dense regions of the feature space receive a high score, corresponding to *normal* behavior. Conversely, profiles that are rare and isolated receive a low score, flagging them as *anomalous*. This produces a continuous spectrum of normalcy, allowing for a ranked analysis of all operational events.

3.4 LLM-Powered State Naming and Interpretation

To translate abstract states into meaningful concepts, an LLM is utilized. For each granular state discovered by the StreamStory model, its statistical profile (e.g., sensor value distributions) and context about the machine type were formatted into a descriptive prompt. The LLM was then tasked with generating a concise, intuitive name for each state (e.g., "Peak Production - High Flow and Heat"). This process, conducted once per model, creates a semantic layer that is then used to interpret the sequences associated with the highest-ranked normal and lowest-ranked anomalous events.

This approach offers two key advantages. First, the LLM-generated names provide a layer of transparency, offering an immediate hypothesis about what each abstract state represents. Second, it shifts the role of the domain expert from the arduous task of initial interpretation to the more efficient step of validating or refining the LLM-generated labels, accelerating the process of gaining actionable insights.

4 Experiment

To validate the proposed framework, an experiment was conducted using a real-world industrial dataset from an oil refinery pump. This section details the dataset, implementation, and results.

4.1 Dataset

The experiment was conducted on a proprietary dataset from the oil and gas sector, spanning approximately one month of operation (March-April 2017) with a 15-minute data resolution. The multivariate time-series data was collected from a suite of IoT sensors monitoring a pump's core functions. Key measurements included fluid flow rate (Kg/h), suction and discharge pressure

(Kg/cm2), and temperatures monitoring the process fluid and mechanical components (degC).

4.2 Implementation Details

The methodology was implemented in a Python environment. The underlying Markov Chain model was built using the entire historical dataset provided, as the goal is to interpret the complete, learned dynamics of the process rather than to perform a predictive task that would require a train/test split. Behavior profiles were constructed using sliding windows of multiple sizes (3, 5, 7, and 10 steps). The resulting profiles were analyzed using the Scikit-learn implementation of Isolation Forest. The ‘contamination‘ parameter was set to 5% for the primary analysis, a common heuristic for industrial processes. State descriptions were generated using the GPT-4o model, which was prompted with the statistical profiles of each state to generate intuitive names.

4.3 Experimental Results and Discussion

The application of the framework yielded a ranked list of operational events, characterized by the Isolation Forest decision score. This score serves as a robust indicator of how typical or anomalous a given time window is. Table 1 details the top five most anomalous events identified. These events are characterized by scores that are more than 3 standard deviations below the mean, signifying extreme statistical rarity.

The true explanatory power of the method is revealed when the abstract state sequences are translated into their LLM-generated names. For instance, the most anomalous event culminates in a sequence of "... -> ‘Startup or Shutdown Transition‘ -> ‘Machine Idle or Shutdown‘ -> ‘Startup or Shutdown Transition‘." This provides a clear, human-readable narrative of the pump entering a period of instability and stoppage. This is a marked improvement over black-box models that simply flag a time point as anomalous without providing a temporal context for the "why." An engineer, seeing this semantic sequence, can immediately infer a potential cause for investigation, such as an attempted restart or a stuttering shutdown process.

Table 1: Top 5 Most Anomalous Events

Rank	Timestamp	Score (Std.)	Final State (LLM Name)
1	2017-04-03 14:30	-0.096 (-3.88)	Startup...Transition
2	2017-03-28 10:00	-0.071 (-3.45)	Startup...Transition
3	2017-03-30 00:00	-0.066 (-3.35)	High-Flow, Cool Op.
4	2017-04-03 12:30	-0.061 (-3.26)	Machine Idle
5	2017-04-03 15:00	-0.056 (-3.18)	Weekday Low-Flow...

Conversely, Table 2 presents the five most normal events, which have high positive scores. Their sequences reveal a stable operational loop between states like "Peak Production," "Weekend Peak-Load Production," and "Extreme Temperature Peak Performance." This recurring pattern defines the pump's healthy operational "heartbeat," providing a data-driven "golden standard" for normal behavior under demanding conditions. This semantic understanding is crucial for operators, as it validates that the system is performing as expected.

To ensure the robustness of the findings, a sensitivity analysis was conducted on the Isolation Forest ‘contamination‘ parameter, testing values of 1%, 5%, and 10%. While the number of points labeled ’Anomalous’ changed as expected, the relative ranking of the most extreme events remained highly consistent, confirming that the core findings are not sensitive to this hyperparameter.

Table 2: Top 5 Most Normal Events

Rank	Timestamp	Score (Std.)	Final State (LLM Name)
1	2017-03-23 22:00	0.192 (1.22)	Weekend Peak-Load
2	2017-03-31 06:00	0.192 (1.22)	Peak Production
3	2017-04-01 00:00	0.191 (1.20)	Peak Production
4	2017-03-31 23:30	0.191 (1.19)	Weekday Peak Perf.
5	2017-03-31 07:30	0.190 (1.17)	Weekday Peak Perf.

The claims in this paper are demonstrated on a single, representative dataset. While the framework is designed to be general, further studies on diverse industrial processes are required to fully validate its broader applicability. The LLM-generated labels were not validated in a formal user study with domain experts; such a study is a valuable next step.

5 Conclusion

This paper presented a complete, self-contained framework for increasing the interpretability of complex industrial process models. By creating behavior profiles of system states and using an LLM to assign semantic names, the approach successfully translates abstract data analysis into practical domain knowledge. The method provides a robust process for ranking and explaining individual operational events in a transparent manner, as demonstrated on a real-world industrial dataset. This work establishes a strong foundation for a new type of explainability, moving beyond feature importance to provide narrative, context-rich descriptions of system dynamics.

The representation of system dynamics as behavior profiles opens a wide array of possibilities for future research. The current work successfully identifies and presents the raw temporal sequences leading to key events. Future work will focus on applying formal pattern mining techniques to automatically discover recurring and significant sequential patterns within these events. Such an analysis could reveal if distinct "families" of anomalous behavior exist, each with its own characteristic temporal signature. This promises a more nuanced description of system operations and provides a stronger foundation for developing targeted predictive maintenance strategies. Finally, to address current limitations, two key areas will be prioritized. First, formal user studies with domain experts will be conducted to validate the utility and accuracy of the LLM-generated explanations, moving beyond the promising initial results. Second, the framework's generalizability will be tested through broader empirical evaluation across diverse industrial sectors and sensor types to boost its credibility and applicability.

6 Acknowledgments

This work was supported by the Slovenian Research Agency and the European Union's Horizon 2020 project FAME (Grant No. 101092639).

References

[1] Liat Antwarg, Ronnie Mindlin Miller, Bracha Shapira, and Lior Rokach. 2019. Explaining anomalies detected by autoencoders using shap. *arXiv preprint arXiv:1903.02407*.

[2] Levli Citron, Kobi Cohen, and Qing Zhao. 2025. Searching for a hidden markov anomaly over multiple processes. *arXiv preprint arXiv:2506.17108*.

[3] Kareth M Leon-Lopez, Florian Mouret, Henry Arguello, and Jean-Yves Tourneret. 2021. Anomaly detection and classification in multispectral time series based on hidden markov models. *IEEE transactions on geoscience and remote sensing*, 60, 1–11.

[4] Sebastian Schmidl, Phillip Wenig, and Thorsten Papenbrock. 2022. Anomaly detection in time series: a comprehensive evaluation. *Proceedings of the VLDB Endowment*, 15, 9, 1779–1797.

[5] Charalampos Shimillas, Kleanthis Malialis, Konstantinos Fokianos, and Marios M Polycarpou. 2025. Transformer-based multivariate time series anomaly localization. In *2025 IEEE Symposium on Computational Intelligence on Engineering/Cyber Physical Systems (CIES)*. IEEE, 1–8.

[6] Luka Stopar, Primoz Skraba, Marko Grobelnik, and Dunja Mladenic. 2018. Streamstory: exploring multivariate time series on multiple scales. *IEEE transactions on visualization and computer graphics*, 25, 4, 1788–1802.

[7] Fengling Wang, Yiyue Jiang, Rongjie Zhang, Aimin Wei, Jingming Xie, and Xiongwen Pang. 2025. A survey of deep anomaly detection in multivariate time series: taxonomy, applications, and directions. *Sensors (Basel, Switzerland)*, 25, 1, 190.