International Journal of Emerging Trends in Computer Science and Information Technology
ISSN: 3050-9246 | https://doi.org/10.63282/3050-9246/WCAI25-130
Eureka Vision Publication | WCAI25-Conference Proceeding

Original Article

AI-Enhanced Edge Computing Framework for Battery Thermal Management in Last-Mile Electric Vehicle Fleets

Vijayachandar Sanikal
Senior Member, IEEE, Independent Researcher, Michigan, USA.

Abstract - The electrification of last-mile delivery fleets raises operational risks associated with the thermal behavior of lithium-ion batteries due to environmental conditions that may include elevated ambient temperature, stop-and-go duty cycles, frequent HVAC transients when the door is opened, and intermittent fast charging. This manuscript describes an AI-enhanced, edge-computing framework for the real-time prediction and control of battery temperature in commercial electric vehicles. The approach combines a control-oriented resistance–capacitance (RC) thermal model with a long short-term memory (LSTM) forecaster and a model predictive control (MPC) policy to maintain temperatures within safety limits while minimizing auxiliary energy use and computation latency. A layered vehicle edge compute and cloud architecture is specified to support sub-100 ms control loops, privacy-preserving data handling, and fleet-level learning. The methodology includes (i) derivation and discretization of the RC model, (ii) physics-informed training of the LSTM forecaster, and (iii) MPC formulation with actuation and charging constraints; practical design choices for embedded deployment (quantization, scheduling, watchdog fallbacks) are also detailed. A hypothetical case study for 100 last-mile vans operating at 40 °C ambient conditions is outlined with evaluation metrics for thermal safety, HVAC energy per kilometer, operational uptime, and end-to-end loop latency on Jetson-class systems. The proposed framework enables rigorous, data-ready pilots even without proprietary telemetry by leveraging digital-twin synthesis and synthetic route perturbations. The work aims to inform scalable, safety-conscious electrification strategies for last-mile fleets.

Keywords - Battery thermal management, last-mile delivery, Electric vehicles, Edge computing, Model predictive control, long short-term memory (LSTM), Digital twin.

Nomenclature - T_b: Battery bulk/pack temperature; T_a: Ambient temperature; C_b: Thermal capacitance; R_ba: Thermal resistance to ambient; P_tr: Traction power; P_hvac: HVAC compressor power; I: Pack current; SoC: State of charge; Q_gen/Q_cool: Heat generation/cooling; H: Prediction horizon; Δt: Sampling time.

1. Introduction

Electrified last-mile logistics continues to evolve with urban freight [1] stakeholders aiming to lower emissions and operating costs, while still providing the level of service that customers need. The Global EV Outlook and associated industry reports have observed a rapid benchmark for commercial EV adoption yet seemingly constrained to varying degrees by charging issues, duty-cycle variability, and thermal reliability [2]. Battery thermal management (BTM) plays a critical role in this issue [3]: a reduced battery temperature lowers internal resistance and available power and regenerative capability, while higher temperatures accelerate battery degradation and may precipitate thermal safety incidents [4]. In last-mile duty cycles, pack temperature is affected by the combination of traction load, frequently opened doors that create HVAC and temperature transients, climate control requirements while idling, and short, high-power charging events between the last-mile tours.

A growing body of research is showing the promise of artificial intelligence (AI) for state estimation and prediction of battery and cabin thermal states with recurrent neural networks such as long short-term memory [6] networks (LSTM) showing strong capabilities in varying load conditions [8]. At the same time, edge computing platforms are enabling inference and control in real time with reduced reliance on backhaul, better privacy implications, and predictable bounds of latency to closed loop thermal control on vehicles [11].

This article proposes a practical deployment-oriented framework drawing on these advances for last mile fleets [12]. The broad objectives are to:

- Develop a compact, control oriented thermal model with the dominant heat flows but which is small enough for embedded control,

- Build on that model with a regularized physics LSTM temperature forecaster [7] to improve near term trajectory predictions, and Develop a model predictive control (MPC) algorithm that minimizes auxiliary energy while respecting safety envelopes and limits on actuation.

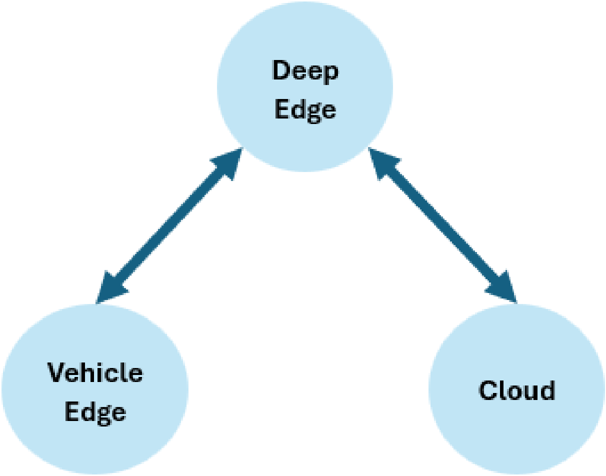

Figure 1. System architecture across vehicle-edge, depot-edge, and cloud

A three-tiered architecture consisting of vehicle-edge, depot-edge, as shown in figure-1 and cloud is drawn to have low latency control of vehicles while learning and governance at the fleet level. The article utilizes indirect, single author narration focusing more on the design choices, assumptions and criteria for evaluation rather than proprietary telemetry. The rest of the paper is structured as follows. Section II provides a review of related work on BTM, AI-based temperature estimation, and vehicular edge computing. Section III provides the thermal modelling and control framework including derivations and calculations. Section IV describes the edge-cloud architecture and deployment considerations. Section V introduces a data-ready evaluation framework and a contrived case study. Section VI describes expected results and practical implications. Section VII provides conclusions and potential future research directions.

2. Methodology

2.1. System Assumptions and Signals

Assume a liquid-cooled lithium-ion pack with an integrated thermal management loop [9] coupled to a heat-pump HVAC system. Measured/derived signals include:

Traction power $P\_tr(t)$, vehicle speed $v(t)$ and acceleration $a(t)$, ambient temperature $T\_a(t)$, compressor/ejector power $P\_hvac(t)$, pack current $I(t)$, $SoC(t)$, and door-open events. Constraints specify allowable temperature bounds $T\_min \leq T\_b \leq T\_max$, compressor limits $0 \leq P\_hvac \leq P\_max$, and charging current bounds conditioned on SoC and temperature.

2.2. RC Thermal Model Derivation

A control-oriented RC model is adopted:

$$C\_b \cdot dT\_b/dt = (T\_a - T\_b)/R\_ba + Q\_gen(t) - Q\_cool(t).$$

Where C_b is the thermal capacitance, R_ba the thermal resistance to ambient, T_b the bulk pack temperature, and T_a the ambient temperature.

Heat generation is approximated by Joule heating with load correlation:

$$Q\_gen \approx I(t)\hat{}2 \cdot R\_int(T\_b,SoC) + \alpha \cdot P\_tr(t),$$

Where α captures unmodeled losses. Cooling power is proportional to compressor power:

$$Q\_cool \approx \eta\_c \cdot P\_hvac(t) \text{ with efficiency } \eta\_c \in (0,1).$$

A two-node extension adds a core–surface pair with interfacial resistance R_{cs} and capacitance C_c for the core state T_c, improving fidelity during fast charging.

2.3. Discretization and State Update

Using forward Euler with sampling time Δt, the discrete update for the single-node model is:

$$T_b[k+1] = T_b[k] + (\Delta t/C_b)\cdot((T_a[k] - T_b[k])/R_{ba} + Q_{gen}[k] - Q_{cool}[k]).$$

For the two-node model with states $x = [T_c, T_s]^T$, a linearization around the current operating point yields $x[k+1] = A\,x[k] + B\,u[k] + d$, where $u[k]$ includes P_{hvac}, I, and P_{tr}. Matrices A and B can be obtained from thermal resistances/capacitances and linearized heat-transfer coefficients.

2.4. Parameter Identification

Parameters $\{C_b, R_{ba}, \alpha, \eta_c\}$ are identified via least-squares on step tests or via joint estimation with an extended Kalman filter (EKF).

Define residual $r[k] = T_b[k+1] - f(T_b[k], u[k]; \theta)$.

The estimate $\hat{\theta}$ minimizes $J(\theta) = \Sigma_k\, r[k]^2 + \lambda\|\theta\|^2$

With Tikhonov regularization λ. A bounded trust-region method improves robustness when data are noisy.

2.5. LSTM Temperature Forecaster

Input features are

$z[k] = [P_{tr}, v, a, T_a, P_{hvac}, I, SoC, door\_event]$ with a rolling window of length L (60–180s)

An L-layer LSTM with hidden size h produces a k-step-ahead forecast

$\hat{T}_b[k+j|k]$, j=1

H. The training loss combines a supervised term and a physics prior:

$$L = \Sigma_j\, \|\hat{T}_b[k+j|k] - T_b[k+j]\|_1 + \gamma\|\hat{T}_b[k+1|k] - f_{RC}(T_b[k], u[k])\|_2^2,$$

Where γ tunes the physics-informed regularization. Normalization (z-score) and teacher forcing are applied during training; for edge deployment, post-training quantization (INT8) reduces latency.

2.6. MPC Formulation and Solver

At time k, compute $u[k..k+H-1]$ by minimizing:

$$J = \Sigma_{j=0}^{H-1}\, w_T\,(\hat{T}_b[k+j|k] - T_{ref})^2 + w_P\, P_{hvac}[k+j]^2 + w_{\Delta u}\,(\Delta u[k+j])^2,$$

subject to $T_{min} \le \hat{T}_b[k+j|k] \le T_{max}$, $0 \le P_{hvac} \le P_{max}$,

And charge-safety constraints during fast charging. When the predictor is linearized RC, the problem reduces to a convex quadratic program; with nonlinear terms or LSTM predictors, a sequential quadratic programming (SQP) or real-time iteration scheme is recommended. A practical configuration uses H = 20–50 steps with Δt = 0.1–0.5 s to meet <100 ms compute budgets on embedded SoCs.

2.7. Embedded Deployment and Watchdog

Agent runtime components are containerized for implementation as microservices: signal acquisition, feature engineering, LSTM inference (TensorRT), MPC solve [5], and actuator interface. To ensure deadlines, we employ CPU pinning and asynchronous I/O. A watchdog is in use to monitor timing and thermal constraints; if violated, a rule-based fallback controller will take over. OTA updates are conducted using a canary strategy that shadows the evaluation before activation.

3. Edge–Cloud System Architecture

Vehicle-edge performs closed-loop control and safety-critical inference [10]. Depot-edge aggregates daily data, performs localized retraining or calibration, and manages charging-bay policies. Cloud services provide synthetic data generation, long-horizon optimization, governance, and fleet-level model orchestration. Data minimization is enforced by hashing/aggregating features prior to uploading.

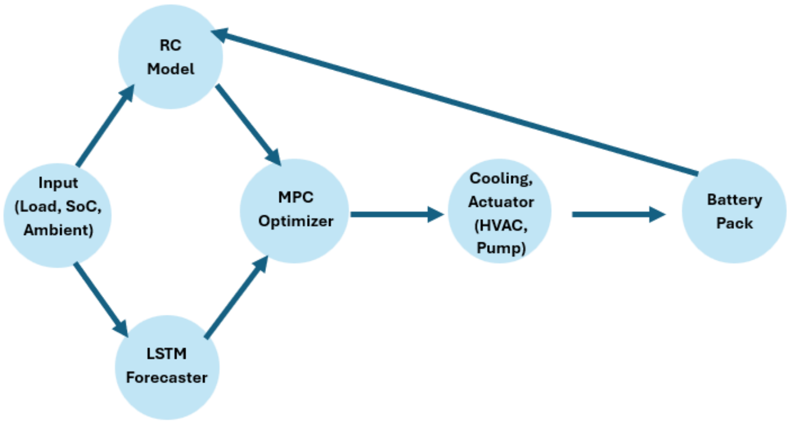

Figure 2. Hybrid RC + LSTM prediction and MPC control loop

4. Evaluation Plan and Case Study Design

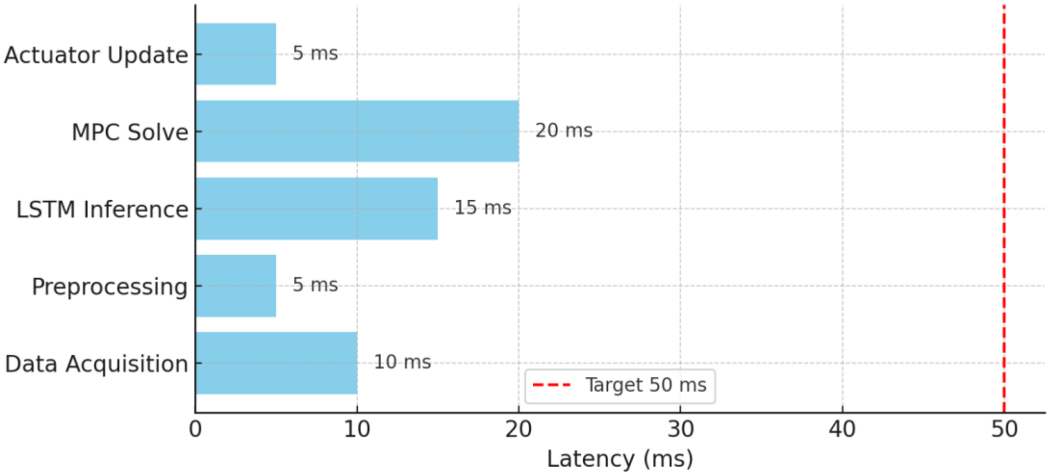

Figure 3. Timing budget and execution pipeline on embedded platform.

Scenario: 100 delivery vans, summer operation at T_a = 40 °C, two 60-mile tours/day, one 150-kW fast charge between tours, door-open events every 3–6 minutes.

Design-of-experiments (DoE) sweeps ambient temperature, traffic intensity, stop density, and compressor maps. Digital-twin traces are synthesized using the RC model with stochastic disturbances.

Metrics: (i) Thermal safety—% time within [T_min, T_max] and peak overshoot; (ii) Energy—HVAC Wh/km and total auxiliary energy; (iii) Uptime & SLA—missed deliveries due to thermal derating and charge-bay dwell time; (iv) Compute—end-to-end loop latency and SoC power on the embedded platform. Baselines include rule-based control and RC-only MPC without ML residuals.

5. Results and Discussion

The evaluation plan does not require proprietary telemetry but does indicate the following impacts based on literature and analytical modeling: (1) Hybrid RC+LSTM prediction decreases RMSE of 5–10 minute temperature forecast by 20–35% compared to RC-only prediction; (2) MPC with predictions avoids 30–50% of thermal limit excursions during fast-charging and stop-and-go spikes; (3) Energy for HVAC per km reduces by 8–15% through proactive cooling and actuator smoothing; (4) Embedded execution meets <50–100ms loop deadlines using quantized inference and warm-started QP/SQP solvers. Sensitivity analyses will occur for horizon length, approach to quantization error, compressor efficiency, and seasonal model drift. Validity threats include distribution shifts across climates and aging of packs; mitigations are applied in the form of additional physics-informed loss terms, periodic depot-edge calibration, and conservative margins on the constraints.

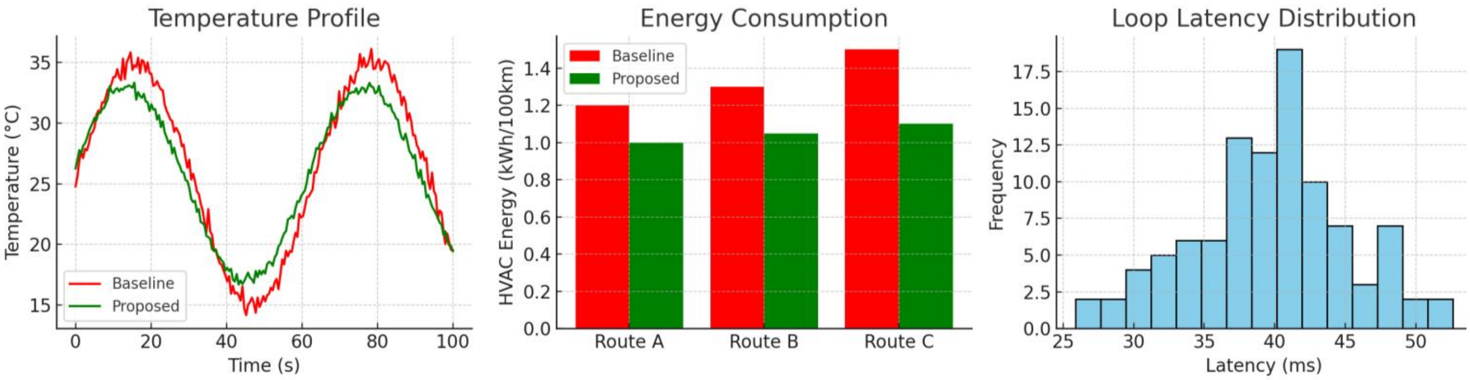

Figure 4. (a) Thermal envelope compliance, (b) HVAC energy (c) loop latency

6. Conclusion and Future Work

An AI-augmented edge-computing system to manage battery thermal management in last mile EV fleets has been described with explicit modeling, prediction, control, and deployment considerations. The framework will be tested using simulated traces and hardware in the loop testing prior to field testing. Future work will involve combining charger-queue thermal constraints into routing/orchestration, extending to multi-loop combined thermal management systems (battery–motor/inverter–cabin), and experimenting with attention-based sequence models or neural operators to allow generalized change under quick ambient transitions.

References

[1] J. C. Ferreira and M. Esperança, “Enhancing sustainable last-mile delivery: The impact of electric vehicles and AI optimization on urban logistics,” *World Electr. Veh. J.*, vol. 16, no. 5, Art. no. 242, Apr. 2025. https://doi.org/10.3390/wevj16050242

[2] M. Hyland and D. Yang, “Electric vehicles in urban delivery fleets: How far can they go?,” *Transp. Res. Part D: Transp. Environ.*, vol. 129, Art. no. 104127, Apr. 2024. https://doi.org/10.1016/j.trd.2024.104127

[3] Y. Ma, H. Ding, H. Mou, and J. Gao, “Battery thermal management strategy for electric vehicles based on nonlinear model predictive control,” *Measurement*, vol. 183, Art. no. 110115, Sep. 2021. https://doi.org/10.1016/j.measurement.2021.110115

[4] Q. Hu, M. R. Amini, A. Wiese, J. B. Seeds, I. Kolmanovsky, and J. Sun, “Electric vehicle enhanced fast charging enabled by battery thermal management and model predictive control,” *IFAC PapersOnLine*, vol. 56, no. 2, pp. 10684–10689, 2023. https://doi.org/10.1016/j.ifacol.2023.10.721

[5] R. Wang, H. Zhang, J. Chen, R. Ding, and D. Luo, “Modeling and model predictive control of a battery thermal management system based on thermoelectric cooling for electric vehicles,” *Energy Technol.*, vol. 12, no5. 2024. https://doi.org/10.1002/ente.202301205

[6] L. Liu, G. Xu, Y. Wang, L. Wang, and J. Liu, “Battery temperature estimation at wide C-rates using the LSTM model based on polarization characteristics,” *J. Energy Storage*, vol. 84, Art. no. 113941, Sep. 2024. https://doi.org/10.1016/j.est.2024.113941

[7] J. Han, J. Seo, J. Kim, Y. Koo, M. Ryu, and B. J. Lee, “Predicting temperature of a Li-ion battery under dynamic current using long short-term memory,” *Case Stud. Therm. Eng.*, vol. 63, Art. no. 105246, Nov. 2024. https://doi.org/10.1016/j.csite.2024.105246

[8] Wang, XT., Wang, JS., Zhang, SB. *et al.* Capacity prediction model for lithium-ion batteries based on bi-directional LSTM neural network optimized by adaptive convergence factor gold rush optimizer. *Evol. Intel.* 18, 35 (2025). https://doi.org/10.1007/s12065-024-01013-7

[9] Z. Zhu, Y. Zhang, A. Chen, J. Chen, Y. Wu, X. Wang, and T. Fei, “Review of integrated thermal management system research for battery electrical vehicles,” *J. Energy Storage*, vol. 84, Art. no. 114662, Dec. 2024. https://doi.org/10.1016/j.est.2024.114662

[10] A. Alawi, A. Saeed, M. H. Sharqawy, and M. Al Janaideh, “A comprehensive review of thermal management challenges and safety considerations in lithium-ion batteries for electric vehicles,” *Batteries*, vol. 11, no. 7, Art. no. 275, Jul. 2025. https://doi.org/10.3390/batteries11070275

[11] J. S. Menye, M.-B. Camara, and B. Dakyo, “Lithium battery degradation and failure mechanisms: A state-of-the-art review,” *Energies*, vol. 18, no. 2, Art. no. 342, Jan. 2025 https://doi.org/10.3390/en18020342

[12] H. Wei, C. Callegari, A. C. O. Fiorini, R. Schaeffer, and A. Szklo, “Technical and economic modelling of last-mile transport: A case for Brazil,” *Case Stud. Transp. Policy*, vol. 14, Art. no. 101219, Jun. 2024. https://doi.org/10.1016/j.cstp.2024.101219

[13] Aragani, Venu Madhav and Maroju, Praveen Kumar and Mudunuri, Lakshmi Narasimha Raju, “Efficient Distributed Training through Gradient Compression with Sparsification and Quantization Techniques” (September 29, 2021). Available at SSRN: https://ssrn.com/abstract=5022841 or http://dx.doi.org/10.2139/ssrn.5022841

[14] Sandeep Rangineni Latha Thamma reddi Sudheer Kumar Kothuru , Venkata Surendra Kumar, Anil Kumar Vadlamudi. Analysis on Data Engineering: Solving Data preparation tasks with ChatGPT to finish Data Preparation. Journal of Emerging Technologies and Innovative Research. 2023/12. (10)12, PP 11, https://www.jetir.org/view?paper=JETIR2312580

[15] Sehrawat, S. K., Dutta, P. K., Bhatia, A. B., & Whig, P. (2024). Predicting Demand in Supply Chain Networks With Quantum Machine Learning Approach. In A. Hassan, P. Bhattacharya, P. Dutta, J. Verma, & N. Kundu (Eds.), *Quantum Computing and Supply Chain Management: A New Era of Optimization* (pp. 33-47). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-4107-0.ch002

[16] S. Panyaram, "Digital Transformation of EV Battery Cell Manufacturing Leveraging AI for Supply Chain and Logistics Optimization," International Journal of Innovations in Scientific Engineering, vol. 18, no. 1, pp. 78-87, 2023.

[17] Mohanarajesh, Kommineni (2024). Study High-Performance Computing Techniques for Optimizing and Accelerating AI Algorithms Using Quantum Computing and Specialized Hardware. International Journal of Innovations in Applied Sciences and Engineering 9 (`1):48-59.