THERMALFAULT: A SESSION-DISJOINT DATASET AND BENCHMARK FOR ONLINE MOTOR FAULT MONITORING

Anonymous authors
Paper under double-blind review

ABSTRACT

Thermal imaging provides a contact-free method for monitoring motor condition. While thermal images can be evaluated individually and effectively, most applications require capturing the event structure needed for online monitoring and/or the correlation of the adjacent thermal frames. This prior information is useful for assessing false alarms, detection delay, post-detection diagnosis, signal recovery, among others. We introduce ThermalFault, an event-annotated thermal dataset and benchmark for motor fault monitoring. The dataset contains 21,556 frames from 18 development sessions covering normal operation, fan faults, single-phase undervoltage, and motor operation at 40 and 50 Hz. Three held-out mixed sessions with fault and recovery periods are included for external evaluation. ThermalFault defines six tasks: frame classification, normal-only anomaly detection, online event detection, post-trigger fault diagnosis, recovery monitoring, and external stress testing. We examine the proposed dataset using classical, deep-learning, time-series, and pretrained vision baselines. We further propose a two-stage pipeline for online fault detection and diagnosis. The detection stage combines deviation from a learned normal thermal trajectory with temporally smoothed fault probabilities from a supervised image classifier. A causal state machine converts the combined evidence into alarm triggers and classifies each event as either a fan fault or single-phase undervoltage. Across six grouped folds, the pipeline detects 11 of 12 fault events at 0.167 false alarms per normal monitoring hour and correctly diagnoses all 11 detected events. These results support session-disjoint, event-level, and recovery-aware evaluation for online thermal fault monitoring.

1 INTRODUCTION

1.1 BACKGROUND, MOTIVATION, AND RESEARCH GAP

Three-phase induction motors are widely used in industry, where unplanned stops can cause high costs. Many faults affect heat generation or removal, making surface temperature useful for condition monitoring. Infrared thermography measures this temperature without contact or service interruption and has long been used for motor monitoring (Bagavathiappan et al., 2013; López-Pérez & Antonino-Daviu, 2017). Early studies compared startup heating curves with healthy baselines (Picazo-Ródenas et al., 2013), while later work showed that spatial and temporal heat patterns can separate faults better than peak temperature alone. For example, ventilation loss caused broad heating, reaching $91.6{,}^{\circ}\mathrm{C}$ with $43{,}399$ pixels above $80{,}^{\circ}\mathrm{C}$, whereas phase unbalance remained concentrated near the stator (Iglesias Martínez et al., 2026). Hotspot location alone cannot distinguish cooling failure from inter-turn faults (Singh & Naikan, 2017), while deviation from a healthy thermal pattern can indicate fan degradation before failure (Jin & Chow, 2013). Recent studies use learned thermal features and report high thermogram classification accuracy (Khanjani & Ezoji, 2021; Chennana et al., 2025; Palomino-Huilcapi et al., 2025; Calderon-Uribe et al., 2024).

Two main issues limit the use of current results in online monitoring. First, random frame splits can place near-duplicate thermal frames from the same session in both training and test sets, so reported scores may reflect seen sessions rather than unseen ones. Many thermal motor datasets also omit class counts, split design, and calibration details, which limits fair comparison (Najafi et al., 2020;

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Janssens et al., 2015); vibration and current datasets are often more fully documented (Jung et al., 2023b). Second, frame-level accuracy does not capture key online measures, including false alarms, detection delay, post-trigger diagnosis, and recovery. Although some studies track temperature over time (Albla et al., 2017), no public thermal motor dataset labels fault and recovery periods as events. These measures therefore cannot be evaluated together under a common protocol.

1.2 Contributions

The main contributions of this work can be summarized as follows.

- **Dataset:** ThermalFault is introduced as an event-annotated thermal motor dataset covering normal operation, fan faults, single-phase undervoltage, two operating frequencies, and mixed external sessions with recovery periods.
- **Benchmark:** A session-disjoint benchmark is provided for frame classification, anomaly detection, online event detection, fault diagnosis, recovery monitoring, and external evaluation.
- **Method and findings:** A two-stage pipeline is proposed that combines normal thermal residuals with supervised visual evidence for fault detection, followed by DenseNet–AIMv2 late fusion for diagnosis. A broad set of baseline models is also evaluated, revealing key limits in undervoltage detection and recovery monitoring.

2 Related Work

2.1 Thermal Fault Monitoring and Datasets

Public thermal motor datasets differ in motor type, fault class, and recording design. Najafi et al. (2020) provide infrared images of a three-phase induction motor in eleven states, including eight stator short-circuit cases, a stuck rotor, and a cooling-fan failure, recorded in one session at fixed ambient temperature (Najafi et al., 2021). Kulkarni (2025) provide 488 thermal images of healthy and corroded bearings under three load levels. Głowacz (2022b) release labelled BLDC motor images with healthy and faulty shafts across several motor models (Głowacz, 2022a), while Janssens et al. (2015) record five bearings per condition for one hour to retain the heating trajectory. Outside thermography, Jung et al. (2023b) provide timestamped vibration, acoustic, temperature, and current signals for several motor faults under three loads (Jung et al., 2023a). Table 1 compares these datasets.

Table 1: Public datasets for motor and rotating-machine fault diagnosis.

Dataset	Modality	Equipment	Conditions	Operating variation	Temporal / event labels	Recov.
Najafi et al. (Najafi et al., 2020; 2021)	Thermal images	Three-phase induction motor	11 states; stator short circuit, stuck rotor, fan failure	Laboratory condition	Not reported	No
Kulkarni (Kulkarni, 2025)	Thermal images	Induction motor	Healthy / corroded bearing (488 images)	No, medium, full load	Not reported	No
Głowacz (Głowacz, 2022b;a)	Thermal images	BLDC motors	Healthy and 1–3 faulty shafts	Multiple motor models	Image classes	No
Janssens et al. (Janssens et al., 2015)	Thermal recordings	Rotating / bearing rig	Lubrication, outer-race fault, imbalance	Five bearings per condition; one-hour runs	Heating sequences	No
Jung et al. (Jung et al., 2023b;a)	Vibration, acoustic, temperature, current	Rotating machine	Bearing, misalignment, unbalance	Three loads	Timestamped signals	No

Public thermographic datasets support fault classification across induction motors, BLDC motors, and bearing systems. However, most benchmarks use image classes or full operating records and do not assess causal fault onset, false alarms, post-trigger diagnosis, or stable recovery. ThermalFault addresses this gap with session-disjoint splits, event annotations, mixed operating streams, and recovery-aware metrics.

2.2 Online Detection and Foundation Models

Online fault detection often uses models trained on normal data, with the difference between observed and expected signals used as an anomaly score (Ruff et al., 2021; Jin & Chow, 2013; Magadán et al., 2020). In thermal monitoring, current behaviour is compared with a healthy heating baseline (Picazo-Ródenas et al., 2013). Alarms can then be generated using thresholds, persistence rules, or change-detection methods such as CUSUM, Page–Hinkley, and Bayesian online change-point detection (Page, 1954; Hinkley, 1971; Adams & MacKay, 2007). Evaluation should also include event recall, false alarms, detection delay, and stable recovery (Liu & Paparrizos, 2024).

Time-series foundation models support forecasting and anomaly detection with little or no extra training (Das et al., 2024). Chronos predicts numerical sequences with a pretrained language-model structure, and forecast errors can serve as anomaly scores (Ansari et al., 2024). However, better forecasting does not always improve fault detection, especially when fault effects develop slowly (Tan et al., 2024; Liu & Paparrizos, 2024). This is relevant to thermal monitoring because electrical faults may occur before clear temperature changes appear (Iglesias Martínez et al., 2026).

Pretrained visual models offer another approach to thermal image analysis. CLIP and SigLIP support zero-shot classification, while EVA, AIMv2, and DINOv2 can serve as frozen feature extractors (Radford et al., 2021; Zhai et al., 2023; Fang et al., 2023; Fini et al., 2024; Oquab et al., 2024). Because these models are mainly trained on natural images, transfer to pseudo-coloured thermal images may be limited. Supervised probes or fusion with task-specific networks can improve performance (Jeong et al., 2023; Zhiyi et al., 2020; Santos et al., 2026). These normal-only, time-series, and visual methods have rarely been compared under a common session-disjoint, causal, and event-level protocol for thermal motor monitoring.

3 ThermalFault Dataset and Benchmark

3.1 Data Collection and Annotation

ThermalFault was collected using the motor test bench shown in Figure 1. The setup included a three-phase induction motor, variable-frequency drive, mechanical loading system, and electrical control panel. Thermal data were recorded with the integrated thermal camera of a Ulefone smartphone mounted on a fixed tripod beside the motor to provide the view consistent across sessions.

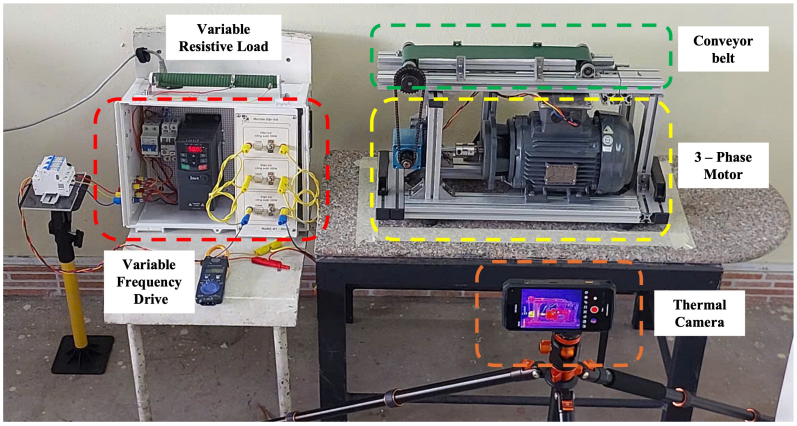

Figure 1: Motor test bench used for thermal data collection.

The motor was operated at 40 and 50 Hz under normal, cooling-fan-fault, and single-phase-undervoltage conditions. The fan fault was created by disabling the cooling fan. Undervoltage was produced by connecting a 100 Ω, 100 W resistor in series with one phase, creating an unbalanced supply without disconnecting the phase. The development set contains 18 sessions, with three repeats for each frequency–condition pair. Each session lasted approximately 60 minutes, with frames sampled every three seconds. Fault sessions comprised 30 minutes of normal operation, approximately 25 minutes of fault operation, and five minutes of recovery. In total, the development set contains 21,556 frames: 15,556 normal, 3,000 cooling-fan-fault, and 3,000 undervoltage frames. Event

labels were assigned from recorded intervention times, and each frame was linked to its session, frequency, elapsed time, and condition. Three mixed streams, M01–M03, were reserved for external evaluation and excluded from training and model selection. M01 and M02 contained both fault types separated by normal periods, whereas M03 contained a longer undervoltage event followed by extended recovery. The model and detector parameters selected using S01–S18 were applied without further tuning. Table 2 summarizes the development set and held-out external streams.

Table 2: Overview of the ThermalFault development set and held-out external streams.

Set/stream	Freq.	Condition	Sessions or event sequence	Frames
Development set				
Dev.	40 Hz	Normal	S03, S07, S15	3,599
Dev.	50 Hz	Normal	S01, S09, S13	3,594
Dev.	40 Hz	Fan fault	S02, S11, S14	3,589
Dev.	50 Hz	Fan fault	S05, S08, S17	3,599
Dev.	40 Hz	Undervoltage	S06, S10, S18	3,587
Dev.	50 Hz	Undervoltage	S04, S12, S16	3,588
Development total			**18 sessions**	**21,556**
Held-out external streams				
M01	40 Hz	Mixed	Normal → fan fault → normal → undervoltage → normal	1,195
M02	50 Hz	Mixed	Normal → undervoltage → normal → fan fault → normal	1,196
M03	50 Hz	Mixed	30-min normal → 60-min undervoltage → 30-min recovery	2,400
External total				**4,791**

3.2 Calibration, Annotation, and Quality Control

Thermal frames were sampled every three seconds and indexed by session and time. The displayed temperature scale was extracted by optical character recognition, and a session-specific color-to-temperature map was built from the thermal palette. Because the images are color-rendered rather than raw radiometric data, the recovered temperatures are treated as calibrated estimates. A fixed motor region was cropped and resized for each model. Metadata and frame order were checked for errors, and event labels were assigned from recorded fault start and removal times. Where required, a 120-second transition buffer was excluded from frame-level evaluation to avoid unstable thermal changes. Calibration results, event boundaries, and transition frames were also checked.

3.3 Benchmark Tasks and Protocol

ThermalFault defines six benchmark tasks that connect frame-level recognition with causal online monitoring. The tasks cover fault classification, normal-only anomaly detection, event detection, post-trigger diagnosis, recovery monitoring, and evaluation on unseen continuous streams. Table 3 summarizes the training setting, evaluation target, and primary metrics for each task.

Table 3: Overview of the six ThermalFault benchmark tasks.

Task	Setting	Target	Metrics
T1. Frame classification	Normal and fault frames	Per-frame normal, fan-fault, or undervoltage class	Accuracy; macro-F1
T2. Normal-only anomaly detection	Normal frames only	Deviation from learned normal thermal behaviour	Event recall; delay; FAR/h
T3. Online event detection	Causal thermal stream	First valid alarm in each fault event	Event recall; delay; FAR/h
T4. Post-trigger diagnosis	Detected fault events	Fan fault or undervoltage after a trigger	Conditional; end-to-end accuracy
T5. Recovery monitoring	Post-fault sequence	Return to a stable alarm-off state	Recovery delay; occupancy; re-triggers
T6. External stress test	Locked models and thresholds	Full pipeline on M01–M03	Detection; delay; FAR/h; diagnosis; recovery

All development experiments use session-disjoint splits, with no session shared across training, validation, and test sets. The 18 sessions form six groups, each containing one normal, one cooling-fan-fault, and one single-phase-undervoltage session (Appendix Table B1). Each test fold contains two fault events, giving 12 session-disjoint test events across six folds. Validation data are used for model and threshold selection, while test data are reserved for final evaluation.

Frame classifiers are trained on all three conditions and evaluated using accuracy and macro-F1. Normal-behaviour models are fitted using normal training samples, with prediction or reconstruction errors used as anomaly scores. Their thresholds are selected on validation sessions. Online detection is causal and uses only current and past observations. A fault event is detected if a valid alarm rising edge occurs between its onset and end; otherwise, it is missed.

For event i, detection delay is $d_i = t_i^{\text{trigger}} - t_i^{\text{onset}}$, where t_i^{onset} and t_i^{trigger} are the fault onset and first valid trigger times. Event recall is $\text{Recall}_{\text{event}} = N_{\text{detected}}/N_{\text{fault events}}$. The false-alarm rate is $\text{FAR/h} = N_{\text{false triggers}}/(T_{\text{normal}}/3600)$, where T_{normal} is the valid normal monitoring time after transition buffers are removed.

Post-trigger diagnosis is evaluated using conditional and end-to-end accuracy. Conditional accuracy includes detected events only, while end-to-end accuracy treats missed events as failures. Recovery delay is measured from fault removal until the alarm remains inactive. M01–M03 were excluded from all training and selection steps. Model weights, fusion weights, thresholds, and alarm rules were fixed before external evaluation. These streams test different event orders, fault durations, and recovery periods; full schedules are given in Appendix Table B2.

4 METHODS

The benchmark evaluates three method groups: supervised condition recognition, normal-only anomaly detection, and causal event detection and diagnosis. Supervised methods include classical classifiers, convolutional networks, frozen visual encoders, and causal temporal models. Normal-only methods include thermal regression, forecasting, and spatial reconstruction. The proposed pipeline combines normal-trajectory residuals with supervised image probabilities for causal detection and diagnosis. Appendix Table B3 summarizes all methods.

4.1 FRAME-LEVEL CONDITION RECOGNITION

Frame-level classification tests whether motor condition can be identified from a single thermal image. For each image x_t, the classifier estimates the probabilities of normal operation, cooling-fan fault, and single-phase undervoltage.

Classical baselines include random forest, support-vector machine, and LightGBM using thermal and regional statistics. Image-based baselines include a shallow CNN, ResNet34, DenseNet121, and DenseNet169 applied to cropped motor images. All models use session-disjoint splits.

Frozen visual encoders are evaluated for transfer from large image datasets to pseudo-coloured thermal images. CLIP, SigLIP, EVA, AIMv2, and DINOv2 are included (Radford et al., 2021; Zhai et al., 2023; Fang et al., 2023; Fini et al., 2024; Oquab et al., 2024). For each encoder, $\mathbf{z} * t = f * \text{enc}(x_t)$ is extracted with fixed weights, and a linear or logistic probe is trained on the session embeddings. Zero-shot image–text classification is also evaluated where supported.

Short temporal models test whether recent thermal changes improve single-frame prediction. Temporal statistics, GRU, and causal TCN models use sequences of visual embeddings,

$$\mathbf{Z} * t = [\mathbf{z} * t - L + 1, \ldots, \mathbf{z}_t], \tag{1}$$

where L is the causal window length and only current and past frames are used. These models output condition probabilities but do not define alarm entry, clearing, or event grouping.

4.2 NORMAL-ONLY THERMAL ANOMALY DETECTION

Normal-only detection tests whether faults can be detected without fault samples during training. A temperature statistic T_t is extracted from the motor center and expressed relative to a session reference as $\Delta T_t = T_t - T_t^{\text{ref}}$.

A regression or forecasting model learns the normal thermal pattern. The anomaly score is the prediction residual,

$$r_t = \left| \Delta T_t - \widehat{\Delta T}_t \right|, \tag{2}$$

where $\widehat{\Delta T}_t$ is the predicted normal value. A large residual indicates deviation from the learned normal pattern.

Four scalar methods are evaluated: thermal regression, naive-last forecasting, and pretrained Chronos-Bolt Small and Base (Ansari et al., 2024). A spatial PCA model is also fitted to normal 32×32 motor images, with reconstruction error used as the anomaly score. Thresholds and persistence settings are selected on validation sessions. These methods detect abnormal behaviour but do not identify the fault type.

4.3 Proposed Two-Stage Causal Detection and Diagnosis

Frame classifiers provide condition probabilities but not alarm timing, while normal-only models provide anomaly scores without fault labels. A two-stage causal pipeline is therefore used (Figure 2). The first stage detects fault events, and the second classifies each detected event as a cooling-fan fault or single-phase undervoltage.

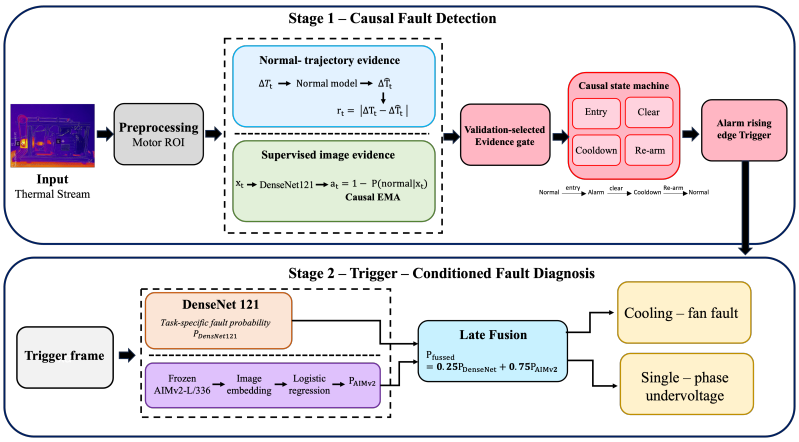

Figure 2: Proposed two-stage causal detection and diagnosis pipeline.

Causal Event Detection

The detector uses two evidence routes. The normal-trajectory route uses the residual r_t from Equation 2. The supervised route uses the DenseNet121 non-normal probability $a_t = 1 - p_t^{\text{normal}}$. Isolated changes are reduced using the causal exponential moving average $\bar{a}_t = \alpha a_t + (1-\alpha)\bar{a}_{t-1}$, which uses no future observations.

The supervised route can trigger directly, while a residual alarm requires support from the DenseNet121 score. Thresholds and support settings are selected on validation sessions. A causal state machine controls alarm activation and clearing. Only an inactive-to-active transition is counted as a trigger; consecutive positive frames form one event, and an alarm already active at fault onset is not counted as a new detection.

Trigger-Conditioned Fault Diagnosis

When an alarm is triggered at time τ, DenseNet121 provides fault probabilities from the trigger frame x_τ. The same frame is encoded by frozen AIMv2-L/336, and a standardized logistic regression probe produces a second set of fault probabilities. The predictions are combined as $p_{\text{fused}} = 0.25 p_{\text{DenseNet}} + 0.75 p_{\text{AIMv2}}$. The fusion weight, one-frame window, and 0.50 confidence threshold are selected on validation data. Diagnosis is made at the trigger frame, adding no extra delay. The detector identifies fault onset, and the fused model determines the fault type.

5 EXPERIMENTS AND RESULTS

5.1 BENCHMARK RESULTS

FRAME-LEVEL CONDITION RECOGNITION

Table 4 reports representative results across the six session-disjoint test folds. The methods cover classical classifiers, image-based networks, frozen visual encoders, and short temporal models. DenseNet121 achieved the best overall performance, with an accuracy of 90.57% and a macro-F1 of 85.03%. It also gave the highest fan-fault F1 (99.12%) and undervoltage F1 (62.43%), although undervoltage remained harder to recognize across methods. Frozen SigLIP and AIMv2 probes performed below DenseNet121. Temporal statistics also gave a lower test macro-F1 than the current-frame AIMv2 probe (64.14% vs. 67.33%). AIMv2 was retained for post-trigger fusion because its event-level predictions complemented DenseNet121.Overall, DenseNet121 provided frame-level recognition, especially for cooling-fan faults, but undervoltage remained less reliable and frame predictions did not define causal event triggers. Full results are given in Appendix Tables B4, B5, and B6.

Table 4: frame-classification results across six session-disjoint test folds.

Family	Method	Accuracy (%)	Macro-F1 (%)	Fan F1 (%)	Under. F1 (%)
Classical	Random forest	76.62	58.20	55.77	34.52
Image-based	Shallow CNN	83.24	72.65	97.08	31.78
	ResNet34	85.98	74.46	97.10	35.90
	DenseNet121	**90.57**	**85.03**	**99.12**	**62.43**
Frozen encoder	SigLIP + logistic	79.98	68.03	62.09	55.79
	AIMv2 + logistic	78.45	67.33	73.49	43.59
Short temporal	Temporal statistics	74.88	64.14	65.03	46.51

NORMAL-ONLY ANOMALY DETECTION

Normal-behaviour models use only normal samples, with thresholds selected on validation data. The outputs are binary alarms and do not identify fault type. Table 5 reports results by event class. The spatial model achieved the highest recall (7/12) and the lowest false-alarm rate (2.00 FAR/h), but detected only 1/6 undervoltage events, the same as thermal regression. Chronos-Bolt Small detected 5/12 events, had the highest false-alarm rate (2.83 FAR/h), and missed all undervoltage events. Thus, normal-only models provide fault-label-free detection and capture cooling-fan events well, but remain limited for undervoltage detection and false-alarm control. Full results are given in Appendix Table B7.

Table 5: Representative Stage 1 timelines for detected and missed fault events.

Detector	Detected	Fan events	Under. events	False alarms	FAR/h
Thermal regression	6/12	5/6	1/6	15	2.50
Chronos-Bolt Small	5/12	5/6	0/6	17	2.83
Spatial normal-only model	7/12	6/6	1/6	12	2.00

5.2 PROPOSED PIPELINE AND ABLATIONS

The benchmark results reveal complementary limitations. DenseNet121 provides strong frame-level fault evidence but does not define event-level alarm timing, whereas normal-only models detect deviations without fault samples but miss most undervoltage events and produce frequent false alarms. The proposed pipeline combines these two sources of evidence to generate causal event triggers with improved recall and false-alarm control, followed by trigger-conditioned fault diagnosis.

Stage 1: Causal event detection. The first stage combines the DenseNet121 non-normal probability with the thermal-residual alarm. DenseNet121 can trigger directly, while a residual alarm requires support from the DenseNet121 score. A causal state machine controls alarm activation and clearing using only current and past frames. Across the six test folds, Stage 1 detected 11 of 12

fault events: all six cooling-fan faults and five of six undervoltage events. It produced one false alarm (0.167 FAR/h), and its mean delay across the 11 detected events was 208.6 s. S16 was the only missed event. Figure 3 illustrates three representative outcomes. S02 shows strong residual and supervised evidence, while S10 is detected after the supervised score remains high. In S16, neither route satisfies the alarm-entry condition, resulting in the only missed event. These results show that Stage 1 converts frame-level evidence into causal event triggers while limiting false alarms. Its remaining failure case is an undervoltage event with insufficient thermal and visual evidence.

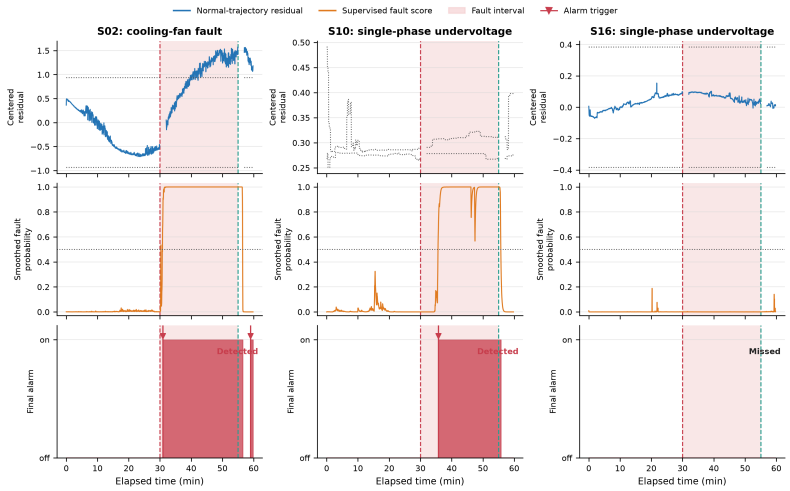

Figure 3: Representative Stage 1 timelines for detected and missed fault events.

Stage 2: Post-trigger fault diagnosis. After each valid trigger, the second stage classifies the event as a cooling-fan fault or single-phase undervoltage. DenseNet121 correctly diagnosed 10/11 detected events, AIMv2 9/11, and late fusion all 11 events (Table 6). Late fusion therefore achieved 100% conditional accuracy and 91.7% end-to-end accuracy, with the missed S16 event counted as a failure. These results show that the two visual representations provide complementary diagnostic information, while end-to-end performance remains limited by Stage 1 recall. Overall, the two-stage pipeline detected and correctly diagnosed 11/12 fault events, with one false alarm (0.167 FAR/h) and a mean detection delay of 208.6 s.

Table 6: Stage 2 post-trigger fault-diagnosis results.

Diagnoser	**Conditional accuracy**	**End-to-end accuracy**
DenseNet121	10/11 (90.9%)	10/12 (83.3%)
AIMv2 + logistic probe	9/11 (81.8%)	9/12 (75.0%)
DenseNet121–AIMv2 late fusion	**11/11 (100%)**	**11/12 (91.7%)**

5.3 External and Recovery Evaluation

The models were evaluated on M01–M03 without retraining or parameter adjustment. The event-level results are shown in Table 7.

Table 7: Stage 1 event-detection results on the held-out external streams.

Stream	**Fault**	**Duration**	**Detection**	**Delay**
M01	Cooling-fan fault	10 min	Yes	111 s
M01	Single-phase undervoltage	10 min	No	—
M02	Single-phase undervoltage	15 min	No	—
M02	Cooling-fan fault	15 min	Yes	60 s
M03	Single-phase undervoltage	60 min	Yes	1,605 s

The cooling-fan events in M01 and M02 were detected and diagnosed correctly, while the 10- and 15-minute undervoltage events were missed. The 60-minute undervoltage event in M03 was detected

after 1,605 s and diagnosed correctly with 0.8566 confidence. This may reflect slow thermal change, but one long event is not enough to link fault duration with detection. Recovery also remained difficult. Of the 11 detected development events, 9 reached a stable alarm-off state. All five detected undervoltage sessions stabilized within 0–144 s after fault removal, while cooling-fan sessions S02 and S05 did not stabilize before recording ended. M03 further showed slow undervoltage detection and unstable recovery. The first valid trigger occurred 1,605 s after fault onset. After the supply returned to normal, 16 additional alarm rising edges occurred, and the detector required 1,377 s to become stably inactive (Figure 4).

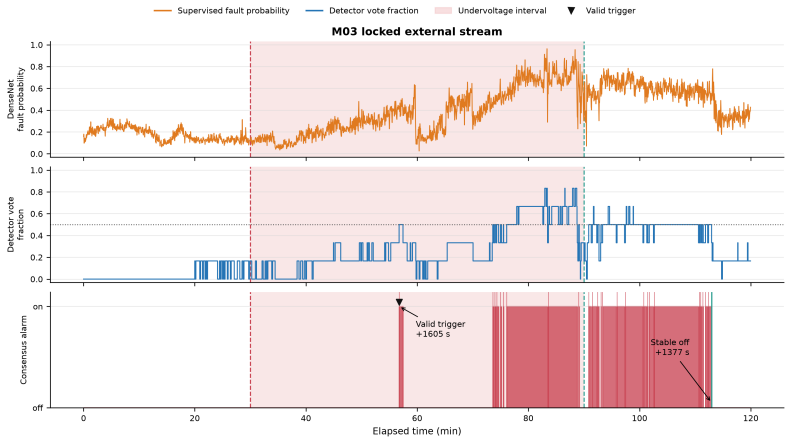

Figure 4: Detection and recovery timeline for the held-out M03 stream.

6 Discussion and Limitations

ThermalFault provides event-annotated thermal streams across three motor conditions, two operating frequencies, and recovery periods. The session-disjoint benchmark covers frame classification, anomaly detection, causal event detection, diagnosis, recovery, and external evaluation. High frame-level accuracy did not ensure reliable online monitoring. DenseNet121 performed well overall, but undervoltage recognition varied across sessions, and normal-only methods also detected cooling-fan faults more reliably than undervoltage. Event recall, detection delay, and FAR/h are therefore important alongside frame-level metrics. The two-stage pipeline improved event detection and diagnosis. Stage 1 detected 11/12 events, compared with 7/12 for the best normal-only baseline, Stage 2 correctly diagnosed all 11 detected events with no extra delay, giving 91.7% end-to-end accuracy.

External results remained limited. The 10- and 15-minute undervoltage events were missed, while the 60-minute M03 event showed long detection delay and unstable recovery. Main limitations include one motor and test bench, few independent fault events, and only three external streams. Future work should include more motors, loads, recording conditions, and external sites, with undervoltage tested at different severities and durations.

7 Conclusion

ThermalFault provides an event-annotated thermal dataset and benchmark for session-disjoint motor monitoring. Its six tasks cover frame classification, normal-only anomaly detection, causal event detection, post-trigger diagnosis, recovery monitoring, and external evaluation. The results show that frame accuracy alone can hide missed events, long detection delays, and unstable recovery. The proposed two-stage pipeline combines normal-trajectory residuals with supervised image evidence, followed by DenseNet–AIMv2 late fusion for diagnosis. Across the development test folds, it detected and correctly diagnosed 11/12 fault events, with one false alarm (0.167 FAR/h). External evaluation exposed remaining limitations for short undervoltage events and post-fault recovery. These findings support session-disjoint, event-level, and recovery-aware evaluation of thermal motor monitoring systems.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, manuscript organization, and LaTeX formatting.

ETHICS STATEMENT

This study uses thermal images of an induction motor collected under controlled laboratory conditions. It involves no human participants, personal data, or animal subjects.

REFERENCES

Ryan Prescott Adams and David J. C. MacKay. Bayesian online changepoint detection. *arXiv preprint arXiv:0710.3742*, 2007. URL https://arxiv.org/abs/0710.3742.

Alauddin Adel Hamoodi Albla, Bogdan M. Brkovic, Miloš M. Ječmenica, and Zoran Mileta Lazarević. Online temperature monitoring of a grid connected induction motor. *International Journal of Electrical Power and Energy Systems*, 93:276–282, dec 1 2017. doi: 10.1016/j.ijepes.2017.06.007. URL https://doi.org/10.1016/j.ijepes.2017.06.007.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024. URL https://arxiv.org/abs/2403.07815.

Subramaniam Bagavathiappan, Barid Baran Lahiri, Thangavelu T. Saravanan, John Philip, and T. K. Jayakumar. Infrared thermography for condition monitoring - A review. *Infrared Physics and Technology*, 60:35–55, may 7 2013. doi: 10.1016/j.infrared.2013.03.006. URL https://doi.org/10.1016/j.infrared.2013.03.006.

Uriel Calderon-Uribe, Rocío A. Lizarraga-Morales, and Igor V. Guryev. Fault Diagnosis in Induction Motors through Infrared Thermal Images Using Convolutional Neural Network Feature Extraction. *Machines*, 12(8), aug 1 2024. doi: 10.3390/machines12080497. URL https://doi.org/10.3390/machines12080497.

Ahmed Chennana, Noureddine Bessous, Ahmed Chaouki Megherbi, Mounir Sayadi, and Ali Teta. Deep-Shallow Features Fusion for Induction Motor Fault Diagnosis through Infrared Thermography Images. *2025 International Conference on Control, Automation and Diagnosis, ICCAD 2025*, jan 1 2025. doi: 10.1109/ICCAD64771.2025.11099458. URL https://doi.org/10.1109/ICCAD64771.2025.11099458.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2310.10688.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. EVA: Exploring the limits of masked visual representation learning at scale. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. URL https://arxiv.org/abs/2211.07636.

Enrico Fini, Mustafa Shukor, Xiujun Li, Philipp Dufter, Michal Klein, David Haldimann, Sai Aitharaju, Louis Béthune, Zhe Gan, Alexander Toshev, et al. Multimodal autoregressive pre-training of large vision encoders. *arXiv preprint arXiv:2411.14402*, 2024. URL https://arxiv.org/abs/2411.14402.

Adam Głowacz. Diagnostyka uszkodzeń mechanicznych silnika BLDC z zastosowaniem termowizji, 2022a. URL https://doi.org/10.58032/AGH/H4ETKQ. Dataset.

Adam Głowacz. Thermographic Fault Diagnosis of Shaft of BLDC Motor. *Sensors*, 22(21):8537, 2022b. doi: 10.3390/s22218537. URL https://doi.org/10.3390/s22218537.

D. V. Hinkley. Inference about the change-point from cumulative sum tests of normal mean. *Biometrika*, 58(3):509–523, 1971. doi: 10.1093/biomet/58.3.509.

Miguel Enrique Iglesias Martínez, José Alfonso Antonino-Daviu, Larisa Dunai, Humberto J. Michinel Álvarez, and Pedro J. Fernández De Córdoba. Explainable Thermographic Fault Diagnosis of Three-Phase Induction Motors Using Transient Thermal Signatures: A Case Study. *Machines*, 14(8), aug 1 2026. doi: 10.3390/machines14080843. URL https://doi.org/10.3390/machines14080843.

Olivier Janssens, Raiko Schulz, Viktor Slavkovikj, Rik van de Walle, and Sofie van Hoecke. Thermal image based fault diagnosis for rotating machinery. *Infrared Physics and Technology*, 73:78–87, nov 1 2015. doi: 10.1016/j.infrared.2015.09.004. URL https://doi.org/10.1016/j.infrared.2015.09.004.

Jongheon Jeong, Yang Zou, Taewan Kim, Dongqing Zhang, Avinash Ravichandran, and Onkar Dabeer. WinCLIP: Zero-/few-shot anomaly classification and segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. URL https://arxiv.org/abs/2303.14814.

Xiaohang Jin and Tommy Wai Shing Chow. Anomaly detection of cooling fan and fault classification of induction motor using Mahalanobis-Taguchi system. *Expert Systems with Applications*, 40(15): 5787–5795, jun 17 2013. doi: 10.1016/j.eswa.2013.04.024. URL https://doi.org/10.1016/j.eswa.2013.04.024.

Wonho Jung, Seong Hu Kim, Sung Hyun Yun, Jaewoong Bae, and Yong Hwa Park. Vibration, Acoustic, Temperature, and Motor Current Dataset of Rotating Machine Under Varying Load Conditions for Fault Diagnosis, 2023a. URL https://doi.org/10.17632/ztmf3m7h5x.6. Dataset.

Wonho Jung, Seong Hu Kim, Sung Hyun Yun, Jaewoong Bae, and Yong Hwa Park. Vibration, acoustic, temperature, and motor current dataset of rotating machine under varying operating conditions for fault diagnosis. *Data in Brief*, 48:109049, 2023b. doi: 10.1016/j.dib.2023.109049. URL https://doi.org/10.1016/j.dib.2023.109049.

Majid Khanjani and Mehdi Ezoji. Electrical fault detection in three-phase induction motor using deep network-based features of thermograms. *Measurement: Journal of the International Measurement Confederation*, 173, mar 1 2021. doi: 10.1016/j.measurement.2020.108622. URL https://doi.org/10.1016/j.measurement.2020.108622.

Vaishnavi V. Kulkarni. Induction Motor-Thermal Images, 2025. URL https://doi.org/10.17632/pymtdhfzbj.1. Dataset.

Qinghua Liu and John Paparrizos. The elephant in the room: Towards a reliable time-series anomaly detection benchmark. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=R6kJtWsTGy.

David López-Pérez and José Alfonso Antonino-Daviu. Application of Infrared Thermography to Failure Detection in Industrial Induction Motors: Case Stories. *IEEE Transactions on Industry Applications*, 53(3):1901–1908, may 1 2017. doi: 10.1109/TIA.2017.2655008. URL https://doi.org/10.1109/TIA.2017.2655008.

Luis Magadán, F. J. Suárez, Juan C. Granda, and D. F. García. Real-time monitoring of electric motors for detection of operating anomalies and predictive maintenance. In *Lecture Notes of the Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering (LNICST)*, volume 323, pp. 301–311, 2020. doi: 10.1007/978-3-030-51005-3_25. URL https://doi.org/10.1007/978-3-030-51005-3_25.

Mohamad Najafi, Yasser Baleghi, Sayyed Asghar Gholamian, and Seyyed Mehdi Mirimani. Fault Diagnosis of Electrical Equipment through Thermal Imaging and Interpretable Machine Learning Applied on a Newly-introduced Dataset. *6th Iranian Conference on Signal Processing and Intelligent Systems, ICSPIS 2020*, dec 23 2020. doi: 10.1109/ICSPIS51611.2020.9349599. URL https://doi.org/10.1109/ICSPIS51611.2020.9349599.

Mohamad Najafi, Yasser Baleghi, Sayyed Asghar Gholamian, and Seyyed Mehdi Mirimani. Thermal image of equipment (Induction Motor), 2021. URL https://doi.org/10.17632/m4sbt8hbvk.2. Dataset.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024. URL https://arxiv.org/abs/2304.07193.

E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1–2):100–115, 1954. doi: 10.1093/biomet/41.1-2.100.

Jimmy Palomino-Huilcapi, Mahmoud Kassas, and Mohammad A. Abido. Lightweight Thermal Fault Diagnosis in Induction Motors: A Comparative Study of CNNs and Depthwise-Separable CNNs. *2025 6th International Conference on Data Analytics for Business and Industry, ICDABI 2025*, pp. 231–235, jan 1 2025. doi: 10.1109/ICDABI67967.2025.11547495. URL https://doi.org/10.1109/ICDABI67967.2025.11547495.

Maria Jose Picazo-Ródenas, Rafael Royo-Pastor, José Alfonso Antonino-Daviu, and José Roger-Folch. Use of the infrared data for heating curve computation in induction motors: Application to fault diagnosis. *Engineering Failure Analysis*, 35:178–192, dec 15 2013. doi: 10.1016/j.engfailanal.2013.01.018. URL https://doi.org/10.1016/j.engfailanal.2013.01.018.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021. URL https://arxiv.org/abs/2103.00020.

Lukas Ruff, Jacob R. Kauffmann, Robert A. Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G. Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021. doi: 10.1109/JPROC.2021.3052449.

Luiz Fillipe Dahmer Dos Santos, Jorge Luiz Dos Santos Canuto, Linnyer Beatrys Ruiz Aylon, Viviana Cocco Mariani, and Rodrigo Clemente Thom De Souza. A Multimodal Model for Fault Diagnosis in Induction Motors Using Thermal Images and Large Language Models. *IEEE Access*, 14:89964–89976, jan 1 2026. doi: 10.1109/ACCESS.2026.3703729. URL https://doi.org/10.1109/ACCESS.2026.3703729.

Gurmeet Singh and Vallayil Narayana Achutha Naikan. Infrared thermography based diagnosis of inter-turn fault and cooling system failure in three phase induction motor. *Infrared Physics and Technology*, 87:134–138, dec 1 2017. doi: 10.1016/j.infrared.2017.10.007. URL https://doi.org/10.1016/j.infrared.2017.10.007.

Mingtian Tan, Mike A. Merrill, Vinayak Gupta, Tim Althoff, and Thomas Hartvigsen. Are language models actually useful for time series forecasting? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2406.16964.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. URL https://arxiv.org/abs/2303.15343.

He Zhiyi, Shao Haidong, Zhong Xiang, Yang Yu, and Junsheng Cheng. An intelligent fault diagnosis method for rotor-bearing system using small labeled infrared thermal images and enhanced CNN transferred from CAE. *Advanced Engineering Informatics*, 46, oct 1 2020. doi: 10.1016/j.aei.2020.101150. URL https://doi.org/10.1016/j.aei.2020.101150.

A Test Bench and Fault Implementation

The data were collected using a laboratory test bench consisting of a three-phase induction motor, a variable-frequency drive (VFD), a belt-driven mechanical load, an electrical control panel, and a tripod-mounted thermal smartphone. The main equipment specifications are summarized in Table A1.

Table A1: Main equipment used in the thermal motor test bench.

Component	Specification
Motor type	Three-phase induction motor
Rated power	0.75 kW (1 hp)
Number of poles	4
Rated connection	Star connection, 380 V
Shaft length	70 mm
Shaft diameter	20 mm
VFD model	INVT GD20-1R5G-S2
VFD control mode	V/F control
VFD input voltage	220 V AC
VFD output range	0–50 Hz
VFD frequency steps	10 Hz
VFD communication	Analog I/O, 0–10 V
Thermal device	Ulefone Armor 29 Pro Thermal
Tested frequencies	40 and 50 Hz
Fault resistor	100 Ω, 100 W

B Dataset Organization and Additional Results

B.1 Dataset Organization and Labels

The released dataset is organized by data partition and recording session. The development partition contains S01–S18, while M01–M03 form a separate external partition.

```
ThermalFault/
|-- development/
|   |-- S01/
|   |   ...
|   `-- S18/
|-- external/
|   |-- M01/
|   |-- M02/
|   `-- M03/
`-- metadata/
    `-- events.csv
```

B.2 Development Fold Composition

The 18 development sessions are divided into six non-overlapping groups, each with one normal, one cooling-fan-fault, and one single-phase-undervoltage session. In each fold, four groups are used for training, one for validation, and one for testing. Validation data are used to select model settings, thresholds, and stopping points, while test data are used only after these settings are fixed.

Table B1: Training, validation, and test sessions in the six grouped folds.

Fold	Training sessions	Validation sessions	Test sessions
1	S07–S18	S03, S05, S06	S01, S02, S04
2	S01, S02, S04, S09, S11–S18	S07, S08, S10	S03, S05, S06
3	S01–S06, S13–S18	S09, S11, S12	S07, S08, S10
4	S01–S08, S10, S15, S17, S18	S13, S14, S16	S09, S11, S12
5	S01–S12	S15, S17, S18	S13, S14, S16
6	S03, S05–S14, S16	S01, S02, S04	S15, S17, S18

B.3 External-Stream Metadata

M01 and M02 contain two fault events separated by normal operation. M03 contains a longer single-phase-undervoltage event followed by an extended recovery period.

Table B2: Event schedules for the held-out external streams.

Stream	Frequency	Frames	Segment	Interval	Duration
M01	40 Hz	1,195	Normal	$[0, 1200)$ s	20 min
			Cooling-fan fault	$[1200, 1800)$ s	10 min
			Normal	$[1800, 2400)$ s	10 min
			Undervoltage	$[2400, 3000)$ s	10 min
			Normal recovery	$[3000, 3585)$ s	≈9.8 min
M02	50 Hz	1,196	Normal	$[0, 600)$ s	10 min
			Undervoltage	$[600, 1500)$ s	15 min
			Normal	$[1500, 2100)$ s	10 min
			Cooling-fan fault	$[2100, 3000)$ s	15 min
			Normal recovery	$[3000, 3588)$ s	≈9.8 min
M03	50 Hz	2,400	Normal	$[0, 1800)$ s	30 min
			Undervoltage	$[1800, 5400)$ s	60 min
			Normal recovery	$[5400, 7200)$ s	30 min

B.4 Summary of Evaluated Methods

Table B3 summarizes the evaluated method families, their training settings, and their outputs.

Table B3: Summary of the methods evaluated in the ThermalFault benchmark.

Setting	Method family	Models	Training setting	Output
Supervised recognition	Classical classifiers	Random forest, SVM, LightGBM	Thermal features and fault labels	Three-class probability
	Convolutional networks	Shallow CNN, ResNet34, DenseNet121/169	Motor ROI and fault labels	Three-class probability
	Frozen visual encoders	CLIP, SigLIP, EVA, AIMv2, DINOv2	Frozen encoder and labelled probe	Three-class probability
	Causal temporal models	Temporal statistics, GRU, causal TCN	Current and past embeddings	Three-class probability
Normal-only detection	Thermal trajectory	Thermal regression, naive-last, Chronos-Bolt	Normal sessions only	Prediction residual
	Spatial reconstruction	PCA reconstruction	Normal motor ROIs only	Reconstruction residual
Proposed pipeline	Causal event detector	Thermal residual, DenseNet121, state machine	Validation-selected alarm rules	Alarm state and trigger
	Trigger-conditioned diagnosis	DenseNet121–AIMv2 late fusion	Frame at each valid trigger	Fault-type probability

B.5 Classical and CNN Classification

Table B4 reports one configuration for each classical and convolutional model family. DenseNet121 achieved the highest accuracy and macro-F1, with cooling-fan faults recognized more consistently than single-phase undervoltage.

Table B4: Classical and convolutional frame-classification results.

Model	Accuracy	Macro-F1	Fan F1	Under. F1
Random forest	0.7662	0.5820	0.5577	0.3452
SVM	0.6984	0.5538	0.6105	0.2754
LightGBM	0.7260	0.5293	0.4801	0.2962
Shallow CNN	0.8324	0.7265	0.9708	0.3178
ResNet34	0.8598	0.7446	0.9710	0.3590
DenseNet121	0.9057	0.8503	0.9912	0.6243
DenseNet169	0.8346	0.7407	0.8593	0.4682

B.6 Frozen Visual Encoders

Each pretrained encoder was frozen and used to extract embeddings from the cropped motor region. The embeddings were standardized using the training sessions, and a logistic-regression probe was trained for condition classification; its regularization was selected on validation sessions. Table B5 reports fold-mean results across the six session-disjoint test folds. All frozen encoders performed below the task-specific DenseNet121 classifier.

Table B5: Frame-classification results for frozen visual encoders with supervised logistic probes. All results are fold means (%).

Encoder	Model variant	Accuracy	Macro-F1	Under. F1
CLIP	Base; patch 32	78.55	67.41	47.63
SigLIP2	Base; patch 16	79.98	68.03	55.79
EVA-02	Base; patch 16	78.61	62.35	36.94
AIMv2	Large; patch 14;	78.45	67.33	43.59
DINOv2	Base; patch 14	74.01	62.43	43.73

B.7 Temporal Representations

All methods use frozen AIMv2-L/336 embeddings. AIMv2 with logistic regression on a single-frame embedding serves as the non-temporal reference. The temporal methods use the same current embedding together with embeddings from previous frames. Temporal statistics combine the current embedding with validation-selected means, changes, and standard deviations from recent frames, followed by logistic regression. The GRU processes the embedding sequence recurrently, while the causal TCN applies causal temporal convolutions.

All temporal methods use only current and previous frames; no future frames are included. Window lengths are selected from 5, 10, and 20 frames using validation data, corresponding to 15, 30, and 60 s. Table B6 compares the single-frame reference with the three temporal methods.

The GRU achieved the highest mean undervoltage F1, but its overall macro-F1 was lower than that of the single-frame reference and varied substantially across folds. Temporal statistics and the causal TCN also did not improve overall performance. Thus, the evaluated short temporal representations provided no consistent advantage over the single-frame AIMv2 probe.

Table B6: Single-frame and causal temporal frame-classification results across six session-disjoint folds. Values are mean ± standard deviation (%).

Method	Accuracy	Macro-F1	Under. F1
AIMv2 + logistic (single frame)	$\mathbf{78.45 \pm 8.22}$	$\mathbf{67.33 \pm 10.19}$	43.59 ± 36.83
Temporal statistics	74.88 ± 15.79	64.14 ± 18.06	46.51 ± 37.01
Causal TCN	68.80 ± 10.57	54.64 ± 11.92	45.27 ± 24.56
GRU	77.10 ± 9.12	59.66 ± 20.83	$\mathbf{50.40 \pm 37.28}$

B.8 Normal-Only Event Detection

All methods in Table B7 use no fault samples. Mean delay is computed only for detected events and should therefore be considered together with the number of missed events. Normal-only methods

detected cooling-fan faults more reliably than single-phase undervoltage. Neither naive-last nor Chronos-Bolt detected any undervoltage events. The spatial model improved total detection to 7/12 but detected only 1/6 undervoltage events and still produced frequent false alarms.

Table B7: Normal-only event-detection results.

Method	**Detected**	**Fan**	**Under.**	**False alarms**	**FAR/h**	**Mean delay**
Thermal regression	6/12	5/6	1/6	15	2.50	488.5 s
Naive-last	6/12	6/6	0/6	9	1.498	245.0 s
Chronos-Bolt Small	5/12	5/6	0/6	17	2.829	222.0 s
Chronos-Bolt Base	5/12	5/6	0/6	18	2.996	216.0 s
Spatial normal-only	7/12	6/6	1/6	12	1.997	450.0 s