Libyan Open University Journal of Applied Sciences (LOUJAS)
Volume 2 - Issue 2 - 2026 - Pages 72-86

Cross-Dataset Evaluation of Multisensor Self-Supervised Anomaly Detection for Early Thermal-Runaway Warning in Electric Vehicle Battery Cells

Abdussalam Ali Ahmed [1 *], Rafat S. A. Abumandil [2], Omar Ahemed Mohamed Edbeib [3]
[1,2,3] Mechanical and Industrial Engineering Department, Bani Waleed University, Libya
***Email (for reference researcher): abdussalam.a.ahmed@gmail.com**

التقييم عبر مجموعات البيانات لكشف الشذوذ ذاتي الإشراف متعدد المستشعرات للإنذار المبكر بالانفلات الحراري في خلايا بطاريات المركبات الكهربائية

عبد السلام علي أحمد 1 *، رأفت سالم أبو منديل 2، عمر أحمد محمد الدبيب 3
1، 2، 3 قسم الهندسة الميكانيكية والصناعية، جامعة بني وليد، ليبيا

Received: 08-05-2026; Accepted: 07-07--2026; Published: 10-07-2026

Abstract

Thermal runaway can develop quickly and create severe hazards in electric vehicle battery systems. Early warning is difficult because public fault records are scarce and sensor sets differ across experiments. This study evaluates a multisensor self-supervised anomaly detector using three public cell-level datasets. Healthy fast-charge records supplied unlabeled normal data from six lithium iron phosphate cells. Five target records covered dynamic operation, local heating, uniform heating, and two prismatic NMC abuse tests. A masked denoising autoencoder learned 42 thermal, electrical, pressure, and availability features. We compared it with a conventional autoencoder, Isolation Forest, PCA, one-class SVM, and temperature features. Evaluation used physical-cell splits, leakage-safe prefix calibration, real-time persistence, and block-bootstrap confidence intervals. The masked model detected all three runaway events. Corrected warning leads were 43.77, 34.52, and 1.25 minutes. However, its pooled AUROC was 0.744, and its AUPRC was 0.492. The temperature score performed better, reaching 0.924 AUROC and 0.812 AUPRC. The masked model's negative-window false-positive rate was 0.546. It also produced one persistent warning during the healthy dynamic-load record. Masked training did not improve the conventional autoencoder significantly. The paired AUROC difference was -0.001, with p = .294. These findings show that event detection alone can hide severe calibration problems. Simple thermal features remained more portable under strong domain shift. Multisensor self-supervision still offers useful representations, but it needs better domain adaptation and event labeling. The study provides a reproducible benchmark for safer future validation.

Keywords: Thermal Runaway; Electric Vehicle Battery; Self-Supervised Learning; Anomaly Detection; Multisensor Fusion; Early Warning; Domain Shift.

الملخص

يمكن أن يتطور الانفلات الحراري بسرعة، مما يشكل مخاطر جسيمة على أنظمة بطاريات المركبات الكهربائية. ويُعد الإنذار المبكر بهذه الحالة تحديًا كبيرًا بسبب ندرة سجلات الأعطال المتاحة للعامة واختلاف مجموعات المستشعرات المستخدمة بين التجارب المختلفة. تهدف هذه الدراسة إلى تقييم كاشف للشذوذ متعدد المستشعرات يعتمد على التعلم الذاتي الإشراف باستخدام ثلاث مجموعات بيانات عامة على مستوى خلايا البطاريات. وقد استُخدمت سجلات الشحن السريع السليمة لتوفير بيانات طبيعية غير موسومة من ست خلايا من بطاريات فوسفات حديد الليثيوم (LFP). كما شملت خمس سجلات مستهدفة حالات تشغيل ديناميكي، وتسخينًا موضعيًا، وتسخينًا منتظمًا، واختبارين للإجهاد على خلايا منشورية من بطاريات النيكل والمنغنيز والكوبالت (NMC). وتعلّم نموذج المشفر التلقائي لإزالة الضوضاء المقنّع (Masked Denoising Autoencoder)من خلال 42 سمة حرارية وكهربائية، إضافة إلى بيانات الضغط وتوافر المستشعرات. وقورن أداء هذا النموذج مع كل من المشفر التلقائي التقليدي، وخوارزمية Isolation Forest، وتحليل المكونات الرئيسية (PCA)، وآلة المتجهات الداعمة أحادية الفئة(One-Class SVM) ، إضافة إلى نموذج يعتمد على السمات الحرارية فقط. واعتمد التقييم على تقسيمات مستقلة للخلايا الفيزيائية، ومعايرة آمنة تمنع تسرب البيانات باستخدام البادئات الزمنية، وآلية استمرارية للإنذار في الزمن الحقيقي، وفواصل ثقة محسوبة بطريقة Block Bootstrap.
أظهر النموذج المقنّع قدرةً على اكتشاف جميع حالات الانفلات الحراري الثلاث، حيث بلغت فترات الإنذار المبكر المصححة 43.77 و34.52 و1.25 دقيقة. ومع ذلك، بلغ متوسط قيمة مساحة منحنى خصائص التشغيل للمستقبل (AUROC) عبر جميع البيانات 0.744، في حين بلغت مساحة منحنى الدقة–الاستدعاء (AUPRC) 0.492. في المقابل، حقق النموذج المعتمد على السمات الحرارية أداءً أفضل، إذ سجل قيمة AUROC بلغت 0.924 وقيمة AUPRC بلغت 0.812، كما بلغ معدل الإنذارات الكاذبة في الفترات السليمة للنموذج المقنّع 0.546، وأصدر أيضًا إنذارًا مستمرًا واحدًا أثناء سجل التشغيل الديناميكي السليم. ولم يُظهر التدريب المقنّع تحسنًا ذا دلالة إحصائية مقارنةً بالمشفر التلقائي التقليدي، حيث بلغ الفرق المزدوج في قيمة AUROC مقدار −0.001 مع قيمة **p = 0.294**
وتشير هذه النتائج إلى أن الاعتماد على اكتشاف الأحداث وحده قد يُخفي مشكلات جوهرية في معايرة النماذج. كما أظهرت السمات

الحرارية البسيطة قابلية أفضل للنقل والتعميم في ظل الاختلاف الكبير بين المجالات (Domain Shift). ورغم ذلك، يظل التعلم الذاتي الإشراف متعدد المستشعرات قادرًا على توفير تمثيلات مفيدة للبيانات، إلا أنه يتطلب تحسين تقنيات التكيف مع اختلاف المجالات وتعزيز جودة توصيف الأحداث. وتوفر هذه الدراسة معيارًا مرجعيًا قابلًا لإعادة الإنتاج لدعم عمليات التحقق المستقبلية وتحسين سلامة أنظمة الإنذار المبكر.

الكلمات المفتاحية: الانفلات الحراري؛ بطاريات المركبات الكهربائية؛ التعلم الذاتي الإشراف؛ كشف الشذوذ؛ دمج المستشعرات متعددة المصادر؛ الإنذار المبكر؛ اختلاف المجالات.

Introduction

Lithium-ion cells provide high energy and power for modern electric vehicles. Their safety remains a central engineering concern. Thermal runaway can follow electrical, mechanical, or thermal abuse. Heat then activates linked exothermic reactions inside the cell. The reaction rate can exceed heat removal. Temperature, pressure, and gas generation may then rise very quickly (Finegan et al., 2015).

Fast charging adds another challenge for battery monitoring. High current changes voltage, temperature, and internal resistance during normal operation. These changes may resemble early fault signals. Fast-charge records therefore offer valuable normal examples for anomaly learning. Severson et al. (2019) released a large dataset from commercial lithium iron phosphate cells. Attia et al. (2020) later used related data for closed-loop charging optimization.

Early warning must occur before uncontrollable reactions develop. A practical system must also avoid frequent false alarms. An alert that always activates during normal heating has little operational value. Event detection alone is therefore an incomplete measure. Warning lead, false-alert burden, calibration stability, and domain transfer require joint assessment.

Many experimental warning methods use temperature, voltage, impedance, pressure, gas, strain, or acoustic signals. Klink et al. (2021) detected simulated thermal faults through temperature-sensitive electrical behavior. Appleberry et al. (2022) identified overcharge failure through ultrasound changes. Zhang et al. (2023) combined voltage, temperature, and gas stages during overcharge. These studies show that several sensor types can provide useful evidence.

Public thermal-runaway datasets remain small and heterogeneous. Cell formats, chemistries, triggers, sample rates, and sensors often differ. A model trained on one source may fail on another source. This domain shift is especially important for self-supervised learning. The model may learn the source identity instead of transferable safety behavior.

Self-supervised anomaly detection offers a possible response to scarce fault labels. The model learns normal relationships from unlabeled data. Unexpected reconstruction errors then become anomaly scores. Audibert et al. (2020) used adversarial autoencoders for multivariate monitoring. Fu and Xue (2022) showed that random masking can improve normal representation learning. Transformer models also capture wider temporal relations (Tuli et al., 2022; Xu et al., 2022).

This study tests whether masked multisensor learning transfers across public battery experiments. The work uses healthy fast-charge data and five target records. Three target records end in confirmed thermal runaway. One record contains a non-runaway local fault. Another record provides healthy dynamic-load validation.

The contribution is a strict real-data evaluation framework. It uses leakage-safe preprocessing, physical-cell splits, real-time alarm timing, and conservative negative labels. It also reports block-bootstrap intervals and threshold sensitivity. The results do not assume that model complexity guarantees better safety performance.

The final evidence is mixed but informative. The masked model detects every runaway event. However, temperature features achieve better discrimination. The masked model also raises substantial negative-window alarms. These findings define the limits of the present approach and guide stronger future systems.

Scientific Background and Related Work

Thermal-runaway development

Thermal runaway describes uncontrolled self-heating inside an electrochemical cell. The process involves several coupled reactions. Solid-electrolyte interphase decomposition can begin at elevated temperature. Anode, electrolyte, separator, and cathode reactions then add heat and gas. The exact sequence depends on chemistry, state of charge, aging, and cell design.

Golubkov et al. (2014) compared commercial cells with different cathode materials. Their experiments reached cell temperatures near 850 °C. Gas production also varied with chemistry. Finegan et al. (2015) later observed gas pockets, delamination, collapse, and venting through high-speed tomography. These internal changes may occur before surface measurements show catastrophic failure.

Pressure provides another view of degradation. Ayayda et al. (2024b) measured pressure and temperature in prismatic NMC111 and NMC811 cells. Their NMC111 cell vented at 152 °C. Thermal runaway followed 35 minutes later. The NMC811 cell vented at 185 °C. Runaway followed after only 104 seconds. This difference shows why one fixed warning delay cannot describe every chemistry.

Early-warning signals

Surface temperature is simple, cheap, and common in battery management systems. Yet surface sensors can lag internal reactions. Sensor position also affects observed peaks. Temperature remains valuable because heat is a direct failure outcome. The present results confirm its strong cross-dataset value.

Electrical signals can respond before extreme temperature appears. Klink et al. (2021) linked heating with impedance-related voltage changes under dynamic load. Their method detected faults before venting or voltage collapse. Voltage and current are already available in vehicle battery systems. Their use can reduce added hardware.

Pressure and gas measurements can identify internal decomposition or venting. Ayayda et al. (2024b) showed clear pressure growth before rupture-disc opening. Zhang et al. (2023) used gas, voltage, and temperature for staged warning. Acoustic monitoring can reveal mechanical changes during overcharge (Appleberry et al., 2022). These modalities offer different warning times and integration costs.

Multisensor fusion should combine complementary evidence. However, missing modalities create practical difficulties. The public records used here do not share one sensor layout. Pressure is absent from Kokam records. Current is absent from both NMC records. NMC111 also lacks voltage. Explicit availability masks are therefore necessary.

Self-supervised anomaly detection

Anomaly labels are rare because destructive battery testing is expensive. Self-supervision creates a learning task from normal data. A model may reconstruct masked inputs or predict future samples. Large errors suggest that the new sample differs from learned normal behavior.

Autoencoders provide a compact normal representation. USAD uses related autoencoders with adversarial training (Audibert et al., 2020). MAD masks random time-series values during training (Fu & Xue, 2022). TranAD and the Anomaly Transformer use attention for wider temporal structure (Tuli et al., 2022; Xu et al., 2022). These methods perform strongly on common industrial benchmarks.

Benchmark success does not guarantee battery transfer. Battery signals depend strongly on chemistry, format, trigger, and test protocol. A reconstruction model may react to harmless source differences. It may also reconstruct faults too well. Careful baselines and false-alert reporting are therefore essential.

Research gap and questions

Published battery studies often test one chemistry or one abuse method. Machine-learning studies often use benchmark labels that are already aligned. Few studies combine fast-charge normal data with independent thermal-abuse experiments. Fewer still report operational persistence and source-specific false alerts.

This study addresses three questions. Can masked multisensor learning detect runaway across different public datasets? Does masking improve a matched conventional autoencoder? Does the complex model outperform simple thermal features under domain shift?

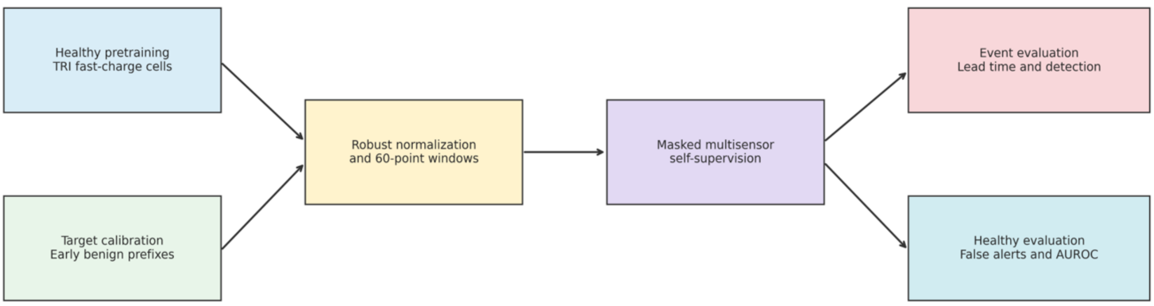

Figure 1 Leakage-safe cross-dataset experimental workflow.

Materials and Methods

Public datasets

Three public sources supported the experiment. Checksums verified every small downloaded file. Six structured TRI files were extracted through remote ZIP ranges. This avoided downloading the complete 27.9 GB archive. The selected files contained 480 charge cycles.

The TRI records describe commercial LFP/graphite cells under fast-charge protocols (Severson et al., 2019). Their trajectories contain voltage, current, and casing temperature. They do not contain physical seconds. We therefore used normalized charge-phase position only for learning. No lead time was calculated from TRI records.

The Kokam source contains a 46 Ah pouch cell under dynamic vehicle load (Klink, 2024). Uniform heating eventually caused thermal runaway. Local heating produced a fault without runaway. A separate WLTC record remained healthy. Heater sensors supported annotation but never entered model features.

Libyan Open University Journal of Applied Sciences (LOUJAS)
Volume 2 - Issue 2 - 2026 - Pages 72-86

The third source contains NMC111 and NMC811 prismatic abuse experiments (Ayayda, 2024a). Temperature and internal pressure were available for both cells. NMC811 also provided voltage. Constant external heating triggered each event under different rates and starting conditions.

Table 1 Public datasets and experimental roles

Dataset	Cells or tests	Available signals	Study role
TRI fast-charge	6 cells	Current, voltage, casing temperature	Healthy training, validation, and held-out testing
Kokam 46 Ah	3 tests	Current, voltage, surface temperatures	Uniform runaway, local fault, and healthy WLTC
NMC prismatic	2 tests	Temperature, pressure, NMC811 voltage	External abuse validation across two chemistries

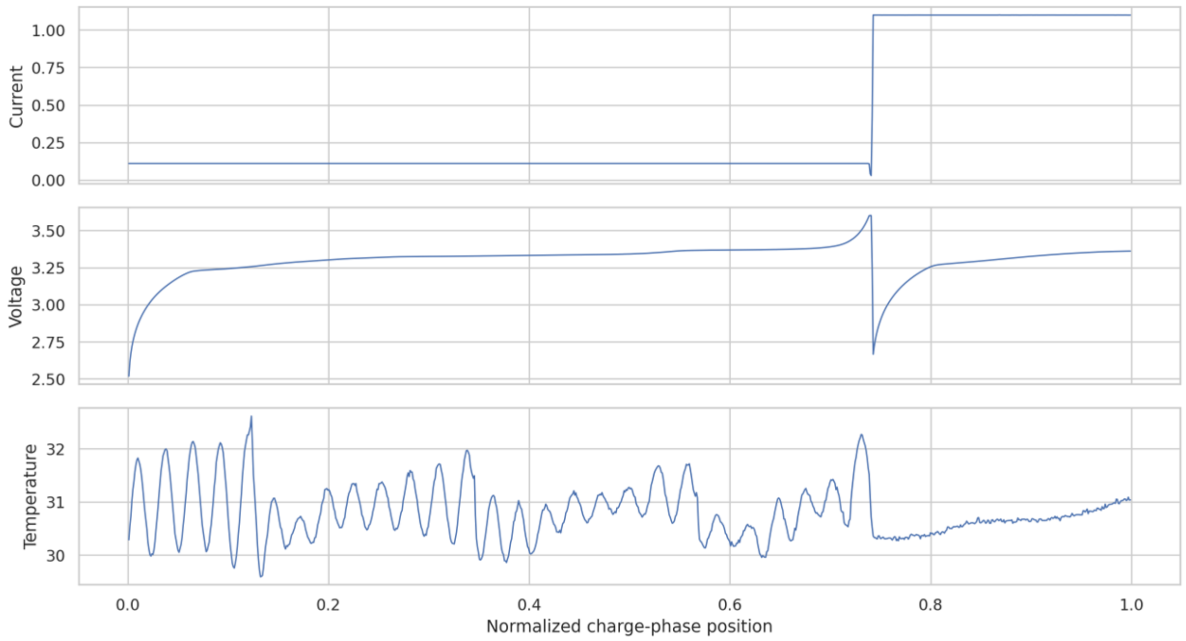

Figure 2 Representative healthy fast-charge trajectory from the TRI source.

Event definitions

Event definitions were frozen before model scoring. Kokam uniform runaway began at 5,255 seconds. This rule required two surface sensors above 100 °C. Their heating rates also exceeded 1 °C per second for three seconds. The excluded heater channel marked a fault proxy at 2,411 seconds.

The local Kokam record used 2,460 seconds as its fault reference. No thermal runaway occurred in that experiment. The healthy Kokam record had no fault reference. Its post-calibration period measured false-alert behavior.

The NMC event times came directly from the source article. NMC111 first vented at 10,260 seconds. Thermal runaway began at 12,360 seconds. NMC811 first vented at 1,350 seconds. Its runaway began at 1,454 seconds (Ayayda et al., 2024b).

Table 2 Frozen target-event definitions

Experiment	Outcome	Calibration end	Precursor or vent	Runaway onset
Kokam Uniform	Runaway	2,200 s	2,411 s	5,255 s
Kokam Local	Non-runaway fault	2,200 s	2,460 s	Not observed
Kokam Validation	Healthy load	800 s	Not applicable	Not observed
NMC111	Runaway	3,600 s	10,260 s	12,360 s
NMC811	Runaway	300 s	1,350 s	1,454 s

Libyan Open University Journal of Applied Sciences (LOUJAS)
Volume 2 - Issue 2 - 2026 - Pages 72-86

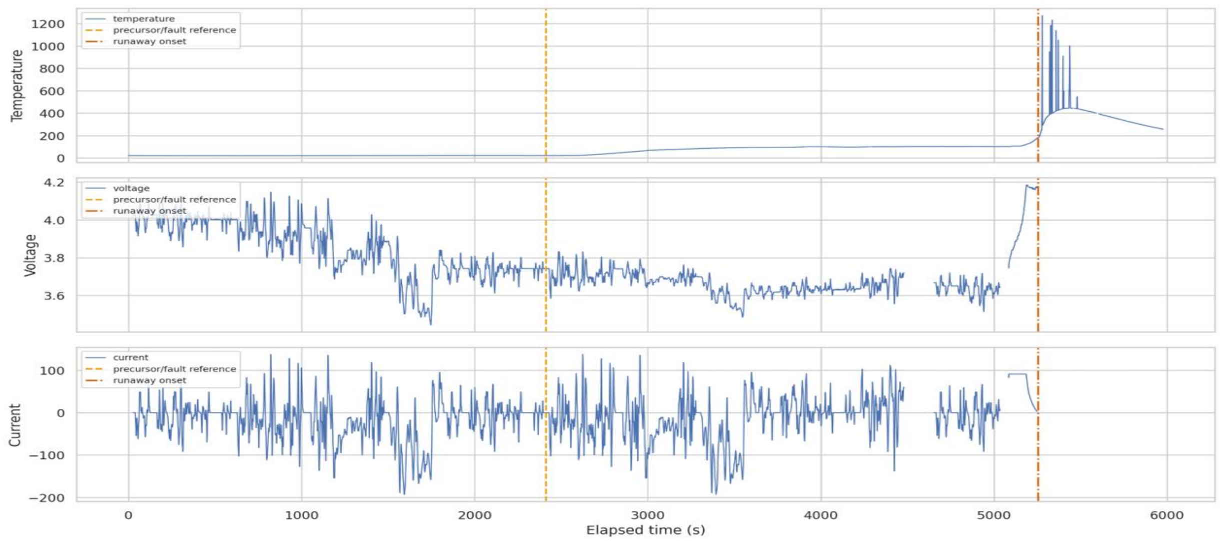

Figure 3 Kokam uniform-heating record with frozen fault and runaway references.

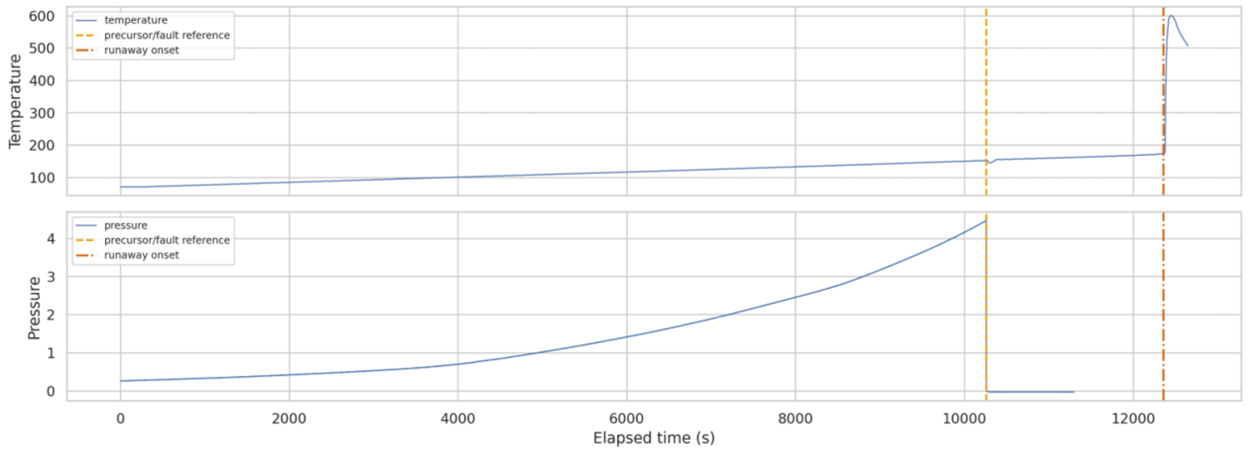

Figure 4 NMC111 temperature and pressure before venting and thermal runaway.

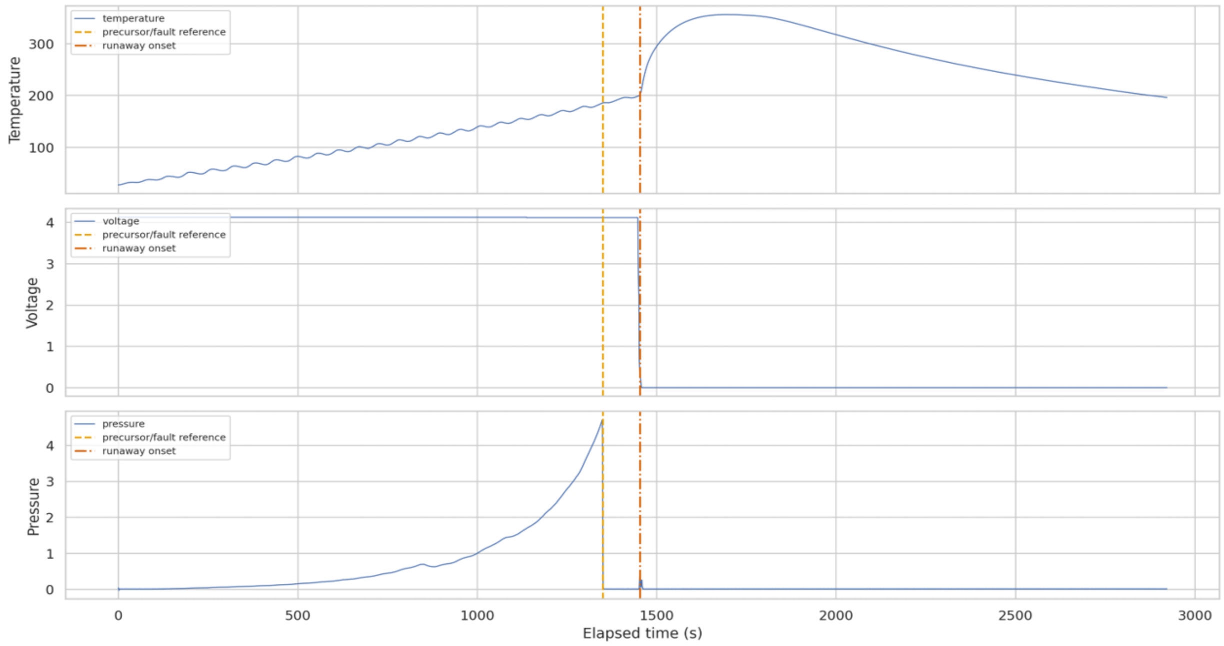

Figure 5 NMC811 temperature, voltage, and pressure under asymmetric heating.

Leakage-safe preprocessing

All physical-time target records were aligned to one-second samples. Short internal gaps were interpolated for at most three seconds. Temperature represented the maximum allowed surface sensor. Temperature spread represented the difference between allowed surface sensors.

Each target record supplied an early benign calibration prefix. The first 70% fitted its robust scaler and supported training. The remaining 30% calibrated thresholds. Future target samples never entered preprocessing or training. Robust scaling used the median and median absolute deviation. A lower scale bound prevented unstable division. The normalized signal was calculated as follows.

$$z = \frac{[x - median(x_{\mathrm{train}})]}{\max[1.4826\, MAD(x_{\mathrm{train}}), s_{\mathrm{min}}]} \tag{1}$$

Target windows covered 60 seconds with a ten-second stride. TRI windows covered 128 interpolated charge points with a 64-point stride. Each modality produced nine summary statistics. They covered mean, spread, bounds, change, slope, and local differences.

The final vector contained 42 features. Thirty-eight features were numeric. Four binary fields described modality availability. Missing numeric modalities were zeroed after standardization. Their presence masks prevented loss or score contributions.

Table 3 Window allocation and model inputs

Source	Split	Windows	Purpose
TRI	Training	2,724	Healthy representation learning
TRI	Validation	640	Global validation
TRI	Healthy test	640	Held-out fast-charge testing
Targets	Calibration training	617	Prefix adaptation
Targets	Calibration validation	268	Threshold calibration
Targets	Evaluation	2,411	Future event and healthy evaluation

Masked multisensor autoencoder

The proposed network used a fully connected encoder and decoder. The encoder widths were 128, 64, and 16. The decoder reversed those widths. GELU activations, layer normalization, and 10% dropout supported stable training. During training, 20% of observed numeric features were randomly masked. Small Gaussian noise was also added. The model reconstructed the original observed features. The loss combined masked and total observed errors.

$$L = 0.80\, MSE(masked) + 0.20\, MSE(observed) \tag{2}$$

Three masked models used seeds 17, 41, and 73. Three conventional autoencoders used the same seeds. Their architecture and data were identical. Only masked corruption changed. Three more masked models used TRI training data alone.

AdamW used a learning rate of 0.001 and weight decay of 0.0001. Batch size was 256. Training allowed 100 epochs. Early stopping used twelve stale epochs. The final anomaly score averaged reconstruction errors across three seeds.

$$a_{\mathrm{i}} = \left(\frac{1}{K}\right) \Sigma_{\mathrm{k}} \left[\left(\frac{1}{O_{\mathrm{i}}}\right) \Sigma_{\mathrm{j}} \in O_{\mathrm{i}}\, (x_{\mathrm{ij}} - \hat{x}_{\mathrm{ijk}})^2\right] \tag{3}$$

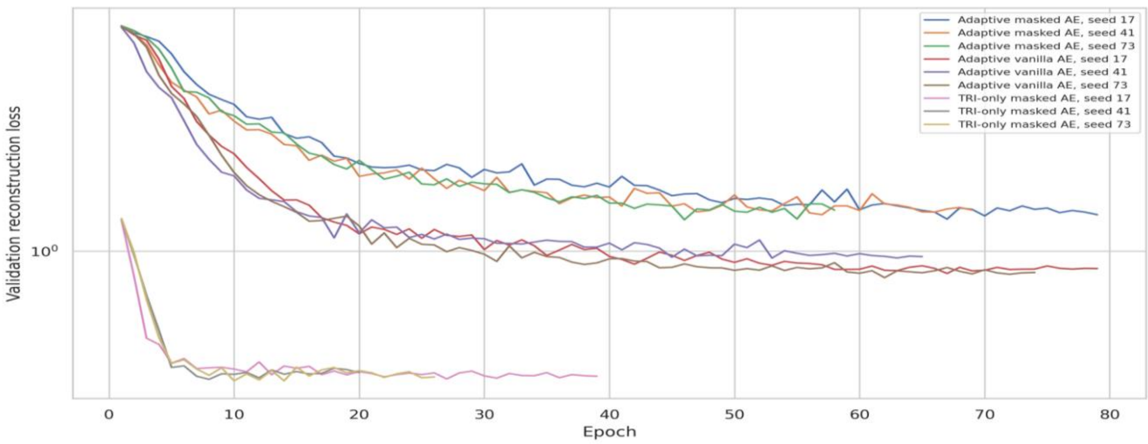

Figure 6 Validation reconstruction loss across adaptive and TRI-only models.

Baselines

Five baselines tested whether model complexity added value. The conventional autoencoder measured the effect of masking. PCA measured linear reconstruction error. Isolation Forest isolated rare feature patterns (Liu et al., 2008). One-class SVM estimated the normal support boundary (Schölkopf et al., 2001).

The temperature-feature score used only standardized temperature statistics. It was not a fixed temperature threshold. The score was the mean squared magnitude across available thermal features. This baseline represented a simple, low-cost detector.

Alert rule and evaluation

Each method used the 99.5th validation-score percentile. Three consecutive windows were required for an alert. The alert time was the third window's end. Earlier versions backdated this time. The final code corrected that problem.

Positive windows began at the frozen precursor. They ended at runaway onset. Negative windows included healthy records and post-calibration pre-reference windows. The latter NMC windows occurred during continued external heating. Their negative status is deliberately conservative.

The pooled evaluation contained 1,711 negative windows and 504 positive windows. Overlapping windows were not treated as independent for uncertainty. Six adjacent stride windows formed each bootstrap block. One thousand resamples produced 95% intervals (Politis & Romano, 1994).

AUROC measured ranking across both labels (Hanley & McNeil, 1982). AUPRC emphasized the positive class. We also measured event detection, lead time, alert episodes, and alerts per hour. Paired bootstrap differences compared masked and conventional autoencoders.

Table 4 Primary model and evaluation settings

Parameter	Value	Interpretation
Target window	60 s	One-minute physical context
Target stride	10 s	Sixfold overlap
Latent width	16	Compressed normal representation
Mask probability	0.20	Self-supervised corruption
Threshold	99.5th percentile	Validation anomaly cutoff
Persistence	3 windows	Real-time alarm confirmation
Training seeds	17, 41, 73	Matched stability assessment
Bootstrap	1,000 blocks	Conditional uncertainty interval

Results

Corrected event detection

The masked ensemble detected all three runaway events after the frozen precursors. Its corrected Kokam lead was 43.77 minutes. NMC111 produced 34.52 minutes. NMC811 produced 1.25 minutes. The median was 34.52 minutes.

The temperature score also detected all three events. Its leads were 43.93, 34.52, and 1.25 minutes. PCA produced 44.93, 34.52, and 1.25 minutes. Isolation Forest missed every runaway event.

The first operational alert requires separate interpretation. Masked scoring raised no earlier alert in Kokam Uniform. It raised alerts before first venting in both NMC tests. Those alerts occurred during continuing external heating. The frozen protocol counted them as pre-reference false alerts.

The local Kokam fault did not reach runaway. The masked model alerted 10.15 minutes after its fault reference. The temperature score alerted after 8.15 minutes. These results show sensitivity to heating that remained below runaway.

Table 5 Event-level warning results

Method	Kokam lead	NMC111 lead	NMC811 lead	Events detected
Masked AE	43.77 min	34.52 min	1.25 min	3/3
Vanilla AE	43.77 min	34.52 min	1.25 min	3/3
Temperature features	43.93 min	34.52 min	1.25 min	3/3
PCA	44.93 min	34.52 min	1.25 min	3/3
One-class SVM	43.93 min	34.52 min	1.25 min	3/3
Isolation Forest	Not detected	Not detected	Not detected	0/3

Libyan Open University Journal of Applied Sciences (LOUJAS)
Volume 2 - Issue 2 - 2026 - Pages 72-86

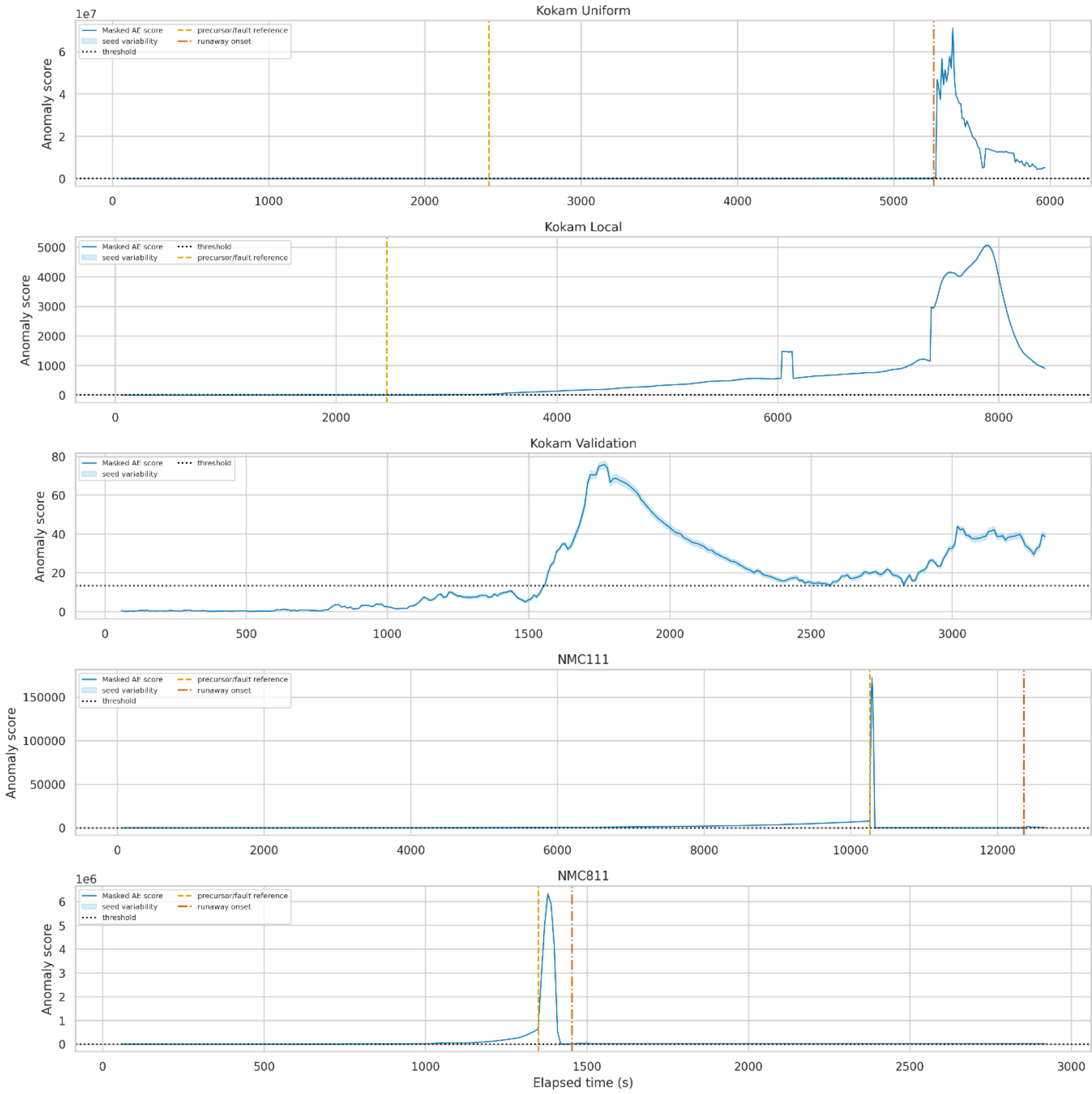

Figure 7 Masked-autoencoder anomaly scores across all target experiments.

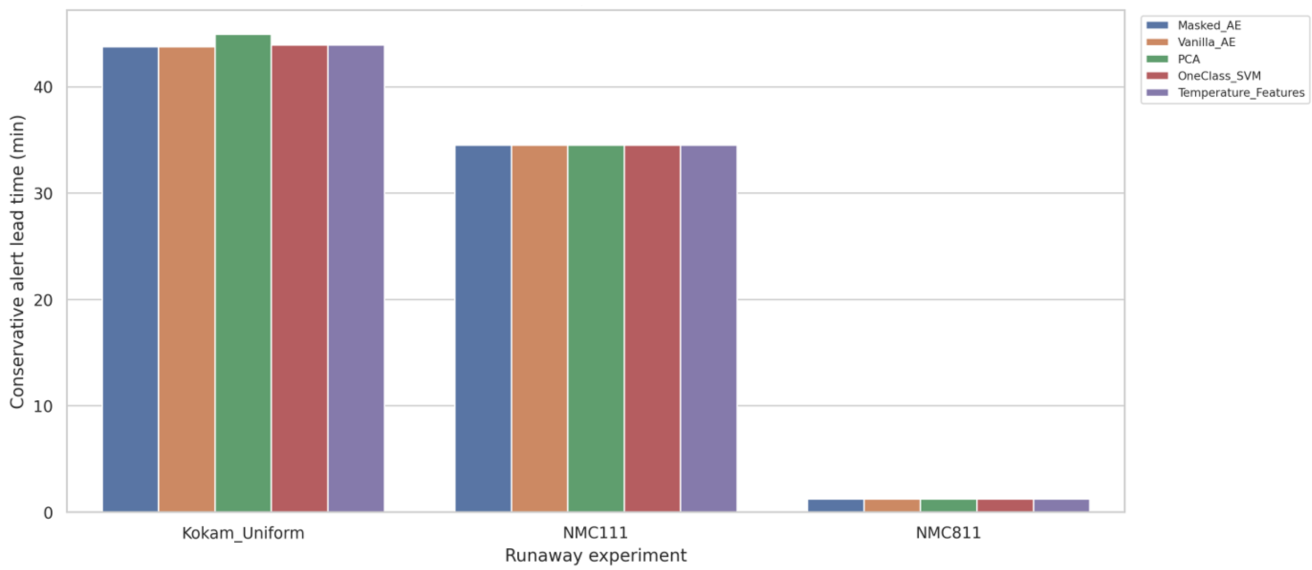

Figure 8 Corrected event-level warning lead across methods.

Window discrimination

The temperature-feature score achieved the best pooled discrimination. Its AUROC was 0.924. Its AUPRC was 0.812. The masked ensemble reached 0.744 AUROC and 0.492 AUPRC. The conventional ensemble produced nearly identical values.

PCA reached 0.713 AUROC and 0.499 AUPRC. Its negative-window false-positive rate was 0.131. Isolation Forest had the lowest rate at 0.059. However, it missed every runaway event. This result shows the tradeoff between sensitivity and alert burden.

Per-event performance was uneven. The masked AUROC was 0.888 for Kokam Uniform. It fell to 0.545 for NMC111. It reached 0.847 for NMC811. Sensor patterns and heating profiles therefore changed ranking quality strongly.

Table 6 Corrected aggregate model results

Method	AUROC	AUPRC	Macro AUROC	Negative FPR	Detected
Temperature features	0.924	0.812	0.929	0.272	3/3
Vanilla AE	0.745	0.493	0.760	0.543	3/3
Masked AE	0.744	0.492	0.760	0.546	3/3
PCA	0.713	0.499	0.753	0.131	3/3
One-class SVM	0.652	0.291	0.659	0.557	3/3
Isolation Forest	0.444	0.189	0.575	0.059	0/3

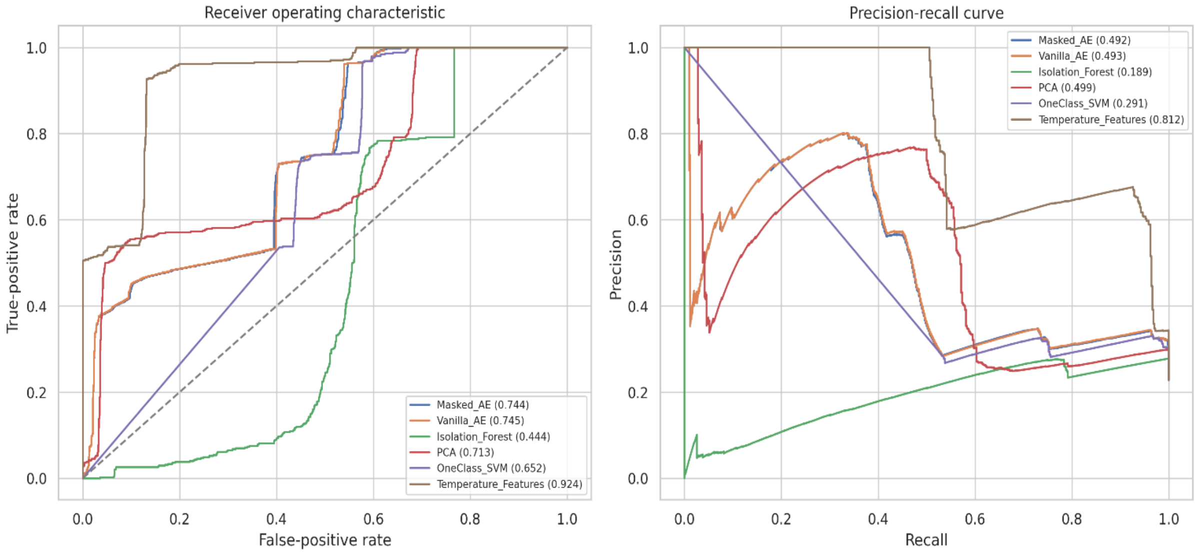

Figure 9 ROC and precision-recall curves under conservative window labels.

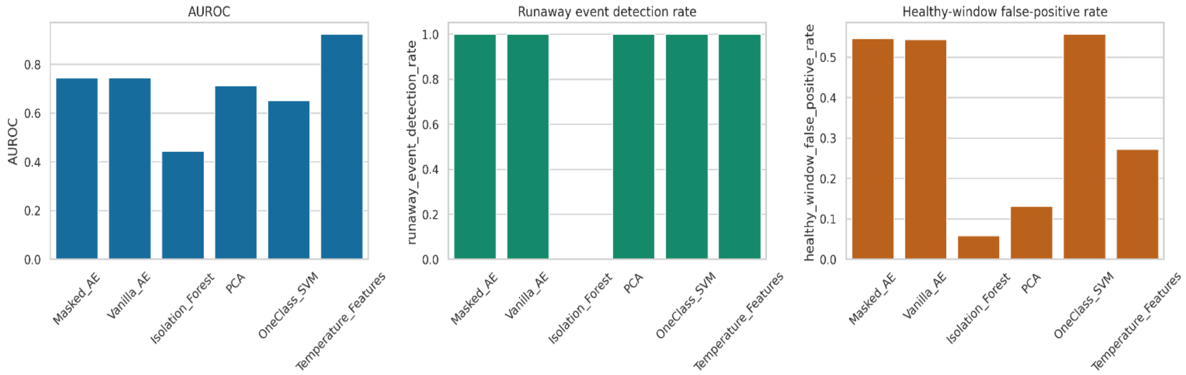

Figure 10 Aggregate discrimination, event detection, and negative-window alarms.

Bootstrap uncertainty and seed stability

Block bootstrap intervals confirmed the performance gap against temperature features. The masked AUROC interval was 0.688 to 0.797. Its AUPRC interval was 0.405 to 0.596. Temperature intervals were 0.894 to 0.950 and 0.748 to 0.869.

Masking did not improve the matched conventional autoencoder. The paired AUROC difference was -0.001. Its 95% interval ranged from -0.003 to 0.001. The two-sided bootstrap p-value was .294. AUPRC and false-positive differences were also not significant.

Libyan Open University Journal of Applied Sciences (LOUJAS)
Volume 2 - Issue 2 - 2026 - Pages 72-86

Seed variation was small. Masked AUROC averaged 0.744 with a 0.002 standard deviation. AUPRC averaged 0.492 with a 0.001 standard deviation. All six neural models detected all three events. Training randomness was therefore not the main weakness.

Table 7 Block-bootstrap intervals for selected methods

Method	AUROC estimate	AUROC 95% interval	AUPRC estimate	AUPRC 95% interval
Masked AE	0.744	0.688-0.797	0.492	0.405-0.596
Vanilla AE	0.745	0.687-0.798	0.493	0.406-0.596
PCA	0.713	0.644-0.779	0.499	0.408-0.610
Temperature features	0.924	0.894-0.950	0.812	0.748-0.869

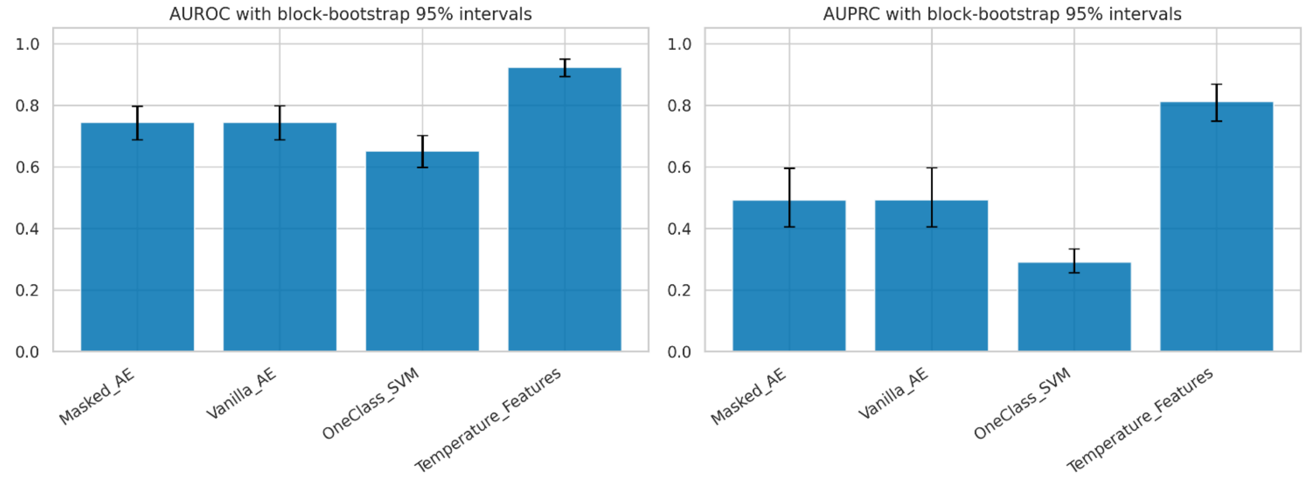

Figure 11 Block-bootstrap 95% intervals for AUROC and AUPRC.

False-alert burden

The masked model performed well on held-out TRI fast-charge cycles. Only one of 640 windows crossed the raw threshold. No charge cycle produced a persistent alert. This result shows strong within-source specificity.

The healthy Kokam WLTC record exposed a transfer problem. About 70.4% of its evaluation windows exceeded the threshold. These windows formed one persistent episode over 0.703 hours. The episode began at 1,579 seconds. Conservative pre-reference labeling produced one episode in NMC111 and one in NMC811. The combined physical-time rate was 1.008 episodes per hour. These periods were externally heated. They should not be interpreted as confirmed healthy operation.

A high window rate can represent one long warning rather than many alarms. Episode counts therefore complement window metrics. Both measures remain important for battery-management design.

Table 8 Masked-model false-alert burden by source

Source	Negative-window rate	Persistent episodes	Observation basis	Episode rate
TRI healthy test	0.002	0 cycles	80 charge cycles	0.000 per cycle
Kokam Validation	0.704	1	0.703 h	1.423 per hour
Kokam Uniform pre-reference	0.000	0	0.059 h	0.000 per hour
Kokam Local pre-reference	0.000	0	0.072 h	0.000 per hour
NMC111 pre-vent	0.977	1	1.850 h	0.541 per hour
NMC811 pre-vent	1.000	1	0.292 h	3.429 per hour

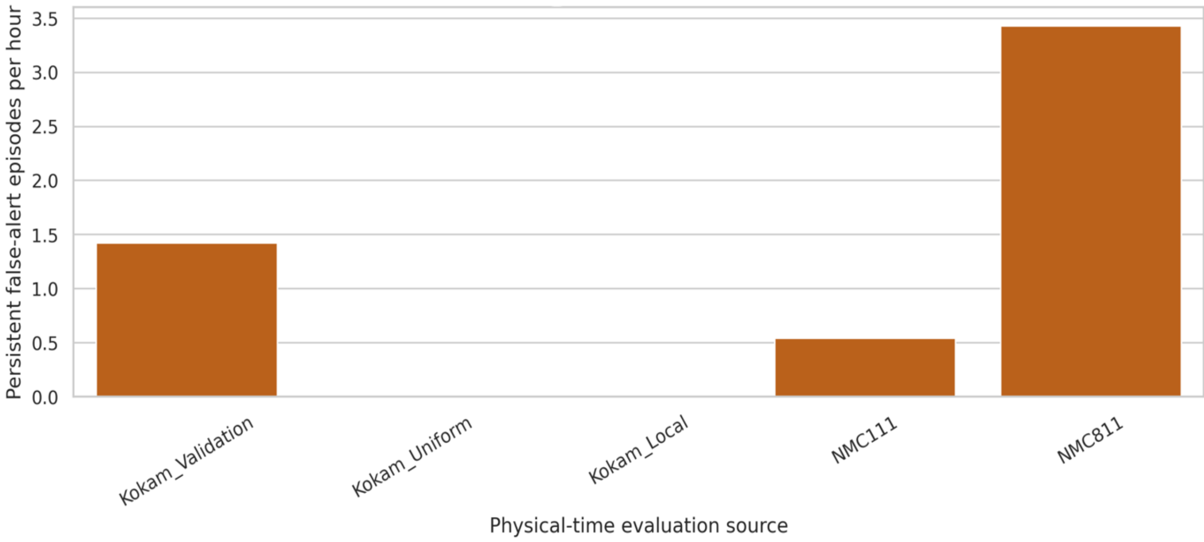

Figure 12 Source-specific persistent alert episodes before frozen references.

Sensitivity, sensors, and deployment

All twenty threshold and persistence settings detected all three runaway events. Higher persistence shortened the minimum NMC811 lead. It fell from 1.58 minutes to 0.75 minutes. False episodes varied because higher thresholds split long score runs.

Thermal-only scoring detected every event. Pressure-only scoring detected both NMC events. Electrical-only scoring detected only Kokam Uniform. This pattern matches sensor availability. It also shows that useful modalities depend on the target experiment.

Removing pressure did not reduce event detection. Full multisensor scoring produced median lead of 34.52 minutes. Thermal-only scoring produced the same median. Pressure remained valuable for NMC111 because its temperature changed gradually before venting.

The TRI-only ensemble reached 0.764 AUROC and 0.622 AUPRC. Those values exceeded the adaptive masked ensemble. Its negative-window rate also rose to 0.629. Prefix adaptation reduced alarms but weakened overall ranking.

Table 9 Sensor ablation and deployment comparison

Configuration	Kokam lead	NMC111 lead	NMC811 lead	Main finding
Full multisensor	43.77 min	34.52 min	1.25 min	Detected all events
No pressure	43.93 min	34.52 min	1.25 min	No event loss
Thermal only	43.93 min	34.52 min	1.25 min	Detected all events
Electrical only	28.43 min	Not detected	Not detected	Kokam only
Pressure only	Not detected	34.52 min	1.25 min	NMC events only
TRI-only ensemble	43.77 min*	34.52 min*	1.25 min*	Higher FPR: 0.629

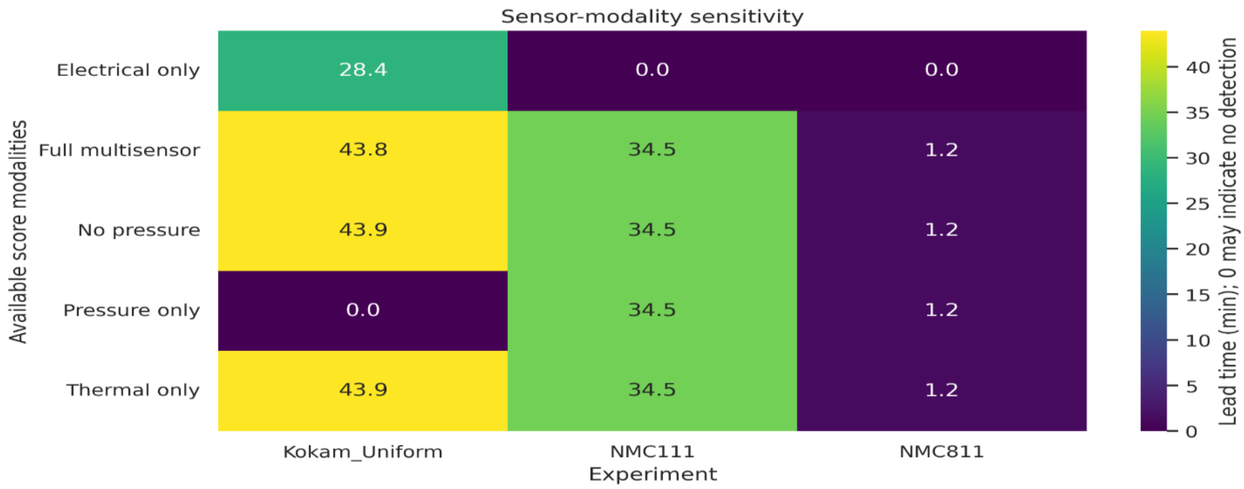

Figure 13 Event lead under sensor-modality ablation.

Libyan Open University Journal of Applied Sciences (LOUJAS)
Volume 2 - Issue 2 - 2026 - Pages 72-86

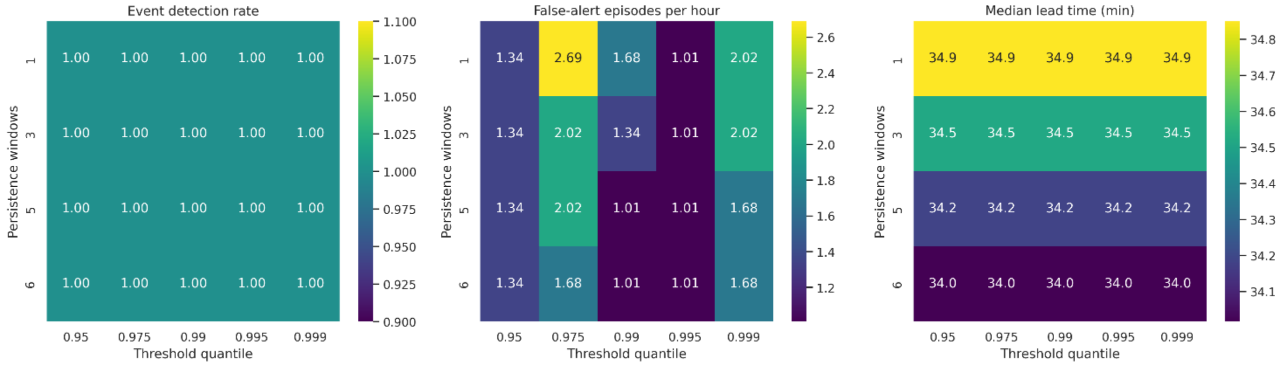

Figure 14 Detection, false-alert, and lead-time sensitivity across alert settings.

Discussion

Main interpretation

The masked model achieved the most visible goal. It detected all three runaway events before frozen onset. Yet this result alone overstates practical quality. The model also showed weak ranking and substantial negative-window activation.

The temperature-feature score was the strongest overall detector. It used far fewer inputs than the masked model. Its AUROC exceeded the masked result by 0.180. Its AUPRC exceeded the masked result by 0.320. The simple baseline therefore remained more portable.

Masked learning also failed to beat conventional reconstruction. Their paired intervals covered zero. Their event leads were identical. The masked objective learned stable representations, but those representations did not improve this transfer task.

The result is not evidence against self-supervised learning generally. It is evidence against uncritical deployment across unmatched battery sources. Strong source differences can dominate reconstruction scores. A model may detect chemistry, heating rate, or sensor layout instead of fault progression.

Why the temperature score performed well

Thermal runaway is defined by accelerating heat generation. Temperature features therefore remain close to the physical outcome. Their slopes, ranges, and local differences changed strongly near the events. Standardized thermal magnitude also avoided pressure and electrical missingness.

The temperature score did not require cross-modal reconstruction. Missing pressure could not distort it. Missing current could not distort it. This simplicity reduced sensitivity to source-specific availability masks.

However, temperature features also raised one alert in the healthy Kokam record. Their window-level rate reached 0.842 there. Temperature alone is therefore not deployment-ready. Better thresholds, spatial confirmation, and operating-state context remain necessary.

Domain shift and label ambiguity

The three sources differ sharply. TRI uses LFP cells and normalized charge trajectories. Kokam uses a pouch cell under dynamic load. NMC records use prismatic cells under continuous external heating. Their sensors and sample structures also differ.

These differences complicate the negative label. NMC pre-vent periods contain sustained heating and pressure growth. Calling every pre-vent window healthy would be incorrect. The study instead calls them conservative negative windows. Early alerts in those periods may reflect real abuse progression.

Exact early-warning claims therefore require stronger onset labels. Venting is clear and reproducible. It is not necessarily the first hazardous change. Future experiments should record internal temperature, gas, pressure, impedance, and intervention points together.

Relation to earlier experiments

Klink et al. (2021) reported electrical fault detection before runaway. Our electrical-only result also detected Kokam Uniform. It failed on NMC records because current was unavailable. This confirms that modality conclusions depend on sensor coverage.

Ayayda et al. (2024b) showed that pressure drops sharply at venting. Our pressure-only score detected both NMC events. The NMC111 lead matched the 35-minute interval between venting and runaway. The NMC811 interval was much shorter.

Appleberry et al. (2022) showed that ultrasound can identify overcharge before failure. Zhang et al. (2023) supported staged fusion of electrical, thermal, and gas evidence. These approaches use signals tied to specific failure mechanisms. The present generic reconstruction model lacked that physical structure.

Industrial anomaly studies often report gains on aligned benchmarks (Audibert et al., 2020; Fu & Xue, 2022). Battery deployment introduces harder domain changes. Our findings show why source-aware validation must accompany benchmark performance.

Practical implications

A vehicle warning system should not rely on one anomaly score. A layered design is more suitable. The first layer can monitor robust temperature change. A second layer can confirm spatial spread, voltage response, pressure, or gas evidence. A final layer can trigger intervention.

Calibration should also follow operating state. Fast charging, rest, driving load, and cold operation create different normal patterns. One global threshold cannot cover every state reliably. Conformal or state-conditioned thresholds may reduce false alarms.

Real-time timing must remain strict. Persistence should never be backdated. Every reported lead must use the confirmation time available to the controller. This correction reduced each three-window lead by twenty seconds.

False-alert episodes should accompany window rates. A long threshold crossing may create one operator warning. Many short crossings may create repeated alarms. Both patterns require different mitigation.

Limitations

The study includes only three confirmed runaway events. This count cannot support population-level safety claims. The bootstrap intervals describe dependent windows within observed records. They do not describe all cells, packs, or abuse causes.

The target experiments use thermal triggers. They do not reproduce runaway caused directly by fast charging. Fast-charge data support normal representation learning only. The paper title therefore avoids claiming fast-charge-induced runaway detection.

Cell formats and chemistries are not balanced. LFP supplies healthy fast-charge records. Kokam and NMC supply faults. Chemistry and label are partly confounded. More matched normal and abuse records are needed.

Precursor labels have different meanings. Kokam uses a heater-related fault proxy. NMC uses first venting. These labels support reproducibility, but they do not share one physical onset definition.

The model uses engineered window statistics. It does not learn directly from raw sequences. A larger transformer or recurrent model might capture wider temporal context. Such models would require more independent events.

The healthy Kokam test contains less than one hour after calibration. False alerts need much longer fleet-like evaluation. Temperature, load, aging, and seasonal conditions should be represented.

Future Work

Future experiments should pair normal and abuse records within each chemistry. They should use shared sensor layouts and synchronized physical time. Repeated cells should support event-level confidence intervals.

The next model should combine physics and self-supervision. Temperature-rate limits can constrain anomaly scores. Pressure and gas changes can confirm internal decomposition. Electrical residuals can provide low-cost context.

Domain adaptation should be evaluated without future leakage. Leave-one-chemistry-out testing would measure transfer. Online conformal calibration could control false-warning probability. It should use only past benign observations.

Event labeling also needs improvement. Researchers should mark heating start, self-heating onset, gas onset, venting, separator failure, and runaway. Intervention studies should identify the last actionable time.

Pack-level validation is essential. Cell spacing, cooling, compression, and neighboring heat change warning signals. A pack controller must also consider propagation risk and evacuation time.

Conclusion

This study provides a reproducible cross-dataset test of multisensor self-supervised warning. The masked model detected all three public runaway events. Corrected leads ranged from 1.25 to 43.77 minutes.

The broader results were less favorable. Masked AUROC was 0.744, and AUPRC was 0.492. Temperature features reached 0.924 and 0.812. Masking did not improve a matched conventional autoencoder.

False alerts remained the main barrier. The masked detector formed one warning during the healthy Kokam record. Conservative NMC labeling also exposed strong pre-vent activation. Event detection alone would have hidden these weaknesses.

The main contribution is therefore methodological and diagnostic. Real-time timing, source-specific alerts, dependent-window uncertainty, and strong simple baselines change the conclusion. Complex models should earn their added cost through independent transfer tests.

Multisensor self-supervision remains promising, but the current design is not deployment-ready. Future systems need physics-aware fusion, better labels, and broader normal testing. These steps can turn early detection into reliable safety action.

Compliance with ethical standards

Disclosure of conflict of interest

The authors declare that they have no conflict of interest.

References

1. Appleberry, M. C., Kowalski, J. A., Africk, S. A., Mitchell, J., Ferree, T. C., Chang, V., Parekh, V., Xu, Z., Ye, Z., Whitacre, J. F., & Murphy, S. D. (2022). Avoiding thermal runaway in lithium-ion batteries using ultrasound detection of early failure mechanisms. Journal of Power Sources, 535, 231423. https://doi.org/10.1016/j.jpowsour.2022.231423
2. Attia, P. M., Grover, A., Jin, N., Severson, K. A., Markov, T. M., Liao, Y.-H., Chen, M. H., Cheong, B., Perkins, N., Yang, Z., Herring, P. K., Aykol, M., Harris, S. J., Braatz, R. D., Ermon, S., & Chueh, W. C. (2020). Closed-loop optimization of fast-charging protocols for batteries with machine learning. Nature, 578, 397-402. https://doi.org/10.1038/s41586-020-1994-5
3. Audibert, J., Michiardi, P., Guyard, F., Marti, S., & Zuluaga, M. A. (2020). USAD: UnSupervised anomaly detection on multivariate time series. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 3395-3404. https://doi.org/10.1145/3394486.3403392
4. Ayayda, M. (2024a). Modeling thermal runaway mechanisms and pressure dynamics in prismatic lithium-ion batteries [Data set]. Zenodo. https://doi.org/10.5281/zenodo.13981390
5. Ayayda, M., Benger, R., Reichrath, T., Kasturia, K., Klink, J., & Hauer, I. (2024b). Modeling thermal runaway mechanisms and pressure dynamics in prismatic lithium-ion batteries. Batteries, 10(12), 435. https://doi.org/10.3390/batteries10120435
6. Finegan, D. P., Scheel, M., Robinson, J. B., Tjaden, B., Hunt, I., Mason, T. J., Millichamp, J., Di Michiel, M., Offer, G. J., Hinds, G., Brett, D. J. L., & Shearing, P. R. (2015). In-operando high-speed tomography of lithium-ion batteries during thermal runaway. Nature Communications, 6, 6924. https://doi.org/10.1038/ncomms7924
7. Fu, Y., & Xue, F. (2022). MAD: Self-supervised masked anomaly detection task for multivariate time series. 2022 International Joint Conference on Neural Networks, 1-8. https://doi.org/10.1109/IJCNN55064.2022.9892218
8. Golubkov, A. W., Fuchs, D., Wagner, J., Wiltsche, H., Stangl, C., Fauler, G., Voitic, G., Thaler, A., & Hacker, V. (2014). Thermal-runaway experiments on consumer Li-ion batteries with metal-oxide and olivin-type cathodes. RSC Advances, 4, 3633-3642. https://doi.org/10.1039/C3RA45748F
9. Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic curve. Radiology, 143(1), 29-36. https://doi.org/10.1148/radiology.143.1.7063747
10. Klink, J. (2024). Measurement dataset of thermal fault emulation of a 46 Ah high-power Kokam nano pouch cell via uniform and local heating [Data set]. Zenodo. https://doi.org/10.5281/zenodo.13903526
11. Klink, J., Grabow, J., Orazov, N., Benger, R., Börger, A., Ahlberg Tidblad, A., Wenzl, H., & Beck, H.-P. (2021). Thermal fault detection by changes in electrical behaviour in lithium-ion cells. Journal of Power Sources, 490, 229572. https://doi.org/10.1016/j.jpowsour.2021.229572
12. Ahmed, A. A., & ASubie, H. Y. (2024). *Electric vehicle technology and vehicle-to-grid (V2G) integration: Opportunities and challenges*. *Bani Waleed University Journal of Humanities and Applied Sciences, 9*(3), 574–590. https://doi.org/10.58916/jhas.vi.420
13. Ahmed, A. A., Abdulla, N. M., & Abuali, T. M. (2025). *Performance optimization and battery health analysis of electric vehicles under real-world driving conditions: A data-driven experimental approach*. *Journal of Libyan Academy Bani Walid, 1*(2), 1–21.
14. Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation Forest. 2008 Eighth IEEE International Conference on Data Mining, 413-422. https://doi.org/10.1109/ICDM.2008.17
15. Ahmed, A. A. (2025). *Data-driven simulation framework for electric vehicle charging and energy management using real driving and grid data*. *Al-Mutawassit Journal for Basic and Applied Sciences, 1*(1), 13–19.
16. Oubelaid, A., Belkhier, Y., Ahmed, A. A., et al. (2024). *Fuzzy power management for low carbon footprint vehicles*.
17. Ahmed, A. A., Alsharif, A., & Asubie, H. Y. (2024). *Lifecycle analysis of electric vehicles: Assessing environmental and economic implications*.
18. Khaleel, M., Yusupov, Z., Kilic, H., Moftah, A., Hesri, A., Ahmed, A. A., Nassar, Y., Yilmaz, M., & Martinez-Morales, A. A. (2025). *Battery technologies in electrical power systems: Pioneering secure energy transitions*. *Journal of Power Sources*.
19. Politis, D. N., & Romano, J. P. (1994). The stationary bootstrap. Journal of the American Statistical Association, 89(428), 1303-1313. https://doi.org/10.1080/01621459.1994.10476870
20. Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. Neural Computation, 13(7), 1443-1471. https://doi.org/10.1162/089976601750264965
21. Severson, K. A., Attia, P. M., Jin, N., Perkins, N., Jiang, B., Yang, Z., Chen, M. H., Aykol, M., Herring, P. K., Fraggedakis, D., Bazant, M. Z., Harris, S. J., Chueh, W. C., & Braatz, R. D. (2019). Data-driven prediction of battery cycle life before capacity degradation. Nature Energy, 4, 383-391. https://doi.org/10.1038/s41560-019-0356-8

22. Tuli, S., Casale, G., & Jennings, N. R. (2022). TranAD: Deep transformer networks for anomaly detection in multivariate time series data. Proceedings of the VLDB Endowment, 15(6), 1201-1214. https://doi.org/10.14778/3514061.3514067
23. Xu, J., Wu, H., Wang, J., & Long, M. (2022). Anomaly Transformer: Time series anomaly detection with association discrepancy. International Conference on Learning Representations. https://openreview.net/forum?id=LzQQ89U1qm_
24. Zhang, Y., Li, S., Mao, B., Shi, J., Zhang, X., & Zhou, L. (2023). A multi-level early warning strategy for the LiFePO4 battery thermal runaway induced by overcharge. Applied Energy, 347, 121375. https://doi.org/10.1016/j.apenergy.2023.121375
25. Ahmed, A. A., & Gomah, M. (2023). *Electric vehicle impact on the sustainable development goals considering renewable energy sources integration*. *African Journal of Advanced Pure and Applied Sciences*, *2*(2), 227–234.
26. Alzayani, I. I., Nassar, Y. F., Adraider, Y. A., Alsharif, A., El-Khozondar, H., & Ahmed, A. A. (2024). *Challenges and novel strategy for electric vehicle integration in power system: Case study Libya*. *African Journal of Advanced Pure and Applied Sciences*. https://doi.org/10.65418/ajapas.vi.1024
27. Belrzaeg, M., Kayfeci, M., & Ahmed, A. A. (2026). *An artificial intelligence-optimized energy management strategy for smart grids with high penetration of electric vehicle charging*. *Lecture Notes on Data Engineering and Communications Technologies*.