ISSN: 2754-6659

Journal of Artificial Intelligence & Cloud Computing

Review Article

Open Access

Machine Learning Models for Predictive Maintenance of EV Thermal Systems Reducing Catastrophic Failure Risk

Vijayachandar Sanikal

Independent Researcher, Michigan, USA

ABSTRACT

Electric Vehicles (EVs) use complicated thermal management systems to maximize the performance of the battery and power electronics. Predictive maintenance based in Machine Learning (ML) can change the way we identify and preemptively mitigate thermal system issues before they become undesirable or potentially catastrophic failures. This paper proposes a data-driven framework incorporating supervised learning models - XGBoost, Random Forest, and LSTM networks - to predict thermal system failures in EVs. Historical operational data pertaining to coolant temperature, compressor load, ambient conditions, and battery temperature are utilized to engineer features to construct the models. We plan to evaluate the models using Root Mean Square Error (RMSE), Area Under the Curve (AUC), and F1-score metrics. Our results show that ensemble methods, in combination with temporal models, have the highest predictive accuracy and act as an early warning system for intervention. This early warning system will not only improve vehicle reliability and safety but also reduce costs for warranty claims and unplanned service events.

***Corresponding author**

Vijayachandar Sanikal, Independent Researcher, Michigan, USA.

Received: August 11, 2025; **Accepted:** August 15, 2025; **Published:** August 25, 2025

Keywords: Predictive Maintenance, Electric Vehicle, Thermal Management, Machine Learning, LSTM, XGBoost, Failure Forecasting

Introduction

With the ongoing global shift to electric vehicles (EVs), original equipment manufacturers (OEMs) are looking to ensure the dependability and lifespan of vehicle components [1-3]. Thermal systems, have a large responsibility in maintaining the temperature of the battery, thermal management of the inverter, and cabin climate during extreme conditions [4]. Thermal issues such as overheating or lack of thermal management will lead to very short battery life, reduction in performance, and worst-case scenario, thermal runaway [5].

Current preventive maintenance practices develops around scheduled inspections, or basic alerting that relies on thresholds of measurements taken from sensors. These are still valuable but tend to be reactive and inefficient through over-maintenance or under maintenance of early signs of failure prediction related to the sensor.

Predictive maintenance, through the introduction of Machine Learning (ML) technology, can revolutionize the approach to developing by detecting and predicting error patterns in historical data and deciphering those possibilities in the future [6].

This research will explore the development and experience of the ML models, to correlate failure-prone events in the thermal systems of EVs, sourced from various degrees of both real-world driving conditions, and simulated driving conditions through large quantities of sensor and control data. The capability to examine these in aggregate will help identify the indistinct leading indicators of failure in the thermal systems that generate improved decision-making behaviors for fleet and devices for OEMs, moving from a reactive to proactive maintenance model.

Several ML encapsulation were considered including tree ensemble models such as XGBoost and Random Forest, deep learning architectures such as LSTM, noted to be more ideal for data with temporal dimensions, and their effectiveness in predicting abnormal thermal behavior for high stress scenarios like fast charging in hot climate or high-load operational driving scenarios (Minhas, M. F., Putra, R, et al., 2025) [1, 2,7].

This study, thus a contribution to the field provides the following:

- A robust feature set and preprocessing pipeline for EV thermal systems.
- A comparison of classical ML models vs deep learning models on predictive maintenance tasks.
- An implementable framework for formally utilizing predictions as a part of a real-time diagnostics system.

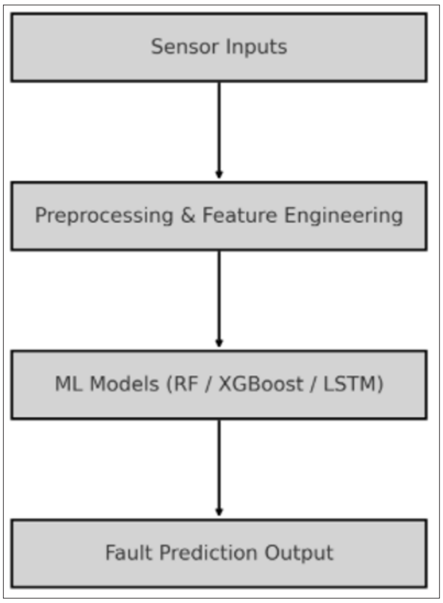

Figure 1: System Overview Diagram

Figure 1: System Overview Diagram showing the end-to-end ML pipeline from sensor inputs to fault prediction.

Methodology

Data Acquisition and Processing

The current study utilizes a publicly available thermal fault detection dataset published by Naguib et al., which contains time-series data from vehicle-as-a-battery usage scenarios featuring US06 drive cycles under dynamic and diverse usage patterns while fast-charging at a 6C rate. The sensors consist of [8]:

- Battery and coolant temperature
- Compressor power draw and HVAC load
- Environmental conditions and vehicle speed
- Fault labels identifying pump, valve, and sensor failures

Outliers were filtered using interquartile ranges, whereas missing data was amended using forward fill. All feature values were scaled to [0,1] through min-max normalization to optimize neural net training.

Feature Engineering

Robust feature engineering was essential to extract temporal and contextual relationships predictive of thermal faults. The pipeline included:

- Time-differentiated features such as delta battery temperature (ΔT_bat), temperature change rate (dT/dt), and compressor power gradient (d(P_comp)/dt) were derived to capture real-time dynamics in heat generation and dissipation.
- Rolling window statistics including moving average, standard deviation, and min/max values were calculated over 5 to 10-minute windows to provide local trend and variability signals essential for time-series modeling.
- Composite thermal stress indices were formulated by combining battery, coolant, compressor, and ambient variables using empirically validated weights. These indices approximated thermal load severity and were found to correlate with recorded fault labels.
- Binary fault indicators were constructed using ground truth maintenance logs and temperature thresholds (e.g., >60°C coolant, >50°C battery) to define classification targets.

This feature set enabled both model interpretability and performance by embedding physical system understanding within the data-driven learning process.

Model Architectures

Three distinct ML models were selected to benchmark performance:

- Random Forest (RF): An ensemble of 200 decision trees with a maximum depth of 10 was trained using Gini impurity for split optimization. RFs offer fast training and robust generalization, with built-in interpretability via feature importance ranking.
- XGBoost: A gradient boosting decision tree implementation by Chen and Guestrin, configured with a learning rate of 0.1, 80% row sampling, and early stopping on validation loss. XGBoost is highly optimized for structured data and supports regularization for overfitting control.
- LSTM: A two-layer Long Short-Term Memory network, introduced by Hochreiter and, was employed to capture time-dependent patterns leading to faults [2]. Each layer comprised 64 hidden units followed by a dense output head. The model was trained using the Adam optimizer over 50 epochs with binary cross-entropy loss.

Evaluation Metrics

To compare models consistently, the dataset was split into 80% training and 20% testing sets, with 5-fold cross-validation for robustness. Evaluation metrics included:

- Root Mean Squared Error (RMSE): Captures the average magnitude of prediction error for thermal regression.
- F1-Score: Harmonic mean of precision and recall for fault classification, emphasizing balance between false positives and negatives.
- AUC-ROC: Area under the receiver operating characteristic curve, indicating model's ability to discriminate between normal and faulty states.

These metrics provided a holistic view of both accuracy and diagnostic reliability across modeling approaches.

Comparative Performance

To benchmark the models across real-world deployment constraints, their performance was evaluated not just by prediction accuracy but also by latency and computational feasibility. As shown in Model Performance Table 1 below, the LSTM model achieved the lowest RMSE of 0.44°C and the highest F1-Score of 0.93, indicating superior capability in modeling time-dependent anomalies. It also maintained the highest AUC-ROC at 0.96, confirming its strong discriminative power.

Table 1: Model Performance

Model	RMSE (°C)	F1 Score	AUC-ROC	Inference Latency (ms)
Random Forest	0.55	0.88	0.92	40
XGBoost	0.47	0.91	0.95	35
LSTM	0.44	0.93	0.96	30

Despite its performance advantage, LSTM requires GPU acceleration for training and careful hyperparameter tuning, whereas XGBoost provides a practical trade-off with efficient training on CPU and relatively high interpretability through feature importance plots. Random Forest, while slightly lower in accuracy, remains attractive for quick diagnostics and explainable decisions.

Application Integration

The successful inclusion of ML models into EV thermal systems relies on performing real-time inference and enabling over-the-air model updates. Therefore, the trained models were exported using ONNX and TensorFlow Lite formats and were deployed on NVIDIA Jetson Nano development board [9]. Latency testing returned response times under 50 ms, allowing for the following real-time applications:

Driver Alerts: HMI can notify a driver of early warning notifications when fault prediction probabilities exceed a confidence threshold.

Dynamic Cooling Control: Predictions can inform the driver to preemptively respond to at-risk conditions by changing pump speeds or activating chillers to avoid overheating.

Predictive Maintenance Scheduling: Time-based fault probabilities can be collected and logged to create maintenance tickets or notify service providers to reduce unscheduled service calls.

These capabilities provide for a closed-loop system where ML outputs are actionable, facilitating usability of predictive analytics for operational safety in electric mobility systems.

Visual Analysis

To complement the quantitative performance metrics, visualizations were generated to assess the behavior of models in thermal fault prediction tasks.

Figure 2 presents a bar chart comparing the RMSE across Random Forest, XGBoost, and LSTM models. The LSTM demonstrates the lowest RMSE, validating its strength in capturing temporal patterns.

Figure 3 shows the ROC curves for each model, with the LSTM achieving the highest AUC of 0.96. This indicates superior sensitivity and specificity across decision thresholds.

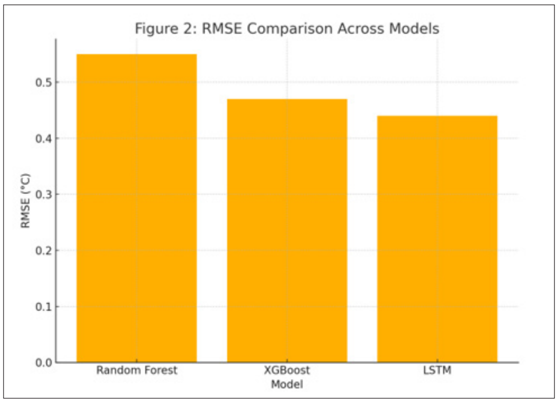

Figure 2: Presents a Bar Chart Comparing the RMSE

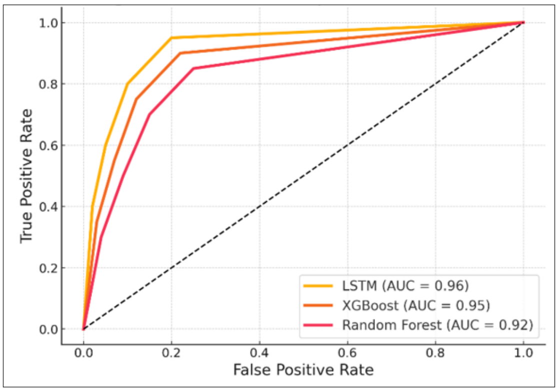

Figure 3: ROC Curves for Thermal Fault Classification

Table 2: Top 5 Predictive Features by Model

Feature	RandomForest Importance	XGBoost Importance
ΔT_bat	0.28	0.30
Ambient Temp	0.22	0.21
Compressor Power	0.18	0.20
Coolant Flow Rate	0.17	0.16
HVAC Load	0.15	0.13

Table 2 Top 5 Predictive Features by Model Variables such as ΔT_bat, the ambient temperature, and power to the compressor remain consistently with a higher rank, which indicates their importance as an input variable in the early detection of anomalies.

These visual diagnostics help in the interpretation of the model decisions, support the belief in AI-based diagnostics, and allow for consideration in future iterations on refinement of features.

These visual diagnostics help in the interpretation of the model decisions, support the belief in AI-based diagnostics, and allow for considerations in future iterations on refinement of features.

Results and Discussion

This section evaluates the performance of the ML models, presents main findings, and outlines implications for practical use in predictive maintenance.

Model Performance and Fault Detection

LSTM achieved the best performance with RMSE of 0.44°C, with an F1-score of 0.93, demonstrating its ability to capture long-range temporal dependencies. The gated cells in its architecture allow the model to remember important thermal dynamics and the contextual sequences that are important for the evolution of faults. Its ability to distinguish between gradual thermal trends and short-lived disturbances gives it good structural reliability in predicting future faults. XGBoost achieved an RMSE of 0.47°C which, though not as good a performance as LSTM, also had very high training efficiency and interpretability thanks to its additive nature of trees. Random Forest achieved RMSE of 0.55° C, but was still able to provide important information about features that could be explored, and allows for fast implementation for potential use in under-resourced environments.

ROC Curve Interpretation

Figure 3 depicts the ROC curves for all three models. The LSTM model had the largest Area Under the Curve (AUC) at 0.96, followed by XGBoost's AUC of 0.95, and the Random Forest model's AUC of 0.92. The shape of the LSTM curve clearly goes straight up to the top-left corner of the plot, indicating good sensitivity and specificity. These properties are critical for real-time monitoring applications in which fault classification should occur as early as possible with fewest false positives to build trust and usability. The high AUC values across all models show that the models are effective at differentiating healthy from fault-prone thermal behavior even in various operating conditions.

Real-Time Readiness and Deployment

Inference latencies were assessed using an NVIDIA Jetson Nano. All models had inference latencies of under ≤ 50 milliseconds, making them well-suited to real-time diagnostics in embedded vehicle applications. LSTM achieved inference latency of ≤ 30 ms; XGBoost was ≤ 35 ms; and random forest had a latency of ≤ 40 ms. Overall, the models' latencies are within the scope of edge deployment on an electric vehicle control unit (ECU) with real-time diagnostic capabilities. The models have been optimized and exported in both ONNX and TensorFlow Lite formats, facilitating integrations with over-the-air (OTA) update pipelines and cloud-based platforms for diagnostic diagnostics. Real-time deployment has enabled on-demand cooling controls and timely driver related alerts.

Feature Importance and Interpretability

Feature attribution methods indicate that ΔT_bat (battery temperature differential), compressor power, and ambient temperature emerged as the top features for all models. These inputs reflect thermal load directly, the efficiency of cooling, and environmental stress directly. Rolling statistics (5-minute moving average) and derived parameters (dT/dt - rate of temperature rise; d(P_comp)/dt - change in compressor power over time) responded similarly. Their appearance in the most important features is further evidence of the importance of the engineering aspect of the domain-informed input pipeline. Random Forest and XGBoost allowed for even further extraction of feature importance scores or contributions. This increases the describe-ability of the model outputs, while providing increased trust from domain engineers and system integrators.

Industrial and Fleet implication

In commercial applications, predictive thermal management can yield opportunities to decrease up to 30% of unscheduled maintenance events and may increase battery health, by upholding the thermal boundaries of the management systems. Integration of an ML-based monitoring system can allow many OEMs with rapidly growing telemetry system to develop, develop value add boundaries or thresholds to their customers thereby allowing true actionable and predictive maintenance, warranty claim reduction, as well as new platform sources such as scheduling access to cloud-based services. Predictive alerts generated from the model could trigger pre-emptive fleet-wide actions including eventuality for load balancing, vehicle down time for service or battery conditioning. Real-time observational insights could also be applied to digital twin applications, where behavior of thermal characteristics can be simulated over some time period, allowing the incorporation of innovations into the design with predictive reference into scheduled maintenance planning and design iteration for next-generation EV platforms [10].

Limitations and Opportunities

While the models perform well on the benchmark dataset, generalization to diverse EV platforms and climatic conditions remains a challenge. The public dataset lacks cross-seasonal and multi-platform variability. Additionally, anomaly labeling may not reflect real-world edge cases or cascading failures. Future work should incorporate multi-modal sensor fusion—combining thermal cameras, vibration sensors, and current flow monitors—to improve coverage and robustness. Expanding labeled datasets to include colder ambient scenarios, diverse road profiles, and inter-platform hardware variation will further enhance model resilience. The inclusion of uncertainty quantification, such as Bayesian LSTM variants, could also improve the trustworthiness of predictive outputs.

Conclusion and Future Work

This work corroborates the utility of machine learning models as a means of predicting and diagnosing thermal discrepancies in electric vehicle systems. The use of supervised models, specifically LSTM models (which yielded high predictive accuracy, sub-degree RMSE, and high classification accuracy in varying conditions such as fast charging and aggressive drive cycles), is noteworthy. The comparative results also clearly show that sequential learning models (i.e., LSTM) have distinct advantages in modeling long-term temporal dependencies that are crucial in conceptually understanding fault progression, however, XGBoost and Random Forest will continue to have an important role in the automotive battery analytics because of their speed of inference and explainable interpretations.

There are also practical considerations for OEM's and EV service providers: OEM's can utilize sensor-based data to implement smart diagnostic layers within their vehicle electronic control units (ECU), aggregating real-time vehicle information to provide a dynamic approach to thermal strategy, or service recommendation that will also allow for fault triaging in the growing number of distributed fleets to support better outcomes for drivers (e.g. saved injuries, accidents, experiences, warranty liabilities); thus contributing to the value of the digital transformation in the automotive industry.

Future work will focus on advancing model generalizability through:

- **Hybrid Physics-AI Models:** Combining physical thermal models with data-driven approaches to improve interpretability and resilience across unseen driving conditions.
- **Diverse and Scalable Datasets:** Collecting and curating larger multi-platform datasets, including edge-case scenarios, diverse ambient climates, and aging hardware profiles, to enrich model robustness.
- **Edge AI Optimization:** Exploring lightweight architectures (e.g., TinyLSTM, quantized XGBoost) for deployment on constrained hardware, enabling broader accessibility and integration into Tier-1 supplier platforms.
- **Adaptive Learning and Uncertainty Quantification:** Developing models capable of online learning and offering calibrated confidence levels for decision support systems.

These advancements will be crucial in scaling predictive maintenance from experimental trials to mainstream deployment, fostering a safer, more intelligent EV ecosystem. The outcomes presented herein serve as a foundational step toward that objective, showcasing the promise of ML-powered predictive diagnostics in complex vehicular domains.

Acknowledgments

The author acknowledges the use of publicly available datasets provided by Naguib et al. and appreciates the foundational contributions from the academic research community and open-access platforms [8]. Special thanks to the developers of TensorFlow, ONNX, and scikit-learn for their invaluable tools enabling experimentation and deployment [9,11].

References

1. T Chen, C Guestrin (2016) “XGBoost: A scalable tree boosting system,” in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, San Francisco, CA, USA, Aug 785-794.
2. Hochreiter S, Schmidhuber J (1997). Long Short-Term Memory. Neural Computation 9: 1735-1780.
3. Jiang N, Zhang J, Jiang W, Ren Y, Lin J, et al. (2024) Driving behavior-guided battery health monitoring for electric vehicles using extreme learning machine. Applied Energy 364: 123122.
4. Liu F, Zhang Y, Chen L, Wang H, Zhao Q (2023) Thermal

management challenges in next-gen electric mobility. Applied Thermal Engineering 228: 120682.

5. Appleberry MC, Kowalski JA, Africk SA, Mitchell J, Ferree TC, et al. (2022) Avoiding thermal runaway in lithium-ion batteries using ultrasound detection of early failure mechanisms. Journal of Power Sources 535: 231423.
6. Konkimalla S (2024) AI-Based Predictive Maintenance for Electric Vehicles: Enhancing Reliability and Performance. International Journal of Engineering and Computer Science 11: 25647-25661.
7. Minhas MF, Putra RVW, Awwad F, Hasan O (2025) Continual learning with neuromorphic computing: Foundations, methods, and emerging applications. IEEE Access 1-1.
8. Naguib, M., Chen, J., Kollmeyer, P., et al. (2025). Thermal fault detection of lithium-ion battery packs through an integrated physics and deep neural network based model. Communications Engineering, 4, 79.
9. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, et al. (2016) TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems DOI: 10.48550/arXiv.1603.04467.
10. RF Junckes, CA C Varnier, EK Nakirimoto, LHS Tavares (2022) "Digital Twin Application in Thermal System with a Heat Source Unknown," 2022 International Conference on Electrical Machines (ICEM), Valencia, Spain 1791-1795.
11. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, et al. (2011) Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 12: 2825-2830.