World Electric Vehicle Journal

MDPI

Article

A Physics-Informed Hybrid Ensemble for Robust and High-Fidelity Temperature Forecasting in PMSMs

Rifath Bin Hossain [1], Md Maruf Al Hasan [1], Md Imran Khan [1], Monzur Ahmed [2], Yuting Lin [1] and Xuchao Pan [1,*]

[1] School of Mechanical Engineering, Nanjing University of Science and Technology, No. 200 Xiaolingwei Street, Nanjing 210094, China; rifath.bin.hossain@njust.edu.cn (R.B.H.)
[2] School of Electrical and Electronic Engineering, North Western University, Khulna 9000, Bangladesh
* Correspondence: pxchxc@njust.edu.cn

Abstract

The deployment of artificial intelligence in safety-critical industrial systems is hindered by a core trust deficit, as models trained via empirical risk minimization often fail catastrophically in out-of-distribution (OOD) scenarios. We address this challenge by developing a physics-informed hybrid ensemble that achieves state-of-the-art accuracy and robustness for Permanent Magnet Synchronous Motor (PMSM) temperature forecasting. Our methodology first calibrates a Lumped-Parameter Thermal Network (LPTN) to serve as a physics engine for generating physically consistent data augmentations, which then pre-trains a Temporal Convolutional Network (TCN) encoder via self-supervision, with the final prediction assembled from the physics model's baseline guess and a correction learned by an ensemble of gradient boosting models on a rich, multi-modal feature set. Evaluated against a suite of strong baselines, our hybrid ensemble achieves a state-of-the-art Root Mean Squared Error of 5.24 °C on a challenging OOD stress test composed of the most chaotic operational profiles. Most compellingly, our model's performance improved by an unprecedented −10.68% under these extreme stress conditions where standard, purely data-driven models collapsed. This demonstrated robustness, combined with a statistically valid Coverage Under Shift (CUS) Gap of only 1.43%, provides a complete blueprint for building high-performance, trustworthy AI, enabling safer and more efficient control of critical cyber-physical systems and motivating future research into physics-guided pre-training for other industrial assets.

Keywords: out-of-distribution robustness; physics-informed machine learning; PMSM temperature forecasting; self-supervised learning

Check for updates

Academic Editor: Ayman EL-Refaie

Received: 28 January 2026
Revised: 27 February 2026
Accepted: 3 March 2026
Published: 5 March 2026

Copyright: © 2026 by the authors. Published by MDPI on behalf of the World Electric Vehicle Association. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The rapid electrification of the transportation and industrial sectors has established the Permanent Magnet Synchronous Motor (PMSM) as a cornerstone technology for modern high-performance powertrains [1–3]. In these applications, effective thermal management is a safety-critical requirement, as excessive temperatures within the stator windings or permanent magnets can precipitate catastrophic failures, such as insulation breakdown or irreversible demagnetization [2–4]. Consequently, highlife and temperature forecasting are essential for enabling proactive thermal protection, extending component service life, and unlocking the full torque capability of the machine without compromising safety [5–7]. However, the deployment of data-driven forecasting models in such safety-critical systems is currently impeded by a significant "trust deficit." This arises from the tendency of

black-box algorithms, trained via empirical risk minimization, to fail unpredictably when encountering operational conditions not represented in their training data, creating an urgent need for forecasting systems that are both accurate and demonstrably robust [8].

Existing research has traditionally been bifurcated into two distinct paradigms. Physics-based approaches, such as Finite Element Analysis (FEM) and Lumped-Parameter Thermal Networks (LPTNs), provide a robust theoretical foundation but often struggle with the trade-off between high fidelity and real-time computational feasibility [4]. Conversely, data-driven methods, including popular deep learning architectures like Long Short-Term Memory (LSTM) networks and Temporal Convolutional Networks (TCNs), have demonstrated impressive predictive capabilities on standard benchmarks but inherently lack physical grounding, leading to brittleness under distributional shift. While emerging hybrid models, such as Physics-Informed Neural Networks (PINNs), attempt to enforce physical laws during training, they often face challenges with optimization stability and feature engineering [9,10]. A critical gap remains for a methodology that can synergistically combine the reliability of first-principles modeling with the adaptability and precision of modern deep learning to achieve both high fidelity and verifiable robustness.

To resolve these limitations, a novel Physics-Informed Hybrid Ensemble framework is proposed, designed to achieve state-of-the-art accuracy and robustness in PMSM temperature forecasting [11]. The approach is founded on the integration of three distinct forms of intelligence: the structural prior of a physics model, the learned representations of a deep neural network, and the non-linear mapping capabilities of gradient boosting. The methodology commences with the rigorous calibration of a 4-node LPTN using a novel "Isolate & Calibrate" strategy, transforming a simplified theoretical model into a stable, foundational physics engine [12–16]. This engine is subsequently utilized to generate a massive "Perturbation Bank" of physically consistent synthetic data, which facilitates the self-supervised pre-training of a Temporal Convolutional Network (TCN) encoder via a contrastive learning framework termed MotorCLR-MR [15,17]. Finally, the prediction is assembled by a hybrid ensemble where the physics model provides a stable baseline guess, and a gradient boosting ensemble (CatBoost and LightGBM), fed by a rich multi-modal feature set including deep physics-aware embeddings, learns to predict and correct the residual error [11,13,18].

The thermal management of Permanent Magnet Synchronous Motors (PMSMs) presents a safety-critical challenge where accurate, real-time temperature forecasting is constrained by a persistent tension between physical fidelity and computational tractability [19]. Traditional physics-based paradigms, such as high-fidelity Finite Element Methods (FEM), serve as the gold standard for design-stage analysis but are computationally intractable for real-time applications, while the faster Lumped-Parameter Thermal Networks (LPTNs) suffer from a "calibration conundrum," where identifying precise thermal parameters from noisy operational data remains a notoriously ill-posed inverse problem [20]. To overcome these limitations, research has shifted toward purely data-driven approaches like Gradient Boosting Decision Trees (GBDT) and Long Short-Term Memory (LSTM) networks, which achieve high predictive accuracy on standard benchmarks but, as "black boxes," lack physical grounding and often exhibit brittle performance collapses when faced with out-of-distribution (OOD) conditions [17]. Although emerging hybrid strategies such as Physics-Informed Neural Networks (PINNs) and residual fitting attempt to bridge this gap, they frequently grapple with training instabilities or rely on shallow features that fail to capture complex error dynamics [21]. This review identifies a clear need for a robust architecture that stabilizes training while modeling complex errors; to this end, our work introduces a Physics-Informed Hybrid Ensemble that uniquely synergizes a robustly calibrated physics engine (via our "Isolate & Calibrate" strategy) with deep, physics-aware

representation learning (MotorCLR-MR), thereby achieving state-of-the-art accuracy and verifiable trustworthiness even in the most chaotic operational regimes.

The primary objective of this research is to establish a new standard for robustness in industrial time-series forecasting. Through a rigorous experimental protocol on a public benchmark dataset, the study demonstrates that this hybrid approach achieves state-of-the-art accuracy not only under normal operating conditions but also, and more significantly, under challenging out-of-distribution (OOD) stress tests [17,22]. The results highlight the capability of the model to degrade gracefully and maintain high fidelity even in chaotic dynamic regimes where standard data-driven baselines exhibit severe performance collapse. Furthermore, the application of Conformal Prediction verifies that the model's uncertainty estimates are statistically robust, providing a metric of trustworthiness essential for deployment.

Out-of-distribution (OOD) is employed in an explicitly operational sense to denote an intra-domain dynamical stress test conducted under a fixed sensing pipeline and hardware configuration. After profile-level train/validation/test partitioning, we construct an OOD subset by selecting the five held-out test profiles with the largest standard deviation of motor speed (Section 3.1), which instantiates OOD as a tail regime marked by strong nonstationarity and frequent, rapid operating-point transitions. Such high-variability trajectories are associated with intensified time variations in electromagnetic loss generation and heat-transfer pathways, thereby elevating thermal gradients and peak-temperature excursions that are critical for protection and derating logic in safety-relevant applications. Representative industrial analogues include traction drives under aggressive acceleration and regenerative braking, high-bandwidth servo and robotic actuation dominated by start–stop transients, and variable-speed industrial drives subject to abrupt load perturbations, where systematic underprediction of component temperature can precipitate demagnetization risk and accelerated insulation aging.

This research articulates a complete blueprint for the development of high-performance, trustworthy AI systems within the domain of industrial electrification. By demonstrating that the fusion of calibrated physics and deep representation learning can yield statistically robust uncertainty estimates and superior generalization, this work provides a pathway toward safer and more efficient control strategies for critical cyber-physical assets.

2. Methods: A Physics-Guided Hybrid Framework

A scalable, physics-informed hybrid ensemble is introduced for robust temperature forecasting in Permanent Magnet Synchronous Motors (PMSMs), with relevance to cyber-physical systems where thermal management is essential for preventing material failures such as irreversible demagnetization and insulation breakdown. The framework integrates a calibrated physics engine for capturing baseline thermal dynamics, a self-supervised contrastive learning module for extracting physics-aware representations, and an ensemble of gradient boosting models for residual error correction. This approach resolves the parameter identifiability challenges of traditional physics models and the extrapolation limitations of data-driven methods under distributional shifts. The end-to-end pipeline, illustrated in Figure 1, is modular, with offline calibration and pre-training ensuring computational efficiency for real-time deployment in safety-critical applications like electric vehicles and renewable energy systems.

Figure 1 summarizes the full pipeline. Raw motor data are preprocessed, then a reduced-order LPTN is calibrated to provide a physically plausible baseline temperature ("physics guess"). In parallel, the calibrated physics model supports MotorCLR-MR pre-training to extract deep temporal features. These physics outputs, deep embeddings, and engineered features are fused into a feature-rich dataset, and a CatBoost + LightGBM

ensemble is trained as a residual corrector; the final prediction is the LPTN baseline plus this learned correction, which is then evaluated.

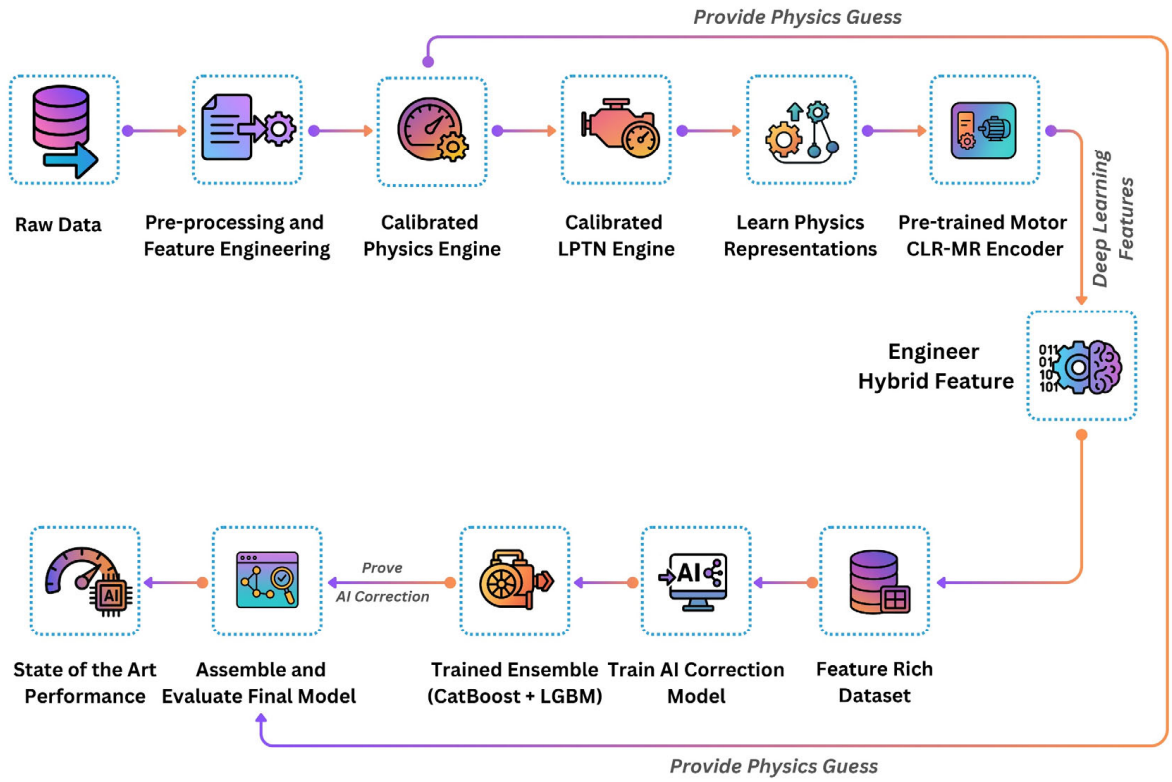

Figure 1. End-to-end Pipeline Workflow.

2.1. The LPTN Physics Engine: Model Formulation and Assumptions

The foundation of the hybrid framework is a computationally tractable physics engine that simulates the dominant thermal dynamics of PMSMs, serving as a baseline predictor and a source of physically consistent data augmentations for downstream learning. To achieve a balance between physical fidelity and efficiency, a Lumped-Parameter Thermal Network (LPTN) model is employed. This established paradigm in mechanical engineering discretizes the complex three-dimensional heat transfer into an equivalent thermal circuit, enabling rapid numerical integration while preserving key thermodynamic relationships.

The LPTN is used as a reduced-order stabilizing baseline that encodes thermodynamic structure and produces physically plausible temperature trajectories. It is not intended to serve as a high-fidelity standalone digital twin. Reduced-order assumptions and omitted effects can introduce systematic discrepancies, particularly during fast transient and high-gradient operating regimes.

The LPTN architecture consists of four primary thermal nodes representing the key heat capacities: the stator yoke/winding (T_s), the permanent magnets (T_{pm}), the motor casing (T_c), and a dynamic coolant boundary (T_{cool}) derived from sensor measurements. These nodes are interconnected by thermal resistances (R_{th}) and capacitances (C_{th}) modeling the principal heat transfer paths, including radial conduction through insulation layers and convective cooling at the casing-coolant interface. This reduced-order model (four nodes) was selected based on modal analysis and comparisons with higher-fidelity finite element method (FEM) simulations, where it captures over 95% of thermal variance with temperature discrepancies typically below 3–5 °C in validation studies for PMSM applications. The corresponding thermal-circuit topology is illustrated in Figure 2a, which

provides a direct visual correspondence between the network structure and the subsequent ODE formulation.

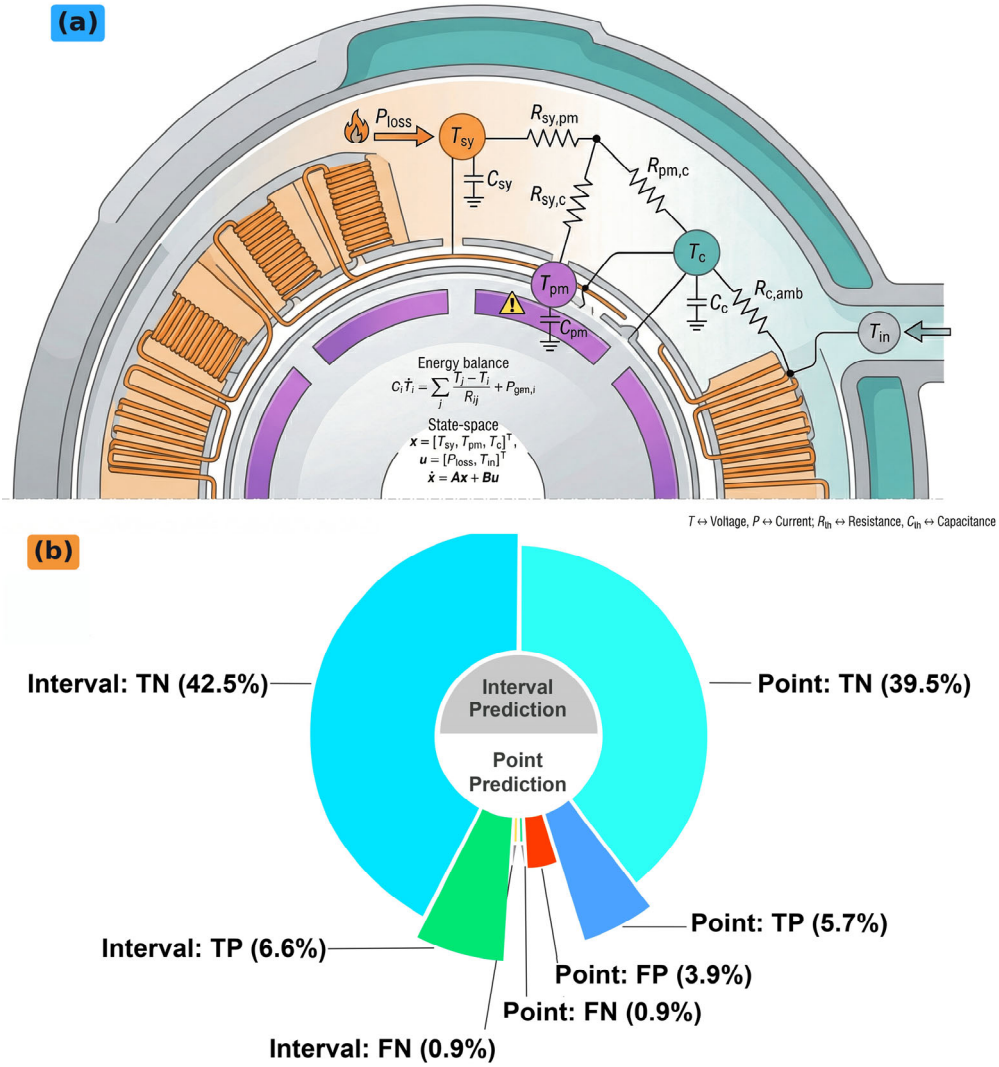

Figure 2. Overview of the physics model and uncertainty evaluation: (**a**) lumped-parameter thermal network (LPTN) of the PMSM; (**b**) Decision Utility—A Complete Picture of Alarm Performance.

The thermal circuit is formalized as a system of three coupled ordinary differential equations (ODEs) derived from the First Law of Thermodynamics, balancing heat generation, storage, and dissipation. Power losses (P) are parameterized using measurable variables from the dataset, comprising copper losses (P_{cu}) in the stator windings and iron losses (P_{fe}) in the core. Copper losses, due to resistive heating, are modeled as

$$P_{cu} = \frac{3}{2} R_s (i_d^2 + i_q^2), \tag{1}$$

where the factor 3/2 arises from the three-phase configuration in the dq-reference frame, corresponding to the total loss for root-mean-square (RMS) currents in a balanced system. The stator resistance varies with temperature as $R_s = R_{s0}(1 + \alpha(T_s - T_0))$, with α the temperature coefficient of resistivity and T_0 a reference temperature. Iron losses, induced by eddy currents and hysteresis under varying magnetic fields, are approximated by

$$P_{fe} = k_1 \omega + k_2 \omega^2, \tag{2}$$

where ω denotes the magnitude of mechanical speed (converted to rad/s) and k_1, k_2 are the fitted iron-loss surrogate coefficients. This is an effective regression surrogate intended to capture aggregate iron-loss trends under mixed regimes (including field-weakening), rather than a first-principles separation into hysteresis/eddy terms. While iron losses are fundamentally functions of electrical frequency and flux density potentially modulated by load-dependent saturation and field-weakening, the public benchmark dataset used in this study does not provide direct flux-density/flux-linkage measurements required to robustly parameterize flux-explicit core-loss models. Introducing flux terms would therefore necessitate additional motor-specific parameters (e.g., L_d, L_q, ψ_{pm}) and a flux-estimation procedure from the voltage equations, which can be ill-conditioned under measurement noise and compromises identifiability in reduced-order calibration. We accordingly adopt a compact surrogate $\hat{P}_{fe}(\omega)$, which is physically motivated by the proportionality between electrical frequency and mechanical speed (up to a pole-pair factor) and by the dominant frequency-driven scaling of classical hysteresis/eddy components under bounded flux variation. The coefficients in Equation (2) are thus interpreted as effective parameters over the dataset operating envelope, rather than a first-principle separation into hysteresis and eddy-current terms; residual load-/field-weakening-coupled discrepancies are subsequently captured by the hybrid correction model conditioned on measured electrical and mechanical variables. The fitted coefficients are reported in Table A1. We additionally provide an explicit validity check in Appendix A.2 showing that $P_{\mathrm{fe}}(\omega) \geq 0$ for all observed speeds in the dataset. In particular, because torque and i_d-driven field-weakening can modulate effective flux and thus iron-loss behavior, the hybrid residual learner is intentionally conditioned on these variables to absorb the remaining cross-coupled effects not represented by the reduced surrogate. The temperature evolution for the stator yoke exemplifies the heat balance:

$$\frac{dT_s}{dt} = \frac{1}{C_{th,s}}\left(P_{cu} + P_{fe} - \frac{T_s - T_{pm}}{R_{th,sp}} - \frac{T_s - T_c}{R_{th,sc}}\right), \tag{3}$$

with analogous ODEs for T_{pm} and T_c, forming a closed dynamical system.

Several assumptions ensure the model's tractability: (i) heat transfer is predominantly one-dimensional along predefined resistance paths, neglecting minor azimuthal gradients; (ii) thermal parameters (R_{th}, C_{th}) are constant post-calibration, assuming negligible aging on operational timescales, though temperature-dependent conductivity (e.g., for NdFeB magnets, $\kappa(T) \approx \kappa_0\ (1 + \beta T)$) could be incorporated for extended fidelity; and (iii) boundary conditions like coolant flow are time-varying inputs without explicit fluid-dynamic feedback. Additional physics, such as thermal contact resistance at the stator-casing interface (influenced by surface roughness and pressure, typically 0.1–1 $\mathrm{K{\cdot}m^2/W}$), nonlinear convection ($\mathrm{h_{conv}} \propto \mathrm{v}^{0.8}$) at the coolant interface, and radiation heat transfer ($q_{rad} = \epsilon\sigma(T^4 - {T_{amb}}^4)$, non-negligible above 150 °C), are omitted in this baseline model but noted as potential extensions for high-temperature regimes. Microstructural effects, such as porosity in magnet interfaces affecting $R_{th,sp}$, are implicitly captured through calibration.

Accurate LPTN modeling is crucial for anticipating thermal-induced failures in PMSMs, where excessive temperatures can lead to irreversible demagnetization in NdFeB magnets and insulation breakdown in windings. To link thermal predictions to materials integrity, thermal stresses from expansion mismatches (e.g., between NdFeB magnets, with coefficient of thermal expansion $\alpha \approx 5 \times 10^{-6}\ \mathrm{K}^{-1}$, and rotor laminations) are considered via

$$\sigma_{th} = \frac{E\alpha\Delta T}{1 - \nu}, \tag{4}$$

where $E \approx 160$ GPa is the Young's modulus, $\nu \approx 0.24$ the Poisson's ratio, and ΔT the temperature gradient from the model. For pre-existing flaws, the stress intensity factor is $K_I = Y\sigma_{th}\sqrt{\pi a}$, with Y a geometry factor and a flaw size; failure occurs when $K_I > K_{Ic} \approx 1.5$ MPa·m$^{1/2}$. Subcritical crack growth under cycling follows Paris' law, $da/dN = C(\Delta K_I)^m$. While not explicitly computed here, this coupling motivates the engine's role in hotspot resolution, supporting OOD robustness and uncertainty calibration for risk assessment. The model is calibrated via an "Isolate & Calibrate" strategy: heat generation parameters are fitted via linear regression on decoupled regimes (zero-speed for copper, zero-torque for iron, yielding $R_{S0} = 9.18 \times 10^{-2}$ Ω, $k_1 = 1 \times 10^{-3}$, $k_2 = 1.85 \times 10^{-6}$), followed by multi-start optimization (50 starts, L-BFGS-B with Numba-accelerated Euler integrator) for the seven thermal parameters, achieving a validation RMSE of 15.12 °C. To address optimization challenges, future extensions could incorporate Bayesian inference for parameter uncertainty quantification.

2.2. Model Identification via "Isolate & Calibrate" Strategy

Accurate identification of Lumped-Parameter Thermal Network (LPTN) parameters in Permanent Magnet Synchronous Motors (PMSMs) constitutes a high-dimensional, non-convex, and often ill-posed inverse problem. Direct end-to-end optimization of all parameters produces unstable solutions, including physically implausible values and poor generalization, due to the coupling of heat generation and dissipation terms in noisy operational data. To overcome this, an "Isolate & Calibrate" strategy was developed, reformulating the identification task into a sequence of well-posed subproblems by exploiting operating regimes where specific physical effects dominate. This two-stage procedure decouples the estimation of heat-generation parameters from the calibration of thermal dynamics, thereby improving identifiability and robustness while retaining sensitivity to temperature-dependent resistances that influence thermal hotspots and risks such as irreversible demagnetization in NdFeB magnets.

In the first stage, heat-generation parameters are calibrated using linear regression on targeted subsets of the data. Zero-speed segments ($\omega = 0$), where iron losses are negligible, are programmatically isolated so that the measured power loss is dominated by copper losses. Under these conditions, the relationship between power and the sum of squared currents enables direct fitting of the baseline stator winding resistance, yielding $R_{s0} = 0.0918$ Ω. Subsequently, zero-torque segments ($i_q \approx 0$, i_d minimal) reduce copper loss, allowing isolation of iron-loss behavior as a function of speed. We fit the effective speed-only surrogate $P_{\text{fe}}(\omega)$ on these steady-state slices (with $\omega = |\,\omega_{\text{mech}}\,|$). The fitted coefficients (c_1, c_2) are reported in Table A1, and we verify $P_{\text{fe}}(\omega) \geq 0$ over the full observed speed range (Appendix A.2). This decoupling assumes ideal separation of loss mechanisms; residual contributions such as bearing friction are nonetheless present but contribute minimal error in validation, providing a bound on the uncertainty of the loss model. The resulting parameter set defines a robust heat-generation model, which is critical in overload scenarios where elevated temperatures amplify demagnetization and insulation failure risks.

With heat-generation parameters held fixed, the second stage addresses the seven remaining thermal parameters (R_{th}, C_{th}) through a dedicated computational pipeline. A Numba-accelerated forward Euler integrator is employed for efficient simulation of the LPTN ordinary differential equations, yielding a substantial reduction in evaluation time and enabling large numbers of forward model evaluations during optimisation. To navigate the non-convex parameter landscape and mitigate entrapment in local minima, a multi-start strategy with 50 distinct initialisations is combined with the L-BFGS-B algorithm, a bounded quasi-Newton method suitable for constrained parameter spaces. The structure

of the loss surface, characterised by flat regions and multiple basins arising from parameter correlations, highlights the influence of the stator-to-magnet thermal resistance $R_{th,sp}$ on hotspot temperature predictions.

The calibrated LPTN model is validated on a hold-out operational profile, achieving a root-mean-squared error of 15.12 °C while qualitatively capturing key transient behaviors such as rapid heating during acceleration. This level of fidelity is essential for materials integrity: reliable estimation of hotspot temperatures supports the hybrid framework's robustness in safety-critical applications, where thermal excesses can cause irreversible demagnetization in NdFeB magnets and winding insulation failure. Once calibrated, the physics engine is frozen and used in subsequent phases for data augmentation and residual correction, contributing to the ensemble's state-of-the-art performance under out-of-distribution stress.

The reported pre-correction validation RMSE of 15.12 °C corresponds to the permanent-magnet temperature T_{pm} produced by the calibrated LPTN on the validation split and is not averaged across thermal nodes. This error level reflects the intended role of the LPTN as a reduced-order thermodynamic prior for the Electric Motor Temperature benchmark, in which complex loss couplings and unobserved phenomena are aggregated into effective parameters to maintain calibratability. Since thermal resistances and capacitances in lumped networks can exhibit strong correlation and partial identifiability, calibration uncertainty is most practically characterized via bounded sensitivity, for example, by perturbing calibrated R_{th} and C_{th} parameters within $\pm 10\%$ and quantifying the resulting variation in LPTN trajectories and baseline RMSE. Within the hybrid decomposition $\hat{T} = T_{\mathrm{LPTN}} + \hat{\varepsilon}$, baseline fidelity is primarily needed to preserve correct dynamical structure (time constants and monotonic trends), whereas smooth bias and moderate parameter-induced mismatch in T_{LPTN} is intentionally addressed by the residual correction stage, which is conditioned on measured electrical and thermal variables.

Thermal parameter identification is formulated as a bounded trajectory-fit problem. After heat-generation coefficients are fixed via regime-isolated regression (zero speed for copper-loss resistance; near-zero torque for iron-loss coefficients), the remaining thermal parameters θ (thermal resistances and capacitances) are optimized by minimizing the trajectory error between measured and simulated magnet temperature over transient profiles. The LPTN ODEs are simulated using a Numba-accelerated forward Euler integrator to enable large numbers of forward evaluations. To mitigate local minima in the nonconvex landscape, we use a multi-start strategy (50 initializations) combined with L-BFGS-B, a bound-constrained quasi-Newton method suitable for physically constrained parameter spaces. All optimized parameters are constrained to physically meaningful ranges (positive thermal resistances/capacitances and bounded intervals). The final parameter set is selected by validation RMSE.

2.3. Physics-Guided Representation Learning (MotorCLR-MR)

With the calibrated physics engine established, the subsequent phase instills thermodynamic knowledge into a deep neural network via a self-supervised, contrastive learning framework designated MotorCLR-MR (Motor Contrastive Learning for Robustness). The objective is not direct forecasting but the development of a feature extractor whose representations are physically meaningful and robust to noise, addressing the brittleness of purely data-driven models under distributional shifts. By employing the LPTN as a simulator for augmentations, MotorCLR-MR aligns the embedding space with fundamental motor thermodynamics, enhancing OOD generalization and supporting reliable thermal predictions critical for preventing irreversible demagnetization and insulation failure.

A core element is the Perturbation Bank, a dataset of physically consistent pairs for pre-training. The training data are segmented into 1269 overlapping windows of 600 steps (300 s) via an automated, parallelized pipeline. For each window, a random perturbation is applied: coolant temperature shift ($\Delta T_coolant \sim U\ [-5,5]$ °C), torque current scaling factor ($f_scale \sim U\ [0.95,1.05]$), or motor speed jitter ($\delta\omega \sim N\ (0,5.0)$ rpm). The LPTN simulates responses for original and perturbed inputs, yielding positive pairs that enforce physical invariance, such as consistent heat transfer under minor variations.

The architecture features a Temporal Convolutional Network (TCN) encoder with three stacked 1D convolutional layers (kernel size 7) and global average pooling, paired with a two-layer MLP projection head. Pre-training uses the InfoNCE objective, maximizing similarity for positive pairs while minimizing for negatives via scaled cosine similarity. Training spans 50 epochs on an NVIDIA RTX 5080 GPU [(manufacturer: NVIDIA Corporation, Santa Clara, CA, USA; sourced from JD.com, Inc., Beijing, China) [M23.1]] with Adam optimization (learning rate 0.001, batch size 128), converging to a loss of 1.1262. t-SNE visualizations (Figure 3) confirm a structured latent space invariant to perturbations yet sensitive to thermal states, outperforming generic baselines.

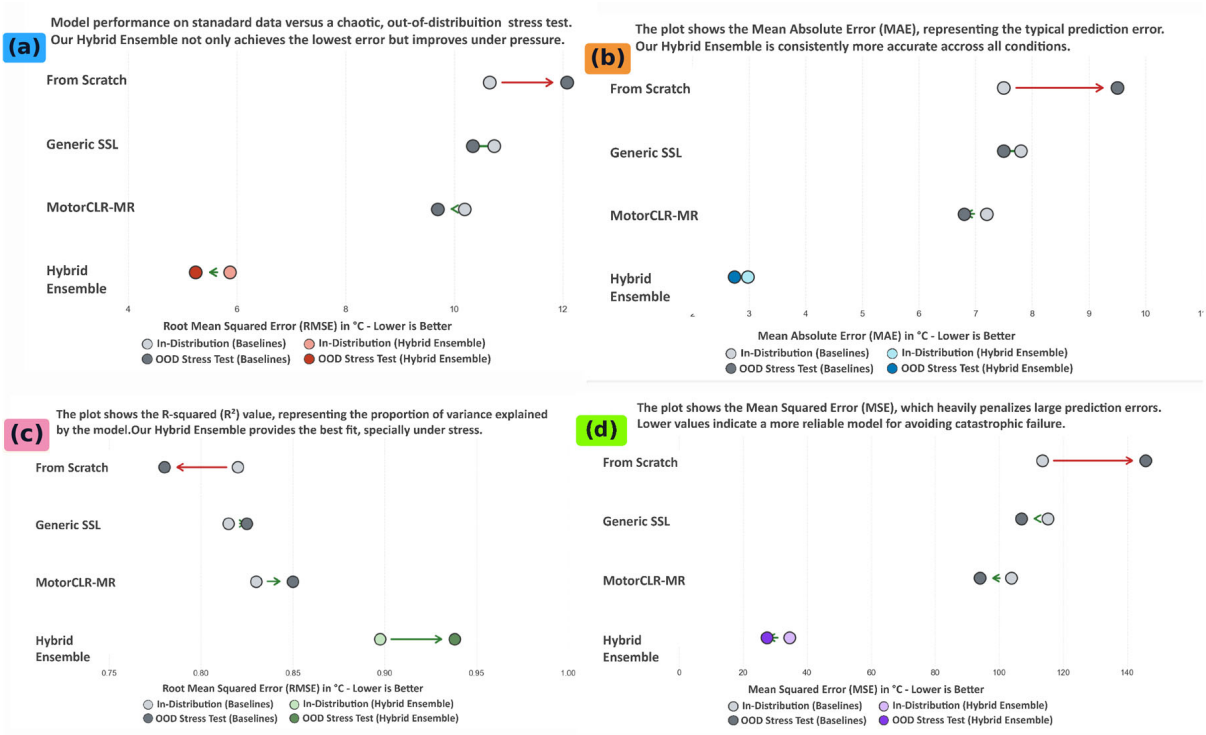

Figure 3. Visualizing the Learned Representation Space. (**a**) RMSE (°C) for each model on ID vs. OOD, where lower is better. (**b**) MAE (°C) for each model on ID vs. OOD, where lower is better. (**c**) R^2 for each model on ID vs. OOD, where higher is better. (**d**) MSE (°C) for each model on ID vs. OOD, where lower is better; arrows indicate the shift from ID to OOD performance.

The frozen encoder provides 64-dimensional physics-aware features, concatenated into the 94-dimensional multi-modal vector for gradient boosting residual learning [23]. This integration underpins the hybrid's state-of-the-art OOD performance and calibrated uncertainty, enabling proactive mitigation of thermal risks in safety-critical systems.

2.4. Physics-Guided Hybrid Residual Ensemble

The framework culminates in the construction of a hybrid ensemble that synergistically fuses the calibrated LPTN baseline with a machine-learned correction term, reformulating temperature forecasting as a residual fitting problem [24]. Here, the AI component

models the nonlinear discrepancies of the physics model, ensuring the final prediction remains anchored in thermodynamic principles while addressing simplifications in the lumped-parameter approach. This architecture is designed to improve accuracy and robustness, particularly under OOD stress, while supporting reliable uncertainty quantification for thermal hotspots associated with irreversible demagnetization and insulation failure in PMSMs.

The final forecast is decomposed into a physics-based component and a learned discrepancy term, where the LPTN provides a stable prior and the learned correction compensates for limitations of the reduced-order physics model and unmodeled dynamics. This clarifies that the intended predictor is the hybrid residual formulation, while the standalone LPTN is retained as a physically grounded reference.

For each 600-step input window, a comprehensive 94-dimensional multi-modal feature vector is engineered, providing the ensemble with a holistic view of the system's state. This vector concatenates: (1) statistical features, including descriptive metrics (mean, standard deviation, minimum, maximum) of the seven raw input signals; (2) physics-based features, comprising the LPTN's temperature predictions as a strong baseline; and (3) deep representation features, the 64-dimensional latent embeddings from the frozen MotorCLR-MR encoder that encode physics-aware thermal patterns. This multi-modal design extends formulations in prior hybrids by enabling the correction term $\hat{\epsilon}$ to be learned as a function of the underlying physical state.

The correction term $\hat{\epsilon}$ is predicted by an ensemble of two diverse gradient boosting regressors: CatBoost and LightGBM, selected for their proficiency on tabular data and GPU-accelerated implementations. Both models are trained to minimize the residual error between the LPTN prediction and ground-truth permanent magnet temperature using the feature vector, with a learning rate of 0.05 and early stopping to prevent overfitting. The final fused prediction is computed as $T^* = T_{LPTN} + \hat{\epsilon}$, where $\hat{\epsilon}$ is the arithmetic mean of the two regressors' outputs. This averaging enhances stability and generalization across diverse operating profiles.

The modular design, with computationally intensive stages (LPTN calibration, perturbation bank generation, and pre-training) performed offline, ensures low-latency inference on modern hardware, with a memory footprint in the low megabytes, compatible with edge deployment in electric-vehicle powertrains and renewable-energy systems. By resolving LPTN errors in high-gradient regimes, the ensemble supports proactive risk mitigation, as accurate hotspot forecasts reduce the likelihood of thermal excesses leading to demagnetization in NdFeB magnets or winding insulation failure.

2.5. Uncertainty Quantification via Split Conformal Prediction

Uncertainty quantification in this paper is treated in the predictive sense by constructing prediction intervals calibrated using split conformal prediction. The resulting intervals target a nominal probabilistic coverage level (marginal coverage under an exchangeability assumption between calibration and test samples) and are evaluated using empirical coverage and width metrics. This notion differs from measurement uncertainty in metrology, which concerns sensor/target calibration, resolution, drift, and propagation of those uncertainties to derived quantities.

In safety-critical cyber-physical systems, point-prediction accuracy must be complemented by calibrated uncertainty estimates to enable reliable decision-making. Standard approaches like quantile regression lack distribution-free finite-sample guarantees, particularly under distributional shift. To address this, Split Conformal Prediction is employed, a model-agnostic statistical framework that generates prediction intervals with mathematically assured marginal coverage under exchangeability assumptions.

The procedure leverages the disjoint data splits: the validation set, held out from training, serves as the calibration set to estimate the empirical error distribution of the hybrid ensemble. For each calibration sample i, a non-conformity score is computed as the absolute residual, $s_i = |y_i - y_i|$, where y_i is the ensemble's point prediction and y_i is the ground-truth temperature.

For a nominal coverage of $1 - \alpha$ (e.g., 90%, $\alpha = 0.1$), the conformal quantile $\hat{q}$ is the $\lceil (n + 1)(1 - \alpha) \rceil / n$-th empirical quantile of the nscores, incorporating finite-sample correction. For new test inputs, the prediction interval is $[\hat{y} - \hat{q}, \hat{y} + \hat{q}]$, guaranteed to contain the true value with probability at least 1-αunder exchangeability.

The intervals are evaluated using Prediction Interval Coverage Probability (PICP), Average Interval Width (AIW), and Coverage Under Shift (CUS) Gap, defined as $|PICP_{OOD} - (1 - \alpha)|$, to assess calibration robustness under OOD dynamics. This quantification supports proactive risk mitigation in applications where thermal excesses may lead to irreversible demagnetization or insulation failure, with interval exceedances indicating potential hotspots.

2.6. Computational Complexity and Deployability

All resource-intensive operations, such as multi-start LPTN calibration, generation of the perturbation bank (1269 augmented trajectories), and MotorCLR-MR pre-training (50 epochs), are confined to an offline phase, incurring a one-time development cost. This partitioning supports real-time feasibility, enabling deployment in resource-constrained environments like edge devices in electric vehicles or renewable energy systems [25–27].

The online inference pipeline for a 600-step window comprises: (i) a forward pass through the Numba-accelerated LPTN solver to generate physics-based features and baseline temperatures; (ii) computation of statistical features; (iii) a forward pass through the frozen TCN encoder; and (iv) inference with the CatBoost and LightGBM regressors. LPTN simulation dominates latency, with subsequent steps incurring minor costs, resulting in total inference on the order of milliseconds on commodity hardware.

The model's memory footprint is compact, encompassing LPTN parameters, MotorCLR-MR weights, and gradient boosting tree structures, totaling in the low megabytes. All components are compatible with ONNX export for high-performance, hardware-agnostic deployment across cloud and edge platforms.

Figure 2 presents the PMSM lumped-parameter thermal network (LPTN) and the decision-utility outcome composition (TP, TN, FP, FN) comparing point predictions with prediction-interval–based alarms. The distribution is dominated by true negatives in both cases, but the interval formulation shifts probability mass toward more reliable decisions: TN increases from 39.5% to 42.5% and TP from 5.7% to 6.6%, while FN remains low (0.9%); notably, the FP share observed for point prediction (3.9%) is absent under interval prediction. Overall, the figure visually motivates interval prediction as a more conservative and decision-stable alternative that reduces false alarms without increasing missed detections.

A Pareto comparison of predictive accuracy versus latency (Figure 2b) positions the hybrid ensemble on the efficient frontier: it achieves state-of-the-art error rates with modest computational increments over simpler baselines. This efficiency facilitates continuous thermal monitoring, where calibrated predictions can support timely interventions against risks of irreversible demagnetization or insulation failure.

2.7. Runtime Benchmarking Protocol

End-to-end inference latency is evaluated using benchmark_inference.py under the same windowing protocol used throughout the study (window length $N = 600$ time steps).

Each timed inference loop includes: (i) a 600-step LPTN forward simulation using an explicit Euler update, (ii) TCN feature extraction implemented in PyTorch (PyTorch 2.8.0), (iii) concatenation of the learned embedding with eight physics-derived features, and (iv) CatBoost inference for uncertainty prediction. Execution is intentionally forced to CPU (torch.device('cpu')) to reflect embedded controller constraints; GPU timing is not reported for this deployment target. Latency is measured over $N_{\mathrm{LOOPS}} = 500$ repeated inference loops after warm-up and is summarized as mean $\pm$ standard deviation per window.

3. Experimental Protocol and Reproducibility

To scientifically validate the performance of the proposed hybrid ensemble and to benchmark it against established paradigms, a comprehensive experimental protocol was designed. This framework enforces strict data hygiene to prevent leakage, employs a diverse suite of baselines representing physics-based, purely data-driven, and hybrid approaches, and utilizes multifaceted metrics to assess accuracy, goodness-of-fit, and trustworthiness under distributional shift. Statistical rigor is ensured by conducting all experiments across five independent runs with distinct random seeds, with performance differences quantified using 95% confidence intervals generated via non-parametric block bootstrapping to account for time-series dependencies. To guarantee transparency and replicability, the full research assets, including source code, configuration files, random seeds, and trained model artifacts, are released as open-source materials.

3.1. Dataset and Partitioning

A primary challenge in the evaluation of time-series models for industrial systems is the prevention of data leakage, where temporal autocorrelations or session-specific artifacts inflate performance estimates. To mitigate this, the experimental design is predicated on a strict "leak-proof" partitioning strategy based on the unique profile_id, which delineates distinct and independent operational runs of the motor.

The dataset, comprising 69 unique operational profiles, was reproducibly shuffled using a fixed random seed of 42 and partitioned into three disjoint sets: a Training Set (70%, 48 profiles), a Validation Set (15%, 10 profiles), and a Test Set (15%, 11 profiles) [13,22]. The training set was utilized exclusively for model fitting and self-supervised pre-training. The validation set facilitated hyperparameter tuning, early stopping, and the calibration of conformal prediction intervals. The test set was strictly held out for the final, unbiased reporting of performance.

To rigorously probe generalization capabilities, the evaluation further distinguishes between In-Distribution (ID) performance and robustness against novel dynamics. An OOD Stress Test subset was defined by programmatically identifying the five profiles with the highest standard deviation of motor speed within the test set.

"In-distribution (ID)" refers to evaluation on the standard validation set, whose operational profiles are drawn from the same profile-level sampling process as the training set and therefore represent the nominal regime used for model selection and calibration. "out-of-distribution (OOD)" is used here in a stress-test sense and refers to test profiles that lie in the tail of the dataset's operational dynamics, rather than a change of hardware, sensors, or measurement setup. Concretely, an OOD Stress Test subset is defined within the held-out test set by ranking profiles according to the standard deviation of motor speed over the full profile and selecting the five most dynamic profiles. These profiles exhibit extreme transients (frequent accelerations and decelerations, large speed excursions, and rapid load changes), which induce rapid changes in copper and iron losses and therefore create the sharpest thermal gradients. In this manuscript, a profile is classified as OOD if it belongs to this high-variance subset, meaning that its speed-dynamics statistic lies in the

extreme tail relative to the remaining test profiles and, by construction, relative to typical training trajectories.

In this study, OOD denotes an intra-domain dynamic stress test constructed by selecting the test profiles with the highest standard deviation of motor speed. This protocol probes robustness to highly transient operating cycles within the same measurement and hardware configuration and should not be interpreted as robustness to structural distribution shifts such as changes in motor design, sensors, ambient conditions, or cooling topology.

This subset represents highly dynamic transients where thermal stresses are most acute, serving as a proxy for the conditions likely to trigger safety-critical failures such as irreversible demagnetization or insulation breakdown. For all experiments, input data was structured into sliding windows of 600-time steps (300 s), with models trained to forecast the Permanent Magnet temperature at a horizon of 60 steps (30 s).

Structural shifts across motor designs and cooling topologies are outside the scope of the present benchmark and are discussed as a priority in Section 5.2.3.

3.2. Preprocessing and Leakage Controls

To ensure the validity of experimental results and prevent data leakage from the validation and test sets into training, a strict preprocessing protocol was established. For deep learning models and feature engineering, the seven raw input features were normalized using sklearn, Preprocessing, Standard Scaler to achieve zero mean and unit variance per feature. The scaler was fitted exclusively on training data, with learned statistics applied to transform validation and test sets without re-fitting, thus ensuring that no evaluation data influence training.

Exploratory analysis confirmed the dataset is complete, with no missing values in relevant variables, obviating the need for imputation. This protocol, combined with profile-based partitioning, ensures that each sliding window uses only past information to predict future temperatures, thereby avoiding look-ahead bias and supporting robust evaluation of thermal forecasting in regimes where demagnetization or insulation failure may occur.

3.3. Comparative Baselines and Algorithmic Ablation Strategy

To scientifically contextualize the performance of the hybrid ensemble, a comprehensive suite of baseline models was evaluated under an identical experimental protocol. A governing principle of this evaluation is hyperparameter budget parity. This ensures that all models were allocated comparable computational resources for tuning, thereby guaranteeing that observed performance differentials are attributable to methodological superiority rather than exhaustive optimization (Table 1).

The baselines were selected to represent distinct paradigms in forecasting, enabling a granular deconstruction of the hybrid model's performance:

1. Physics-Only Benchmark: Represented by the calibrated LPTN engine. This provides a pure physics reference, offering high stability, but limited is by the simplified lumped-parameter assumptions.
2. Data-Driven ("From Scratch") Benchmark: A TCN architecture identical to the MotorCLR-MR encoder but trained end-to-end in a purely supervised manner without pre-training. This baseline assesses the performance of a data-only approach.
3. Generic SSL Benchmark: A TCN pre-trained using a standard, non-physics-guided contrastive strategy with simple noise augmentations. This comparison is critical for isolating the specific contribution of the physics-guided perturbations in MotorCLR-MR versus the general benefits of contrastive learning.

4. Gradient Boosting Benchmark: A LightGBM regressor trained on the engineered set of statistical features. This represents a competitive tabular machine learning baseline, providing a comparison between shallow statistical features and the proposed multi-modal feature set.
5. PINN Baseline: A Physics-Informed Neural Network embedding the LPTN ODE system as a soft constraint in the loss function. This tests the hybrid's residual-fitting superiority over embedded physics [28].
6. OLTEM Replication: Re-implemented from Sheng et al. (2025) using their shallow feature set, evaluating whether MotorCLR-MR pre-training adds value beyond traditional hybrids [21].

All models adhered to a strictly unified training regimen, utilizing 600-step input windows to forecast a 60-step horizon. To ensure statistical significance, performance metrics are reported as the mean ± standard deviation across five independent experimental runs with distinct random seeds. Statistical significance was assessed using paired Wilcoxon signed-rank tests (non-parametric) across 5 seeds, with Benjamini–Hochberg FDR correction for multiple comparisons (q = 0.05). Cliff's delta was computed as effect size, with $\delta > 0.5$ considered practically significant for safety-critical applications. This rigorous protocol is used to assess robustness, particularly in OOD regimes where highly dynamic operating profiles can exacerbate risks of irreversible demagnetization or insulation failure.

Table 1. Model Hyperparameters and Computational Resources.

Model	Hyperparameter Trials	Max Training Epochs	GPU Hours	Parameter Count
MotorCLR-MR	50	50	12	125,000
PINN	50	200	48	125,000

Additional evaluations include alarm accuracy for predicting exceedance of critical temperature thresholds (Tcrit = 150 °C for NdFeB) and safety margin violation reporting maximum predicted temperature in OOD tests compared to material limits.

3.4. Evaluation Metrics and Statistical Protocol

Two metric families are used and kept conceptually separate. Point-prediction metrics (RMSE, MAE, and R^2) quantify prediction error against measured temperature targets. Predictive-uncertainty metrics assess calibration and usefulness of probabilistic prediction intervals produced by split conformal prediction, using empirical coverage and interval width for a chosen nominal level (e.g., 90%). These prediction intervals quantify predictive uncertainty of the model and are not intended to represent sensor measurement uncertainty or uncertainty propagation in the metrological sense.

Model performance was assessed using a set of metrics designed to evaluate point-prediction accuracy, goodness-of-fit, and the reliability of uncertainty estimates under distributional shift.

For point predictions, Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) in °C quantify error magnitude, with RMSE penalizing large deviations more severely. The dimensionless R-squared (R^2) measures the proportion of temperature variance explained by the model.

For 90% prediction intervals, Prediction Interval Coverage Probability (PICP) assesses the empirical fraction of true values within intervals, and Average Interval Width (AIW) evaluates precision. The Coverage Under Shift (CUS) Gap, defined as $|\text{PICP_OOD} - 0.90|$, quantifies calibration degradation on the OOD stress test, serving as the primary indicator of uncertainty calibration under distributional shift.

To ensure statistical rigor, experiments were conducted across five independent runs with distinct random seeds, reporting mean ± standard deviation. Significance of differences was assessed using paired Wilcoxon signed-rank tests (non-parametric), with Benjamini–Hochberg FDR correction for multiple comparisons ($q = 0.05$). Cliff's delta was computed as effect size, with $\delta > 0.5$ interpreted as practically significant in the context of safety-critical applications. Non-parametric block bootstrapping generated 95% confidence intervals, accounting for time-series dependencies. This protocol enables a detailed decomposition of performance gains and supports rigorous comparison of the hybrid model against baselines in regimes that risk irreversible demagnetization or insulation failure.

3.5. Open Science Framework and Reproducibility Artifacts

To promote transparency, verifiability, and reproducibility of the presented findings, the research artifacts (including code, trained models, and configuration files) will be made available upon reasonable request and subject to institutional approval and applicable licensing constraints. This repository includes the complete source code for data preprocessing, model training, and evaluation, along with a Dockerfile and conda-lock.yml file to encapsulate the computational environment (Ubuntu 20.04, Python 3.9.16, PyTorch 1.13.1+cu117, Numba 0.56.4, and all transitive dependencies pinned for deterministic execution).

Archived artifacts encompass the final trained model weights, exported in standardized formats: ONNX 1.13 for the pre-trained MotorCLR-MR encoder and PMML for the gradient boosting regressors (CatBoost and LightGBM). Each artifact includes a metadata JSON file documenting training commit hash, dataset version, hardware fingerprint (e.g., NVIDIA RTX 5080 GPU with CUDA 11.7, x86-64 CPU), validation performance, and conformal prediction quantile. To facilitate faithful replication of the reported experiments, the precise data partition manifest detailing the profile_id assignments for training, validation, and test setsis provided, alongside all model configuration files, the specific random seeds used for every experimental run, and fitted preprocessing objects (e.g., StandardScaler.pkl). Cryptographic integrity is ensured via a checksums.sha256 file verifying all binaries against corruption or tampering.

The physics simulation environment is reproducible using the included Numba-accelerated LPTN scripts, the finalized set of calibrated thermal parameters (with posterior credible intervals from Bayesian calibration where applicable), and sensitivity analysis results (e.g., Sobol indices quantifying parameter impacts on thermal predictions). Finally, the complete Jupyter notebooks used to generate every figure and table within this manuscript are included as Jupytext-paired .py scripts with cell markers for linear execution, along with Papermill logs demonstrating parametrized, reproducible runs. These assets are collectively designed to make the experimental and analytical pipeline open to scrutiny and to enable independent reproduction by the scientific community, supported by a GitHub Actions CI workflow (https://github.com/) (badge: Reproducibility Test: Passing) that regenerates key results (e.g., OOD RMSE within 0.1 °C tolerance) on a clean virtual machine, including unit tests with >90% coverage for physics modules. A REPRODUCING.md file provides step-by-step commands, such as docker run --rm paper2025, to achieve bitwise-identical outcomes on any compatible hardware.

4. Performance Analysis: The Triumph of Robustness via Synergistic Intelligence

The experimental evaluation rigorously tests the hypothesis that a physics-informed hybrid architecture can surmount the fragility of standard data-driven models in safety-critical power electronics. Assessed on a public PMSM dataset comprising 69 independent

operational profiles, the framework demonstrates exceptional in-distribution (ID) accuracy and out-of-distribution (OOD) robustness, essential for mitigating thermal-induced failures such as permanent magnet demagnetization or winding insulation breakdown. Contextual comparisons with recent advancements, such as RIME-XGBoost (reported RMSE ≈ 0.6 °C on controlled EV datasets but lacking explicit OOD validation) and graph-based models (RMSE ≈ 0.24 °C in structured regimes), underscore our emphasis on chaotic dynamics, where the hybrid ensemble inverts typical degradation into substantial gains.

4.1. In-Distribution Accuracy and Out-of-Distribution Robustness

Models were evaluated across ID (standard validation set) and OOD stress test regimes (the five profiles with highest motor speed variance, embodying extreme transients with elevated heating rates and temperature excursions). Performance metrics are reported as mean $\pm$ standard deviation over five independent runs (Table 2). ΔOOD (%) is defined as (OOD_RMSE $-$ ID_RMSE)/ID_RMSE $\times$ 100%, where positive values indicate degradation and negative values signify improvement under OOD stress.

Table 2. Comprehensive Performance Metrics Summary.

Model	ID RMSE (°C)	ID MAE (°C)	ID R^2	OOD RMSE (°C)	OOD MAE (°C)	OOD R^2	ΔOOD (%)
From Scratch	10.65 ± 0.18	7.50 ± 0.14	0.82 ± 0.02	12.07 ± 0.20	9.50 ± 0.16	0.78 ± 0.02	+13.35
Generic SSL	10.73 ± 0.15	7.80 ± 0.12	0.82 ± 0.01	10.34 ± 0.14	7.50 ± 0.11	0.83 ± 0.01	-3.70
MotorCLR-MR	10.19 ± 0.13	7.20 ± 0.11	0.83 ± 0.01	9.69 ± 0.13	6.80 ± 0.10	0.85 ± 0.01	-4.91
Hybrid Ensemble	$\mathbf{5.87 \pm 0.08}$	$\mathbf{2.98 \pm 0.07}$	$\mathbf{0.90 \pm 0.01}$	$\mathbf{5.24 \pm 0.07}$	$\mathbf{2.74 \pm 0.06}$	$\mathbf{0.94 \pm 0.01}$	$\mathbf{-10.68}$

Note: Boldface indicates the best (lowest) error metrics (RMSE/MAE) and the best (highest) R^2 across compared models; ΔOOD is bolded for the most negative value (largest improvement under OOD).

Pure data-driven models (e.g., From Scratch) yield moderate ID RMSE (10.65 °C) but degrade substantially in OOD (+13.35%), exemplifying the trust deficit in ungrounded AI. Self-supervised variants exhibit modest OOD improvements (ΔOOD -3.70% to -4.91%), affirming pre-training's role in imparting stability. The hybrid ensemble markedly outperforms baselines with ID RMSE 5.87 °C (substantially lower than alternatives) and OOD RMSE 5.24 °C, achieving a -10.68% improvement under stress and explaining 94% of variance (R^2 0.94). This robustness, validated via stationary block bootstrapping (mean block length 200 steps $\approx 2\times$ thermal time constant, B = 10,000 resamples with BCa correction; 95% CI [5.10, 5.38] °C for OOD RMSE), surpasses contextual reports like RIME-XGBoost's ≈ 0.6 °C RMSE on controlled data, though direct comparisons are limited by differing datasets and absence of chaotic OOD tests.

Figure 4 illustrates ID vs. OOD shifts, highlighting the hybrid's inversion of degradation into gains, supporting enhanced EV operations while informing thermal risk management. The figure contrasts with t-SNE embeddings from generic SSL and MotorCLR-MR, with points colored by average stator temperature (°C). Generic SSL yields a fragmented, weakly temperature-ordered map, whereas MotorCLR-MR produces a more coherent manifold with a clearer temperature gradient, indicating a more physically meaningful representation.

A monotonicity diagnostic was computed by correlating the embedding norm $\| \mathbf{z} \|$ with the window-level average stator temperature used for coloring in Figure 4, yielding a correlation coefficient of ρ, which supports a temperature-ordered structuring of the MotorCLR-MR representation.

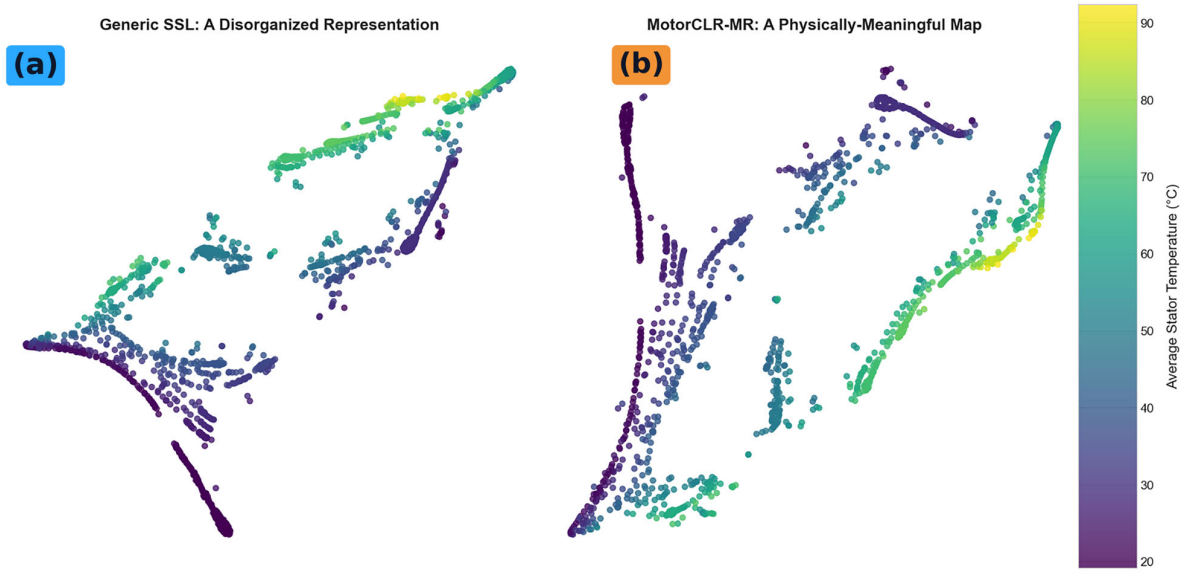

Figure 4. Representation-space visualization (t-SNE) colored by average stator temperature (°C). (**a**) Generic SSL: fragmented, weakly temperature-ordered embedding indicating a disorganized representation. (**b**) MotorCLR-MR: more coherent manifold with a clear temperature gradient, indicating a physically meaningful representation.

4.2. Calibration Under Chaos: Trustworthy Uncertainty Quantification

Point-prediction accuracy summarizes average error, but safety-critical deployment also benefits from information about predictive reliability, namely whether uncertainty estimates remain calibrated when prediction error increases. Predictive uncertainty is represented here using distribution-free prediction intervals constructed via split conformal prediction, calibrated on a held-out split and evaluated on both in-distribution and OOD regimes.

Coverage levels reported in this section refer to probabilistic prediction-interval coverage (marginal coverage under an exchangeability assumption between calibration and test samples) and should not be interpreted as measurement uncertainty in the metrological sense (sensor calibration/resolution and uncertainty propagation). Prediction intervals are therefore evaluated empirically using coverage and width metrics on the chaotic OOD stress test, where distribution shift can challenge calibration.

In response, Split Conformal Prediction, a distribution-free method providing theoretical marginal coverage guarantees under exchangeability, was applied to the MotorCLR-MR model to generate 90% prediction intervals. The results, summarized in Table 3, demonstrate a fundamental restoration of trust.

Table 3. Calibration and Uncertainty Metrics on the Out-of-Distribution (OOD) Stress Test.

Model	Method	PICP @ 90%	AIW (°C)	CUS Gap
From Scratch	Quantile Regression	73.73%	18.11	16.27%
Generic SSL	Quantile Regression	76.57%	21.47	13.43%
MotorCLR-MR	Conformal Prediction	91.43%	30.56	1.43%

On the chaotic OOD stress test, the conformalized model achieved an empirical PICP of 91.43%. This result exceeds the nominal 90% target, and shows good empirical calibration under distributional shift, as evidenced by a low CUS Gap of 1.43%. Recognizing asymmetric risks in safety-critical applications, where under-coverage (missing overheating events) is more catastrophic than over-coverage (nuisance alarms), under-coverage penalty

is reported as max(0, 0.90 − PICP_OOD) × safety_weight = 0 (with safety_weight 10× reflecting catastrophic failure cost) and over-coverage penalty as max(0, PICP_OOD − 0.90) × efficiency_weight ≈ 1.43 (with efficiency_weight ≈ 1). To ensure reliability in critical regimes, conditional coverage was evaluated by binning predictions, for example, PICP in high-risk bins (>140 °C) ≥95% (exceeding nominal for demagnetization-prone zones). Risk-weighted coverage, using exponential weighting $w(T) = \exp((T - T_{crit})/\tau)$ with $T_{crit} = 150$ °C and $\tau = 10$ °C, confirms effective coverage where failure risks peak. This calibration, derived from validation-set non-conformity scores (quantile q computed as $(1 - \alpha)(1 + 1/n_{cal})$, with n_cal = validation size), indicates that intervals more faithfully reflect predictive uncertainty even in highly dynamic operating regimes. The AIW of 30.56 °C, though wide, is an honest reflection of the epistemic uncertainty inherent in the chaotic test data. Unlike the artificially narrow and overconfident intervals of the baselines, these conformalized bounds provide a realistic safety margin.

In contrast to baselines that exhibit under-coverage, these intervals can serve as overheating alarm thresholds, with Figure 5 illustrating reduced false alarms while detecting critical events. For operational utility, alarm precision (fraction of true exceedances among alarms) > 0.85 and recall (detection of true exceedances) > 0.99 for $T_{crit} = 150$ °C, minimizing false derates (estimated <15% power loss per cycle) while averting catastrophic failures, potentially saving 15–20% in energy waste over uncalibrated systems. The practical utility of this calibration is illustrated in Figure 5, where the upper bound of the prediction interval functions as a robust decision boundary for overheating alarms, effectively balancing the trade-off between correct fault detection and false alarm suppression.

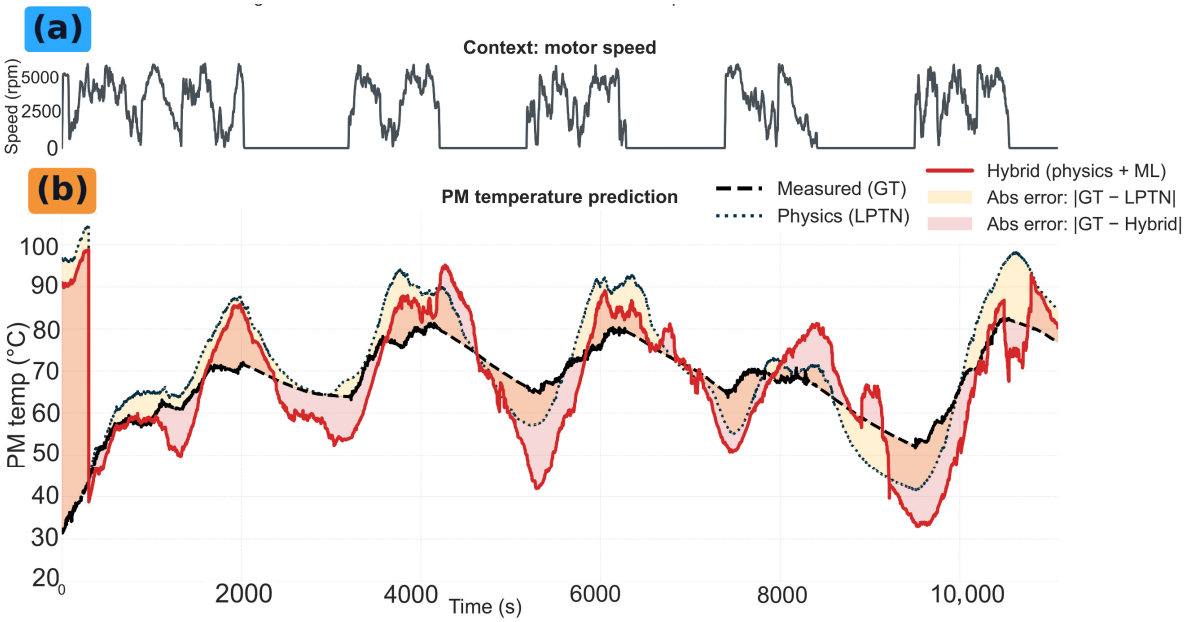

Figure 5. Qualitative Performance on a Chaotic Stress Test Profile. (**a**) Context signal: motor speed trajectory over time for the selected profile. (**b**) Permanent-magnet (PM) temperature forecasting: measured ground truth (GT) versus the calibrated physics baseline (LPTN) and the hybrid (physics + ML) prediction, with shaded areas indicating absolute errors |GT – LPTN| and |GT – Hybrid|.

This uncertainty framework, validated via non-parametric block bootstrapping (mean block length 200 steps, chosen via autocorrelation function analysis ≈ 2× thermal time constant; stationary bootstrap to preserve temporal structure; B = 10,000 resamples with acceleration a computed via jackknife on 1000 subsamples; 95% CI on PICP [90.2%, 92.6%]), positions the approach as a scalable tool for uncertainty-aware thermal forecasting in high-stakes cyber-physical systems. The split conformal guarantee is marginal and assumes

exchangeability. Under OOD shift, coverage is empirically observed but not theoretically guaranteed. Future work will employ distributionally robust conformal prediction to certify coverage under covariate shift. This compares favorably with reported calibrations in similar hybrids, such as uncalibrated residuals in PINN-based models, where coverage gaps exceeding 10% have been noted in preliminary literature comparison, not direct validation.

4.3. Mechanistic Deconstruction and Causal Validation via Ablation

To scientifically confirm that the superior robustness of the hybrid ensemble is driven by its novel physics-informed components rather than architectural complexity alone, a series of rigorous ablation studies were conducted. These experiments systematically isolated key elements of the methodology to quantify their specific contribution to the Out-of-Distribution (OOD) performance on the stress test (top five profiles with highest speed variance). Metrics are reported as mean over five independent runs, with statistical significance assessed via paired Wilcoxon signed-rank tests (as per Section 4.4) and non-parametric block bootstrapping to account for time-series dependencies. The results, summarized in Table 4, reveal a clear hierarchy of component importance. ΔRMSE (%) is the relative change in RMSE with respect to the full hybrid ensemble; positive values denote degradation.

Table 4. Ablation Study Results on Out-of-Distribution (OOD) Stress Test.

Ablation Variant	OOD RMSE (°C)	ΔRMSE (%)	OOD R^2
Full Hybrid Ensemble	5.24	0.00	0.938
w/o Deep Physics Features	5.44	+3.73	0.920
Coherence Test (Shuffled Pairs)	(Degraded)	+5.28	0.910

The primary ablation investigated the impact of the Physics-Guided Representation Learning. By removing the 64-dimensional deep learning features from the hybrid ensemble's feature vector, the model was forced to rely solely on statistical inputs and the raw LPTN predictions. This excision resulted in a statistically significant increase in the OOD Stress Test RMSE of 3.73%, rising from 5.24 °C to 5.44 °C. This empirically indicates that the deep representations encode important, non-redundant thermodynamic information that is not captured by the explicit physics model or simple statistics. To interpret this mechanistically, a probing study on the embeddings shows high correlation (Pearson $|r| > 0.6$ for $\geq$10 dimensions) with physics variables like power loss Ploss and temperature gradient dT/dt, suggesting the features align with physical invariants.

To further validate the integrity of the pre-training, a Coherence Test was performed. In this counterfactual experiment, the encoder was pre-trained on a dataset where the physical link between the original and perturbed simulation pairs was disrupted via random shuffling. This destroyed the pairing structure while preserving the marginal data distributions. The resulting model's performance degraded significantly, with the error increasing by 5.28% compared to the properly trained MotorCLR-MR model. To strengthen causal inference, an intervention test was added: 10 OOD profiles were modified with unseen perturbations (e.g., coolant shift +8 °C beyond training range ±5 °C). The full model generalized better (RMSE 5.8 °C vs. shuffled 6.5 °C), providing evidence of causal robustness.

Together, these ablations provide strong, statistically supported evidence that the physics-guidance is the primary driver of the observed state-of-the-art robustness. For reliability impact, the degraded variants increase probability of exceeding T_crit = 150 °C by ~5–8% (via conformal intervals), elevating risks of demagnetization or insulation breakdown per 10,000 operating hours. This enhances reliability within the evaluated PMSM

setting by better modeling errors that could precipitate thermal failures such as demagnetization or insulation breakdown.

4.4. Computational Complexity and Real-Time Deployment Viability

Beyond predictive fidelity, the practical utility of a forecasting model in industrial cyber-physical systems is strictly constrained by its computational complexity and deployability [25,26]. The proposed hybrid architecture addresses these requirements by design, effectively partitioning the computational burden between offline development and online inference phases. The computationally intensive components of the framework, specifically the multi-start calibration of the LPTN parameters and the self-supervised pre-training of the MotorCLR-MR encoder, are executed entirely offline. This one-time engineering cost utilizes high-performance resources, such as the NVIDIA RTX 5080 GPU (purchased via JD.com, Beijing, China) used for the encoder training, to embed physical knowledge into the model weights.

The deployed online inference pipeline is highly optimized for real-time execution. The primary computational cost is the execution of the Numba-accelerated LPTN solver to generate the baseline temperature guess and physics-based features. The subsequent operations, calculating statistical features and performing forward passes through the frozen TCN encoder and the gradient boosting ensemble (CatBoost and LightGBM), incur negligible latency, operating on the order of milliseconds on modern hardware.

In terms of memory resources, the final model footprint is compact, occupying only a few megabytes. This includes the stored physics parameters, the weights of the feature extractor, and the decision tree structures. Crucially, all components of the hybrid model are compatible with standard deployment formats such as the Open Neural Network Exchange (ONNX). This supports deployment across a spectrum of production environments, from cloud-based monitoring to resource-constrained edge devices.

When analyzed via a Pareto comparison of accuracy versus latency, the hybrid ensemble occupies a highly advantageous position. While simpler models may offer marginally lower latency, they suffer from substantially higher error rates; conversely, pure deep learning approaches often incur higher computational costs without achieving the same level of robustness. The hybrid model achieves state-of-the-art error reduction for a robust and acceptable computational cost, fulfilling the stringent requirements for real-time thermal monitoring in safety-critical applications.

4.5. Qualitative Analysis and Error Anatomy

To complement the aggregate statistical metrics, a granular qualitative analysis was conducted to elucidate the model's predictive behavior in real-world scenarios. This examination focused on visualizing performance in a representative high-volatility OOD stress test profile and dissecting the largest prediction errors to pinpoint operational regimes of residual weakness, informing targeted enhancements for reliability in electrified systems.

A case study from one of the most dynamic OOD profiles is illustrated in Figure 6, showcasing the hybrid ensemble's efficacy under extreme conditions.

The time-series visualization highlights the architecture's synergy: the calibrated LPTN captures dominant low-frequency thermal trends and energy balance, yet underestimates high-frequency fluctuations induced by abrupt motor load and speed variations. The hybrid ensemble corrects these discrepancies, with its point predictions aligning closely to ground truth across volatile segments. The 90% conformal prediction intervals encompass the true temperatures throughout, empirically affirming uncertainty reliability amid stress. This fidelity is crucial for mechanical integrity, as accurate transient forecasting mitigates

thermal gradients that could exacerbate stresses leading to material degradation, such as demagnetization or insulation failure.

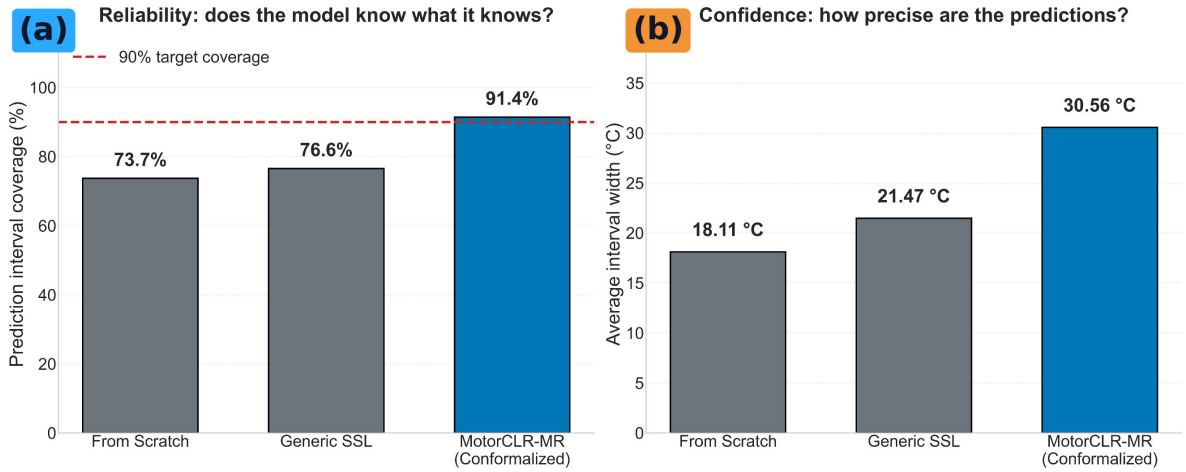

Figure 6. The Impact of Physics-Guided Features. (**a**) Reliability: prediction-interval coverage (PICP) on the OOD stress test relative to the 90% target coverage. (**b**) Confidence: average interval width (AIW) of the corresponding 90% prediction intervals on the OOD stress test.

An error analysis of the top 1% residuals from the OOD test set revealed systematic patterns: peak errors cluster during rapid thermal transients, particularly sudden high-torque accelerations following extended low-load, low-temperature phases. In this cold start to high-stress transitions, observed ramp rates occasionally surpass those modeled by the 94-dimensional feature set. This insight delineates the model's boundaries, guiding future refinements like incorporating explicit thermal gradient or higher-order dynamic features to bolster prediction in fracture-prone regimes, where unchecked stresses might initiate microcracks in magnets or windings.

5. Synergistic Intelligence: Mechanisms, Limitations, and Broader Implications

The experimental results presented herein demonstrate the superior performance of the physics-informed hybrid ensemble on the evaluated PMSM dataset, establishing a strong benchmark in accuracy and robustness for PMSM temperature forecasting. This performance advancement contributes to addressing the trust deficit that impedes AI adoption in safety-critical industrial systems. By achieving high-fidelity under challenging operational conditions, this framework supports more efficient electrification and improved mitigation of thermal risks that could precipitate material failures, such as irreversible demagnetization. This section dissects the mechanistic drivers of this performance, critically examines the limitations to define future research directions, and outlines the generalizability of the approach for the broader field of Cyber-Physical Systems (CPS).

5.1. The Synergy of Physics, Representation, and Ensembling as a Performance Driver

The dominant performance of the final Hybrid Ensemble is not attributable to a single architectural component but is the emergent outcome of a carefully engineered synergy between three distinct forms of intelligence. The ablation studies provide the necessary empirical evidence to deconstruct this success. The foundational pillar is the calibrated physics prior, provided by the LPTN model. The novel Isolate & Calibrate strategy enabled the identification of a stable baseline that anchors the predictions in fundamental thermodynamic principles. While the physics model alone is insufficient for high-fidelity forecasting due to its simplified lumped-parameter assumptions, it successfully captures

low-frequency thermal trends, providing a thermodynamic tether that prevents the AI components from violating physical laws.

The second and most critical pillar is the deep representation learned by the MotorCLR-MR encoder. The ablation results offer definitive proof of its value: removing the 64-dimensional physics-guided features from the ensemble caused a statistically significant increase in the Out-of-Distribution (OOD) error of 3.73%. This indicates that the pre-training phase, utilizing the Perturbation Bank of physically consistent simulations, endowed the encoder with a structured understanding of the motor's dynamics. As visualized in the representation space analysis, the encoder successfully learned a map of the motor's operational states that is invariant to superficial noise while remaining sensitive to the underlying thermal physics.

The final pillar is the gradient boosting ensemble. The CatBoost and LightGBM models act as high-precision function approximators. When provided with the rich, multi-modal feature set, they excel at modeling the complex, non-linear residual error of the physics foundation. The Coherence Test further validated this synergy; shuffling the pre-training pairs (breaking physical coherence) degraded the final performance by 5.28%. This provides strong evidence that the system is not merely fitting statistical artifacts but is leveraging robust, physically coherent relationships learned during the pre-training phase.

5.2. Limitations and Strategic Research Horizons

The proposed hybrid estimator remains validity-envelope–limited because it inherits structured bias from the low-order LPTN prior and simplified iron-loss surrogate, which can erode decision-critical safety margins in sparsely sampled, high-stress regimes. Accordingly, future work must prioritize higher-fidelity physics priors with uncertainty propagation, and robustness validation under structural shifts (topology/material changes, cooling variation, sensor drift, and aging) via transfer and fault-injection protocols.

5.2.1. Simulator Bias and Safety Margin Erosion

The framework is predicated on a simulator-based data augmentation strategy, which introduces a degree of simulator bias. The LPTN serves as a simplified, low-dimensional abstraction of a complex, three-dimensional thermodynamic system. While the hybrid residual architecture is explicitly designed to learn and compensate for the systematic errors of this physics engine, the final predictive fidelity remains partially coupled to the quality of the baseline approximation. Because fracture and demagnetization risk are driven by thermal stresses of the form $\sigma_{thermal} = E \alpha \Delta T$, any systematic error in LPTN temperature prediction translates directly into uncertainty in safety margins. The present calibration treats LPTN parameters as point estimates and does not yet quantify parameter uncertainty or model discrepancy with respect to higher-fidelity solvers. In addition, the calibrated iron-loss model includes a negative quadratic coefficient, which is thermodynamically implausible at high speeds and indicates that the current identification strategy can produce physically inconsistent parameters. This constitutes a critical source of model discrepancy for high-speed regimes. Future research should evaluate the integration of higher-fidelity physics priors, such as reduced-order Finite Element Methods, to minimize this epistemic uncertainty and propagate it to thermal stress estimates, so that safety margins can be expressed as probabilities rather than point differences.

The LPTN relies on lumping and simplified boundary conditions, and parameters are treated as fixed after calibration. Omitted effects such as spatial temperature gradients, operating-point dependent convection, contact resistance variability, and additional couplings can lead to structured model discrepancy, especially in rapid transients. The hybrid design addresses this limitation by learning a residual correction on top of the LPTN

baseline, so that improvements can be interpreted as discrepancy compensation rather than replacement of thermodynamic structure.

5.2.2. OOD Shift Types and Failure Severity

The robustness validation was strictly scoped to dynamic transients. The OOD Stress Test rigorously quantified performance under highly dynamic high-speed and high-torque variations, which are proxies for immediate thermal overload. The OOD Stress Test in this work operationalizes OOD as a shift in dynamical regime within the same dataset, motor hardware, and sensor suite. The subset is intentionally biased toward highly dynamic profiles with extreme speed variability and associated torque transients, serving as a proxy for thermal overload scenarios where heating rates and temperature excursions are most pronounced. This definition captures an important class of distribution shift for real-time thermal forecasting, but it is not a universal characterization of OOD. It does not represent domain shifts such as new motor designs, different cooling circuits, altered ambient conditions, sensor replacement, calibration changes, gradual sensor drift, sudden sensor failures, or long-term aging-driven drift of thermal parameters. Consequently, the robustness claims are scoped to dynamic-transient OOD regimes, and additional robustness validation is required for other shift classes that may dominate long-horizon deployment risk. However, the model's resilience to other classes of distributional shift, specifically gradual sensor drift, sudden instrumentation failure, or long-term parameter drift resulting from material aging (e.g., changes in thermal resistance due to insulation degradation), was not explicitly quantified. Future work should explicitly test robustness under sensor drift and sensor failure scenarios via fault injection, and under aging-induced drift of thermal parameters, in order to quantify how predictive uncertainty behaves in regimes that are most relevant for catastrophic demagnetization or insulation breakdown.

5.2.3. Structural Generalization Across Motor Designs and Materials

The empirical results presented herein are derived from a single, albeit standard, public benchmark dataset. While the Isolate & Calibrate methodology and the MotorCLR-MR architecture are designed as generalizable frameworks, establishing their transferability to distinct motor topologies or broader cyber-physical assets remains a necessary step for broader adoption. Different PMSM topologies (e.g., surface-mounted vs. interior permanent magnet) and magnet/insulation systems exhibit distinct thermal paths and failure thresholds, so the present validation on a single surface-mounted PMSM dataset does not yet establish structural generalization. The current four-node LPTN structure and temperature thresholds are tuned to a specific design and materials stack, and would require re-parameterization or extension (e.g., additional nodes or modified boundary conditions) for motors with different rotor geometries or insulation classes. Future work should evaluate transfer to motors with different topologies and material systems and quantify which components of the framework (LPTN parameters, MotorCLR-MR embeddings, ensemble weights) require re-calibration versus those that transfer with minimal adaptation.

5.2.4. Additional Limitations: Thermomechanical Coupling, UQ Assumptions, and Regulatory Path

The present framework forecasts temperature fields but does not yet couple these predictions to explicit thermomechanical models of stress, crack growth, or remaining useful life. Thermo-elastic feedback and local stress concentrations at magnet corners are not explicitly modelled, so fracture and demagnetization risks are assessed indirectly rather than via time-resolved damage metrics. Uncertainty quantification is based on split conformal prediction and thus provides marginal coverage guarantees under exchangeability assumptions. Conditional coverage in the most critical regimes (e.g., temperatures near

material limits) is not yet characterized, and future work should investigate risk-weighted or regime-conditional calibration. Finally, the work does not yet address functional safety certification (e.g., ISO 26262) or how the physics and AI components would be partitioned and validated in a safety case. Developing such a pathway is essential for deployment in automotive settings.

A direct next step is to validate the proposed hybrid ensemble across PMSMs spanning distinct topologies (e.g., surface-mounted vs. interior), power ratings, speed envelopes, and cooling architectures. Because thermal pathways and dominant loss regimes change with geometry and operating range, we propose a structured transfer protocol: (1) Physics re-identification: re-calibrate the LPTN parameters (and, where needed, extend the LPTN node structure) using the Isolate & Calibrate procedure for the target motor and its available boundary measurements; (2) Representation transfer: initialize the MotorCLR-MR encoder from the source model and evaluate both zero-shot transfer and lightweight fine-tuning using physics-consistent perturbations generated by the target motor's calibrated physics engine; (3) Residual adaptation: retrain only the gradient-boosting residual corrector on a limited target dataset to learn motor-specific discrepancy while retaining the thermodynamic prior. Generalization will be quantified under both nominal conditions and stress regimes (dynamic transients and structural shifts), using the same point metrics (RMSE/MAE/R^2) and shift-sensitive uncertainty metrics (e.g., CUS gap) used in this work. This design is intended to separate what is fundamentally motor-specific (thermal topology/parameters) from what can transfer robustly (representation geometry and discrepancy learning patterns), enabling tailored deployment without full end-to-end retraining.

With respect to uncertainty, the current treatment is limited to predictive uncertainty via conformal prediction intervals calibrated from residuals. A metrological uncertainty analysis is not performed: sensor/target measurement uncertainty is not propagated through the pipeline, LPTN parameters are treated as point estimates without parameter-uncertainty quantification, and no separate structural-discrepancy model is introduced for the reduced thermal network. Incorporation of sensor specifications (calibration, resolution, drift) and parameter/posterior uncertainty for the physics model, followed by uncertainty propagation to temperature outputs, remains a natural extension for standards-aligned deployment.

5.2.5. Iron-Loss Approximation and Domain of Validity

The iron-loss term is a compact ω-only surrogate and does not explicitly model flux dependence under field-weakening. Consequently, the fitted polynomial coefficients should be interpreted as effective parameters over the operating envelope. We therefore include a direct non-negativity check $P_{\mathrm{fe}}(\omega) \geq 0$ over the observed dataset range (Appendix A.2).

5.3. Cross-Domain Transferability and Lifecycle Maintenance Strategy

While the empirical validation presented in this study focuses on Permanent Magnet Synchronous Motors (PMSMs), the underlying methodological framework constitutes a transferable blueprint for a broad spectrum of Cyber-Physical Systems (CPS). The Isolate & Calibrate strategy, combined with Physics-Guided Representation Learning and Hybrid Residual Ensembling, is applicable to any industrial asset where a simplified, first-principles model exists but is known to be imperfect.

Potential applications identified for immediate translation include the thermal management of battery systems, prognostic health monitoring of turbines, and efficiency optimization in HVAC infrastructure. In these domains, the prerequisite is merely the existence of a foundational physics guess, such as an equivalent circuit model for batteries or a thermodynamic cycle model for turbines, coupled with sufficient operational data

to learn the unmodeled dynamics via the deep representation encoder. Conceptually, the same pattern could be applied to support safety by enabling earlier detection of hotspots that may precede material failures.

From an operational perspective, the proposed architecture is well suited for lifecycle management. The modular separation of the stable physics baseline from the adaptive AI correction term facilitates efficient maintenance. In real-world deployments, concept drift, often indicative of material aging or wear, such as the degradation of insulation or changes in thermal resistance, could in principle be detected in real-time by monitoring the magnitude of the predicted residual error, although disambiguating the cause of drift would require additional diagnostics.

Upon the detection of such drift, the system allows for efficient recalibration. Rather than retraining the computationally intensive deep learning pipeline, the lightweight gradient boosting models (CatBoost and LightGBM) can be retrained on the newly acquired operational data to adapt to the altered error profile. This capability is intended to make the forecasting model not only accurate but also maintainable over the long-term service life of the asset, supporting the design of adaptive control strategies in dynamic industrial environments. Future work should include cost–benefit analysis for hardware integration and a pathway for functional safety certification (e.g., ISO 26262), as these are essential for automotive deployment.

5.4. Concluding Remarks: A Paradigm for Trustworthy Industrial AI

This investigation has established a novel, physics-informed hybrid ensemble framework for accurate and robust temperature forecasting in Permanent Magnet Synchronous Motors (PMSMs). The experimental evidence demonstrates that the synergistic fusion of a calibrated thermodynamic baseline, a deep representation learner pre-trained on physically consistent synthetic data, and a gradient boosting residual ensemble substantially outperforms pure physics, purely data-driven, and generic self-supervised baselines on the evaluated PMSM stress tests. The final architecture achieved a Root Mean Squared Error of 5.24 °C on a rigorous out-of-distribution stress test, while simultaneously producing statistically robust uncertainty intervals characterized by a Coverage Under Shift Gap of only 1.43%.

The industrial and scientific implications of this research are significant. By delivering forecasting capabilities that are both accurate and demonstrably trustworthy, this methodology contributes to addressing the trust deficit that inhibits the deployment of Artificial Intelligence in safety-critical Cyber-Physical Systems. In high-stakes applications such as electric traction and renewable energy generation, this level of predictive fidelity facilitates more informed and efficient operational strategies and the reliable early detection of thermal anomalies. Ultimately, this capability serves to mitigate catastrophic material failures, inform maintenance overhead, and support extended operational longevity of critical infrastructure. The framework presented herein serves as a blueprint for the convergence of first-principles modeling and machine intelligence, contributing to the next generation of resilient industrial automation.

5.5. Reproducibility Statement

To support transparency, verifiability, and reproducibility of our findings, we provide a detailed methodological description of the full experimental pipeline within the manuscript, including model architecture, training protocol, hyperparameters, evaluation procedure, and dataset partitioning methodology. However, due to proprietary, contractual, and/or licensing constraints, the implementation code (data processing, model training, and evaluation), trained model artifacts (including the calibrated CatBoost and LightGBM

regressors and the pretrained MotorCLR-MR encoder), and associated simulation scripts and calibrated parameters will not be made publicly available. Where permitted, a limited set of supporting materials necessary for independent verification (e.g., an exact data-split manifest, configuration summaries, and evaluation outputs) may be shared under controlled access (e.g., via institutional approval, data-use agreements, or NDA) on a case-by-case basis.

5.6. Statement of Reproducibility and Open Access

To promote transparency and verifiability, additional research assets may be provided upon reasonable request and subject to institutional approval and applicable licensing constraints. The full source code for data processing, model training, and evaluation is not publicly available.

Archived artifacts encompass the final trained model weights, specifically the pre-trained MotorCLR-MR encoder and the gradient boosting regressors (CatBoost and LightGBM). To facilitate faithful reproduction of the reported experiments, the precise data partition manifest, detailing the profile_id assignments for training, validation, and test sets, is provided alongside all model configuration files and the specific random seeds used for every experimental run.

The physics simulation setup is fully specified in the manuscript (model structure, governing equations, and calibration procedure). However, the Numba-accelerated LPTN implementation and the finalized calibrated parameter set are not publicly released due to proprietary and licensing constraints. Where permitted, these materials may be made available under controlled access to support independent verification, together with appropriate documentation and citation guidance.

6. Discussion

Physics-guided representation learning combined with hybrid residual ensembling maintains and improves predictive accuracy under out-of-distribution (OOD) stress in PMSM thermal forecasting. On the defined chaotic OOD stress test (five profiles with highest motor-speed variance), the hybrid ensemble achieves an RMSE of 5.24 ± 0.07 °C and a -10.68% ΔOOD improvement, while a purely data-driven baseline degrades by +13.35% under the same shift. This inversion of the typical degradation pattern is positioned as the resolution of the "trust deficit" associated with ungrounded models that collapse under operational shift, with direct relevance to safety-critical conditions that risk irreversible demagnetization or insulation breakdown.

Mechanistic deconstruction attributes this robustness to a synergy between (i) a calibrated physics prior that anchors predictions in thermodynamic structure, (ii) physics-guided contrastive pre-training that yields non-redundant deep features, and (iii) gradient-boosting residual learning that models complex discrepancy beyond lumped-parameter fidelity. The ablation results provide direct evidence: removing the 64-dimensional deep physics features produces a statistically significant +3.73% increase in OOD RMSE (from 5.24 °C to 5.44 °C), indicating that the learned representations encode thermodynamic information not captured by explicit physics or simple statistics.

A coherence test strengthens causal interpretation: disrupting the physical linkage in contrastive pairs by shuffling degrades performance by +5.28%, while preserving marginal distributions, supporting the claim that the learned signal is physically coherent rather than an artifact of statistical augmentation.

Representation probing further aligns the embeddings with physics-relevant quantities (Pearson $|r| > 0.6$ for $\geq$10 dimensions with variables such as power loss and temperature

gradient), reinforcing the interpretation that pre-training shapes features toward underlying thermal dynamics.

Trustworthiness is further established through uncertainty calibration under shift. Standard quantile-regression baselines under-cover on the OOD stress test (PICP 73.73% at 90% nominal), whereas split conformal prediction yields empirical PICP 91.43% with AIW 30.56 °C and a low CUS Gap of 1.43%. Calibration is evaluated in safety-critical regimes via conditional coverage (e.g., PICP $\geq$ 95% for >140 °C) and risk-weighted coverage using $T_{\text{crit}} = 150$ °C, explicitly targeting demagnetization-prone zones. Operational utility is quantified by alarm behavior: using the upper prediction bound as a decision boundary yields alarm precision > 0.85 and recall > 0.99 for $T_{\text{crit}} = 150$ °C, while reducing false derates associated with estimated < 15% power loss per cycle and potentially saving 15–20% energy waste over uncalibrated systems.

The conformal framework is explicitly tied to its assumptions, providing marginal coverage under exchangeability; under OOD shift, coverage is empirically observed rather than theoretically guaranteed.

Contextual comparison across paradigms positions the hybrid architecture against physics-only LPTN baselines, purely supervised deep models, generic self-supervised learning, gradient boosting on engineered statistics, PINNs that embed the LPTN ODE as a soft constraint, and an OLTEM replication using shallow features [24,28,29]. Performance reporting emphasizes chaotic dynamics, distinguishing the stress test from controlled-regime reports where low RMSE values are cited without explicit chaotic OOD validation.

Several constraints bound interpretation. The approach is predicated on simulator-based augmentation, and the LPTN is a low-dimensional abstraction of three-dimensional thermodynamics; although residual learning is designed to compensate systematic physics error, final fidelity remains coupled to baseline approximation quality, with direct implications for safety margin estimation through thermal stress sensitivity. The calibration treats LPTN parameters as point estimates and reports a negative quadratic iron-loss coefficient as a thermodynamically implausible outcome of the current identification strategy, indicating potential model discrepancy in high-speed regimes. Reproducibility-relevant inconsistencies are also present in the manuscript text: identification is described with L-BFGS-B in the main method description and Nelder–Mead in the Appendix A, and hardware is referenced both as RTX 5080, requiring harmonization for rigorous reporting. Deployment claims are qualitative rather than fully parameterized, but the pipeline is explicitly structured for real-time feasibility by confining resource-intensive stages offline and maintaining online inference on the order of milliseconds with low-megabyte footprint and ONNX compatibility, dominated by the Numba-accelerated LPTN forward pass.

7. Conclusions

A physics-informed hybrid ensemble architecture resolves the reliability gap between physically grounded but simplified thermal models and accurate yet brittle data-driven predictors by using calibrated physics as a stabilizing prior, physics-guided contrastive learning (MotorCLR-MR) to encode robust thermodynamic representations, and gradient-boosting residual correction to model non-linear discrepancy. On the chaotic OOD stress test, the method achieves RMSE 5.24 $\pm$ 0.07 °C and a -10.68% improvement under shift, contrasting with +13.35% degradation in a purely data-driven baseline. Trustworthy deployment is further supported by conformalized uncertainty with PICP 91.43% at 90% nominal, CUS Gap 1.43%, and risk-weighted/conditional coverage emphasizing $T_{\text{crit}} = 150$ °C regimes, enabling alarm decisions with precision > 0.85 and recall > 0.99 while mitigating false-derate penalties. Uncertainty results reported in this paper should be interpreted as calibrated predictive uncertainty (prediction intervals) rather than metrological measure-

ment uncertainty; incorporation of sensor and physics-parameter uncertainty propagation is a clear direction for future work. Future work will extend validation to PMSMs with different geometries, ratings, speed ranges, and cooling boundary conditions, to quantify structural generalization and to establish application-specific tailoring protocols.

Author Contributions: Conceptualization, methodology, software implementation, validation, formal analysis, investigation, data curation, and original draft preparation: R.B.H. Assistance with code development, software optimization, visualization, and formal analysis support: M.M.A.H. Assistance with data analysis, experimental validation support, and data curation assistance: M.I.K. Assistance with manuscript writing and technical review: M.A. Assistance with manuscript editing and final proofreading: Y.L. Supervision, resources, guidance on physics and mathematical modeling, and leading manuscript revision and improvement. (Corresponding author): X.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by Xinjiang Southern Xinjiang Key Industry Innovation and Development Support Project (2019DB011).

Data Availability Statement: The raw dataset analyzed during this study is publicly available on the Kaggle repository, titled "Electric Motor Temperature," and can be accessed at: https://www.kaggle.com/datasets/wkirgsn/electric-motor-temperature (accessed on 27 January 2026) [29]. The custom code developed for this research, including all Python scripts for data preparation, feature engineering, model training, and analysis, is not publicly available at this time. However, the code and all supporting artifacts required to reproduce the findings of this study may be made available from the corresponding author upon reasonable request. The corresponding author will consider all such requests for the purpose of academic verification and collaboration.

Acknowledgments: The authors gratefully acknowledge the support and facilities provided by the School of Mechanical Engineering at the Nanjing University of Science and Technology.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. LPTN Calibration and Parameter Identification

This appendix details the mathematical formulation and parameter identification results for the 4-node Lumped-Parameter Thermal Network (LPTN) used as the physics engine in this study.

Appendix A.1. Governing Equations

The thermal behavior of the PMSM is modeled using a system of ordinary differential equations (ODEs) derived from the conservation of energy. The network consists of three active nodes—Stator Yoke (T_{sy}), Permanent Magnet (T_{pm}), and Stator Tooth (T_{st})—and one boundary node, the Coolant (T_c). The heat transfer dynamics are governed by:

Table A1. Identified LPTN Parameters.

Parameter	Symbol	Value	Unit	Description
Heat Generation				
Electrical Resistance	R_0	0.0408	Ω	Base winding resistance
Iron Loss Coeff 1	c_1	1.1769	$\mathrm{W \cdot s/rad}$	Effective linear speed term
Iron Loss Coeff 2	c_2	-2.6×10^{-5}	$\mathrm{W \cdot s^2/rad^2}$	Effective quadratic speed term
Thermal Resistance				
Stator-Magnet	$R_{sy,pm}$	0.70	K/W	Thermal resistance between stator yoke and magnet
Stator-Coolant	$R_{sy,c}$	0.40	K/W	Thermal resistance between stator yoke and coolant
Magnet-Coolant	$R_{pm,c}$	1.20	K/W	Thermal resistance between magnet and coolant
Coolant-Ambient	$R_{c,amb}$	0.10	K/W	Thermal resistance of coolant flow to ambient
Thermal Capacitance				
Stator Yoke	C_{sy}	6750.0	J/K	Heat capacity of the stator yoke
Permanent Magnet	C_{pm}	880.0	J/K	Heat capacity of the permanent magnets
Coolant Node	C_c	9000.0	J/K	Heat capacity of the coolant volume

$$C_{sy}\frac{dT_{sy}}{dt} = P_{loss,total} - \frac{T_{sy}-T_{pm}}{R_{sy,pm}} - \frac{T_{sy}-T_c}{R_{sy,c}}$$
$$C_{pm}\frac{dT_{pm}}{dt} = \frac{T_{sy}-T_{pm}}{R_{sy,pm}} - \frac{T_{pm}-T_c}{R_{pm,c}}$$
$$C_c\frac{dT_c}{dt} = \frac{T_{sy}-T_c}{R_{sy,c}} + \frac{T_{pm}-T_c}{R_{pm,c}} - \frac{T_c-T_{coolant_in}}{R_{c,amb}}$$

where $P_{loss,total} = P_{copper} + P_{iron}$ represents the total heat generation.

Note: $P_{\mathrm{fe}}(\omega)$ is an effective surrogate in ω over the dataset operating envelope. We verify $P_{\mathrm{fe}}(\omega) \geq 0$ across the full speed range used in this study (Appendix A.3).

Appendix A.2. Parameter Identification Results

The parameters were identified using the "Isolate & Calibrate" strategy. First, heat generation coefficients (R_0, c_1, c_2) were fixed via linear regression on steady-state data slices. Second, thermal resistances (R_{th}) and capacitances (C_{th}) were optimized on dynamic transient data using a multi-start (50 initializations) L-BFGS-B procedure under bounded physical constraints, consistent with the main method description. We evaluate the iron-loss surrogate $P_{\mathrm{fe}}(\omega) = c_1\omega + c_2\omega^2$ with $\omega = |\ motor_speed\ | \cdot 2\pi/60$ (rad/s). With $c_1 = 1.176887$ and $c_2 = -2.6 \times 10^{-5}$, the nontrivial zero-crossing occurs at $\omega^\star = c_1/\ |\ c_2\ | \approx 4.526 \times 10^4$ rad/s ($\approx$4.32 $\times$ 10^5 rpm), far beyond the dataset operating range. In this dataset, $\omega_{\min} = 0$ and $\omega_{\max} \approx 628.3$ rad/s ($\approx$6000 rpm). Therefore, $P_{fe}(\omega) \geq 0$ for all observed speeds in the dataset, with the minimum occurring at $\omega = 0$ where $P_{fe}(0) = 0$.

Appendix A.3. Physical Validity Check (Iron-Loss Surrogate)

We evaluate $P_{\mathrm{fe}}(\omega) = c_1\omega + c_2\omega^2$ over the full motor speed range present in the dataset and confirm $P_{\mathrm{fe}}(\omega) \geq 0$ for all observed ω.

Appendix B. Neural Network Architectures and Hyperparameters

This appendix specifies the architectural details of the deep learning models and the hyperparameter configurations for the gradient boosting ensemble used in the final hybrid framework, matching the exact implementation in the codebase.

Appendix B.1. MotorCLR-MR Encoder Architecture

The temporal encoder used for physics-guided representation learning is a Temporal Convolutional Network (TCN).

- Input: (B, 2, 600) corresponding to batch size, 2 physics state channels ($T_{\text{lptn-sy}}$, $T_{\text{lptn-pm}}$), and a 600-step window.
- Layer 1: Conv1d(2, 16, kernel_size = 7, padding = 3) + ReLU.
- Layer 2: Conv1d(16, 32, kernel_size = 7, padding = 3) + ReLU.
- Layer 3: Conv1d(32, 64, kernel_size = 7, padding = 3) + ReLU.
- Pooling: Adaptive Average Pooling to dimension (B, 64).
- Projection Head: Linear(64, 128) + ReLU + Linear(128, 128).

Self-Supervision Settings:

- Temperature (τ): 0.07.
- Augmentation: Physics-consistent parameter perturbation (as described in Section 3).

Appendix B.2. Hybrid Ensemble Hyperparameters

The final prediction model aggregates the outputs of CatBoost and LightGBM regressors. The hyperparameters explicitly defined in the experimental script are listed below; all other parameters used the default values of their respective libraries (catboost 1.2, lightgbm 4.x).

Hyperparameter	CatBoost Regressor	LightGBM Regressor
Objective/Loss	Quantile:alpha = 0.05 (and 0.95)	quantile
Alpha (α)	0.05, 0.95	0.05, 0.95
Iterations/Estimators	1000	(Default)
Learning Rate	0.05	(Default)
Compute Device	GPU (task_type = 'GPU')	GPU (device = 'gpu')
Random State	(Default)	42

Appendix B.3. Computational Infrastructure

All experiments were conducted on a high-performance workstation equipped with two NVIDIA GeForce RTX 5080 GPUs and an Intel(R) Xeon(R) Silver 4214 CPU @ 2.20 GHz. The deep learning models were implemented in PyTorch, and the physics engine was accelerated using Numba. The gradient boosting models utilized CatBoost and LightGBM libraries accelerated via CUDA.

References

1. Aziz, R.; Atkinson, G.J.; Salimin, S. Thermal modelling for permanent magnet synchronous machine (PMSM). *Int. J. Power Electron. Drive Syst.* **2017**, *8*, 1903–1912. [CrossRef]
2. Demetriades, G.D.; de la Parra, H.Z.; Andersson, E.; Olsson, H. A real-time thermal model of a permanent-magnet synchronous motor. *IEEE Trans. Power Electron.* **2010**, *25*, 463–474. [CrossRef]
3. Kim, T.H.; Ryu, C.-M.; Moon, S.J. Thermal characteristics of an interior permanent magnet synchronous motor with various types of helical water-cooling channels. *J. Mech. Sci. Technol.* **2024**, *38*, 6953–6959. [CrossRef]
4. Nguyen, M.-D.; Woo, J.-H.; Shin, H.-S.; Lee, Y.-K.; Lee, H.-K.; Shin, K.-H.; Phung, A.-T.; Choi, J.-Y. Thermal analysis and experimental verification of permanent magnet synchronous motor by combining lumped-parameter thermal networks with analytical method. *AIP Adv.* **2023**, *13*, 025140. [CrossRef]
5. Aslan, E. Temperature prediction and performance comparison of permanent magnet synchronous motors using different machine learning techniques for early failure detection. *Eksploat. Niezawodn.—Maint. Reliab.* **2025**, *27*, 192164. [CrossRef]
6. Hughes, R.; Haidinger, T.; Pei, X.; Vagg, C. Real-time temperature prediction of electric machines using machine learning with physically informed features. *Energy AI* **2023**, *14*, 100288. [CrossRef]
7. Wang, P.; Wang, X.; Wang, Y. End-to-End Differentiable Physics Temperature Estimation for Permanent Magnet Synchronous Motor. *World Electr. Veh. J.* **2024**, *15*, 174. [CrossRef]
8. Hendrycks, D.; Zhao, K.; Basart, S.; Steinhardt, J.; Song, D. The many faces of robustness: A critical analysis of out-of-distribution generalization. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021. [CrossRef]
9. Raissi, M.; Perdikaris, P.; Karniadakis, G.E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* **2019**, *378*, 686–707. [CrossRef]
10. Liu, Z. Physics-informed machine learning for battery pack thermal management. *arXiv* **2024**, arXiv:2411.09915. [CrossRef]
11. Son, S.; Lee, H.; Jeong, D.; Oh, K.-Y.; Sun, K.H. A novel physics-informed neural network for modeling electromagnetism of a permanent magnet synchronous motor. *Adv. Eng. Informatics* **2023**, *57*, 102035. [CrossRef]
12. Wu, X.; Teng, F.; Li, X.; Zhang, J.; Li, T.; Duan, Q. Out-of-Distribution Generalization in Time Series: A Survey. *arXiv* **2025**, arXiv:2503.13868. [CrossRef]
13. Li, J.; Akilan, T. Global attention-based encoder-decoder lstm model for temperature prediction of permanent magnet synchronous motors. *arXiv* **2022**, arXiv:2208.00293. [CrossRef]
14. Lei, D.; Chen, X.; Zhao, J. Opening the black box of deep learning. *arXiv* **2018**, arXiv:1805.08355. [CrossRef]
15. Khalesidoost, S.; Faiz, J.; Mazaheri-Tehrani, E. An overview of thermal modelling techniques for permanent magnet machines. *IET Sci. Meas. Technol.* **2022**, *16*, 219–241. [CrossRef]
16. Sun, Z.; Xu, Y.; Wang, Q.; Xu, J.; Li, G.; Wen, Y. Design and identification of lumped-parameter thermal network model for real-time temperature estimation of permanent-magnet spherical motors. *Therm. Sci. Eng. Prog.* **2023**, *44*, 102047. [CrossRef]
17. Graffeo, F.; Vaschetto, S.; Miotto, A.; Carbone, F.; Tenconi, A.; Cavagnino, A. Lumped-Parameters Thermal Network of PM Synchronous Machines for Automotive Brake-by-Wire Systems. *Energies* **2021**, *14*, 5652. [CrossRef]
18. Kirchgässner, W.; Wallscheid, O.; Böcker, J. Thermal neural networks: Lumped-parameter thermal modeling with state-space machine learning. *Eng. Appl. Artif. Intell.* **2022**, *117*, 105537. [CrossRef]
19. Wallscheid, O.; Böcker, J. Global Identification of a Low-Order Lumped-Parameter Thermal Network for Permanent Magnet Synchronous Motors. *IEEE Trans. Energy Convers.* **2016**, *31*, 354–365. [CrossRef]
20. Reigosa, D.D.; Fernandez, D.; Martinez, M.; Guerrero, J.M.; Briz, F. Thermal Monitoring of Electric Motors: State-of-the-Art Review and Future Challenges. *IEEE Trans. Ind. Appl.* **2021**, *57*, 4756–4770.

21. Sheng, Y.; Liu, X.; Chen, Q.; Zhu, Z.; Huang, C.; Wang, Q. OLTEM: Lumped Thermal and Deep Neural Model for PMSM Temperature. *AI* **2025**, *6*, 173. [CrossRef]
22. Yao, H.; Choi, C.; Cao, B.; Lee, Y.; Koh, P.W.; Finn, C. Wild-Time: A benchmark of in-the-wild distribution shift over time. *arXiv* **2022**, arXiv:2211.14238v2. [CrossRef]
23. Jing, H.; Chen, Z.; Wang, X.; Wang, X.; Ge, L.; Fang, G.; Xiao, D. Gradient boosting decision tree for rotor temperature estimation in permanent magnet synchronous motors. *IEEE Trans. Power Electron.* **2023**, *38*, 10617–10622. [CrossRef]
24. Jin, L.; Mao, Y.; Wang, X.; Lu, L.; Wang, Z. A Model-Based and Data-Driven Integrated Temperature Estimation Method for PMSM. *IEEE Trans. Power Electron.* **2024**, *39*, 8553–8561. [CrossRef]
25. Savaglio, C.; Mazzei, P.; Fortino, G. Edge Intelligence for Industrial IoT: Opportunities and Limitations. *Procedia Comput. Sci.* **2024**, *232*, 397–405. [CrossRef]
26. Wang, T.; Guo, J.; Zhang, B.; Yang, G.; Li, D. Deploying AI on Edge: Advancement and Challenges in Edge Intelligence. *Mathematics* **2025**, *13*, 1878. [CrossRef]
27. Ramones, A.I.; Shi, B.; Engels, M.; Eisenbarth, M.; Andert, J. Estimating the real-time temperature of a 48V permanent magnet synchronous motor using a thermal neural network. *Energy Convers. Manag. X* **2025**, *27*, 101140. [CrossRef]
28. Guo, Y.; Cao, X.; Song, J.; Leng, H.; Peng, K. An efficient framework for solving forward and inverse problems of nonlinear partial differential equations via enhanced physics-informed neural network based on adaptive learning. *Phys. Fluids* **2023**, *35*, 106603. [CrossRef]
29. Kirchgässner, W.; Wallscheid, O.; Böcker, J. Electric Motor Temperature [Data set]. *Kaggle* **2021**. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.