Toward Real-Time Junction Temperature Estimation with a Physics-Informed Attention LSTM

Alessandro Varaldi*, Maurizio Tranchero†, Lorenzo Giraudi†, Claudio Genta†, Marco Vacca*
*DET, Politecnico di Torino, Torino, Italy
†Ideas & Motion, Cherasco, Italy
{name.surname}@polito.it, {name.surname}@ideasandmotion.com

***Abstract*—Accurate real-time junction temperature estimation for automotive power electronics is essential to guarantee reliability under dynamic operating conditions. Purely data-driven recurrent networks, such as Long Short-Term Memory (LSTM), can capture complex transients but, without physical constraints, may yield biased or poorly calibrated predictions; conversely, purely physics-based lumped models can be robust yet miss nonlinearities and operating-context variability. We propose a physics-informed LSTM that enforces steady-state and transient energy-balance constraints derived from a lumped 1-RC thermal model. The network is optimized through Bayesian hyperparameter search. On proprietary automotive drive cycles, the physics-informed model delivers substantial gains: relative to a plain LSTM, the mean absolute error is reduced by 18.6 %; relative to a calibrated 1-RC ODE, the mean squared error is lowered by 56.6 %. Predictive intervals obtained via Monte Carlo dropout are well calibrated: for nominal 95 % intervals, empirical coverage is 98.25 % for the LSTM and 99.54 % for the physics-informed model, with comparable sharpness. These results show that embedding lightweight physics in sequence models improves prediction reliability without additional computational overhead, making the approach suitable for onboard deployment.**

***Index Terms*—Physics-informed neural network, Junction-temperature estimation, Automotive power electronics, Attention LSTM.**

I. Introduction

Accurate knowledge of the semiconductor *junction temperature* T_j is a cornerstone requirement for the safe, efficient and reliable operation of modern traction inverters used in battery-electric vehicles (BEVs) and industrial drives [1]. Exceeding the maximum allowed T_j accelerates wear-out mechanisms such as bond-wire lift-off and solder fatigue, ultimately shortening device lifetime and risking catastrophic failure [2]. Conversely, conservative guard bands reduce exploitable current capability, hindering power-density targets and escalating cooling-system cost.

Traditionally, T_j is inferred either from slow embedded sensors (thermocouples, on-chip diodes) or from real-time loss observers coupled with *a posteriori* thermal impedances. Such model-based approaches rely on fixed lumped-parameter networks whose parameters must be laboriously identified for each package variant and operating condition, and their linear structure can be challenged by highly dynamic load profiles that produce rapid temperature excursions [3].

Data-driven regression has recently emerged as a flexible alternative: recurrent networks and temporal convolutions demonstrated superior accuracy on power-electronic temperature benchmarks [4]. However, purely data-driven models often over-fit catalog-specific duty cycles and may generate thermodynamically inconsistent predictions when extrapolated. Embedding physics within the learning objective has therefore become a promising strategy to combine accuracy with plausibility [5].

Building on these premises, the present study proposes a physically consistent, resource-efficient observer for T_j reconstruction. Our main contributions are:

1) **Lightweight attention-augmented backbone**: a single-layer LSTM with dot-product attention processes 128-sample windows of power-loss and base-plate temperature, using only ≈ 1.3k parameters and fitting typical automotive MCU budgets.
2) **Physics-informed dual loss**: steady-state and transient energy-balance residuals derived from a lumped 1-RC thermal model enforce physical consistency under both static and fast-varying loads.
3) **Training protocol for calibrated uncertainty**: a three-stage Bayesian hyperparameter search and physically consistent duty-cycle augmentation improve data efficiency, while Monte Carlo dropout provides calibrated predictive intervals for risk-aware control.
4) **Experimental validation on automotive cycles**: on proprietary drive cycles, the model reduces mean absolute error by 18.6 % versus a plain LSTM and mean squared error by 56.6 % versus an optimally calibrated single-RC ODE, with comparable interval sharpness.

The remainder of the paper is organized as follows: Section II reviews the background and technical context; Section III surveys related work; Section IV presents the methodology, including the attention-augmented LSTM architecture, the physics-informed loss, uncertainty evaluation, and the training protocol; Section V reports results and discussion; and Section VI concludes with an outlook toward deployment and future research.

This work was funded by the PNRR-NGEU programme under MUR – DM 117/2023.

II. Background

A. Why Reconstruct the Junction Temperature?

The silicon junction temperature T_j is the highest temperature inside a power device and therefore the dominant reliability driver. If T_j exceeds its limit even briefly, bond wires may lift off, solder layers crack and thermo-mechanical fatigue accelerates. When operation is too conservative, cooling resources are wasted and power density suffers. Accurate real-time knowledge of T_j therefore enables:

- *Dynamic thermal derating* – the inverter can deliver more current when the chip is cool and gently reduce output only when headroom shrinks;
- *Remaining-life estimation* – lifetime-consumption models integrate the actual junction-temperature swing, not the case or ambient temperature;
- *Design optimization* – precise temperature feedback permits smaller heatsinks, pumps and coolant flow.

B. Conventional Estimation Techniques

- *Direct sensing.* Thermocouples or on-chip diodes measure temperature but are slow ($\gtrsim$1 ms) and often placed away from the hotspot.
- *Observer models.* Industrial drives typically embed a lumped RC network driven by online loss estimates; parameters are extracted from step-response experiments and cannot adapt to ageing or mounting tolerances [6].
- *Kalman and sliding-mode observers.* Stochastic and sliding-mode filters reduce noise but remain limited by the fidelity of the gray-box thermal model they build upon [7] [8].

C. Neural Networks, LSTM and Attention

An *artificial neuron* applies a linear transform followed by a non-linear activation. Stacking many neurons yields a feed-forward neural network (FNN) that approximates complex, high-dimensional mappings. Training minimizes a loss between network output and target via gradient descent and back-propagation of derivatives.

FNNs, however, ignore temporal context because they treat each input independently. *Recurrent neural networks* (RNNs) introduce a hidden state that evolves with time, endowing the model with memory. Vanilla RNNs suffer from vanishing or exploding gradients on long sequences; the *Long Short-Term Memory* (LSTM) cell mitigates this problem by gating information flow through *input*, *forget* and *output* gates, and by maintaining a separate cell state that can carry information almost unchanged across many time steps [9] (see Fig. 1). LSTMs therefore underpin most sequence-modeling applications in speech, text and sensor time-series data.

Although LSTMs capture long-range dependencies, they compress the entire history into a single hidden vector. *Attention mechanisms* alleviate this bottleneck by allowing the network to compute a context-dependent weighted sum over all hidden states, effectively letting the model decide *which* past instants are most relevant to the current prediction. Originally introduced in neural machine translation [10], attention layers have since been combined with LSTMs in numerous time-series tasks, yielding improved performance and interpretability.

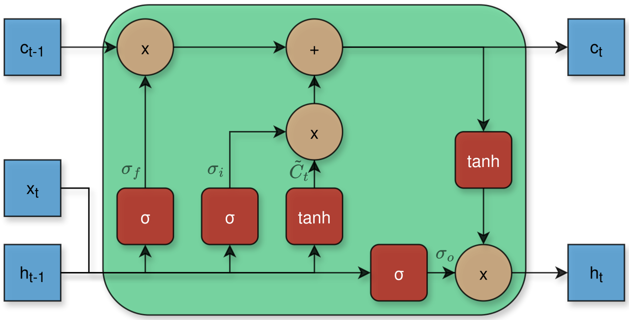

Fig. 1. **Canonical LSTM cell.** At every time-step t the cell receives the input vector $\mathbf{x}_t$ and the previous hidden and cell states $(\mathbf{h}_{t-1}, \mathbf{c}_{t-1})$. The three gate activations (input σ_i, forget σ_f, output σ_o) and the candidate update $\tilde{\mathbf{c}}_t = \tanh(\cdot)$ are computed through affine transformations followed by non-linearities. The updated cell state $\mathbf{c}_t$ is obtained by combining the previous memory (forget gate) with the candidate content (input gate); the hidden state $\mathbf{h}_t$ is produced by modulating the $\tanh$-squashed memory with the output gate. This mechanism enables the network to selectively retain, overwrite, or expose information over long temporal horizons.

D. Physics-informed Neural Networks

Purely data-driven models can fit nonphysical functions if over-parameterized. Physics-informed neural networks (PINNs) embed the governing equations in the loss:

$$\mathcal{L} = \underbrace{\mathrm{MSE}\big(T_\mathrm{pred}, T_\mathrm{meas}\big)}_{\text{data term}} + \lambda \underbrace{\mathrm{MSE}\Big(\mathcal{F}(T_\mathrm{pred}, \dot{T}_\mathrm{pred}, \mathbf{u}), 0\Big)}_{\text{physics term}},$$

where $\mathcal{F}(\cdot) = 0$ represents the physical law, and λ balances data versus physics [11]. Benefits include better extrapolation, data efficiency and interpretability, but determining the right scaling between loss components remains challenging.

III. Related Work

Deep learning has been widely explored for thermal estimation in electrical machines. Kirchgässner *et al.* [4] showed that residual CNN/RNN predictors deliver high-accuracy temperature estimates in PMSMs (errors of a few degrees). Variants include feed-forward models that learn temperature *differences* to reduce bias (e.g., Lee and Ha [12] report average errors around 1 °C and worst-case within 4.5 °C) and sensor-aided estimators for BLDC windings, where Czerwinski *et al.* [13] achieve below 4.5 % MAPE (down to ∼1 % with a casing sensor). To inject domain physics, Rönnberg *et al.* [14] combine machine learning with an LPTN, tuning its parameters to improve robustness across operating conditions and aging. More recently, physics-informed deep models have gained traction: Hughes *et al.* [15] leverage physically motivated features (MSE $\approx$ 2.40 °C^2 on a standard benchmark), while Sheng *et al.* [16] embed an LPTN into an attention-based network to better track fast transients (MSE $<$ 4 °C^2 and peak errors 5 °C to 8 °C). Our work follows this line but targets *junction-temperature* observers for power devices, coupling a lightweight attention-augmented LSTM with explicit energy-balance residuals.

IV. Methodology

A. Problem Formulation

Let P_t (W) denote instantaneous power loss and $T_{\mathrm{bp},t}$ (°C) the base-plate temperature. Given the most recent $L = 128$ samples,

$$\mathbf{x}_t = \left[P_{t-127}, \dots, P_t, T_{\mathrm{bp},t-127}, \dots, T_{\mathrm{bp},t}\right]^\top \in \mathbb{R}^{256},$$

the task is to predict the current junction temperature

$$\hat{T}_{\mathrm{j},t} = f_{\boldsymbol{\theta}}(\mathbf{x}_t), \tag{1}$$

where $f_{\boldsymbol{\theta}}$ denotes a neural mapping with parameters $\boldsymbol{\theta}$. The window slides forward by one step at each instant, so (1) is evaluated without look-ahead.

B. Dataset and Measurement Setup

Data were collected on an ECU featuring top-side–cooled power transistors. The device thermal pad is coupled to an air heatsink through a thermal interface material (TIM) with conductivity $\lambda = 2.5\,\mathrm{W\,m^{-1}\,K^{-1}}$. To foster repeatability, tests were conducted inside a small enclosure hosting only the ECU and the IR camera; airflow was intentionally limited to the heatsink region to emulate the final application. The surface area of interest was coated matte black to homogenize emissivity and mitigate reflections. Transistor activations and camera acquisition were synchronized to minimize timing skew between thermal excitation and measurement. Due to the available instrumentation, only case/surface temperature was acquired by the IR camera. Throughout this study we therefore use case temperature as a proxy for junction temperature and write $T_\mathrm{j} \approx T_\mathrm{c}$; all reported "junction" temperatures should be interpreted accordingly.

The dataset comprises three runs sampled at 0.1 s ($\sim 1 \times 10^5$ samples). We extracted overlapping windows of length $L = 128$ (12.8 s) with stride 1 ($\sim 1 \times 10^5$ sequences). To avoid leakage, each run is assigned to a split (train/val/test) *before* windowing. We apply data augmentation (20× per window): (i) small temporal jitter, (ii) quasi-static temperature offsets to T_bp and T_j, (iii) global power scaling, and (iv) zero-mean power noise. The final split is 70 %/30 % for training/test, with 15 % of training held out for validation.

C. Attention-augmented LSTM Architecture

Figure 2 sketches the complete pipeline. A single-layer LSTM processes each $(L, F{=}2)$ window, producing hidden states $\mathbf{h}_{1:L}$. A *dot-product attention* head computes

$$\alpha_\tau = \frac{\exp(\mathbf{h}_\tau^\top \mathbf{h}_L)}{\sum_{\sigma=1}^{L} \exp(\mathbf{h}_\sigma^\top \mathbf{h}_L)}, \qquad \mathbf{z}_L = \sum_{\tau=1}^{L} \alpha_\tau \mathbf{h}_\tau,$$

and predicts

$$\hat{T}_{\mathrm{j},t} = \mathbf{w}^\top \left[\mathbf{h}_L ; \mathbf{z}_L\right] + b,$$

To tune this architecture we performed three successive Optuna [17] studies (20 trials each, median pruning):

1) *Architecture & learning-rate search.* Physics penalties disabled ($\lambda_\mathrm{ss} = \lambda_\mathrm{tr} = 0$), exploring *hidden_size* $\in \{8, 16, 24, 32\}$, *num_layers* $\in \{1, 2, 3\}$ and lr $\in [10^{-4}, 10^{-2}]$.
2) *Steady-state penalty search.* Best backbone frozen; $\lambda_\mathrm{ss} \in [10^{-7}, 10^{-5}]$, $\lambda_\mathrm{tr} = 0$.
3) *Transient penalty search.* λ_ss fixed; $\lambda_\mathrm{tr} \in [10^{-7}, 10^{-5}]$.

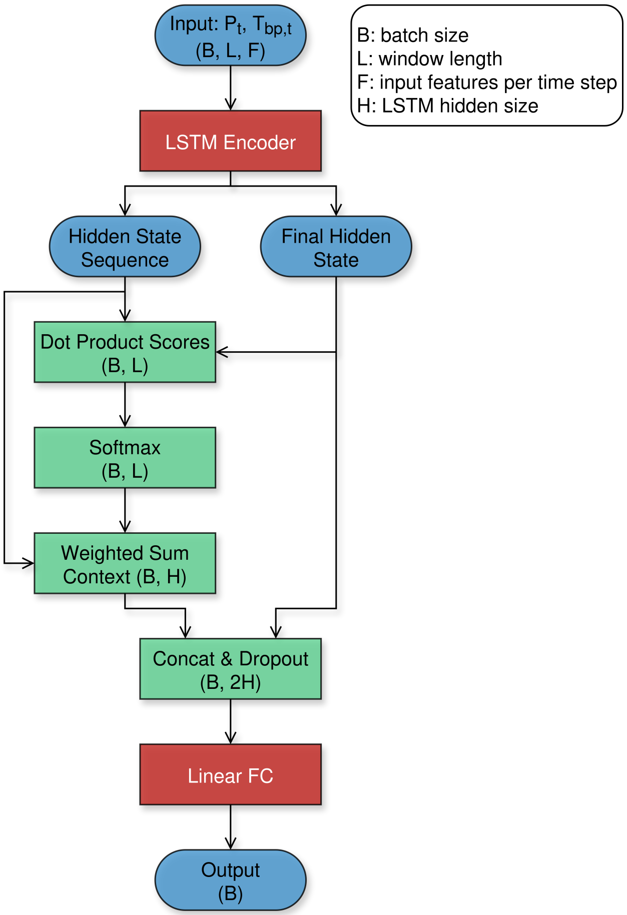

Fig. 2. **Attention-augmented LSTM architecture employed in this study.** A window of L time-steps with $F = 2$ features (net power P_t and base-plate temperature $T_\mathrm{bp}, t)$ is fed to a single-layer LSTM encoder, producing the hidden-state sequence $\{\mathbf{h}_1, \dots, \mathbf{h}_L\}$ and the final hidden state $\mathbf{h}_L$. The final state acts as a query in a dot-product attention mechanism: similarity scores $s_t = \mathbf{h}_t^\top \mathbf{h}_L$ are normalized by a softmax to yield weights α_t. A context vector $\mathbf{c} = \sum_{t=1}^{L} \alpha_t \mathbf{h}_t$ is then concatenated to $\mathbf{h}_L$, followed by dropout ($p = 0.1$) and a fully connected layer that outputs the junction-temperature estimate $\hat{T}_{\mathrm{j},t}$.

The search converged to the following *optimal* hyper-parameters:

$$(H, \mathit{layers}, \mathrm{lr}) = (16,\ 1,\ 1 \times 10^{-4}),$$
$$(\lambda_\mathrm{ss}, \lambda_\mathrm{tr}) = (9.78 \times 10^{-6},\ 2.15 \times 10^{-6}).$$

This observer architecture counts only ≈ 1.3 k parameters, ensuring microcontroller deployability.

D. Physics-informed Loss

To keep predictions thermodynamically plausible, we augment the mean-squared data term with two residuals derived from a single-RC ladder. To ensure *uniformity* between steady-state and transient constraints, both use the *same* RC topology: a conduction branch $R_{\theta,\mathrm{C}}$ ($\mathrm{K\,W^{-1}}$) to the coolant, a nearly-open convective branch $R_{\theta,\mathrm{V}}$ ($\mathrm{K\,W^{-1}}$) to ambient, and an effective capacitance C_θ ($\mathrm{J\,K^{-1}}$). Let $T_{\mathrm{env},t}$ denote ambient temperature.

The steady-state setpoint of this network is

$$T_{\mathrm{j},ss}(t) \;=\; \frac{P_t + \frac{T_{\mathrm{bp},t}}{R_{\theta,\mathrm{C}}} + \frac{T_{\mathrm{env},t}}{R_{\theta,\mathrm{V}}}}{\frac{1}{R_{\theta,\mathrm{C}}} + \frac{1}{R_{\theta,\mathrm{V}}}}\,,$$

which yields the steady-state residual

$$r_{\mathrm{ss},t} = \underbrace{\hat{T}_{\mathrm{j},t}}_{\text{network}} - \underbrace{T_{\mathrm{j},ss}(t)}_{\text{RC steady state}}\;. \tag{2}$$

The transient residual follows from the same RC network:

$$r_{\mathrm{tr},t} = \underbrace{C_\theta \frac{\hat{T}_{\mathrm{j},t} - \hat{T}_{\mathrm{j},t-1}}{\Delta t}}_{\text{thermal inertia}} - \Big(\underbrace{P_t}_{\text{loss input}} + \underbrace{\frac{T_{\mathrm{bp},t} - \hat{T}_{\mathrm{j},t}}{R_{\theta,\mathrm{C}}}}_{\text{conduction}} + \underbrace{\frac{T_{\mathrm{env},t} - \hat{T}_{\mathrm{j},t}}{R_{\theta,\mathrm{V}}}}_{\text{convection}}\Big). \tag{3}$$

Minimising

$$\mathcal{L} = \frac{1}{N}\sum_t \big(\hat{T}_{\mathrm{j},t} - T^{\mathrm{ref}}_{\mathrm{j},t}\big)^2 + \lambda_{\mathrm{ss}}\, r^2_{\mathrm{ss},t} + \lambda_{\mathrm{tr}}\, r^2_{\mathrm{tr},t} \tag{4}$$

balances data fidelity with steady-state consistency ($r_{\mathrm{ss},t}$, in K) and transient energy conservation ($r_{\mathrm{tr},t}$, in W).

E. Thermal Model Parameters Identification

The reduced first-order RC model underlying the residuals in (2)–(3) comprises a *conduction* path $R_{\theta,\mathrm{C}}$ from the chip stack to the coolant and an effective thermal capacitance C_θ. A parallel *convective* path toward ambient, $R_{\theta,\mathrm{V}}$, is included in the topology but, for sealed automotive ECUs, it is set *a priori* to a very large value ($R_{\theta,\mathrm{V}} \approx 10^5\,\mathrm{K\,W^{-1}}$) to reflect negligible internal-air convection over the time-scales of interest, consistent with standard RC thermal-modeling practice for automotive electronics [18].

All parameters are fixed once; in particular, $R_{\theta,\mathrm{V}}$ is *not* identified from data but kept at its nominal large value, whereas $R_{\theta,\mathrm{C}}$ and C_θ are obtained offline via a simple power-step experiment on the target module:

- **Steady state.** Measuring the asymptotic temperature rise ΔT_∞ at several power levels yields $R_{\theta,\mathrm{tot}} = \Delta T_\infty / P$. Subtracting the coolant-side contribution (from an NTC on the ceramic) isolates $R_{\theta,\mathrm{C}}$.
- **Transient.** Fitting $T(t) = T_\infty\big(1 - e^{-t/\tau}\big)$ returns the time constant τ; with $R_{\theta,\mathrm{tot}}$ known, $C_\theta = \tau / R_{\theta,\mathrm{tot}}$.

This procedure leads to

$$\begin{aligned} R_{\theta,\mathrm{C}} &\approx 1.7\,\mathrm{K\,W^{-1}}, \\ R_{\theta,\mathrm{V}} &\approx 1.0 \times 10^5\,\mathrm{K\,W^{-1}} \quad (\text{fixed } \textit{a priori}), \\ C_\theta &\approx 1.5\,\mathrm{J\,K^{-1}}. \end{aligned} \tag{5}$$

Given its magnitude, the convective branch contributes negligibly to the residuals; we keep it in the formulation for completeness, while identification and learning focus on the conductive dynamics in (1).

F. Uncertainty Evaluation

In a safety-critical automotive setting, point forecasts are insufficient; we estimate predictive uncertainty with Monte Carlo (MC) dropout applied only at the model head. A dropout layer with rate $p = 0.1$ is applied to the concatenated vector $[\,\text{query; context}\,] \in \mathbb{R}^{2H}$ immediately before the final linear layer; no dropout is used inside the LSTM (i.e., no recurrent/variational dropout). At test time the same dropout remains active and we perform $T = 30$ stochastic forward passes per input window, producing $\{\hat{T}_j^{(t)}\}_{t=1}^{T}$; the model size is unchanged and compute cost scales linearly with T. The point estimate is the predictive mean $\bar{T}_j = \frac{1}{T}\sum_{t=1}^{T}\hat{T}_j^{(t)}$. Uncertainty intervals are reported as the 95 % empirical quantiles $[\mathrm{q}_{2.5}, \mathrm{q}_{97.5}]$ over the T samples. We report the mean-squared error (MSE) of $\bar{T}_j$ and the empirical 95 % coverage (fraction of targets falling within the reported interval).

G. Training Protocol and Deployment

The pipeline is written in `PyTorch 2.5` and leverages automatic mixed-precision whenever a CUDA-capable device is available. Mini-batches of 32 windows are processed with Adam; gradients are clipped to $\|\nabla\|_2 \le 1$. Early stopping monitors the validation loss with a patience of ten epochs and halts after at most one hundred epochs. During the first $N_{\text{warm}} = 10$ epochs, the physics-informed penalties λ_{ss} and λ_{tr} are linearly annealed from 0 to their target values to avoid unstable gradients at the beginning of training. All experiments were run on a single mid-range GPU (e.g. laptop-grade RTX 3060). Inference on the trained model fits comfortably within the SRAM and compute budget of recent automotive MCUs, enabling on-board deployment without external accelerators.

V. Results and Discussion

A. Computational Footprint

Both the plain LSTM and its physics-informed counterpart (PI-LSTM) share the same lightweight backbone.

$$\#\text{params} = 1\,313 \quad (\approx 5.1\,\text{kB in } \texttt{float32})$$

$$\text{FLOPs} = 1.8 \times 10^5 \text{ per window}$$

B. Test-set Accuracy

Table I compares point-error and calibration metrics on the held-out $\sim$ 30 % test split. Embedding the steady-state and transient energy-balance terms markedly improves every point-error metric and tightens the predictive intervals:

TABLE I
Test-set performance comparison.

Metric	Standard LSTM	PI-LSTM
MSE (°C^2)	6.6350	**4.6030**
MAE (°C)	1.9933	**1.6222**
Huber ($\delta = 1$)	2.3821	**2.1198**
NLL	**2.4150**	2.7327
95 % coverage [↑]	98.25 %	**99.54 %**

For reference, the optimally calibrated lumped-RC ODE reaches an MSE of $10.62\ ^\circ\mathrm{C}^2$, i.e. 57 % worse than the PI-LSTM.

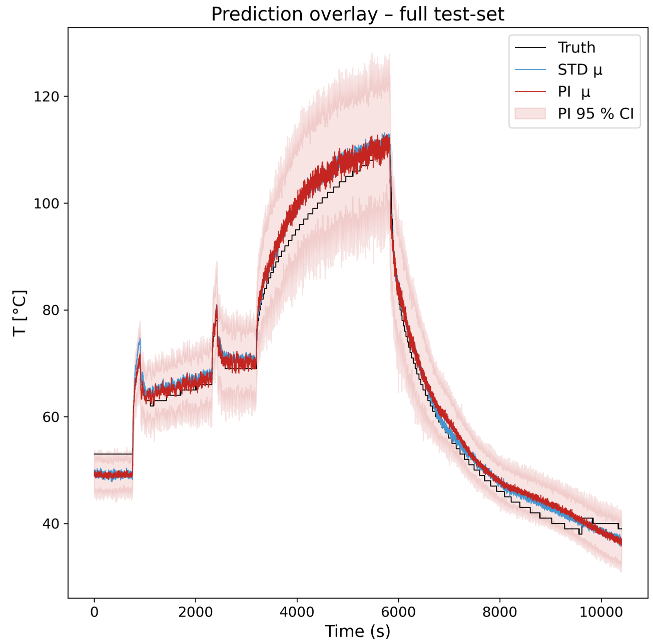

Fig. 3. Overlay of ground truth, mean predictions, and 95 % confidence band for the first test trajectory. The PI-LSTM tracks sharp transients more closely than the standard model while providing well-calibrated uncertainty.

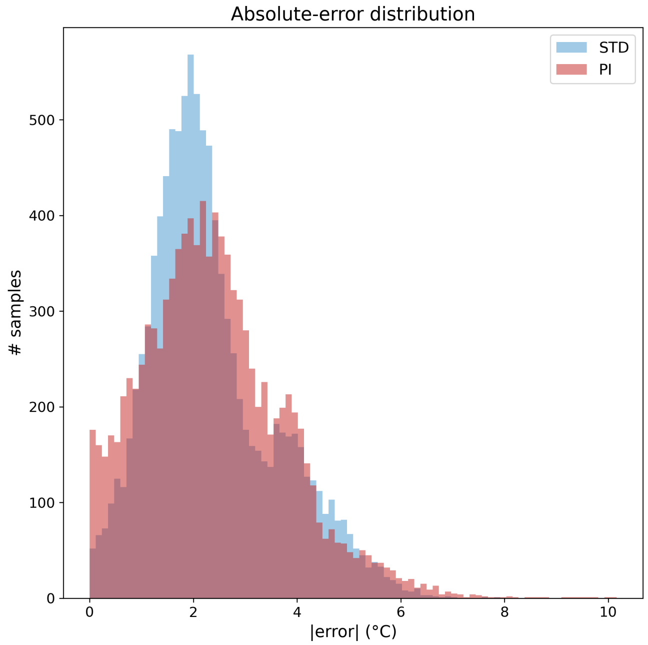

Fig. 4. Absolute error histograms. The PILSTM (red) places visibly more probability mass below 1 °C, whereas the standard model (blue) is more concentrated in the 1–3 °C range. Above 4 °C the two curves overlap closely, but only the PILSTM shows a handful of rare outliers beyond 6 °C.

C. Discussion

Table I highlights the gains from embedding physics-informed constraints in the LSTM, while Fig. 3 (trajectory overlay) and Fig. 4 (error histogram) qualitatively corroborate improved accuracy and calibration.

Across overall error metrics (MSE, MAE, and Huber loss), the PI-LSTM markedly boosts point-wise accuracy, reducing MSE by roughly 31 % and MAE by 19 % relative to the purely data-driven model. In Fig. 3, the red PI-LSTM trajectory closely follows ground truth, especially during sharp transients that typically inflate errors in the standard model. Enforcing energy-balance constraints steers the neural dynamics toward physically consistent temperature trajectories, improving performance in challenging regimes.

The PI-LSTM also clearly outperforms the purely physics-based ODE model (about 57 % lower MSE), showing that while equations encode valuable domain knowledge, a gray-box alone struggles with complex transients. PI-LSTM blends both strengths: neural flexibility for nonlinear dynamics and physical laws for steady-state and transient consistency.

Uncertainty metrics reveal a nuanced trade-off: a slight NLL increase (the PI-LSTM produces marginally wider predictive distributions) paired with higher empirical coverage of the 95 % intervals (from 98.25 % to 99.54 %). As Fig. 3 indicates, intervals remain narrow enough to track the true temperature while containing it, indicating improved calibration despite modestly increased width. This balance is crucial for real-time automotive use, where underestimating uncertainty is risky.

Figure 4 provides a granular view of absolute errors. The PI-LSTM's distribution peaks slightly later, but assigns over 30 % of samples to the sub-1 °C region, confirming superior high-precision performance. Between 1 °C and 3 °C the standard model dominates, while beyond 3 °C the tails converge; both remain light, though the PI-LSTM shows a dozen errors in the 6 °C to 10 °C range absent in the baseline. Together with reliability curves, this suggests the physics-informed model trades a slightly heavier extreme tail for many more near-zero-error predictions, and improved uncertainty calibration, which is desirable for real-time monitoring where tight bounds matter and occasional larger misses are tolerable.

Additional analyses show that larger predicted standard deviations align with larger absolute errors, indicating informative, well-calibrated uncertainty; attention weights focus on the most recent 10–30 samples and adapt during start-up transients, enhancing interpretability.

Crucially, these gains come with no extra complexity or resources: the PI-LSTM retains the same compact architecture (1.3 k parameters) as the standard model, with unchanged inference latency, satisfying stringent memory and compute limits of automotive microcontrollers.

Overall, quantitative metrics and qualitative plots converge to show that the PI-LSTM effectively marries data-driven flexibility with physics-based regularization, delivering reliable temperature predictions well suited to embedded automotive applications.

VI. Conclusions and Outlook

In this work, we presented a lightweight, attention-augmented, physics-informed LSTM neural observer coupled to a single RC thermal model. The architecture attains 1.63 °C MAE on unseen drive cycles with just 1313 parameters and 1.8×10^5 FLOPs per inference window, well within the envelope of modern automotive microcontrollers.

These results demonstrate that physics-informed learning can be effective with modest model capacity. The remaining step is to meet the stringent compute, memory, and energy budgets of ultra-low-power MCUs (e.g., ARM Cortex-M, RISC-V), which motivates targeted optimizations of the LSTM and its inference pipeline:

- **Mixed-precision quantization:** Quantize weights and most intermediates to INT8, while keeping cell/hidden states and critical outputs in higher precision (INT16/INT32) to avoid accuracy loss.
- **Integer-only activation functions:** Replace nonlinearities with integer-friendly approximations (e.g., low-degree polynomials, piecewise-linear LUTs with dyadic scaling) so that `sigmoid` and `tanh` in the LSTM gates are evaluated entirely in INT arithmetic (INT8 multiplies with INT32 accumulation), followed by a single re-quantization per gate.
- **Quantization-aware training (QAT):** Fine-tune with simulated quantization to preserve performance under reduced precision of weights and activations.

Further gains may come from structured pruning to remove redundancy and MCU-specific inference kernels that reduce memory traffic and power; combining these strategies can deliver robust, accurate, and energy-efficient thermal state estimation suitable for real-time, safety-critical automotive deployment on highly constrained MCUs.

VII. Acknowledgements

Funded by the European Union (grant agreement No 101056760, HiPE). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Climate, Infrastructure and Environment Executive Agency (CINEA). Neither the European Union nor the granting authority can be held responsible for them.

References

[1] M. A. Eleffendi and M. Johnson. "Application of Kalman Filter to Estimate Junction Temperature in IGBT Power Modules". In: *IEEE Transactions on Power Electronics* 31 (Mar. 2015), pp. 1–1. DOI: 10.1109/TPEL.2015.2418711.

[2] A. Abuelnaga, M. Narimani, and A. S. Bahman. "A Review on IGBT Module Failure Modes and Lifetime Testing". In: *IEEE Access* 9 (2021), pp. 9643–9663. DOI: 10.1109/ACCESS.2021.3049738.

[3] H. Lim, J. Hwang, S. Kwon, H. Baek, J. Uhm, and G. Lee. "A Study on Real Time IGBT Junction Temperature Estimation Using the NTC and Calculation of Power Losses in the Automotive Inverter System". In: *Sensors* 21.7 (2021). DOI: 10.3390/s21072454.

[4] W. Kirchgässner, O. Wallscheid, and J. Böcker. "Estimating Electric Motor Temperatures With Deep Residual Machine Learning". In: *IEEE Transactions on Power Electronics* 36.7 (2021), pp. 7480–7488. DOI: 10.1109/TPEL.2020.3045596.

[5] X. Jia et al. "Physics Guided RNNs for Modeling Dynamical Systems: A Case Study in Simulating Lake Temperature Profiles". In: *Proceedings of the 2019 SIAM International Conference on Data Mining (SDM)*, pp. 558–566. DOI: 10.1137/1.9781611975673.63. eprint: https://epubs.siam.org/doi/pdf/10.1137/1.9781611975673.63.

[6] A. Lamanuzzi, M. Tranchero, A. Pastore, C. Romano, and P. Santero. "A Simple and Effective Power Derating Strategy Based on Junction Temperature Estimation Improving Both Performance and Reliability". In: *2023 29th International Workshop on Thermal Investigations of ICs and Systems (THERMINIC)*. 2023, pp. 1–4. DOI: 10.1109/THERMINIC60375.2023.10325912.

[7] X. Han and M. Saeedifard. "Junction temperature estimation of SiC MOSFETs based on Extended Kalman Filtering". In: *2018 IEEE Applied Power Electronics Conference and Exposition (APEC)*. 2018, pp. 1687–1694. DOI: 10.1109/APEC.2018.8341244.

[8] B. Rodríguez, E. Sanjurjo, M. Tranchero, C. Romano, and F. González. "Thermal Parameter and State Estimation for Digital Twins of E-Powertrain Components". In: *IEEE Access* 9 (2021), pp. 97384–97400. DOI: 10.1109/ACCESS.2021.3094312.

[9] S. Hochreiter and J. Schmidhuber. "Long Short-Term Memory". In: *Neural Computation* 9.8 (1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

[10] D. Bahdanau, K. Cho, and Y. Bengio. *Neural Machine Translation by Jointly Learning to Align and Translate*. 2016. arXiv: 1409.0473 [cs.CL].

[11] M. Raissi, P. Perdikaris, and G. Karniadakis. "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations". In: *Journal of Computational Physics* 378 (2019), pp. 686–707. DOI: 10.1016/j.jcp.2018.10.045.

[12] J. Lee and J.-I. Ha. "Temperature Estimation of PMSM Using a Difference-Estimating Feedforward Neural Network". In: *IEEE Access* 8 (2020), pp. 130855–130865. DOI: 10.1109/ACCESS.2020.3009503.

[13] D. Czerwinski, J. Gęca, and K. Kolano. "Machine Learning for Sensorless Temperature Estimation of a BLDC Motor". In: *Sensors* 21.14 (2021), p. 4655. DOI: 10.3390/s21144655.

[14] K. Rönnberg, P. Kakosimos, Z. Kolondjovski, and E. Nordlund. "Machine Learning-Based Adjustments of Thermal Networks". In: *Proc. 11th Int. Conf. on Power Electronics, Machines and Drives (PEMD)*. 2022, pp. 424–428. DOI: 10.1049/icp.2022.1087.

[15] R. Hughes, T. Haidinger, X. Pei, and C. Vagg. "Real-Time Temperature Prediction of Electric Machines Using Machine Learning with Physically Informed Features". In: *Energy and AI* 14 (2023), p. 100288. DOI: 10.1016/j.egyai.2023.100288.

[16] Y. Sheng, X. Liu, Q. Chen, Z. Zhu, C. Huang, and Q. Wang. "OLTEM: Lumped Thermal and Deep Neural Model for PMSM Temperature". In: *AI* 6.8 (2025), p. 173. DOI: 10.3390/ai6080173.

[17] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. "Optuna: A Next-Generation Hyperparameter Optimization Framework". In: *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2019, pp. 2623–2631.

[18] *Automotive LED Side Light SEPIC DC-to-DC Converter Design Example*. Application Note AN50002. Rev. 2.0. Nexperia, 2021.