Transformer-Based Anomaly Detection for Mobile Robots

Davide Villaboni[1,2] Francesco Bazzani[1] Alberto Castellini[1] Alessandro Farinelli[1]

2025 European Conference on Mobile Robots (ECMR) | 979-8-3315-2705-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ECMR65884.2025.11163081

***Abstract*— Anomaly detection and model interpretation are key components for robots deployed in safety-critical scenarios. In this paper, we propose to use Sentinel —a Transformer-based architecture for multivariate time series forecasting —to improve anomaly detection performance for mobile robots, and we investigate whether the model's attention mechanisms faithfully reflect the underlying statistical structure of the data. Our results on the ALFA dataset (a widely used aerospace benchmark) demonstrate that Sentinel achieves good anomaly detection performance when compared to state-of-the-art approaches. Moreover, the empirical evaluation shows that sentinel's attention mechanisms capture relevant dependencies among the features hence offering key insight for early warning indicators. These findings highlight the potential of attention-based interpretability in complex, sensor-rich robotic environments and pave the way towards explainable and resilient anomaly detection frameworks.**

I. Introduction

Anomaly detection is a critical task in modern artificial intelligence, enabling the identification of patterns that deviate from expected norms. This capability is especially important in safety-critical systems, such as autonomous vehicles, where timely detection of faults can prevent failures and improve system reliability.

Mobile robots, particularly Unmanned Aerial Vehicles (UAVs), are increasingly deployed in complex and dynamic environments. These systems generate large volumes of multivariate time series data from sensors and control systems, making manual rule-based monitoring impractical. An effective anomaly detection method for such systems must not only identify deviations with high accuracy and low latency but also offer interpretable outputs to support human decision-making. In complex industrial systems, simply identifying operational deviations is insufficient; precise localization of anomalous components facilitates the implementation of specific corrective measures.

Traditional approaches—ranging from statistical models [1], [2] to autoencoders and recurrent networks—have shown promise, but face limitations. Autoencoder-based methods, while effective in capturing the structure of normal data, often lack interpretability; Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks (LSTMs) [3], widely used for sequential data such as time series, struggle to process long sequences and capture long-term dependencies. These limitations affect performance in complex, high-dimensional scenarios, where classical statistical and machine learning methods often perform suboptimally..

[1] Computer Science Department, University of Verona davide.villaboni@univr.it
[2] UniCredit davide.villaboni@unicredit.eu

The Transformer architecture, introduced by [4], has emerged as a powerful alternative for modeling sequential data. Initially developed for natural language processing, its architecture, based on attention mechanisms, enables efficient parallelization and excels at capturing long-term dependencies. These properties have made Transformers increasingly popular for time series modeling tasks [5], [6].

Building on this, the Sentinel architecture [7] extends the Transformer framework specifically for multivariate time series forecasting and anomaly detection. Sentinel introduces a multi-patch attention mechanism that structures temporal and inter-channel dependencies more effectively than standard attention. Its encoder captures spatial (sensor/channel-wise) patterns, while the decoder focuses on temporal dynamics, making it well suited to model complex robotic systems.

In this work, we propose and evaluate the use of Sentinel for anomaly detection in mobile robotics applications. Specifically, we focus on the ALFA dataset [8], which contains sensor logs from flights under various fault conditions. We also propose a statistical framework to analyze Sentinel's attention maps, enhancing the explainability and aiding in the localization of failure sources within the system.

The main contributions of this paper can be summarized as:

- Application of the Sentinel Transformer model for UAV anomaly detection, in particular to the ALFA dataset, demonstrating competitive performance in identifying diverse fault conditions.
- Introduction of a statistical framework for analyzing attention maps to provide interpretable insights and infer causal relationships between anomalous behaviors and system components.

II. Related Work

The Transformer architecture, introduced by [4] marked a pivotal shift in sequential data processing. Prior to this, recurrent neural networks (RNNs), such as Long Short-Term Memory (LSTM) [9] and Gated Recurrent Units (GRU) [10] were the dominant models for sequence-to-sequence tasks. These architectures processed inputs sequentially, maintaining a hidden state that carried information through time steps. While effective, RNNs suffered from inherent limitations: they struggled with long-range dependencies due to the vanishing gradient problem [11], and their sequential nature hindered parallelization during training.

The Transformer architecture addressed these limitations by completely removing recurrence and relying instead on attention mechanisms to capture dependencies within the sequence. Their self-attention mechanism allows for efficient

parallel computation and robust long-term temporal modeling, driving their adoption across various time series tasks including forecasting [12], [13], classification [14], [13], [15], and anomaly detection [16], [17].

Several Transformer-based models have been proposed for anomaly detection in multivariate time series. TranAD [16] integrates adversarial training to improve detection accuracy; the Anomaly Transformer [18] adapts the Transformer architecture for unsupervised anomaly detection by explicitly modeling relationships between time points; InterFusion [19] that employs hierarchical Variational Auto-Encoder with explicit low-dimensional inter-metric and temporal embeddings to jointly learn robust multi-variate representations, proposing also a novel anomaly interpretation method; Autoformer [6] replaces traditional self-attention with an auto-correlation mechanism to better exploit time series periodicity; FEDformer [20] operates in the frequency domain and incorporates seasonal-trend decomposition to improve efficiency and accuracy; TimesNet [5] introduces a novel perspective on time series analysis by transforming one-dimensional temporal data into two-dimensional representations; Sentinel [7], is a fully Transformer-based architecture designed for multivariate time series forecasting and anomaly detection. It consists of an encoder that extracts contextual information across both temporal and channel dimensions, and a decoder that captures causal dependencies over time. A key innovation of Sentinel is the introduction of a multi-patch attention mechanism, which restructures the input sequence through patching instead of the standard multi-head division, allowing better integration of long-range and inter-channel dependencies.

Given Sentinel's capacity to model complex, long-horizon multivariate sequences and its interpretability through attention mechanisms, we adopt it as the core architecture for our anomaly detection framework in this work.

III. Transformer for anomaly detection

The time series anomaly detection task can be formally defined as follows: given a multivariate time series $\mathbf{X} = \{x_1, x_2, \ldots, x_L\} \in \mathbb{R}^{L \times C}$, where L represents the sequence length and C the number of channels or features, the objective is to identify anomalous points or segments within the series. These anomalies are defined as observations that deviate significantly from the predicted normal behavior of the system.

The anomaly detection approach follows a reconstruction-based paradigm, which is particularly effective for multivariate time series where complex inter-channel dependencies exist. This approach can be formalized as follows:

1) A model $f_\theta(\cdot)$ is trained to reconstruct normal patterns in the time series data.
2) For a given segment $\mathbf{X}_t = \{x_{t-L+1}, \ldots, x_t\}$, the model predicts a reconstruction $\hat{\mathbf{X}}_t = f_\theta(\mathbf{X}_t)$ based on this historical context.
3) A reconstruction error e_t is computed using a predefined distance function $d(\cdot, \cdot)$:

$$e_t = d(\mathbf{X_t}, \hat{\mathbf{X}}_t) \tag{1}$$

We use the mean squared error (MSE) across all features as the distance metric.

4) An anomaly score s_t is derived, averaging the reconstruction error across all channels C and time steps L:

$$s_t = \frac{1}{L \cdot C} \sum_{i=t-L+1}^{t} \sum_{j=1}^{C} (x_{i,j} - \hat{x}_{i,j})^2 \tag{2}$$

5) By applying a threshold τ to the anomaly score, each time point is classified as normal or anomalous:

$$y_t = \begin{cases} 1 & \text{if } s_t > \tau \text{ (anomaly)} \\ 0 & \text{otherwise (normal)} \end{cases} \tag{3}$$

The core intuition behind this method is that a model trained exclusively on normal data will accurately reconstruct known patterns but poorly reconstruct unseen anomalous ones, resulting in higher reconstruction errors for anomalies.

A. Sentinel Architecture

Sentinel employs an encoder-decoder structure similar to the original Transformer, but with fundamental modifications tailored to time series analysis. The architecture processes input sequences of multivariate time series data $X \in \mathbb{R}^{L \times C}$, where L represents the length of the look-back window and C represents the number of channels or features. The model is designed to capture both temporal relationships and inter-channel dependencies, which are crucial for accurate time series forecasting and anomaly detection.

Unlike the original Transformer, which was designed primarily for natural language processing tasks, Sentinel incorporates several key innovations:

- A **Patching mechanism** that reorganize time series data into overlapping segments
- **Multi-patch attention** that replaces traditional multi-head attention
- **Specialized encoder** focused on capturing inter-channel dependencies
- **Specialized decoder** focused on temporal relationships

IV. Attention Map analysis

The self-attention mechanism in the Sentinel encoder provides valuable insights into how different features interact across time. In the context of anomaly detection, analyzing changes in attention patterns between normal and anomalous segments can reveal which feature relationships shift significantly under failure conditions.

The encoder self-attention in Sentinel reveals how different features interact within the time series. By comparing these patterns across normal and failure conditions, we can identify shifts in inter-feature dependencies that signal anomalies; the decoder self-attention models temporal dependencies, while the cross-attention mechanism links temporal queries from the decoder to contextual embeddings from the encoder,

enabling the model to integrate feature-wise context with temporal patterns.

To effectively analyze attention patterns, attention weights are treated as part of a distribution and robust statistical methods are employed to identify significant differences between normal and anomaly states. Two key statistical measures form the foundation of this approach: the Mann-Whitney U test and Cohen's d effect size.

a) Mann-Whitney U Test: Unlike parametric tests such as the t-test, the Mann-Whitney U Test does not require the assumption of normally distributed data, making it particularly suitable for analyzing attention weights which often follow complex, non-Gaussian distributions. The Mann-Whitney U Test has been applied to analyze attention maps in [21].

Given two samples X and Y of sizes n and m respectively, the Mann-Whitney U statistic is calculated by:

$$U = \sum_{i=1}^{n} \sum_{j=1}^{m} S(x_i, y_j) \tag{4}$$

where $S(x_i, y_j) = 1$ if $x_i > y_j$; 0.5 if $x_i = y_j$; and 0 if $x_i < y_j$. The corresponding p-value is derived from the distribution of U under the null hypothesis that the two samples come from the same distribution. A low p-value (e.g., $p < 0.05$) indicates that the two distributions differ significantly, suggesting a shift in attention dynamics between normal and anomalous states.

b) Cohen's d: While the Mann-Whitney U Test determines whether distributions differ, it does not quantify the *magnitude* of the difference. For this reason, we use Cohen's d, which standardizes the mean difference between two groups:

$$d = \frac{\bar{X}_1 - \bar{X}_2}{s_p} \tag{5}$$

where $\bar{X}_1$ and $\bar{X}_2$ are the means of the two groups, and s_p is the pooled standard deviation, calculated as:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}} \tag{6}$$

where n_1 and n_2 are the sample sizes, and s_1 and s_2 are the standard deviations of the two groups.

Cohen's guidelines for interpreting the magnitude of d:

- $|d| < 0.2$: Negligible effect
- $0.2 \leq |d| < 0.5$: Small effect
- $0.5 \leq |d| < 0.8$: Medium effect
- $|d| \geq 0.8$: Large effect

Combining these two measures enables the identification of statistically and practically meaningful differences in feature interactions during anomalous states.

c) Attention Pattern Analysis Methodology: The proposed approach treats the attention weights as distributions, where, in the case of the encoder, they reflect the relationships between different features in the input time series. By comparing these distributions between normal and failure states, we aim to identify relationships that change significantly during anomalies—potentially revealing diagnostic patterns or causal structures.

The core idea is that self-attention weights represent how feature j influences feature i. We compare the attention distributions across samples from normal and failure classes. If these relationships change consistently between normal and failure conditions (e.g., becoming weaker or absent) they can provide valuable insight into the nature of the anomaly.

d) Extracting Attention Distributions: Let $\mathbf{A}^{\text{normal}} \in \mathbb{R}^{B\times P\times C\times C}$ and $\mathbf{A}^{\text{failure}} \in \mathbb{R}^{B\times P\times C\times C}$ represent the attention maps for normal and failure states respectively. Here, $\mathbf{W}_{attn} \in \mathbb{R}^{P\times C\times C}$ denotes the attention weights, where B is the batch size, P is the number of patches, C is the number of channels (features). For each patch $p \in \{1, 2, \ldots, P\}$ and each feature pair (i, j) where $i, j \in \{1, 2, \ldots, C\}$, two distributions are extracted:

- $D^{\text{normal}}_{p,i,j} = \{\mathbf{A}^{\text{normal}}_{b,p,i,j} \mid b \in \{1, 2, \ldots, B\}\}$
- $D^{\text{failure}}_{p,i,j} = \{\mathbf{A}^{\text{failure}}_{b,p,i,j} \mid b \in \{1, 2, \ldots, B\}\}$

These distributions represent the attention weights from feature i to feature j at patch p across all samples in the batch, for normal and failure states respectively.

V. ALFA DATASET

The Air Lab Fault and Anomaly (ALFA) dataset [8] provides comprehensive real-world flight data for evaluating Fault Detection and Isolation (FDI) and Anomaly Detection (AD) algorithms in Unmanned Aerial Vehicles (UAVs). The dataset was collected using a modified Carbon Z T-28 fixed-wing UAV platform equipped with a Holybro PX4 2.4.6 autopilot, GPS module, Pitot tube, and Nvidia Jetson TX2 onboard computer. The processed portion of the dataset encompasses 47 autonomous flight sequences, incorporating 23 sudden full engine failure scenarios and 24 scenarios covering seven distinct control surface faults. The data captures 66 minutes of normal flight conditions and 13 minutes of post-fault flight time. Additionally, the dataset includes several hours of raw flight data from autonomous, autopilot-assisted, and manual flights containing multiple fault scenarios. The data is recorded through a modified MAVROS system that interfaces with the autopilot using the MAVLink protocol. Each sequence contains high-frequency sensor measurements (20-25 Hz) of critical flight parameters including roll, pitch, velocity, airspeed, and yaw, along with their corresponding autopilot commands. The dataset also provides GPS information, local and global state estimates, and wind measurements at varying frequencies between 1-50 Hz. A key feature of the dataset is the inclusion of precise ground truth information for fault occurrence times and types, enabling rigorous evaluation of detection methods. The fault scenarios encompass various control surface failures including engine power loss, rudder hardover, elevator malfunction, and aileron issues in different configurations. The controlled testing environment ensures safety while maintaining realistic flight conditions, as the implemented faults allow for recovery by the safety pilot.

This dataset addresses a critical gap in the field, as it represents one of the first extensive collections of real flight data with actuator faults, whereas most existing methods rely solely on simulation for validation. Its structure support research in fault detection and flight safety.

VI. Empirical Analysis

A. Experimental setting

a) Standard Datasets: Experiments were run on standard anomaly detection benchmarks: **SMAP**: The Soil Moisture Active Passive (SMAP) dataset from NASA provides multivariate telemetry data including sensor and instrument readings; **MSL**: The Mars Science Laboratory (MSL) dataset contains annotated telemetry from NASA's Curiosity rover; **SWaT**: The Secure Water Treatment (SWaT) dataset originates from a functional water treatment plant; **SMD**: The Server Machine Dataset (SMD) features real-world operational data from production servers, including CPU, memory, and I/O metrics; **PSM**: The Pooled Server Metrics (PSM) dataset, released by eBay, includes synthetic anomalies embedded in industrial time series data.

b) Baselines: To evaluate the performance of Sentinel, the following models were selected as benchmarks: Transformer [4], TimesNet [5], Autoformer [6], FEDformer [20]. The reasoning behind the usage of these models is that Transformer is the original transformer architecture, while TimesNet, Autoformer and FEDformer have been shown to rank the highest in terms of performance by [22]. It should be noted that TimesNet employs a CNN architecture.

c) Hardware: Experiments are run on a NVIDIA GeForce RTX 4070 Ti (12 GiB).

d) Settings: For each dataset multiple seeds have been used reporting the average f1-score. A patch size of 16 and a stride of 8 are used in all experiments. The dropout is set to 0.2 and AdamW [23] is used as optimizer with a learning rate of 0.001 and an L1 loss. For each dataset, various runs with a variable number of encoder layers $N_{\text{enc}} = 1, \ldots, 4$, decoder layers $N_{\text{dec}} = 1, \ldots, 4$, and different model dimensions: $d_{\text{model}} = \{16, 32, 64, 128, 256, 512\}$ were performed. The configuration yielding the best performance was selected. A fixed sequence length $L = 100$ and a prediction (reconstruction) length $L = 100$ were used to align with baseline performance in the literature.

B. Metrics on standard benchmarks

The F1-Score is the harmonic mean of precision and recall, and is one of the most important metrics to assess performance. When compared to the other benchmark models, Sentinel has the best (highest) average F1-Score in the analysed datasets, and is consistently shows either the best F1 score or is within 1% of the best F1-Score for each dataset.

Datasets	Models				
	Sentinel	**TimesNet**	**Transformer**	**Autoformer**	**FEDformer**
SMD	**89.16**	85.81	83.04	84.72	85.08
MSL	81.31	**85.15**	84.86	77.50	78.57
SMAP	**81.92**	70.85	71.52	71.09	70.76
SWaT	**94.72**	92.10	91.78	79.88	93.19
PSM	96.72	97.24	82.08	**97.29**	97.23
Average	**88.62**	86.23	82.57	82.08	84.97

TABLE I

Comparison of F1-scores (%) for anomaly detection across different models and datasets. Sentinel ranked first on average, proving state-of-the-art performance in anomaly detection tasks.

C. Attention maps on ALFA

This section will explore the results of the statistical analysis proposed in IV. The ALFA dataset presents 4 different types of failure: ***Engine Failure, Bilateral Aileron Failure, Left Aileron Failure, Right Aileron Failure***. In the next part Engine Failure will be analyzed, with the purpose of identifying emerging patterns that could indicate the failure.

a) Encoder Self-Attention Patterns Under Engine Failure Conditions: The encoder self-attention analysis, as shown in Figure 1, reveals distinctive patterns during engine failure conditions that align with previous statistical findings. This analysis examines how the model's attention mechanisms focus on specific feature interactions when processing engine failure data.

- **Velocity Measurement and Command Coupling**: The relationship between velocity measurements and their commanded values was found to be significantly enhanced during engine failure conditions. The *vfr_groundspeed* emerged as the most active feature during failure states, being involved in 147 patterns compared to only 86 in normal operational states. Strong attention patterns were also observed between *vfr_groundspeed* and various velocity components, as exemplified in Patch 7 where a notable relationship between *vfr_groundspeed* and *vel_des_y* was identified with 68.11% confidence and an effect size of 2.14.
- **Quaternion and Velocity Interdependencies** Quaternion components, particularly *quat_z* and *quat_w*, exhibited remarkably strong attention patterns with velocity components during failure conditions. In Patch 5, the relationship between *quat_z* and *vel_des_x* was characterized by a 69.35

b) Decoder Self-Attention Patterns Under Engine failure Conditions: Decoder attention exhibits temporal redistribution under failure scenarios, shifting focus toward earlier time steps—often indicative of the model prioritizing early signs of system degradation.

- **Temporal Shift in Attention**: A consistent pattern emerges across multiple features: during engine failure, attention shifts predominantly toward earlier temporal patches (particularly Patch_0 and Patch_1). This pattern is especially evident in *yaw_measured* and *quat_w*, where the strongest failure-state patterns connect later patches to earlier ones. This temporal redistribution of attention reflects the model's learned understanding that in engine failure scenarios, the initial conditions and early reactions become highly predictive of subse-

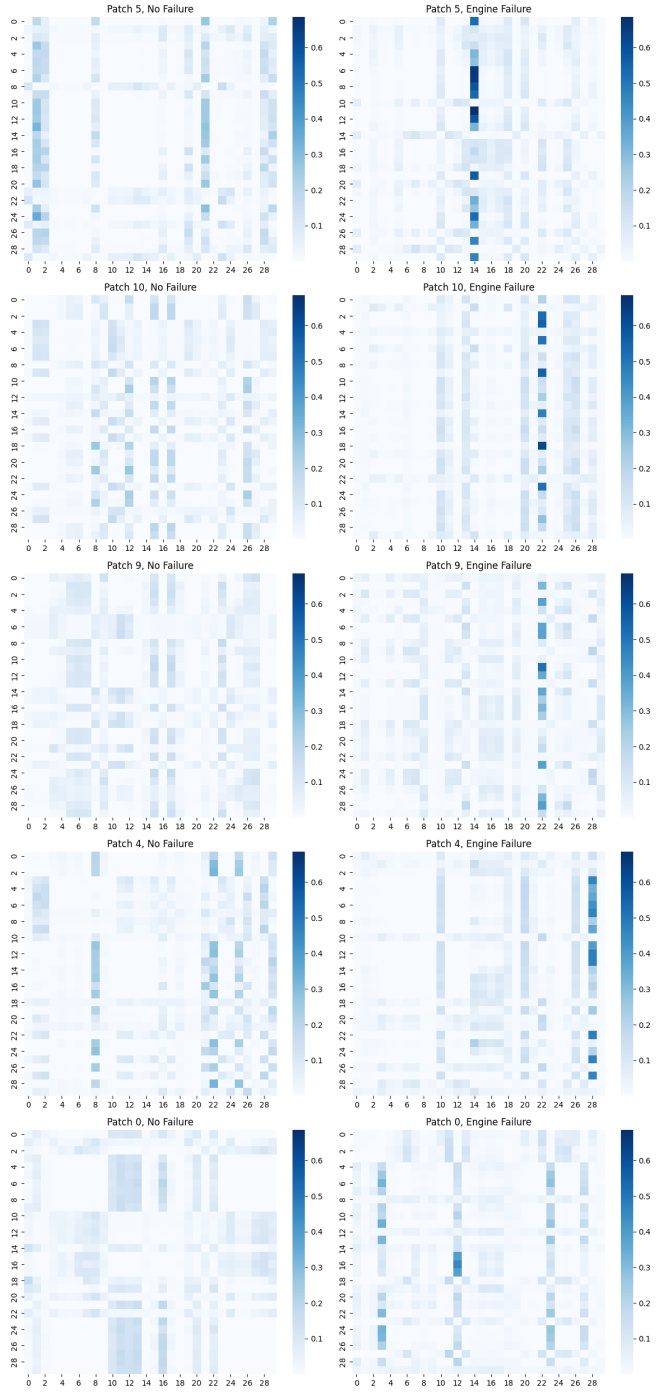

Fig. 1. Comparative analysis of the five most active patches in the Encoder Self-Attention layer under normal (left) and engine failure (right) scenarios. Each heatmap shows the attention weights between features, where darker color indicate stronger dependencies. Feature indices (0–29) correspond to sensor and control variables and remain consistent across both conditions. The comparison reveals how attention patterns shift under failure, highlighting changes in inter-feature relationships.

quent behavior. The flight control system likely exhibits characteristic responses immediately after engine failure detection, and these early indicators become crucial reference points.

- **Control System Response Signatures**: The presence of both *yaw_measured* and *yaw_commanded* among the top influential features reveals the importance of yaw control during engine failure. The attention patterns for these features exhibit a telling difference: while normal operation shows strong self-attention within the same temporal patches (e.g., Patch_3 $\rightarrow$ Patch_3), failure conditions redirect attention toward the beginning of the sequence (e.g., Patch_3 $\rightarrow$ Patch_0).
- **Integration of Attention Patterns with Statistical Measures**: The mutual information analysis revealed exceptionally high information sharing between velocity desired components (exceeding 4.0) during engine failure—the highest observed across all conditions. This aligns with the attention patterns for *wind_linear_x* and *vfr_airspeed*, which show significant redistribution during failure. The high mutual information reflects complex non-linear interdependencies, while the attention patterns reveal how these interdependencies manifest through specific temporal relationships. The model's attention to yaw dynamics during engine failure is particularly noteworthy when considered alongside the strong negative correlation (-0.880) between vertical velocity and climb rate. A possible explanation is that single-engine aircrafts experience asymmetric thrust loss, creating yaw moments as the aircraft rotates around its vertical axis. The attention patterns show the model has learned to focus on this physical relationship, identifying early yaw measurements as crucial predictors of subsequent flight behavior during engine failure. This effectively captures the causal chain where initial yaw disturbances propagate to affect vertical control, explaining the statistical correlation between vertical parameters.

c) Decoder Cross-Attention Patterns Under Engine failure Conditions: Decoder cross-attention captures direct links between temporal queries and encoded feature relationships. Under failure conditions, this mechanism highlights key inter-feature dependencies.

- **Velocity Component Relationships**: The decoder attention analysis identified *vel_des_x* and *vel_des_z* as significant features with identical pattern distributions. Both features exhibited 27 significant patterns each, with an average effect size of 1.00 and a maximum effect size of 1.55. This remarkable similarity in attention patterns corresponds directly with the statistical finding of perfect correlation (1.0) between these desired velocity components. In normal flight conditions, both features demonstrated similar attention patterns, particularly focusing on Patch_2 $\rightarrow$ Patch_6 connections (effect size: 1.10). However, during engine failure, attention shifted dramatically to Patch_10 $\rightarrow$ Patch_11 connections (effect size: 1.55), indicating a substantial redistribution of computational focus. This attention shift aligns with the exceptionally high mutual information values (exceeding 4.0) observed between these velocity components, confirming that the

model prioritizes information exchange between these parameters during failure conditions. The near-identical attention patterns confirm that the model processes these velocity components as a coordinated unit during engine failure compensation attempts, supporting the statistical evidence of their perfect correlation.

- **Vertical Control Mechanisms**: The negative correlation (-0.880) between vertical velocity and climb rate during engine failure is reflected in the attention patterns observed for *yaw_measured*, which emerged as the third most significant feature. Under failure conditions, *yaw_measured* exhibited increased attention between Patch_7 $\rightarrow$ Patch_1 (effect size: 1.41) and Patch_11 $\rightarrow$ Patch_7 (effect size: 1.32), a substantial shift from normal operations where Patch_6 $\rightarrow$ Patch_10 (effect size: 1.48) connections dominated. This redistribution of attention illustrates the model's adaptation to compromised vertical control dynamics. The attention flow toward lower-numbered patches (Patch_0 and Patch_1) during failure conditions suggests a computational focus on fundamental flight parameters, consistent with the statistical evidence of enhanced coupling between velocity measurements and commanded values.

VII. Conclusion

We explored transformer-based anomaly detection in robotic time series. A key aspect was the application to the ALFA dataset, which provides real UAV flight data under various fault scenarios. Unlike simulated benchmarks, ALFA enables realistic and reliable evaluation of detection methods. The model showed strong performance in identifying anomalies by leveraging long-range dependencies through self-attention. Attention map analysis offered valuable insights into the model's internal reasoning, aligning with statistical relevance and revealing complex, higher-order patterns. Results confirm the potential of transformers for accurate detection and interpretability in robotic contexts. Future work will explore comparisons with interpretability methods like SHAP[24], introduce detection-timeliness metrics, and extend the model to online adaptive settings.

Acknowledgments

This work has been supported by PNRR MUR project PE0000013-FAIR and PNRR iNEST project ECS00000043.

References

[1] D. Azzalini, A. Castellini, M. Luperto, A. Farinelli, and F. Amigoni, "HMMs for anomaly detection in autonomous robots," in *Proceedings of 2020 International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS 2020)*, 2020, p. 105–113.

[2] A. Castellini, F. Masillo, D. Azzalini, F. Amigoni, and A. Farinelli, "Adversarial data augmentation for hmm-based anomaly detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 12, pp. 14 131–14 143, 2023.

[3] K. Hundman, V. Constantinou, C. Laporte, I. Colwell, and T. Soderstrom, "Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding," in *Proceedings of the 24th ACM SIGKDD*, 2018, pp. 387–395.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[5] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long, "Timesnet: Temporal 2d-variation modeling for general time series analysis," in *International Conference on Learning Representations*, 2023.

[6] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," in *Advances in Neural Information Processing Systems*, 2021.

[7] D. Villaboni, A. Castellini, I. L. Danesi, and A. Farinelli, "Sentinel: Multi-patch transformer with temporal and channel attention for time series forecasting," *arXiv preprint arXiv:2503.17658*, 2025.

[8] A. Keipour, M. Mousaei, and S. Scherer, "Alfa: A dataset for uav fault and anomaly detection," *The International Journal of Robotics Research*, vol. 40, no. 2-3, pp. 515–520, 2021.

[9] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, conference Name: Neural Computation. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6795963

[10] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[11] Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, Mar. 1994, conference Name: IEEE Transactions on Neural Networks. [Online]. Available: https://ieeexplore.ieee.org/document/279181

[12] S. Li, X. Jin, Y. Xuan, X. Zhou, W. Chen, Y.-X. Wang, and X. Yan, "Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting," *Advances in neural information processing systems*, vol. 32, 2019.

[13] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 12, 2021, pp. 11 106–11 115.

[14] G. Zerveas, S. Jayaraman, D. Patel, A. Bhamidipaty, and C. Eickhoff, "A transformer-based framework for multivariate time series representation learning," in *Proceedings of the 27th ACM SIGKDD*, 2021, pp. 2114–2124.

[15] Y. Liu, H. Wu, J. Wang, and M. Long, "Non-stationary transformers: Exploring the stationarity in time series forecasting," *Advances in neural information processing systems*, vol. 35, pp. 9881–9893, 2022.

[16] S. Tuli, G. Casale, and N. R. Jennings, "Tranad: Deep transformer networks for anomaly detection in multivariate time series data," *arXiv preprint arXiv:2201.07284*, 2022.

[17] X. Wang, D. Pi, X. Zhang, H. Liu, and C. Guo, "Variational transformer-based anomaly detection approach for multivariate time series," *Measurement*, vol. 191, p. 110791, 2022.

[18] J. Xu, H. Wu, J. Wang, and M. Long, "Anomaly transformer: Time series anomaly detection with association discrepancy," *arXiv preprint arXiv:2110.02642*, 2021.

[19] Z. Li, Y. Zhao, J. Han, Y. Su, R. Jiao, X. Wen, and D. Pei, "Multivariate time series anomaly detection and interpretation using hierarchical inter-metric and temporal embedding," in *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, 2021, pp. 3220–3230.

[20] T. Zhou, Z. Ma, Q. Wen, X. Wang, L. Sun, and R. Jin, "Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting," in *International conference on machine learning*. PMLR, 2022, pp. 27 268–27 286.

[21] V. Ruscio, V. Maiorca, and F. Silvestri, "Attention-likelihood relationship in transformers," 2023. [Online]. Available: https://arxiv.org/abs/2303.08288

[22] Y. Wang, H. Wu, J. Dong, Y. Liu, M. Long, and J. Wang, "Deep time series models: A comprehensive survey and benchmark," *arXiv preprint arXiv:2407.13278*, 2024.

[23] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[24] E. Mosca, F. Szigeti, S. Tragianni, D. Gallagher, and G. Groh, "Shap-based explanation methods: a review for nlp interpretability," in *Proceedings of the 29th international conference on computational linguistics*, 2022, pp. 4593–4603.