Applied Thermal Engineering 288 (2026) 129552

Contents lists available at ScienceDirect

Applied Thermal Engineering

journal homepage: www.elsevier.com/locate/apthermeng

Research Paper

Transformer autoencoder framework for estimating core temperature of lithium-ion battery from pulse discharge dynamics

Mustehsan Beg [a,*], Keith M. Alcock [a], Vishnu Sam [a], Sanjay Rakshit [b], Sambit Paul [b], Hongnian Yu [a], Keng Goh [a]

[a] School of Computing and Engineering & The Built Environment Edinburgh Napier University, Merchiston Campus, EH10 5DT Edinburgh, UK
[b] Albert AI LTD, 40 Harbour Place, KY11 9GD Dunfermline, UK

ARTICLE INFO

Keywords:
Data-driven
Transformer autoencoder
Lithium-ion
Core temperature
Thermal
Pulse discharge

ABSTRACT

This study presents a non-invasive, data-driven approach for estimating the core temperature of lithium-ion batteries using a transformer-based autoencoder model. Pulse discharge data were collected from a Panasonic NCR18650B cell at three different C-rates (0.5C, 1C, and 2C) to train and evaluate the model. The transformer autoencoder leverages its ability to capture long-range temporal dependencies, efficiently reconstructing multivariate battery signals while compressing them into latent representations that serve as proxies for core temperature. Additionally, two-dimensional visualisation using Principal Component Analysis and t-distributed Stochastic Neighbor Embedding of the latent space revealed well-separated and structured clusters, further confirming the model's ability to encode relevant thermal dynamics effectively. These latent features were used as inputs to a random forest regressor, which was trained to predict temperature and validated against a physical-based thermal model. The proposed method achieved high accuracy across all discharge rates, physics-based model, and drive cycle analysis, with R^2 scores of >0.99 outperforming previously reported studies. These results demonstrate the transformer autoencoder's superior ability to extract meaningful temporally structured representation and its robustness in dynamic operating conditions.

1. Introduction

Estimating the core temperature of lithium-ion (Li-ion) batteries is a critical task in the design and operation of advanced battery management systems (BMS), with direct implications for safety, performance, longevity, and overall system reliability. Unlike surface temperature measurements, core temperature provides a more accurate indication of the electrochemical conditions occurring within the cell, including heat generation due to internal resistance, phase transitions, side reactions, and degradation mechanism [1]. Accurate core temperature estimation enables the detection of thermal runaway risks before it reaches the surface, allowing for proactive thermal control strategies. This is particularly vital in high-demand applications such as electric and hybrid vehicles, aerospace systems, and grid storage, where batteries are subjected to dynamic load profiles and elevated thermal stress [2]. Moreover, real-time estimation of core temperature supports more efficient thermal management, optimised charging/discharging protocols, and extended battery life by preventing over-temperature conditions that accelerate degradation [3]. Consequently, the ability to infer core temperature non-invasively, using data-driven models, has become an essential component of next-generation BMS frameworks focused on improving reliability, safety, performance, and longevity in Li-ion energy storage systems.

Pulse discharge protocol involves subjecting a battery to short, current bursts and then rest period, rather than a continuous and steady load [4]. This approach is particularly valuable as it closely replicates real-world operating conditions encountered in applications such as electric and hybrid vehicles, power tools, portable medical devices, and communication systems amongst others, where energy demand is inherently dynamic and intermittent. The use of pulse discharge is essential for evaluating the battery's transient performance characteristics, including internal resistance, voltage recovery, thermal dynamics, and state-of-health under realistic load profiles [4]. Unlike traditional constant current discharge, pulse discharge protocols provide more comprehensive insights into how a battery responds to rapid fluctuations in power demands, which directly affects its operational reliability,

* Corresponding author.
E-mail address: mustehsan.beg@napier.ac.uk (M. Beg).

https://doi.org/10.1016/j.applthermaleng.2025.129552
Received 8 September 2025; Received in revised form 7 December 2025; Accepted 17 December 2025
Available online 19 December 2025

efficiency, and safety. Moreover, the thermal and electrical data obtained during these tests are critical for the development and calibration of advanced BMS, as well as for data-driven modelling approaches.

The transformer architecture represents a significant advancement in deep learning for time-series modelling, particularly in applications requiring the extraction of complex temporal dependencies and robust feature representation [5]. Originally introduced for natural language processing, the transformer's self-attention mechanism allows it to dynamically weigh the importance of different time steps in a sequence, making it exceptionally well-suited for processing sequential battery data where temporal context is critical [6]. When integrated into an autoencoder framework, the transformer serves as a powerful encoder-decoder model that compresses high-dimensional input signals, such as voltage, current, and power into a low-dimensional latent space while preserving essential temporal patterns [6]. Transformer autoencoders offer key advantages over traditional recurrent architectures by providing concurrent processing, improved scalability, and enhanced modelling accuracy of long-range dependencies, which are crucial for capturing the dynamics of electrochemical systems under varying load conditions [7]. The use of a transformer autoencoder is motivated by the need to extract compact, informative features from high-dimensional pulse discharge data to accurately estimate the core temperature of Li-ion batteries. While a standard seq2seq model can learn input-output mappings, it does not enforce dimensionality reduction or focus on isolating the underlying physical structure of the data. In contrast, an autoencoder introduces a bottleneck that compresses signals into a low-dimensional latent space, effectively filtering out noise and redundant information while preserving features that are most relevant to the battery's internal thermal dynamics. This compression is essential not only for computational efficiency but also for enhancing model generalisation and robustness, particularly across varying operating conditions.

In recent years, data-driven techniques have demonstrated significant promise for real-time battery state estimation, particularly in applications where direct sensor measurements are limited due to cost, safety, or spatial constraints [8,9]. This is especially important under dynamic load conditions and high-current pulse discharges, where thermal gradients and electrochemical dynamics evolve rapidly for conventional sensors to track with precision. Similar flexible sensing technologies have also shown strong potential in other domains, such as wearable health monitoring, soft robotics, and environmental, demonstrating their versatility for capturing fast-changing, spatially distributed signals [10]. Long Short-Term Memory (LSTM) networks are a form of RNN that can identify and retain useful information in a hidden state passed from one time step to the next. This hidden state acts as a kind of memory, allowing the network to account for both short- and long-term effects of the modelled system when predicting an output. They have become prominent due to their ability to model nonlinear and long-term dependencies in sequential battery data crucial for extracting internal states from observable quantities such as surface voltage, current, and temperature. Previously, work done demonstrated that LSTM models significantly outperformed backpropagation and Temporal Convolutional Networks (TCNs) in State of Health (SOH) prediction tasks by learning from statistical cycle-level features [11]. Similarly, F across dynamic drive cycles, exceeding the performance of both RNN and BP-based models [12]. These findings confirm LSTM's strength in generalising across varied time resolutions and cycling behaviours. However, LSTM is not universally dominant across all scenarios. It was found that in temperature prediction tasks involving relatively stable conditions, Artificial Neural Networks (ANNs) outperformed LSTM models in terms of both accuracy and computational load [13]. Nevertheless, under more dynamic and safety-critical environments, such as thermal runaway detection and adaptive cooling control, LSTM-based models offer distinct advantages. For instance, a study proposed an LSTM-based thermal system capable of proactively managing heat dissipation to mitigate thermal risks [14], while, another study introduced a hybrid LSTM-AEKF model that improved SOC accuracy under low-temperature and high-load conditions [15]. Furthermore, LSTM models are inherently well-suited for online deployment in modern Battery Management Systems (BMS), adapting in real time to nonlinearities and emerging failure modes without relying on explicit physical models [16]. This makes LSTM not only a strong baseline but also a foundation for more advanced attention-driven architectures like the Transformer Autoencoder proposed in this study.

Convolutional Neural Networks (CNNs) are a type of neural network capable of capturing geometric relationships within a problem. They are commonly used for image-based tasks, such as image classification and object detection. CNNs have emerged as powerful tools for extracting spatial and local temporal features from multivariate battery sensor data. CNNs are particularly well-suited for structured input formats such as time-windowed sensor matrices or encoded thermal maps and have demonstrated impressive performance in feature abstraction, noise filtering, and degradation trend identification. A study proposed a jump-connection multi-scale CNN that effectively captured both high-frequency and long-range degradation characteristics from voltage-current profiles, leading to superior Remaining Useful Life (RUL) prediction compared to baseline models. Their approach demonstrated strong generalisation on various battery chemistries and aging patterns [17]. Similarly, a study validated CNN-based prognostic models on industrial datasets, highlighting their robustness to noisy input and non-stationary loading conditions [18]. These works reinforce CNN's role in capturing nuanced patterns that might be overlooked by traditional statistical or shallow learning models. To further enhance temporal learning, hybrid CNN models have been developed by integrating CNN layers with recurrent architectures. An introducing of an Attention-augmented CNN-BiLSTM framework that achieved R^2 values exceeding 0.99 and Mean Absolute Percentage Error (MAPE) below 1 % for SOH estimation across multiple datasets [19]. Another study, applied a CNN-LSTM hybrid to battery pack-level SOC estimation, leveraging CNN for spatial feature extraction and LSTM for modelling long-term temporal correlations, ultimately outperforming standalone LSTM models [20]. A notable advancement came from Park et al. who innovatively transformed sequential battery sensor data into 2D grayscale image representations, enabling real-time state estimation using 2D CNNs. Their architecture not only improved prediction accuracy but also reduced inference time by over 90 %, making it particularly suitable for real-time embedded deployment in BMS. These results collectively demonstrate CNN's scalability, noise tolerance, and suitability for low-latency applications, reinforcing its value in battery diagnostics especially when spatial encoding or edge-based deployment is a requirement [21].

Autoencoders are a type of neural network that learn to compress input data into a lower-dimensional representation and then reconstruct it, enabling efficient feature extraction, denoising, and dimensionality reduction. Previous studies have applied autoencoders to battery state estimation; one such study proposed a stacked denoising autoencoder–based deep learning framework that extracts and selects key features from battery discharge curves using clustering by fast search (CFS), enabling more accurate and stable prediction of lithium-ion battery lifetime [22]. A study introduces a denoising-autoencoder-enhanced GRU framework (DAE-GRU) that pre-processes noisy battery measurements to extract cleaner, higher-quality features, enabling significantly more accurate and robust SOC estimation across diverse driving cycles [23]. Another study presents an SAE-GMR framework that extracts and fuses indirect health indicators using a stacked autoencoder and then applies Gaussian mixture regression to deliver more accurate and reliability-aware RUL predictions for lithium-ion batteries [24]. Also, a paper proposes a deep neural network with memory features (DNNwMF), enhanced by an optional autoencoder, to capture multi-cycle degradation patterns and accurately predict lithium-ion battery RUL, demonstrating improved accuracy and reduced model complexity across multiple benchmark datasets [25] and a study develops a hybrid

denoising autoencoder architecture combining convolutional and bidirectional GRU layers to learn nonlinear degradation features and improve RUL prediction accuracy and robustness for lithium batteries, validated on NASA benchmark datasets [26].

Previous work done on data-driven Li-ion core temperature estimation, introduces a multi-step ahead thermal warning LSTM network for Li-ion batteries to predict core temperature overruns, achieving over 97 % accuracy and harmonic average of precision and recall score by incorporating surface temperature differences as input, with a well-tuned dropout improving accuracy by about 2 %, and enabling faster thermal protection through efficient 10-step forecasting. [27]. Another study estimates core temperature for lithium polymer battery (LiPo) and lithium iron phosphate ($LiFePO_4$) batteries using a Kalman Filter and MATLAB/Simscape modelling, evaluates multiple regression models, finds linear regression to be optimal due to core temperature's linearity with current, and achieves 96–99 % prediction accuracy [28]. A novel smart Li-ion battery embedded with distributed fibre-optic sensors to detect internal temperature inhomogeneity, and proposes a hybrid lumped-thermal-neural-network (LTNN) model that significantly improves radial and axial temperature prediction accuracy—achieving Root mean square error (RMSE) as low as 0.18 °C (1C) and 0.07 °C (0.3C), reducing modelling error by ~85 %, and demonstrating strong real-time estimation performance when combined with a UKF-based observer [29]. A developed a GRU-RNN model with two stacked layers (256 and 128 neurons) to estimate Li-ion battery core temperature from voltage, current, ambient, and surface temperature inputs, achieving a mean absolute error (MAE) of 0.066 °C and a maximum error of 0.275 °C, with strong predictive performance across various C-rates (1C—6C) and good generalisation to other batteries of the same type [30]. A novel Nonlinear AutoRegressive with eXogenous inputs (NARX) network was developed and tuned for a large 25 Ah prismatic cell and compared to a similarly structured feedforward neural network using the same input data. Both models demonstrated strong performance in training, validation, long-term prediction, and dynamic driving scenarios, achieving temperature prediction accuracy within 1 K over a 10-h period [31]. A developed an 18-layer CNN to predict the internal temperature of a ternary polymer Li-ion battery pack using external temperature inputs from virtual thermal sensors, achieving a mean square error of 0.047 and outperforming linear regression in prediction accuracy based on a dataset of 81,376 samples collected from 128 internal and external sensors over seven discharge cycles [32]. A hybrid method combining an extended Kalman filter (EKF) with a neural network for core temperature estimation in Li-ion batteries, where the EKF uses a physics-based model and the neural network compensates for model noise, dynamically optimised via particle swarm optimization—resulting in improved accuracy by at least 56.8 % at −15 °C and 60.9 % at 5 °C compared to three existing methods [33].

The goal of using a transformer-based deep learning model is to extract meaningful latent thermal features from the pulse discharge dynamics, which can be correlated with the cell's core temperature for a highly accurate model. In this study, we propose an approach that employs a transformer autoencoder to infer the core temperature of Li-ion batteries based solely on standard electrical signals, such as voltage, current, and power. These electrical are strongly coupled with the internal electrochemical and thermal dynamics of Li-ion batteries, making them effective proxies for inferring core temperature. To evaluate the model's performance, the proposed framework was tested using Panasonic NCR18650B Li-ion cell under three different constant current pulse discharge rates: 0.5C, 1C, and 2C. The model's predictions were benchmarked against a physics-based thermal Li-ion model, demonstrating excellent agreement across all scenarios. A statistical analysis of the prediction error showed that the transformer-based approach maintained high accuracy, with RMSE below 0.166 and coefficient of determination (R^2) values exceeding 0.99 in all cases. Additionally, the error distribution was shown to be tightly centred around zero, indicating minimal bias. The contributions of this paper are summarised as follows:

1. A transformer autoencoder framework is proposed to infer core temperature from standard battery signals, eliminating the need for internal or external temperature sensors or detailed thermal models.
2. The latent space of the autoencoder is demonstrated to contain meaningful thermal information, which can be effectively used for core temperature prediction.
3. The model is evaluated using pulse discharge data under multiple discharge conditions and achieves high predictive accuracy, confirming its robustness and potential for real-time BMS integration.
4. An error analysis is performed against core temperature predictions generated by physics-based Li-ion battery model, highlighting the model's reliability and low prediction variance.

2. Data profile

2.1. Experimental set-up

The pulse discharge tests on the Panasonic NCR18650B Li-ion cell were conducted using battery testing facilities that included a temperature-controlled chamber maintained at a constant 25 °C. A B&K PRECISION 9202 programmable DC power supply and a B&K PRECISION 8610 DC electronic load were used for the charging and discharging operations, which provides adjustable current and voltage control, crucial for executing the pulse discharge protocol with high accuracy. During testing, the Li-ion cell was placed inside the temperature chamber and connected to the external instrumentation, allowing for safe and consistent operation. The entire pulse discharging procedure, including timing, current control, and data acquisition was managed using B&K PRECISION's dedicated battery testing software installed on a PC. Fig. 1(a) illustrates the block diagram of the setup, which established the experimental framework for characterising the cell's behaviour under controlled pulse discharge conditions.

2.2. Li-ion battery pulse discharge data

The Panasonic NCR18650B Li-ion cell undergoes pulse discharge testing to establish a comprehensive understanding of its electrical behaviour, serving as a foundation for estimating core temperature by using a transformer autoencoder framework. Tests are conducted at an ambient temperature of 25 °C across three C-rates 0.5C, 1C, and 2C as shown in Fig. 1 (a), (b), (c), and (d) respectively. During each test cycle, a discharge current pulse is applied for 120 s, followed by a 600-s rest period to allow electrochemical stabilisation. This discharge-rest sequence is repeated until the cell reaches full depletion from 100 % to 0 % state of charge (SOC), terminating at the manufacturer-specified cut-off voltage of 2.5 V. The data collected, including time, voltage, current, and power throughout and used as input to a transformer autoencoder model and implemented to reconstruct these same sequences as its output in Python. This approach allows the model to extract latent features from the dynamic pulse discharge behaviour, facilitating the prediction of core temperature. Fig. 1(e) presents a close-up view of a single pulse during a 2C discharge at 100 % SOC. The plot initially shows the open circuit potential (OCP), followed by a sudden pulse discharge, and is then succeeded by a rest period during which the voltage rises again.

3. Data-driven modelling

3.1. The proposed modelling framework

The goal of the framework is to encode complex relationships within the electrical characteristics of the cell during discharge and reconstruct the signals with high fidelity, while also extracting latent representations that can be used for advanced downstream tasks such as core

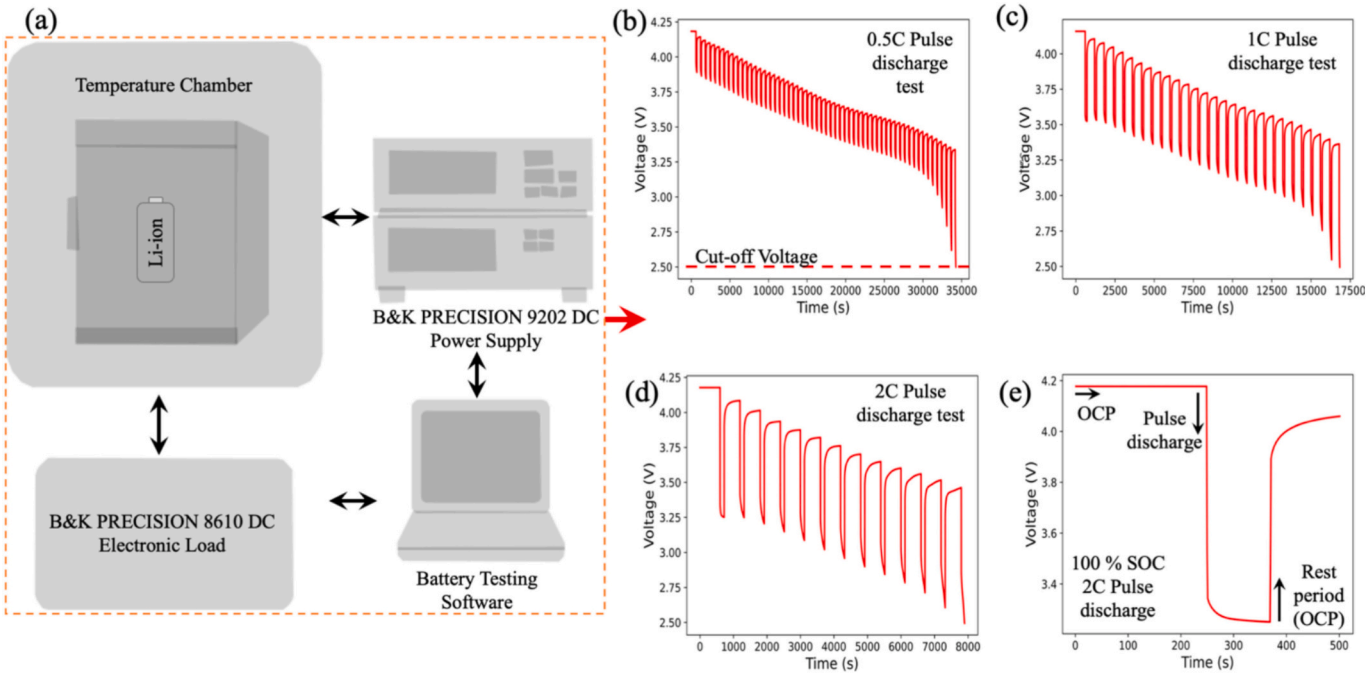

Fig. 1. (a) Block diagram of the experimental set-up for pulse discharge of Panasonic NCR18650B Li-ion cell. (b), (c), and (d) shows the Panasonic NCR18650B Li-ion cell pulse discharge curve at 0.5C, 1C, and 2C respectively, and (e) a detailed view of the 2C discharge process at 100 % state of charge (SOC).

temperature estimation. As can be seen in Fig. 2, the first stage of the framework involves structured data acquisition and pre-processing. The data includes time series records of time, voltage, current, and power. At the core of the framework lies a transformer autoencoder, which is specifically chosen for its ability to handle sequential data and model long-range dependencies without relying on recurrent structures. The encoder component of the Transformer captures the temporal and feature-wise correlations from the input sequences using self-attention mechanisms. It compresses this information into a latent vector, which acts as a distilled representation of the sequence. The decoder then takes

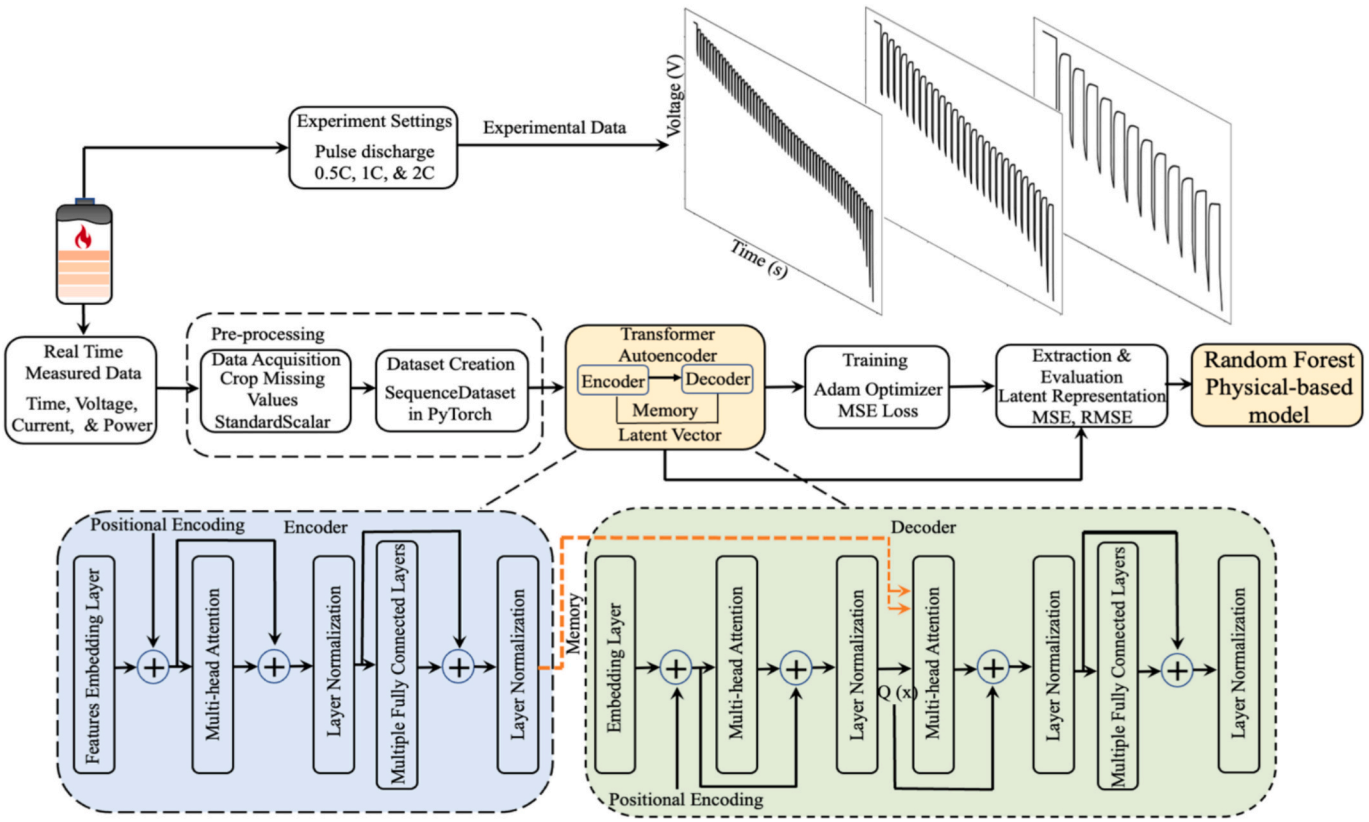

Fig. 2. The proposed transformer autoencoder and random forest regression framework for Li-ion battery core temperature modelling.

the same input sequence along with the encoded memory and attempts to reconstruct the original signal, minimising reconstruction error. The autoencoder is trained in a supervised fashion using Mean Squared Error (MSE) as the loss function, which guides the optimiser to fine-tune the model weights to produce accurate reconstructions. During training, the model is exposed to mini batches of sequences, and its parameters are updated using backpropagation with the Adam optimizer. Training continues for 10 epochs until stability is observed in the loss values. Upon completion of training, the model is transitioned to evaluation mode. It is then used to extract latent features by passing each input sequence through the encoder and capturing the mean of the attention-based memory. These latent features serve as compressed summaries of the battery's dynamic behaviour and are critical for downstream Random Forest regression model for temperature prediction. Furthermore, the effectiveness of the transformer autoencoder is quantitatively evaluated by comparing the reconstructed sequences against the original input using MSE, RMSE, MAE, and the R^2 coefficient.

3.2. The transformer autoencoder network

The encoder of transformer autoencoder receives the embedded input features, which are projected into a latent space via a linear layer. These features are then passed through the attention and feedforward layers, producing a compact latent representation of the time-series data, which is crucial for the task of reconstruction and predicting future sequences. Consisting of multi-head self-attention and a position-wise feedforward network. The multi-head self-attention mechanism allows the encoder to attend to different parts of the input sequence simultaneously, capturing dependencies between different time steps in the time-series data. The feedforward network within the encoder processes the output of the attention mechanism and helps capture complex patterns within the latent representation. This structure is ideal for handling time-series data as it can model both short and long-range temporal dependencies efficiently. The decoder also leverages multi-head attention and feedforward layers as part of a decoder layer.

The decoder receives the latent representation from the encoder and uses cross-attention to attend to the encoder's output (memory) while also applying self-attention on the decoder's previous outputs. This allows the decoder to generate or reconstruct a sequence based on the context of the entire time-series data. The self-attention mechanism in the decoder captures temporal dependencies within the output sequence, while the cross-attention mechanism ensures that the decoder uses information from the encoder's representation of the input sequence. By applying these attention mechanisms, the model can effectively reconstruct the time-series data. The decoder, with its ability to model complex dependencies across time steps, is well-suited to handle the temporal patterns inherent in time-series data. The key equations for the model are presented in Eqs. (1–3), beginning with the self-attention mechanism in Eq. (1) [6].

$$Attention\ (Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{1}$$

where each input attends to others by computing weighted combinations of values (V) based on the similarity between queries (Q) and keys (K), scaled by and normalised with softmax. Multi-head attention splits the model dimension into multiple heads (for parallel attention) shown in Eq. (2) [6].

$$MultiHead\ (Q, K, V) = Concat\ (head_1, \ldots\ldots, head_h)\ W^O \tag{2}$$

where $head_1$ = Attention $\left(QW_i^Q, KW_i^K, VW_i^V\right)$.

where, W_i^Q, W_i^K, W_i^V projection matrices for each attention head and W^O is the output projection matrix and normalisation step. Finally, the two-layer feedforward network applied independently to each time step as shown in Eq. (3) [6].

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \tag{3}$$

where x is the input vector, W_1 and W_2 are weight matrices, b_1 and b_2 are bias vectors and max is an ReLU activation function.

3.3. Physical temperature model

Random Forest regression model was employed to predict temperature based on the latent variables extracted from the transformer autoencoder. The physical temperature estimation was grounded in thermodynamic principles, a 1D lumped-parameter method with variable internal resistance calculated from experimental data that better captures the behaviour of battery pulse discharge and drive cycles, allowing it to accurately predict internal thermal dynamics. The initial model used a simplified, one-dimensional, physics-based approach with a constant internal resistance, ignoring complex thermal gradients and heat-transfer mechanisms, more details and results can be found in Section S4 of supplementary information. The battery mass and specific heat capacity were taken as 0.0475 kg and 776.59 J/kg·°C, respectively [34]. The initial temperature was set to 25 °C. The Random Forest model effectively learned the nonlinear mapping between the latent representations, capturing electrochemical and thermal dynamics and the physical temperature response. Initially, the transient thermal behaviour of a cylindrical 18,650 lithium-ion cell was simulated using a lumped-parameter energy balance approach and the resulting temperature rise of 2C, 1C, and 0.5C pulse discharge are discussed and shown in Section S2 of supplementary information. The R_{int} values of one-dimensional, physics-based approach are taken from the full NCR18650B cell pulse discharge data and calculated as shown in Eq. (4).

$$R_{int} = \frac{V - OCV}{I} \tag{4}$$

where V represents the battery's pulse-discharge voltage, OCV is the open-circuit voltage median value measured during the 10-min rest period following the pulse discharge, and I denotes the applied pulse-discharge current.

$$Q = I^2 \bullet R_{int} \bullet \Delta t \tag{5}$$

Eq. (5) where, Heat is the thermal energy generated, $I^2$is the current in amperes, R_{int} is the internal resistance and Δt is the difference in time (seconds). The method to model the heat transfer of the cell by using a 1D lumped-parameter model which can be described as shown in Eq. (6).

$$\frac{dT}{dt} = \frac{Q - hA(T - T_{amb})}{m\ x\ cp} \tag{6}$$

where Q is the rate of heat generation by the cell (W), T is the cell temperature, T_{amb} is the ambient temperature, h is the heat transfer coefficient including both convection and conduction effects, and A is the equivalent heat transfer area. Table 1 shows the parameters used in the 1D lumped-parameter model.

Table 1
Battery characteristics utilised in the Li-ion core temperature estimation 1D lumped-parameter model.

Parameters	Value	Notes
Specific heat capacity (c_p)	776.59 J/kg °C	[34]
Mass (m)	0.0475 kg	Datasheet
T_{amb}	25 °C	Laboratory
h	10 W/m^2 °C	[34]
A	0.0011895 m^2	Datasheet
R_{int}	Various	Laboratory

3.4. Model training and regression evaluation

The parameters used during the training of the proposed transformer autoencoder for predicting the core temperature of Li-ion battery are detailed in Table 2. The model is designed to handle multivariate time-series input with three features, voltage, current, and power and compress them into a latent space using a linear embedding layer followed by transformer-based encoder and decoder blocks. The input dimension is set to 3, and the latent dimension is set to 16. The model is configured with two layers in both the encoder and decoder, and each transformer layer employs a single attention head and a feedforward network of size 64. The training was carried out for 10 epochs using a batch size of 32 and a learning rate of 0.001. The mean squared error (MSE) between the predicted and actual sequences is used as the loss function, and the 'Adam' optimizer is adopted for efficient and adaptive gradient updates.

To mitigate overfitting, an 70/30 train-test split on the scaled dataset before constructing temporal sequences, ensuring a clear separation between training and evaluation data, and allowing the model's performance to be assessed reliably on unseen test data. The train-test split is done twice to protect it from leakage. The train-test split initially done on the battery pulse discharge is split into training (70 %) and test (30 %) by time sequence per C-rate. This split is used for training the autoencoder and evaluating its reconstruction performance. After extracting latent features from the transformer autoencoder, the dataset was split into training (70 %) and test (30 %) subset with a fixed random seed for reproducibility. The Random Forest regressor was trained solely on the training subset, and the test subset was completely held out during model training ensuring no input variable is created using the target value or any post-processing that leaks information from the test set.

The hyperparameter tuning was performed in two stages. For the Transformer autoencoder, important hyperparameters such as the latent dimension, number of attention heads, number of encoder and decoder layers, feedforward dimension, learning rate, batch size, and sequence length were tuned using a random search strategy. Random search was chosen because the Transformer has a large and nonlinear hyperparameter space, and sampling combinations is more efficient than exhaustively testing all possibilities. Each sampled configuration was trained for a few epochs, and the reconstruction loss on the training data was used to select the best-performing model. After extracting latent features from the optimised Transformer, a Random Forest regressor was tuned using grid search over parameters such as the number of trees, maximum tree depth, minimum samples for splitting, minimum samples per leaf, and maximum features considered for splitting. Each parameter combination was evaluated using cross-validation, and the configuration yielding the lowest mean squared error was selected. This two-stage tuning ensured that the latent features were informative, and the regression model was optimally calibrated for accurate temperature prediction.

The model is built using Python 3.9 and PyTorch 2.1.1, and all experiments were conducted on a system equipped with an Apple M2 processor, 8 GB of RAM, an 8-core GPU, and a 16-core Neural Engine. The total training and prediction process for the 10-epoch transformer autoencoder took ~2.25 min, ~2.10 min for training and ~15 s for testing. This low inference latency suggests the model is highly suitable for real-time applications, such as on-board battery monitoring systems. The evaluation metrics, mean squared error (MSE), RMSE, MAE and R^2 are computed. The mathematical expressions for these are provided in Eqs. (7–10).

$$\mathrm{MSE} = \frac{1}{n}\sum_{k=1}^{n}(yk - \hat{y}k)^2 \tag{7}$$

where n is the number of samples, y is the actual value and $\hat{y}$ is the predicted value of the k^{th} sample. The rest of the evaluation metrics equations are provided below.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(yk - \hat{y}k)^2} \tag{8}$$

where n is the number of samples and $(yk - \hat{y}k)^2$ is the squared error for each prediction.

$$MAE = \frac{\sum_{k=1}^{n}(yk - xk)}{n} \tag{9}$$

where n is the number of samples and $(yk - xk)$ is the absolute error for the data point.

$$R^2 = 1 - \frac{\sum_{k=1}^{n}(\hat{y}k - yk)^2}{\sum_{k=1}^{n}(yk - \overline{y})^2} \tag{10}$$

where $\hat{y}k$ is the predicted value, yk is the actual value and $\overline{y}$ is the mean of all actual values. The numerator $\sum_{k=1}^{n}(\hat{y}k - yk)^2$ is the residual sum of squares, measuring the prediction error and $\sum_{k=1}^{n}(yk - \overline{y})^2$ is the total sum of squares measuring the total variation in the true values.

Table 2
The parameters employed for training the transformer autoencoder model.

	Criteria	Value
Model parameters	Input size	3
	Latent dimension	16
	Encoder/Decoder size	2
	Feedforward dimension	64
	Positional Encoding	Applied
Training settings	Epoch	10
	Batch size	32
	Learning rate	0.001
	Optimizer	Adam
	Loss function	Mean Squared Error (MSE)

3.5. Model assumptions and current limitations

The main assumptions, simplifications, and current limitations of the proposed approach include the reliance on a single cell type and limited temperature range, the use of a simplified thermal model for validation, and the data-driven nature of the method, which requires sufficient experimental data for training. Compared to framework with recent physics-informed deep learning methods such as the Deep Energy Method (DEM) and its neural operator extension (VINO) [35] [36]. These approaches unify modelling and simulation by formulating the learning process directly from thermodynamic and variational principles, thereby enabling both forward and inverse problem-solving without explicit discretisation or extensive data requirements. In contrast, the transformer-based autoencoder offers strong representation learning and scalability advantages, particularly for data-rich battery systems where large experimental datasets are available. However, it lacks the built-in physical consistency of DEM or VINO. Future work will explore hybrid formulations that integrate energy-based constraints or neural operator components to combine the interpretability of physics-informed methods with the flexibility and efficiency of transformer architectures.

The approach proposed in this paper, a Random Forest is trained on latent features rather than directly on the original electrical signals, utilises transformer autoencoder's compressed representation retains all information relevant for temperature prediction; however, because the latent space can be abstract, it becomes difficult to interpret how the encoded features relate to underlying thermal behaviour. The method blends first-principles physics via a simplified dynamic 1D lumped-

parameter method with data-driven components such as the transformer and Random Forest, leveraging both physical intuition and flexible function approximation. By training the temperature regressor on latent features instead of raw inputs, the framework reduces dependence on large, labelled temperature datasets, mitigates overfitting risk, and speeds computation. The transformer autoencoder's role in compressing high-dimensional time series into a lower-dimensional embedding also enables reuse of these features for tasks beyond temperature estimation. Nonetheless, because the transformer primarily learns statistical correlations rather than true causal thermal dynamics, the resulting predictions may become physically inconsistent in scenarios that deviate from the training distribution.

Overall, the approach of hybrid physics-data methodology that leverages transformers for temporal patterns and Random Forests for temperature prediction. Its main strengths are flexibility, compact latent representation, and rapid prototyping. Weaknesses include limited physical fidelity, potential overfitting due to smaller datasets, poor extrapolation, and low interpretability of latent features. More sophisticated physical modelling or physics-guided learning could improve reliability, especially for predictive thermal management.

4. Results and discussion

4.1. Evaluation of model

The evaluation results presented in the Table 3, reflect the performance of the proposed transformer autoencoder in predicting the core temperature of Li-ion batteries across various C-rates using pulse discharge data. At the highest discharge rate of 2C, the model achieved highest loss value of 0.3206, with a MSE of 0.0004 and a RMSE of 0.0205. Furthermore, MAE is 0.0099 and R^2 reached 0.9996, indicating that the model is capable of accurately capturing rapid thermal dynamics under high-stress operational conditions. As the C-rate decreased, the model continued to demonstrate robust performance. At 1C, the autoencoder yielded an MSE of 0.0002 and an RMSE of 0.0131, MAE of 0.0072, accompanied by an R^2 of 0.9998, reflecting even stronger predictive reliability. Notably, at the lowest discharge rate of 0.5C, which represents a more stable and gradual thermal evolution, the model achieved its best performance with an MSE of 0.0001, RMSE of 0.0074, MAE of 0.004, and a near-perfect R^2 of 0.9999. These results indicate that the transformer autoencoder is highly effective at modelling temperature dynamics across a broad range of discharge rates. The strong R^2 values across all C-rates validate the transformer autoencoder's capability to reconstruct complex temporal dependencies and infer core thermal states with high precision. The standard deviation of transformer autoencoder model's performance across multiple training runs with different random seeds, as can be seen in section S1 in the supplementary information.

2C pulse discharge rate of Li-ion battery will be used to evaluate the transformer autoencoder performance for the rest of the paper. Fig. 3 (a) illustrates the input signals and the corresponding latent temperature representation extracted by the transformer autoencoder during a 2C pulse discharge test. The top three subplots display the voltage, current, and power profiles, respectively, which serve as inputs to the model. These signals exhibit the characteristic pulse discharge behaviour, with sharp periodic transitions that reflect the high-frequency, high-stress loading conditions imposed on the battery. Notably, the current and power signals demonstrate rectangular pulse patterns, while the voltage signal reveals both the immediate drop during each pulse and the gradual degradation trend over time. The bottom subplot shows the latent temperature variable inferred by the encoder from the input dynamics. This latent representation captures the temporal evolution of core thermal behaviour, demonstrating clear alignment with the pulsed input patterns. The smooth yet responsive trajectory of the latent temperature signal suggests that the autoencoder effectively encodes both the fast transients and the underlying thermal accumulation associated with repeated high-current discharges. This ability to infer a physically meaningful latent state from non-temperature input variables highlights the potential of the transformer autoencoder to serve as a non-invasive estimator of core battery temperature. Moreover, the structured latent dynamics reflect the model's capacity to learn relevant temporal features, which is critical for applications in real-time battery monitoring and thermal management. Fig. 3(b) shows the epoch-wise training results of the model, illustrating the decrease in the mean squared error (MSE) loss function over the 10 training epochs.

Table 3
The evaluation of the proposed model under three different c-rates of Li-ion battery.

C-Rate	Epoch	Loss	MSE	RMSE	MAE	R^2
2C	10	0.3206	0.0006 ± 0.0001 %	0.0247 ± 0.0025 %	0.0129 ± 0.0016 %	0.9994 ± 0.0001 %
1C	10	0.2789	0.0002 ± 0.0001 %	0.0153 ± 0.0022 %	0.0083 ± 0.0015 %	0.9998 ± 0.0001 %
0.5C	10	0.1292	0.0002 ± 0.0003 %	0.0139 ± 0.0070 %	0.0087 ± 0.0049 %	0.9998 ± 0.0002 %

A heatmap of Pearson correlation coefficients between the original electrical signals of voltage, current, and power—and the learned latent representation, denoted as latent_temp, extracted from the transformer autoencoder model is shown in Fig. 3 (c). The strong positive correlation between latent_temp and voltage (0.62) indicate that the latent space encodes significant information about the voltage dynamics of the system. Conversely, the strong negative correlations with current (−0.75) and power (−0.76) suggest that the latent variable captures inverse trends related to these features. These relationships demonstrate that the latent space effectively encapsulates meaningful underlying structure and temporal dependencies from the multivariate time series data. The clear diagonal values of 1.00 reaffirm perfect self-correlation, while the off-diagonal values reflect inter-feature relationships, including the strong positive correlation between current and power (1.00), which is expected due to their physical coupling. The heatmap affirms the interpretability of the learned features and validates the capability of the transformer encoder to extract relevant latent factors that preserve key relationships from the original sensor inputs.

The residual plots for voltage, current, and power over time, computed as the difference between the original signals and their corresponding reconstructions from the transformer autoencoder shown in Fig. 3 (d). These residuals offer a direct visualisation of the model's reconstruction error across all timesteps. The red curves indicate areas where the autoencoder struggled to accurately capture transient behaviours of pulse discharge dynamics. Notably, high-magnitude residuals often correspond to signal peaks and sharp transitions, which may be more challenging for the model to generalise due to their sparsity or irregular patterns. For the most part, residuals remain close to zero, confirming that the model achieves high reconstruction fidelity across normal operational phases. This residual analysis is crucial for identifying outliers or anomalies, and further demonstrates the temporal sensitivity and reconstruction accuracy of the transformer architecture when applied to electronic signal data.

The distribution of the squared reconstruction errors for the three original signals consisting of voltage, current, and power via a strip plot visualisation is illustrated in Fig. 4 (a). Each strip plot summarises the model's error profile per feature by showing the spread, central tendency, and presence of outliers in the squared error values resulting from the autoencoder's reconstruction process. Most of the error values are densely concentrated near zero, indicating that the model can reconstruct most inputs with high accuracy. However, the presence of numerous outliers, especially in the current and power signals,

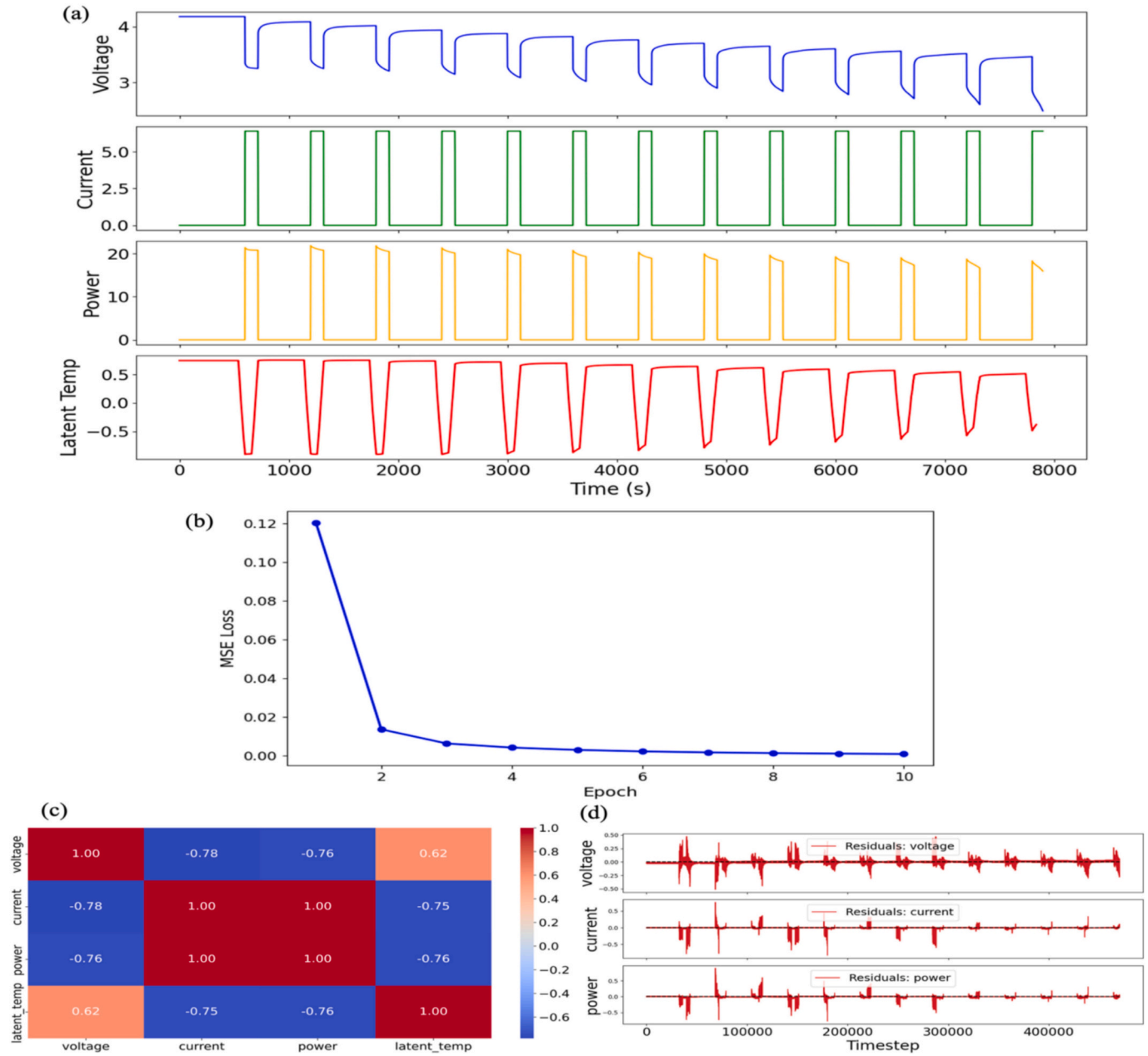

Fig. 3. (a) shows input signals and the latent temperature representation from the transformer autoencoder during a 2C pulse discharge test, (b) Epoch and loss function (MSE) results for the model (c) Shows a heatmap of Pearson correlations between original signals and the learned latent representation, and (d) displays the residual plots for voltage, current, and power over time.

highlights specific instances where the reconstruction deviated significantly from the original input. These deviations may correspond to rare or anomalous events in the input sequence, which the model finds more difficult to generalise. The higher vertical dispersion for current and power, compared to voltage, suggests that these features exhibit more complex temporal dynamics or variability that challenge the autoencoder's learning capacity. Fig. 4 (b) presents a time series plot of MSE computed at each timestep across all features, capturing the point-wise reconstruction quality of the transformer autoencoder over the entire input sequence. The red curve reveals sharp peaks and fluctuations in MSE at various intervals, signifying temporal windows where the model's prediction diverged notably from the actual input. These error spikes reflect underlying physical transitions, transient phenomena, related to pulse discharge of the battery. Importantly, the model maintains a consistently low reconstruction error across large portions of the time series which improves as time steps increases over time, attesting to its robustness in modelling typical system behaviour. Together, the strip plot and time series visualisation provide a comprehensive understanding of the autoencoder's reconstruction performance both at the feature level and across temporal dynamics.

4.2. *Evaluation in two-dimensional space*

A high-dimensional data can be challenging to interpret and analyse directly. To address this, dimensionality reduction techniques such as Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) are commonly used to project complex datasets into two-dimensional space for visualisation and evaluation [37,38]. PCA is a linear method that reduces dimensionality by finding the directions (principal components) that capture the greatest variance

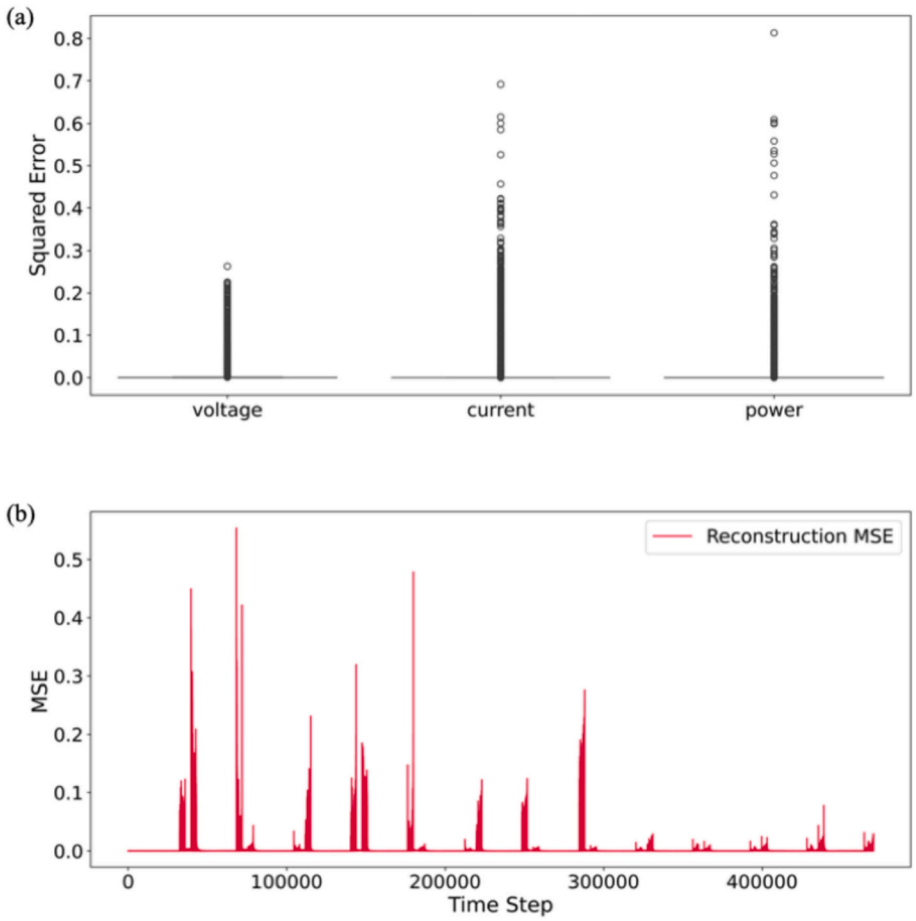

Fig. 4. (a) Strip plot showing squared reconstruction errors for voltage, current, and power signals and (b) time series plot of MSE computed at each timestep across all features.

in the data. This allows for understanding of global patterns and feature importance across the dataset. In contrast, t-SNE is a nonlinear technique that focuses on preserving local neighbourhood structures, making it particularly effective for visualising clusters and subtle relationships in the data. When applied to the latent space of models of transformer autoencoders, these methods help assess how well the model organises input signals, detect separable structures, and provide insights into learned representations in a more intuitive, visual format. Fig. 5 (a) shows the projection of the latent representations derived from the transformer autoencoder onto a two-dimensional space using PCA. Each point in the plot corresponds to a latent vector at a specific timestep, coloured according to its time of occurrence (in seconds), with a gradient from purple (earlier) to yellow (later). The structure of the PCA projection reveals a smooth and continuous manifold, indicating that

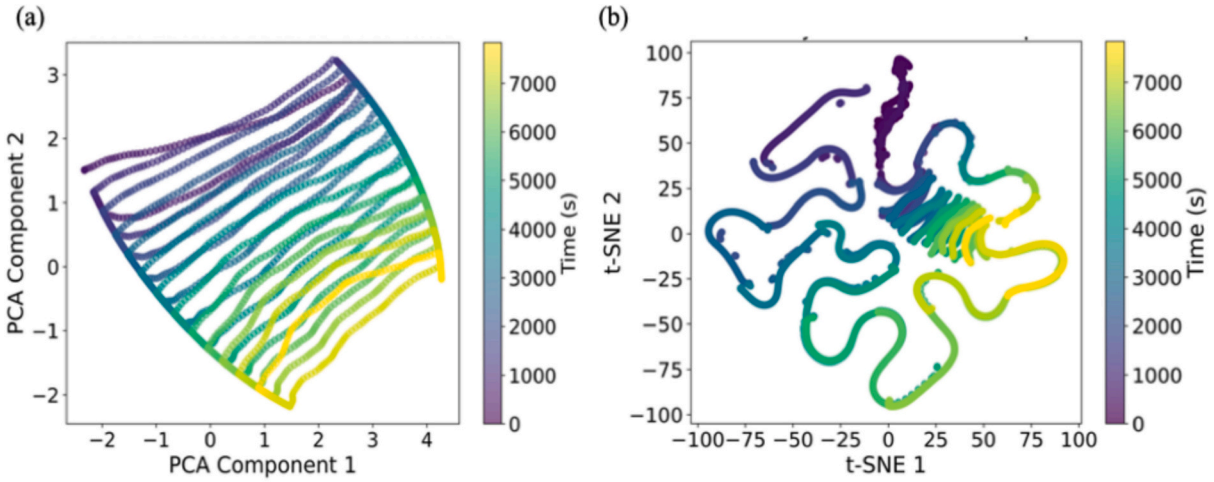

Fig. 5. (a) Shows PCA projection of latent representations from the transformer autoencoder in 2D, shows smooth manifold implies temporal progression is well-captured and (b) 2D t-SNE visualisation of the same latent space, shows ability for disentanglement of complex temporal patterns in the t-SNE latent space for finer detection of transitions, cycles, and deviations.

the learned latent space effectively captures the underlying temporal progression of the input data. The arching, layered appearance of the trajectories suggests that the autoencoder organises temporally adjacent points closely in the latent space, maintaining chronological coherence. This orderly mapping reflects the model's capacity to extract and encode meaningful dynamical patterns from the input sequence, which is essential for both reconstruction quality and downstream tasks like forecasting or anomaly detection.

A two-dimensional t-SNE visualisation of the same latent space, offering a nonlinear perspective of its topology is presented in Fig. 5 (b). Compared to PCA, t-SNE focuses on preserving local neighbourhood relationships rather than global structure. The resulting plot reveals more intricate and separated looping structures, with distinct temporal clusters evident along the trajectory. This visualisation underscores the presence of local variations and micro-patterns within the latent representations, which PCA may compress into linear components. The colour coding by time again emphasises the temporal unfolding of system states, and the continuity of the t-SNE path further validates the coherence of the model's internal representations. Overall, the t-SNE projection demonstrates how the latent space disentangles complex temporal patterns into a rich and interpretable structure, potentially enabling fine-grained detection of transitions, cycles, or deviations in the data stream.

While the latent features extracted by the transformer autoencoder are not directly interpretable as physical quantities, preliminary correlation analyses indicate that certain dimensions of the latent space are associated with key battery parameters such as state of health and aging life estimation [39,40]. Moreover, when visualised using t-SNE, the latent features form distinct clusters that align with variations in operating conditions such as C-rate as can be seen in Fig. 8. These clustering patterns imply that the autoencoder learns to embed subtle electrochemical behaviours, such as reaction kinetics and internal resistance trends, without being directly provided with these labels.

While these findings suggest that the latent representation implicitly captures important aspects of battery dynamics, a more rigorous investigation is needed to establish definitive physical interpretability. As part of our future work, we plan to systematically quantify these correlations through sensitivity analysis, latent-space probing, and controlled perturbation experiments. This will enable us to better understand which latent dimensions correspond to specific physical mechanisms and, ultimately, enhance the transparency and scientific insight provided by the model. Fig. S10 in the supplementary information presents the PCA and t-SNE visualizations for the data collected at 2C, 1C, and 0.5C-rates.

4.3. Evaluation of Li-ion physical-based thermal model at different pulse discharge C-rate

To assess computational efficiency and real-time feasibility across practical operating conditions, we evaluated the transformer autoencoder at multiple discharge rates (0.5C, 1C, and 2C) as shown in Table 4. The model exhibited consistently low inference latency across all C-rates, demonstrating minimal sensitivity to the underlying signal dynamics. Specifically, average inference times were 0.928 ms at 0.5C, 0.737 ms at 1C, and 0.919 ms at 2C, with corresponding throughputs of 1078 FPS, 1356 FPS, and 1088 FPS, respectively. These sub-millisecond latencies confirm that the model can operate within real-time constraints required for battery monitoring and fast dynamic profiles. Notably, the computational footprint is extremely lightweight, requiring only 1.64 million MACs and 15.47 k trainable parameters. This compact architecture combined with stable multi-rate performance demonstrates strong suitability for deployment on embedded battery management hardware and other resource-limited edge platforms where deterministic, low-latency inference is essential.

Fig. 6 (a), (b), and (c) presents an evaluation of the temperature estimates derived from the transformer-based autoencoder model and a physics-based thermal model of Li-ion battery under three C-rate conditions: 2C (reaching the peak temperature of ~28 °C), 1C (reaching the peak temperature of ~25.8 °C), and 0.5C (reaching the peak temperature of ~25.1 °C). In each subplot, the solid blue line represents the reference physics-based temperature profile, while the dashed red line shows the temperature inferred from the latent space of the machine learning model. Visually, the predicted lines align closely with the physics-based signals across all operating rates, indicating that the latent features successfully capture the thermal dynamics of the system. The slight deviations observed at step change, particularly at higher C-rates at 2C, are likely due to the more rapid and complex thermal changes at higher current loads, which present a greater challenge for the model to generalise accurately. Nonetheless, the overall trend, stepwise behaviour, and thermal slope are well replicated in the predicted curves.

In addition to time-series comparisons, a histogram of the prediction error (Fig. 6 (d)) provides further insight into the accuracy and distribution of the model's temperature estimates. The plot shows the frequency of the error values calculated as the difference between the physics-based temperature and the transformer-inferred temperature for 2C pulse discharge. The sharp, narrow peak cantered around zero demonstrates that most predictions are highly accurate, with minimal deviation from the reference signal. The error distribution is symmetrically cantered and shows minimal skew, indicating that the model does not exhibit a systematic overestimation or underestimation bias. The relatively tight spread of the histogram confirms the consistency and reliability of the autoencoder-based predictions across the dataset. Only a small number of predictions fall outside a $\pm$ 0.2 °C error range, further reinforcing the conclusion that the learned latent representation effectively captures the underlying thermal behaviour. Fig. 6 (e) presents the correlation heatmap between the latent variables extracted by the transformer autoencoder and the actual physical core temperature of the Li-ion battery model. As shown in the heatmap, several latent variables (e.g., latent_4, latent_5, latent_6, latent_8, latent_11, latent_15,) exhibit strong positive correlations (above 0.3) with temperature, while others (e.g., latent_1, latent_2, latent_7, latent_9, and latent_10) show strong negative correlations, indicating that the model has effectively learned to encode both direct and inverse thermal dependencies. The diversity in correlation strengths suggests that different latent dimensions capture complementary aspects of the battery's core thermal behaviour. Fig. 6 (f) compares physics-based and transformer autoencoder-predicted battery temperature profiles across three pulse discharge rates (0.5C, 2C, and 1C), reaching the peak temperature of ~28 °C, in a combined dataset, demonstrating the model's accuracy in capturing temperature dynamics under varying C-rates and showing the higher diversion at step change at 2C as discussed previously.

Moreover, Quantitative results provided in Table 5, further confirm the strong performance of the model. At each C-rate, the MSE, RMSE,

Table 4
Computational efficiency and real-time feasibility of transformer autoencoder framework at 2C, 1C, and 0.5C.

Runtime characterisation	Inference latency			FLOPs & Parameters	
	Average inference latency (ms)	Standard deviation (ms)	Throughput (FPS)	MACs (MMac)	Parameters
2C	0.919	0.277	1087.89	1.64	15.47 k
1C	0.737	0.104	1356.17	1.64	15.47 k
0.5C	0.928	0.260	1077.81	1.64	

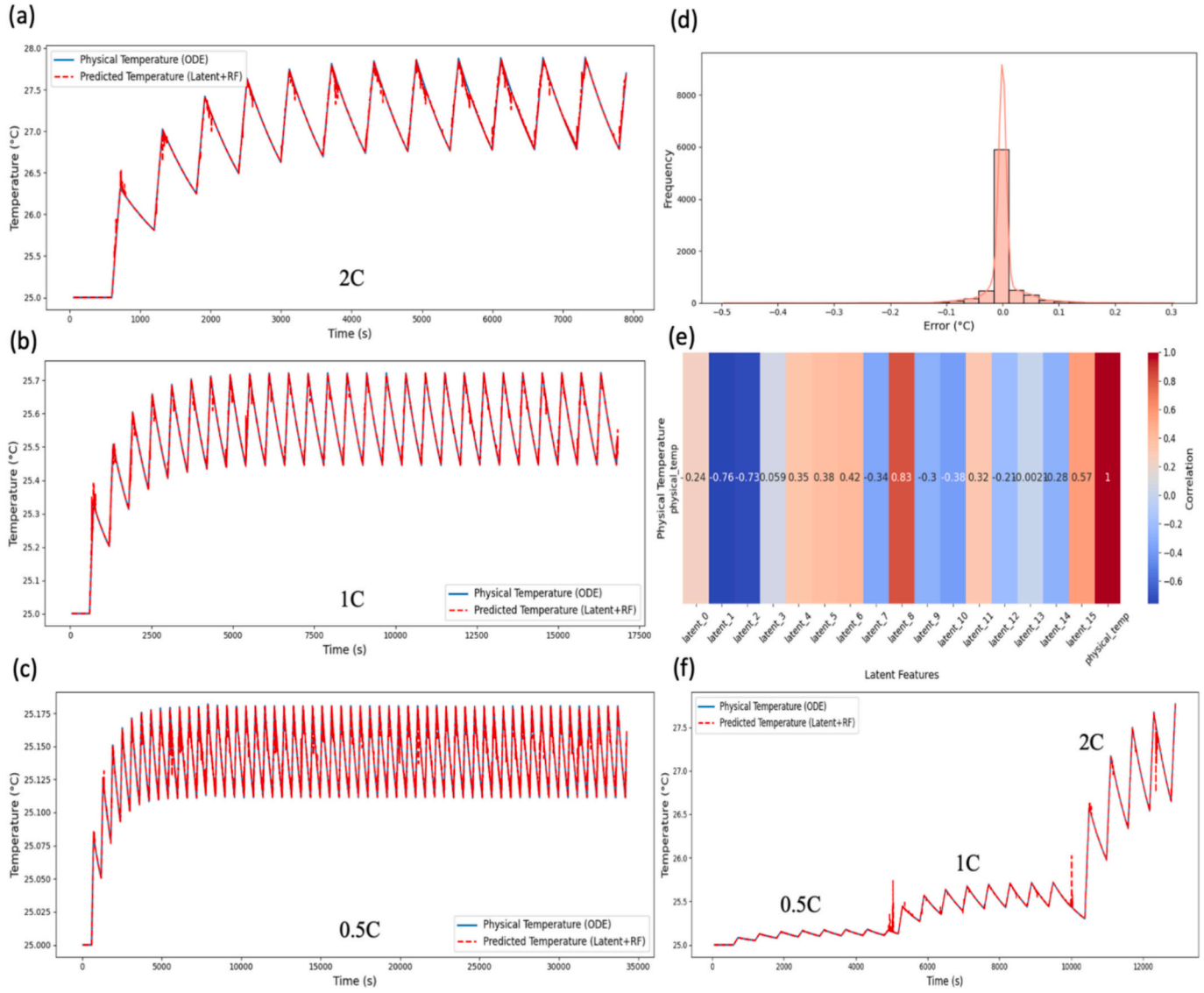

Fig. 6. (a), (b), and (c) compare temperature estimates from the transformer autoencoder and a physics-based model at 2C, 1C, and 0.5C, respectively, (d) accuracy and distribution of the model's temperature estimates, (e) Correlation heatmap between transformer-extracted latent variables and physical-based core battery temperature, and (f) Battery temperature profiles from physics-based and transformer models across 0.5C, 1C, and 2C pulse rates in a combined dataset.

Table 5
Quantitative results physical-based temperature and transformer autoencoder model.

C-Rate	MSE	RMSE	R^2	MAE
2C	0.0276 ± 0.0004 °C	0.1661 ± 0.0013 °C	99.9723 ± 0.0004 %	0.0638 ± 0.0006 °C
1C	0.0031 ± 0.0001 °C	0.0556 ± 0.0002 °C	99.9892 ± 0.0001 %	0.0216 ± 0.0002 °C
0.5C	0.0002 ± 0.0001 °C	0.0128 ± 0.0002 °C	99.9978 ± 0.0001 %	0.0050 ± 0.0001 °C
Mixed	0.0038 ± 0.0001 °C	0.0619 ± 0.0002 °C	99.9884 ± 0.0001 %	0.0132 ± 0.0002 °C

MAE, and R^2 are calculated. The model achieves low error values across the board and notably high R^2 scores: 0.9997 for 2C, 0.9998 for 1C, and 0.9999 for 0.5C. These results indicate that the model preserves most of the variance present in the true temperature data and can accurately predict core thermal estimates, even when subjected to different operational stress levels and combined (mixed) C-rate dataset ($R^2 = 0.9998$). The declining MSE, RMSE, and MAE values with decreasing C-rate also reflect the model's robustness in lower dynamic regimes, where core thermal behaviour is more stable. Collectively, these results support the effectiveness of the transformer-based latent representation in capturing complex thermal characteristics of battery systems.

4.4. Statistical comparison

The current configuration is relatively simple compared to typical transformer usage. This design choice was made to balance training efficiency with the size of the dataset and enabling real-time, on-device deployment. Nevertheless, to validate the necessity of the transformer, we conducted experiments using simpler architectures, including a 1D CNN + Random Forest, standard Autoencoder + Random Forest, and a LSTM + Random Forest, on input features. The results, shown in Table 6, show the results of simpler models when compared to Transformer Autoencoder on the current dataset, the transformer autoencoder provides improved latent-space representation quality and scalability for future extensions to larger and more complex datasets. Transformer Autoencoder combined with Random Forest provides the most accurate

Table 6
2C pulse discharge comparison between the transformer autoencoder and other simple model.

	Epoch	MSE	RMSE	MAE	R^2
Transformer Autoencoder + Random Forest	10	0.000276	0.001661	0.000638	0.999723
1D CNN + Random Forest	300	0.003590	0.059915	0.042155	0.973467
Basic Autoencoder + Random Forest	200	0.028043	0.167459	0.117586	0.792733
LSTM + Random Forest	30	0.013458	0.116007	0.079436	0.856040

and efficient prediction for 2C pulse discharge data, requiring far fewer training epochs and achieving the best scores in all metrics. 1D CNN + Random Forest offers reasonable accuracy, but with higher training cost and error. Both LSTM and Basic Autoencoder models are noticeably less effective for this task, with substantially higher errors and lower R^2 values.

The proposed transformer autoencoder framework was compared to the representative state-of-the-art internal temperature estimation models. Table 7 presents a summary of performance in terms of accuracy, computational cost, generalisation capability, and suitability to embedded or edge deployment. The Transformer autoencoder contains 15,475 trainable parameters, placing it well within the 0–1 M range typically associated with very low-parameter models [41]. Operating in this ultra-small category offers several meaningful advantages for practical deployment. Models of this size are highly efficient, require minimal computational resources, and can run smoothly on hardware with strict memory or processing constraints [42]. They also reduce energy consumption and operational costs, making them particularly suitable for resource-limited environments or cost-sensitive applications. With lightweight architectures are ideal for real-time inference, edge devices, and scenarios where responsiveness and accessibility are critical. According to the recent literature, the data-based, physics-based, and physics-informed strategies are good, and most high-precision strategies are costly to calculate, which is a drawback of embedded battery management systems [3,43–45]. The LSTM-based models can help provide a better approach to modelling the time information, although they can expand to larger memory footprints [43]. Estimators based on CNNs (especially those that make use of virtual sensors where the spatial temperature fields are used to improve the information propagation) have heavier architectures, on the contrary, and are difficult to deploy in, e.g., low-power microcontrollers [3,45]. PINNs are also more realistic to extrapolate and interpretable, though a significant amount of training and domain modelling is needed [32,44,45]. Conversely, the suggested transformer autoencoder has with high accuracy and low complexity of computation. It possesses a minimal architecture in latent space, a shorter inference pathway, and scalability properties and can be particularly applied to real-time applications at the edge. The fact, that the model is easier to apply considering the real-world duty profiles without necessarily introducing extra sensors and more intricate hybrid modelling to achieve the equivalent results also justifies the fact that the model is more pragmatic to generalise the concept of cross-C-rate across various levels, the model is also better suited to be deployed in comparison to the recent compact models [46].

The statistical comparison in Table 8 highlights the superior performance of the proposed transformer autoencoder & random forest method against previous approaches. The direct numerical comparison across studies is inherently limited due to variations in input features, battery types, dataset properties, and ground-truth definitions, therefore, it should be interpreted qualitatively rather than as one-to-one performance. Previous approaches, utilising inputs of voltage, current, and power, the method achieves notably lower error metrics, with MSE < 0.0275, RMSE <0.1659, and MAE < 0.0644, outperforming studies such as the CNN (MSE: 0.047), Hybrid LTNN & UKF (RMSE: 0.18) and GRU-RNN (MAE: 0.074). Additionally, it demonstrates exceptional accuracy (> 99.97 %), surpassing the >97–99 % range reported in earlier works like LSTM and Kalman Filter & Regression. While some methods, such as, NARX & FFNN focus on temperature error (± 0.5 K), the current study provides a more comprehensive evaluation across multiple metrics. The results underscore the efficiency of combining transformer-based architectures with diverse input parameters to enhance predictive precision in energy or temperature-related applications.

4.5. Drive cycle analysis of the transformer autoencoder

To further analyse the performance of the transformer autoencoder, a vehicle model was developed with a battery pack comprising 100 cells connected in series and 50 cells in parallel. Operational data for an individual cell were extracted from the battery pack. The tractive force $F_{(t)}$ required at the wheels must overcome several resistive forces, including aerodynamic drag, rolling resistance, and gravitational forces arising from road gradients, while also providing the force necessary for acceleration. The governing equation for the vehicle's longitudinal dynamics, Vehicle and environmental parameters, and powertrain and battery parameters are presented in section S3 of supplementary information. The parameters are for a typical mid-size saloon and per-cell

Table 7
Comparison of recent core-temperature estimation methods.

Method	Reported Performance	Computational Characteristics	Generalisation Capability	Edge Deployment Potential
Transformer Autoencoder + RF (this paper)	$R^2 \geq 0.99$	low parameter count (<0.2 M); fast inference; short training time (~2 min)	Strong cross-C-rate generalisation; robust latent structure	High — compact, low-latency, highly suitable for MCU-class devices
LSTM fused with numerical electro-thermal model [43]	Significant RMSE reduction vs standalone data-driven models	Moderate complexity: LSTMs require more memory than small transformers	Improved robustness due to physics fusion	Moderate — higher compute and memory overhead on MCUs
2D-CNN joint CT and SOC estimation [3]	High CT and SOC estimation accuracy across operating conditions	CNN architecture heavier (0.5 M–5 M parameters)	Good when surface temperature sensors are available	Moderate — additional sensors and higher compute cost
Physics-informed neural networks (PINNs) [44,45]	Strong extrapolation and physics consistency; typically, <1 °C error	High training cost; inference lighter after model distillation	Excellent extrapolation under off-nominal conditions	Moderate–Low — distillation needed for real-time embedded use
CNN with virtual thermal sensors [32]	MSE reported as low as 0.047 for pack-level mapping	Deep CNN (1 M–10 M parameters); pack-level computations heavy	Strong spatial mapping but dataset-specific	Low–Moderate — compute demands too high for most BMS MCUs
Multi-encoder autoencoders for SOH/thermal representation [46]	Robust latent features under small-sample conditions	Compact AE design (20 k–200 k parameters)	Good generalisation with limited training data	High — lightweight and suitable for embedded systems

Table 8
Core temperature comparison between this research and previous studies.

Method	Battery type	Inputs	Input type/ test conditions	MSE	RMSE	MAE	R^2 / Accuracy	Ref.
Transformer Autoencoder & Random Forest	Li-ion NCR18650	Voltage, Current, Power	C-rates Pulse discharge & Drive cycles/ 25 °C.	< 0.0275	< 0.1659	< 0.0644	> 99 %	This study
LSTM (Multi-step Forecasting)	Li-ion	Surface Temp., SOC, heat Q, Ambient Temp., Temp. difference	–	–	–	–	> 97 %	[27]
Kalman Filter & Regression	LiPo & $LiFePO_4$	Voltage, Surface Temp., Current	Synthetic data of Charge/ discharge voltage, current, & Surface Temp.	–	–	–	> 96–99 %	[28]
Hybrid LTNN & UKF	Li-ion	Fibre-optic Surface Temp. Internal temp. Current & voltage	CCCV charge- CC discharge/ 25 °C.	–	0.18	–	–	[29]
GRU-RNN	$LiFePO_4$	Voltage, current, SOC, Surface Temp., Ambient Temp.	Constant charge/discharge cycles & Drive cycles/ Room temp.	–	–	0.074	–	[30]
NARX & FFNN	Prismatic Li-ion	Current, Voltage, SOC, heat Q, cooling system temp.	Laboratory experiments	–	–	–	± 0.5 K	[31]
CNN	Ternary polymer Li-ion	External Temp.	Constant discharge cycles/ 25 °C.	0.047	–	–	–	[32]

SOC, current, and power were calculated by dividing the battery pack voltage by the number of cells in series, dividing the battery current by the number of cells in parallel, and then multiplying the resulting per-cell voltage and current to obtain the per-cell power. The drive cycles used in the simulation include the WLTC, Artemis Urban, and Artemis Motorway 130 profiles, because these standard cycles are relatively short, an additional custom drive cycle was generated in Python to ensure that the battery state of charge (SOC) reaches zero over the full duration of the simulation (SOC 80 % to 0 %) Fig. 7 presents the input variables for each of the drive cycles and latent temperature representation from the transformer autoencoder.

Fig. 8 presents a comparison of the temperature estimates produced by the transformer-based autoencoder model against those obtained from the physics-based thermal model across four different drive cycles: (a) WLTC, (b) Artemis Urban, (c) Artemis Motorway 130, and (d) a Python-generated drive cycle. The corresponding quantitative performance metrics are summarised in Table 9. For the WLTC drive cycle, the model achieves a R^2 of 0.9958, indicating an excellent fit to the reference thermal behaviour. Similarly, the Artemis Urban and Artemis Motorway 130 cycles yield R^2 values of 0.9948 and 0.9937, respectively,

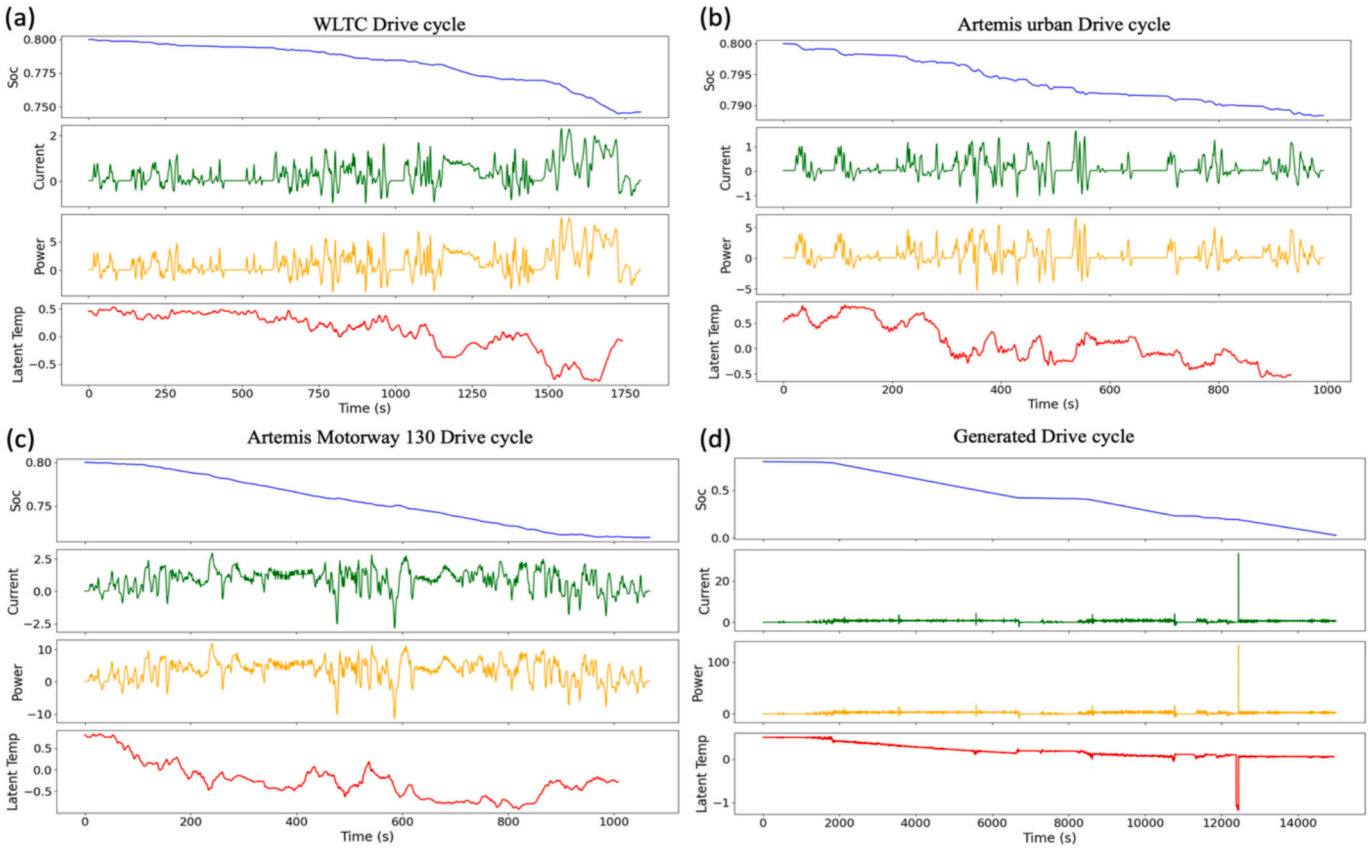

Fig. 7. Presents the final input variables of a cell from the vehicle model to the transformer autoencoder and latent temperature. (a) WLTC, (b) Artemis urban, (c) Artemis motorway 130 drive cycles and (d) Python generated drive cycle.

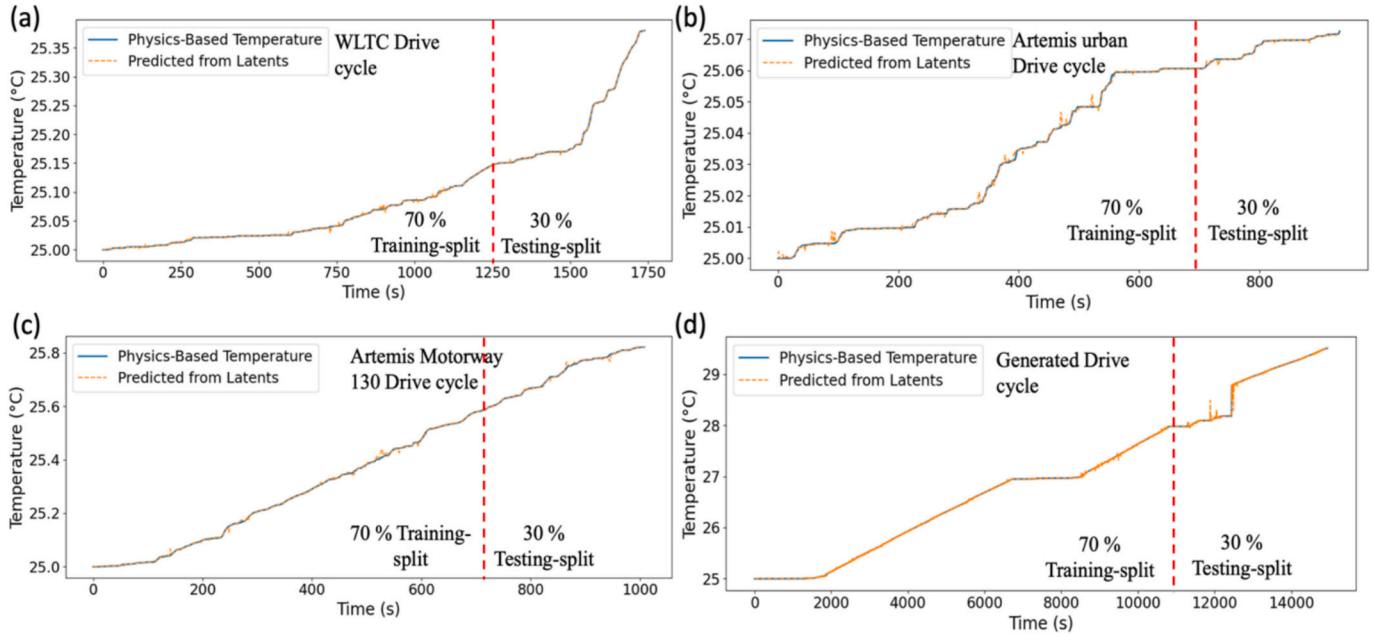

Fig. 8. Resulting prediction from latent temperature for (a) WLTC, (b) Artemis Urban, (c) Artemis Motorway 130, (d) generated profiles, respectively.

Table 9
Quantitative results of various drive cycles with physical-based temperature and transformer autoencoder model.

Drive cycle	Epoch	Loss	MSE	RMSE	MAE	R^2
WLTC	20	0.2094	0.0001	0.0014	0.0007	0.9958
Artemis urban	20	0.2442	0.0001	0.0008	0.0003	0.9948
Artemis motorway 130	20	0.2282	0.0001	0.0048	0.0027	0.9927
Generated	20	0.2724	0.0002	0.0086	0.0036	0.9979

both reflecting consistently high predictive accuracy. The Python-generated drive cycle exhibits the highest agreement, with an R^2 of 0.9979 and the cell highest temperature of ~30 °C, demonstrating the model's strong generalisation capability even for non-standard operating profiles. Overall, these results confirm that the transformer-based autoencoder closely reproduces the temperature dynamics predicted by the physics-based thermal model across a wide range of driving conditions.

5. Conclusions

Non-invasive, data-driven core temperature estimation is a critical enabler for advanced battery management systems (BMS), supporting safer and more reliable operation of Li-ion energy storage in demanding applications. Pulse discharge testing were collected from a Panasonic NCR18650B cell at three different C-rates and used as input to the transformer autoencoder. Which replicates real-world operating conditions, provides valuable insights that aid in the development of intelligent BMS, and machine learning models aimed at enhancing battery safety, reliability, and efficiency. Leveraging its ability to capture long-range dependencies and efficiently process sequential data, the transformer autoencoder emerges as a powerful tool for intelligent, data-driven battery management. In this approach, the autoencoder's encoder compresses multivariate battery signal inputs into a compact latent representation, while the decoder reconstructs the temporal evolution of the signals. This latent vector acts as a temporally structured representation for the cell's core temperature.

Model results show that at the highest pulse discharge rate (2C), the model yielded its highest MSE of 0.0006 ± 0.0001 %, RMSE of 0.0247 ± 0.0025 %, and R^2 of 0.9994 ± 0.0001 % demonstrating accurate tracking of rapid thermal dynamics. As the C-rate decreased, performance improved, at 1C, the MSE was 0.0002 ± 0.0001 %, RMSE was 0.0153 ± 0.0022 %, and R^2 reached 0.9998 ± 0.0001 %; at 0.5C, the model achieved its best performance with an MSE of 0.0002 ± 0.0003 %, RMSE of 0.0139 ± 0.0070 %, and R^2 of 0.9998 ± 0.0002 %. These consistently strong metrics confirm the transformer autoencoder's effectiveness in modelling temperature behaviour across varying discharge rates. Core temperature estimates obtained from the transformer-based model and a physics-based thermal model, using random forest regression, showed closely aligned results. At 2C, the latter achieved an MSE of 0.0276 ± 0.0004 °C, an RMSE of 0.1661 ± 0.0013 °C, an MAE of 0.0638 ± 0.0006 °C, and an R^2 of 0.9997 ± 0.0004 %, outperforming results reported in previous studies. The 2D PCA projection shows smooth temporal progression, while the t-SNE visualisation highlights clear disentanglement of complex patterns for detecting transitions and cycles. This indicates that the transformer autoencoder's ability to extract meaningful latent representations enhances temperature estimation accuracy and overall thermal modelling performance. The model's improved performance with decreasing C-rates further highlights its robustness under low-dynamic regimes, where core thermal behaviour remains more stable. The drive cycle analysis of WLTC, Artemis Urban, Artemis Motorway 130, and a Python-generated drive cycle also showed high accuracy with $R^2 > 0.99$.

Future work will focus on extending the model to cover a broader range of operating conditions, including varying ambient temperatures, additional C rates, cycles over battery lifespan, different cell chemistries, and pulse/rest durations. Additionally, efforts will be made to enable real-time, on-device deployment for practical integration into battery management systems.

CRediT authorship contribution statement

Mustehsan Beg: Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Keith M. Alcock:** Writing – review & editing, Software, Data curation. **Vishnu Sam:** Writing – review & editing, Methodology, Investigation. **Sanjay Rakshit:** Writing – review & editing, Methodology, Formal analysis. **Sambit Paul:** Writing – review & editing,

Software, Methodology. **Hongnian Yu:** Writing – review & editing, Supervision, Resources, Project administration. **Keng Goh:** Writing – review & editing, Supervision, Software, Resources, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the Edinburgh Napier University SCEBE Starter Grant.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.applthermaleng.2025.129552.

Data availability

Data will be made available on request.

References

[1] M.J. Lain, E. Kendrick, Understanding the limitations of lithium ion batteries at high rates, J. Power Sources 493 (2021).
[2] T. Heenan, et al., Mapping internal temperatures during high-rate battery applications, Nature 617 (7961) (2023) 507–512.
[3] Y. Li, C. Ma, K. Liu, L. Chang, C. Zhang, B. Duan, A novel joint estimation for core temperature and state of charge of lithium-ion battery based on classification approach and convolutional neural network, Energy 308 (2024) 132721.
[4] A.G. Li, M. Berecibar, M. Preindl, Nonlinear characterization of Lithium-ion batteries with bipolar pulsing, IEEE Trans. Ind. Electron. 71 (10) (2024) 12983–12990.
[5] W. Li, K.E. Law, Deep learning models for time series forecasting: A review, IEEE Access 12 (2024) 92306–92327.
[6] A. Vaswani, et al., Attention is all you need, Adv. Neural Inf. Proces. Syst. 30 (2017).
[7] K. Berahmand, F. Daneshfar, E.S. Salehi, Y. Li, Y. Xu, Autoencoders and their applications in machine learning: a survey, Artif. Intell. Rev. 57 (2) (2024) 28.
[8] K.M. Alcock, K. Goh, M. Beg, S. Melendi-Espina, M. Hernaez, Encapsulated U-shape lossy mode resonance optical fibre sensor for temperature quantification of lithium-ion batteries, Sensors Actuators A Phys 395 (2025) 117004.
[9] K.M. Alcock, et al., Individual cell-level temperature monitoring of a lithium-ion battery pack, Sensors 23 (9) (2023) 4306.
[10] P. Markapudi, et al., Nitrate pollution mapping for reservoirs using flexible sensors integrated with underwater robot, IEEE Internet Things J 12 (2025) 39172–39180.
[11] W. He, R. Zhang, X. Wang, S. Wang, A.I. Gavrilov, Lithium Battery Health State Prediction Based on LSTM, in: 2025 IEEE 20th Conference on Industrial Electronics and Applications (ICIEA), IEEE, 2025, pp. 1–6.
[12] H. Chen, et al., State of Charge Estimation for Lithium-ion Battery Using Long Short-Term Memory Networks, in: Journal of Physics: Conference Series 2890, IOP Publishing, 2024, p. 012024, no. 1.
[13] N. Ranjan, R.K. Inapakurthi, EV Lithium-Ion Battery Temperature Prediction Using Machine Learning, in: 2024 IEEE 4th International Conference on Sustainable Energy and Future Electric Transportation (SEFET), IEEE, 2024, pp. 1–6.
[14] B. Sridhar, S. Allirani, R.R. Shafi, K.R. Abinav, A. Siddharth, Electric Vehicle Battery Temperature Prediction Using LSTM Algorithm, in: 2024 IEEE 4th International Conference on Sustainable Energy and Future Electric Transportation (SEFET), IEEE, 2024, pp. 1–6.
[15] H. Zhao, C. Liao, C. Zhang, L. Wang, L. Wang, State-of-charge estimation of lithium-ion battery: joint long short-term memory network and adaptive extended Kalman filter online estimation algorithm, J. Power Sources 604 (2024) 234451.
[16] A. Mousaei, Y. Naderi, I.S. Bayram, Advancing state of charge management in electric vehicles with machine learning: a technological review, IEEE Access 12 (2024) 43255–43283.
[17] L. Sun, X. Huang, J. Liu, J. Song, S. Wu, Remaining useful life prediction of lithium batteries based on jump connection multi-scale CNN, Sci. Rep. 15 (1) (2025) 32873.
[18] A. Saxena, A.D. Kumar, Enhancing Lithium-Ion Battery Reliability Through Machine Learning: A CNN-Based Approach for RUL Estimation, in: 2025 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA), IEEE, 2025, pp. 1–6.
[19] F.-M. Zhao, D.-X. Gao, Y.-M. Cheng, Q. Yang, Application of state of health estimation and remaining useful life prediction for lithium-ion batteries based on AT-CNN-BiLSTM, Sci. Rep. 14 (1) (2024) 29026.
[20] S. Luo, et al., State of Charge Estimation of Lithium-ion Battery Pack based on CNN-LSTM, in: 2024 IEEE 10th International Conference on Underwater System Technology: Theory and Applications (USYS), IEEE, 2024, pp. 1–6.
[21] H. Park, Y.S. Kim, Y.-J. Shin, Real-time state estimation of lithium-ion battery using image-based regression with CNN, in: 2024 10th International Conference on Condition Monitoring and Diagnosis (CMD), IEEE, 2024, pp. 390–393.
[22] F. Xu, F. Yang, Z. Fei, Z. Huang, K.-L. Tsui, Life prediction of lithium-ion batteries based on stacked denoising autoencoders, Reliab. Eng. Syst. Saf. 208 (2021) 107396.
[23] J. Chen, X. Feng, L. Jiang, Q. Zhu, State of charge estimation of lithium-ion battery using denoising autoencoder and gated recurrent unit recurrent neural network, Energy 227 (2021) 120451.
[24] M. Wei, M. Ye, Q. Wang, J.P. Twajamahoro, Remaining useful life prediction of lithium-ion batteries based on stacked autoencoder and gaussian mixture regression, J. Energy Storage 47 (2022) 103558.
[25] M.O. Tarar, I.H. Naqvi, Z. Khalid, M. Pecht, Accurate prediction of remaining useful life for lithium-ion battery using deep neural networks with memory features, Front. Energy Res. 11 (2023) 1059701.
[26] W. Xia, J. Xu, B. Liu, H. Duan, A novel denoising autoencoder hybrid network for remaining useful life estimation of lithium-ion batteries, Energy Sci. Eng. 12 (8) (2024) 3390–3400.
[27] M. Li, C. Dong, X. Yu, Q. Xiao, H. Jia, Multi-step ahead thermal warning network for energy storage system based on the core temperature detection, Sci. Rep. 11 (1) (2021) 15332.
[28] S. Surya, A. Chhetri, V. Rao, Kalman filter—machine learning fusion for core temperature estimation in Li-ion batteries, J. Energy Storage 113 (2025) 115656.
[29] Z. Wei, et al., Machine learning-based hybrid thermal modeling and diagnostic for lithium-ion battery enabled by embedded sensing, Appl. Therm. Eng. 216 (2022) 119059.
[30] O. Ojo, X. Lin, H. Lang, Y. Kim, A recurrent neural networks approach for estimating the core temperature in lithium-ion batteries, in: Proc. Can. Soc. Mech. Eng. Int. Congr., 2020, pp. 1–6.
[31] J. Kleiner, M. Stuckenberger, L. Komsiyska, C. Endisch, Real-time core temperature prediction of prismatic automotive lithium-ion battery cells based on artificial neural networks, J. Energy Storage 39 (2021) 102588.
[32] M. Wang, W. Hu, Y. Jiang, F. Su, Z. Fang, Internal temperature prediction of ternary polymer lithium-ion battery pack based on CNN and virtual thermal sensor technology, Int. J. Energy Res. 45 (9) (2021) 13681–13691.
[33] Y. Liu, Z. Huang, Y. Wu, L. Yan, F. Jiang, J. Peng, An online hybrid estimation method for core temperature of Lithium-ion battery with model noise compensation, Appl. Energy 327 (2022) 120037.
[34] P. Saechan, I. Dhuchakallaya, Numerical investigation of air cooling system for a densely packed battery to enhance the cooling performance through cell arrangement strategy, Int. J. Energy Res. 46 (14) (2022) 20670–20684.
[35] E. Samaniego, et al., An energy approach to the solution of partial differential equations in computational mechanics via machine learning: concepts, implementation and applications, Comput. Methods Appl. Mech. Eng. 362 (2020) 112790.
[36] M.S. Eshaghi, C. Anitescu, M. Thombre, Y. Wang, X. Zhuang, T. Rabczuk, Variational physics-informed neural operator (VINO) for solving partial differential equations, Comput. Methods Appl. Mech. Eng. 437 (2025) 117785.
[37] M. Greenacre, P.J. Groenen, T. Hastie, A.I. d'Enza, A. Markos, E. Tuzhilina, Principal component analysis, Nat. Rev. Methods Primers 2 (1) (2022) 100.
[38] M.C. Cieslak, A.M. Castelfranco, V. Roncalli, P.H. Lenz, D.K. Hartline, T-distributed stochastic neighbor embedding (t-SNE): a tool for eco-physiological transcriptomic analysis, Mar. Genomics 51 (2020) 100723.
[39] J. Song, H. Wang, Y. Liu, R. Wang, K. Wang, Enhancing variational autoencoder for estimation of lithium-ion batteries state-of-health using impedance data, Energy 337 (2025) 138739.
[40] Y. Liu, Q. Li, K. Wang, Revealing the degradation patterns of lithium-ion batteries from impedance spectroscopy using variational auto-encoders, Energy Storage Mater. 69 (2024) 103394.
[41] Z. Xu, A. Zeng, Q. Xu, FITS: Modeling time series with $10 k $ parameters, in: arXiv preprint arXiv:2307.03756, 2023.
[42] G. Menghani, Efficient deep learning: a survey on making deep learning models smaller, faster, and better, ACM Comput. Surv. 55 (12) (2023) 1–37.
[43] A. Yuan, T. Cai, H. Luo, Z. Song, B. Wei, Core temperature estimation of lithium-ion battery based on numerical model fusion deep learning, J. Energy Storage 102 (2024) 114148.
[44] K. Shen, et al., Physics-informed machine learning estimation of the temperature of large-format lithium-ion batteries under various operating conditions, Appl. Therm. Eng. 269 (2025) 126200.
[45] F. Wang, Z. Zhai, Z. Zhao, Y. Di, X. Chen, Physics-informed neural network for lithium-ion battery degradation stable modeling and prognosis, Nat. Commun. 15 (1) (2024) 4332.
[46] C. Liu, S. Wang, Z. Ma, S. Guo, Y. Qin, A multi-encoder BHTP autoencoder for robust Lithium battery SOH prediction under small-sample scenarios, Batteries 11 (5) (2025) 180.