Applied Energy 426 (2026) 128639

Contents lists available at ScienceDirect

Applied Energy

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

Uncertainty-aware anomaly detection for wind turbines using a physics-informed spatiotemporal feature modelling model

Hao Duan [a], Qiang Zhang [a,b], Huahao Ou [a], Chun Li [a,*], Wanfu Zhang [a], Zifei Xu [b,c,**]

[a] School of Energy and Power Engineering, University of Shanghai for Science and Technology, Shanghai, China
[b] Liverpool Logistics Offshore and Marine Research Institute, Liverpool John Moores University, Liverpool, UK
[c] Department of Civil and Environmental Engineering, University of Liverpool, Liverpool, UK

HIGHLIGHTS

- Proposed a physics-informed probabilistic deep learning framework.
- Introduced a physics-informed regularization strategy.
- Established An uncertainty-aware multi-sensor anomaly fusion framework.

ARTICLE INFO

Keywords:
Anomaly detection
Physics-informed neural network
Prognostics and health management
Wind turbine
SCADA data

ABSTRACT

Wind turbines are prone to failures in critical subsystems such as the drivetrain and generator, making reliable early anomaly detection essential for safe and stable operation. With the increasing availability of monitored time-series data, data driven methods have emerged as an effective basis for anomaly detection and have demonstrated considerable potential for improving the reliability and safety of wind turbines. Nevertheless, conventional data driven approaches often struggle to adequately characterize the temporal dependencies in wind turbine operational data, the correlations among monitored variables in the feature space, and the physical constraints governing system operation, resulting in limited detection accuracy and robustness under varying operating conditions. To address these limitations, this paper proposes a physics-informed deep learning framework based on multi-sensor data, which integrates spatiotemporal feature modeling with physical consistency to improve the reliability of early anomaly detection and support prognostic functionality for wind turbines. The proposed framework is validated using a real SCADA dataset collected from two wind farms. Experimental results demonstrate that the proposed PITG achieves the best predictive performance across four datasets, with an average RMSE of 0.1397 and an average R^2 of 97.21%. Meanwhile, the physics-informed embedding module effectively reduces prediction uncertainty, decreasing the maximum prediction variance by approximately 27.25% compared with the baseline model. Combined with the inverse-variance multi-sensor fusion strategy, the proposed framework achieves an F1-score of 0.566 for system-level anomaly detection, validating its effectiveness and robustness under complex operating conditions.

1. Introduction

The scale of wind turbines continues to expand, and their deployment is progressively moving from nearshore areas to deep sea regions [1]. Compared to onshore or nearshore wind turbines, offshore wind turbines require more advanced health monitoring and assessment methods to ensure stable operation, as they are subjected to harsh marine environments characterized by high salinity and humidity during

Abbreviations: PHM, Prognostics and Health Management; SCADA, Supervisory Control and Data Acquisition; NBM, Normal Behavior Model; CNN, Convolutional Neural Network; LSTM, Long Short-Term Memory; BiGRU, Bidirectional Gated Recurrent Unit; MC, Monte Carlo; PITG, Physics-Informed Transformer–GRU; TG, Transformer-GRU Deep Neural Network; AE, Autoencoder; MLP, Multilayer Perceptron; MCNN, Multi-channel Convolutional Neural Network; GRU, Gated Recurrent Unit; DeepHPM, Deep Hidden Physics Model.

* Corresponding author.
** Corresponding author at: Liverpool Logistics Offshore and Marine Research Institute, Liverpool John Moores University, Liverpool, UK.
E-mail addresses: lichun@usst.edu.cn (C. Li), zifei.xu@liverpool.ac.uk, z.xu@ljmu.ac.uk (Z. Xu).

https://doi.org/10.1016/j.apenergy.2026.128639
Received 6 April 2026; Received in revised form 20 July 2026; Accepted 4 August 2026
Available online 8 August 2026
0306-2619/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

their service life, making them more prone to failures. Environmental operating conditions significantly affect the reliability of their key subsystems (e.g., the drive train system), leading to increased maintenance costs [2,3]. The effectiveness of health monitoring and assessment methods directly determines whether reliable operation and maintenance strategies can be developed for wind turbine drive trains. Therefore, there is an urgent need to advance research on health monitoring and assessment, particularly in Prognostics and Health Management (PHM) technologies.

As a general diagnostic method, anomaly detection is regarded as a fundamental component of PHM [4]. It provides a binary judgment regarding the health status of equipment, indicating whether the system is operating normally [5]. This capability makes anomaly detection particularly suitable for complex engineering systems, where identifying detailed fault mechanisms can be challenging [6,7]. Anomaly detection methods also align naturally with Supervisory Control and Data Acquisition (SCADA) based wind turbine monitoring [8,9]. Although the low sampling frequency of SCADA data limits its ability to capture fine-grained fault features, the time series still contain coarse-grained health information that can be used to assess system level anomalies [10].

Anomaly detection for mechanical systems has evolved a variety of methods, which can generally be categorized into three main types [11]: 1. Model-based methods: These rely on mathematical [12] or physical representations [13] of system dynamics, identifying anomalies by comparing the deviation between model-predicted responses and actual measured responses. Such methods offer strong interpretability but often require high-precision models and prior knowledge of system parameters. 2. Signal-based methods: These identify anomalies directly from measured signals such as vibration [14], acoustic [15], or current signals [16], employing common techniques including statistical analysis [17], spectral decomposition [18], and time-frequency feature extraction [19]. These methods are relatively simple to implement but are sensitive to noise and environmental changes, leading to limited robustness. In recent years, 3. Data driven methods (such as deep learning [20] and probabilistic models [21]) have gained increasing attention. These approaches leverage large-scale monitoring data to automatically learn underlying patterns of normal and faulty behavior, offering better adaptability to complex nonlinear systems. Consequently, they are particularly suitable for anomaly detection in wind turbine drive trains operating under variable and uncertain conditions.

Under abnormal operating conditions, onshore wind turbines tend to exhibit more distinct patterns in monitoring data, making anomalies easier to identify [22], In contrast, offshore wind turbines are subject to significant environmental uncertainties, leading to frequent and highly non-stationary changes in operational states, making it difficult to distinguish abnormal conditions from normal operating variations. To reduce reliance on prior knowledge of operating conditions and fault mechanisms, it is necessary to construct data-driven models capable of characterizing equipment health status [23]. Therefore, data-driven approaches based on the Normal Behavior Model (NBM) have been widely adopted for wind turbine anomaly detection [19].

For data-driven anomaly detection in wind turbines, detection performance is closely related to the temporal evolution of monitored variables. Consequently, existing studies typically focus on modeling the temporal dependencies inherent in monitoring data. Autoencoder-based reconstruction frameworks [24] and time-series prediction frameworks [22] are widely employed for temporal feature modeling in anomaly detection. Compared with reconstruction-based approaches, prediction frameworks establish temporal relationships between current and future states, making them more suitable for characterizing the dynamic evolution of monitored variables over time. However, because prediction tasks require learning the evolving dependencies among variables over time, achieving good generalization remains challenging.

To model the temporal dynamics of monitored variables, deep learning methods based on MLP [25,26], CNN [27,28], and RNN [29,30] have been applied to capture nonlinear dependencies. Among these, MLP is primarily used for modeling nonlinear relationships between features but lacks the capacity for temporal dependency modeling. Although CNN can extract local temporal patterns, its ability to represent long-range dependencies remains limited. RNNs rely on step-by-step recursive computation, which limits parallel processing efficiency and may lead to insufficient information propagation as sequence length increases, thereby restricting their application in long-sequence SCADA data modelling [31].

Attention-based methods address these limitations by assigning adaptive weights to sequence elements, reducing redundant information and enabling the modelling of long-range temporal dependencies [32–34]. These methods have also been integrated with other deep learning approaches for anomaly detection [27]. However, existing attention mechanisms are typically employed as auxiliary modules for feature weighting, with the spatiotemporal feature extraction process still largely dependent on the underlying network architecture. Consequently, attention-centered structured representation learning frameworks remain insufficiently explored, which may limit the ability of existing methods to capture complex temporal dependency patterns.

In the anomaly detection of wind turbine drivetrain systems, the appropriate selection of monitoring features is of significant importance for enhancing the accuracy and reliability of anomaly identification [52]. Drivetrain faults are typically accompanied by component performance degradation, increased energy losses, and variations in operational states, with different monitoring variables exhibiting distinct response characteristics to the fault evolution process [53]. For critical transmission components such as bearings, wear, lubricant degradation, and localized damage lead to increased mechanical friction losses, converting more mechanical energy into thermal energy, which consequently results in elevated bearing temperatures or deviations from normal temperature variation patterns. Existing studies have demonstrated that temperature signals can comprehensively reflect the thermal effects induced by mechanical friction, lubrication conditions, and load variations, and are therefore widely employed in wind turbine drivetrain health monitoring. Tutivén et al. established an anomaly detection model based on SCADA main bearing temperature data, demonstrating that bearing temperature can effectively reflect the health status of the main bearing [54]. Liu et al. utilized temperature-related features from wind turbine SCADA data to construct a deep autoencoder-based anomaly detection model, which achieved the identification of abnormal wind turbine states by learning the inter-variable correlations among monitoring variables such as temperature under normal operating conditions [55]. Bearing temperature not only serves as a direct response to the thermal state of the equipment but also embeds information regarding the internal friction characteristics and energy conversion processes within the transmission system [56].

Existing studies have adopted various physical information embedding strategies. Schröder et al. employed a transfer learning approach, using simulation data as physical priors during the pre-training phase, which were then transferred to build a NBM for wind turbines based on SCADA data, thereby enhancing anomaly detection performance [57]. Lee et al. selected SCADA monitoring variables associated with wind turbine faults and extracted fault-related features based on the relationships among these variables. These features were then integrated with an optimized XGBoost model to achieve anomaly and fault detection in wind energy systems [58]. Gao et al. incorporated the response relationships among different wind turbine monitoring variables as physical information into the development of a NBM. A heteroscedastic residual model was then established based on the characteristics of error fluctuations under varying operating conditions, enabling anomaly identification through the detection of deviations between the actual state and the expected state derived from physical relationships [59]. However, the aforementioned approaches mainly utilize physical prior data, variable correlations, or operational rules to assist model development, without incorporating physical information as explicit constraints during model training. As a result, their ability to guide feature learning with physical knowledge remains limited.

Furthermore, online anomaly detection for data streams also faces stability issues. As data continuously arrives, uncertainty can accumulate and affect the decision boundary, causing fluctuations in detection results [35,36]. Therefore, it is necessary to assess uncertainty during the detection process. Typically, statistics such as mean and variance are used to characterize it indirectly [37]. Davide et al. conducted anomaly detection experiments using various deep autoencoders, jointly employing statistics like reconstruction error, prediction mean, and prediction variance. They treated prediction variance as the model's uncertainty for that specific sample to improve anomaly discrimination [38]. However, relying solely on output statistics makes it difficult to explain the source of uncertainty. This interpretability can be enhanced by explicitly modelling model uncertainty. Bang et al. obtained the posterior predictive distribution through Bayesian inference and used variational inference to represent network weights as probability distributions, explicitly modelling uncertainty to achieve probabilistic remaining useful life prediction [39]. Yarin et al. demonstrated that using dropout in neural networks can be interpreted as a Bayesian approximation of a probabilistic model, directly providing uncertainty in predictions. This approach addresses the problem of representing model uncertainty without sacrificing computational complexity or test accuracy [40]. However, these methods do not consider incorporating the uncertainty information contained in prediction results into the fusion process of multi-sensor detection outcomes, thereby limiting the stability and reliability of multi-sensor anomaly assessment.

Overall, existing studies still present three major limitations: 1. The dynamic dependencies among multiple monitoring variables remain insufficiently modeled, limiting the generalization capability of prediction-based frameworks. 2. Existing methods mainly incorporate physical information through prior data, variable correlations, or operational rules, while explicit physical constraints are rarely integrated into the model training process, restricting the ability of physical knowledge to guide feature representation learning. 3. The uncertainty information associated with prediction results is generally neglected in the fusion of multi-sensor anomaly detection outcomes, which reduces the stability of multi-sensor anomaly assessment.

In response to the aforementioned issues, this study proposes an intelligent anomaly detection model based on a Physics-Informed Transformer-GRU Neural Network (PITG) and an uncertainty-aware multi-sensor fusion strategy, aiming to achieve reliable intelligent anomaly detection for wind turbine drivetrain systems. The proposed method develops a spatiotemporal feature modelling model, namely the Transformer-GRU neural network (TG), which utilizes the self-attention mechanism and gated recurrent units to capture feature correlations among monitoring variables and temporal evolution patterns, respectively, thereby achieving effective representation of multivariate time-series information. Furthermore, physical constraints are incorporated into the model training process by constructing a physics-informed loss function to constrain the physical consistency and dynamic variation characteristics of monitoring variables in the prediction results, thereby enhancing the capability of the model to characterize wind turbine operating states. Meanwhile, a sensor reliability evaluation indicator is established based on the uncertainty information of prediction errors, and an inverse-variance weighting strategy is employed to fuse multi-sensor anomaly results, improving the stability of anomaly assessment.

The main contributions of this study are summarized as follows:

1. A Transformer-GRU Neural Network (TG) based anomaly detection framework is proposed for wind turbine drivetrain systems. The proposed framework can model spatiotemporal dependencies in multivariate SCADA data by capturing correlations among variables in high-dimensional feature space and temporal evolution patterns across time, thereby improving the representation capability of complex operating characteristics.

2. A physics-informed learning strategy is developed by incorporating a thermal balance function into the model training process. This strategy constrains the predicted temperature responses according to the thermal equilibrium relationship among monitoring variables, enabling the model to learn representations that are more consistent with the thermal behavior of wind turbine drivetrain systems.

3. An uncertainty-aware multi-sensor fusion strategy is proposed to improve the reliability of anomaly assessment. The variance of prediction errors, which reflects both sensor measurement uncertainty and the capability of the normal behavior model to characterize sensor response patterns, is used to quantify sensor reliability. An inverse-variance weighting strategy is then adopted to fuse multi-sensor anomaly scores, reducing the contribution of sensors with higher uncertainty and improving the stability of anomaly detection.

The remainder of this paper is organized as follows. Section 2 describes the basic framework of anomaly detection and formulates the problem. Section 3 presents the details of the proposed methods. Section 4 describes the dataset, parameter settings throughout the detection process, and the supporting hardware. Section 5 provides case studies, including performance evaluation of the NBM, uncertainty analysis, and anomaly detection results. Section 6 concludes the paper.

2. Problem formulation

2.1. Task definition

Given a multivariate SCADA time series:

$$\mathbb{X} = \begin{bmatrix} \mathbb{X}_1^{\mathrm{T}} \\ \vdots \\ \mathbb{X}_T^{\mathrm{T}} \end{bmatrix}, \mathbb{X}_t = \left[x_{t,1}, x_{t,2}, \cdots, x_{t,N}\right]^{\mathrm{T}} \tag{1}$$

where $x_{t,1}$ is the scalar measurement of the i-th sensor at time t.

A sliding window of length L is used to construct the input samples, and the target variable at the future h-th step is taken as the prediction target.

$$\mathbb{X}_{t,L} = \begin{bmatrix} \mathbb{X}_{t-L+1}^{\mathrm{T}} \\ \vdots \\ \mathbb{X}_t^{\mathrm{T}} \end{bmatrix}, \mathbb{Y}_{t+h,L} = \begin{bmatrix} \mathbb{X}_{t-L+h+1}^{\mathrm{T}} \\ \vdots \\ \mathbb{X}_{t+h}^{\mathrm{T}} \end{bmatrix} \tag{2}$$

Thus, the training sample set under normal operating conditions can be obtained as $\mathbb{D} = \{(\mathbb{X}_{1,L}, \mathbb{Y}_{1+h,L}), \cdots, (\mathbb{X}_{T,L}, \mathbb{Y}_{T+h,L})\}$.

For the output of the NBM $\widehat{\mathbb{Y}}_{t+h,L}$, anomaly score is defined based on the prediction error as:

$$\mathbb{S}_{t,L} = \ell\left(\mathbb{Y}_{t+h,L}, \widehat{\mathbb{Y}}_{t+h,L}\right) \tag{3}$$

where $\ell(\cdot)$ can be defined as the absolute error, squared error, or RMSE. By introducing a threshold τ, the detection task can be converted into a binary decision problem:

$$\begin{cases} \tau = Q_{0.95}\left(\mathbb{S}_{t,L}\right) \\ \widehat{\mathbb{A}}_{t,L} = \mathbb{I}\left(\mathbb{S}_{t,L} > \tau\right) \end{cases} \tag{4}$$

where $\mathbb{I}(\bullet)$ denotes the indicator function, which outputs 1 if the condition within the parentheses is true, and 0 otherwise. $\widehat{\mathbb{A}}_{t,L}$ denotes the monitoring results within a time window of length L from $t-L+1$ to t.

2.2. Predictive NBM for anomaly detection

To realize anomaly detection for the wind turbine drivetrain system, the key lies in constructing a NBM capable of characterizing the statistical patterns and dynamic evolution of the system under normal operating conditions. Assume that the time-series signal x_t is drawn from two latent distributions: the normal state distribution $\mathbb{P}_N(x)$ and the abnormal state distribution $\mathbb{P}_{AN}(x)$. In practice, the available training data $\mathbb{D}$ are usually dominated by samples collected under normal operating conditions. A deep neural network is then employed to learn the distribution characteristics of the normal time-series data. In this context, a predictive autoencoder, consisting of an encoder and a decoder, is introduced as follows:

$$\mathbb{u}_{t,L} = f_{\varphi}\left(\mathbb{X}_{t,L}\right), \widehat{\mathbb{Y}}_{t+h,L} = g_{\psi}\left(\mathbb{u}_{t,L}\right) \tag{5}$$

where $f_{\varphi} : \mathscr{X} \to \mathscr{U}$ denotes the encoding mapping, $g_{\psi} : \mathscr{U} \to \mathscr{X}$ denotes the decoding mapping, and φ, ψ represents the network parameters. Let the overall parameter set be $\theta = \{\varphi, \psi\}$.

However, the predictive autoencoder directly maps the current latent representation u_t to the future output, which may neglect the temporal evolution characteristics contained in sequential signals. To capture the dynamic transition of the system state, an additional temporal encoder is introduced to predict the future latent representation $u_{t+h} = \Phi_{\omega}(u_t)$.Then, eq. (5) can be rewritten as:

$$\widehat{\mathbb{Y}}_{t+h,L} = g_{\psi}\left(\Phi_{\omega}\left(f_{\varphi}\left(\mathbb{X}_{t,L}\right)\right)\right) = G_{\theta}\left(\mathbb{X}_{t,L}\right), \theta = \{\varphi, \omega, \psi\} \tag{6}$$

where $\Phi_{\omega} \circ f_{\varphi}$ denotes the spatiotemporal feature encoding module, and g_{ψ} denotes the decoding module.

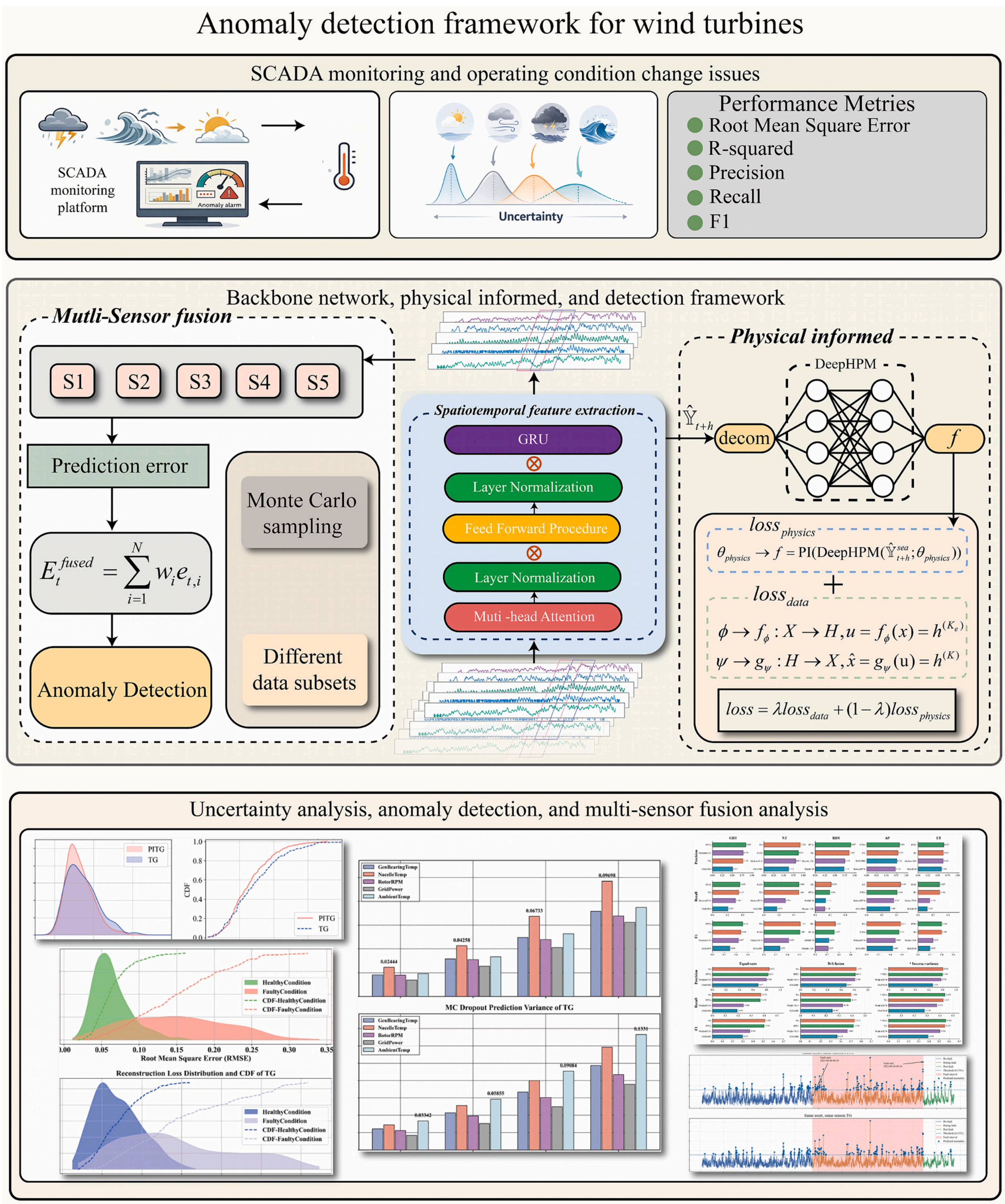


Fig. 1. The proposed anomaly detection framework.

The autoencoder is trained by minimizing the prediction loss on normal data:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \frac{1}{|\mathbb{D}|} \sum_{(\mathbb{X}_{t,L},\mathbb{Y}_{t+h,L})\in\mathbb{D}} \mathscr{L}\left(\mathbb{Y}_{t+h,L}, G_\theta\left(\mathbb{X}_{t,L}\right)\right) \tag{7}$$

where $\mathscr{L}(\cdot)$ denotes the loss function. After training, the anomaly score is defined based on the prediction error as:

$$\mathbb{S}_{t,L} = \ell\left(\mathbb{Y}_{t+h,L}, G_{\theta^*}\left(\mathbb{X}_{t,L}\right)\right) \tag{8}$$

3. Methodology

Fig. 1 illustrates the overall methodological framework proposed in this study for anomaly detection in wind turbine drivetrain systems. To address the issues identified in Section 2, a unified anomaly detection framework is developed, consisting of a spatiotemporal predictive backbone network, physics-informed regularization module, uncertainty analysis, and multi-sensor anomaly fusion framework.

3.1. *Spatiotemporal predictive NBM*

Fig. 2 illustrates the overall architecture of the proposed deep neural network model. To characterize the operational patterns of wind turbines under normal operating conditions, a spatiotemporal joint modeling network, TG is proposed. Based on a parameterized NBM $G_\theta(\cdot) = g_\psi(\Phi_\omega(f_\varphi(\cdot)))$ to capture the nonlinear mapping from historical multivariate monitoring sequences to future target states.

In this network, the self-attention module is responsible for modelling the global relationships among different features, while the GRU module is used to capture the dynamic temporal dependencies along the time axis. The future target state is then generated by the encoder output.

3.1.1. *Spatial feature extraction via self-attention*

In wind turbine SCADA data, different sensor variables often exhibit significant correlations, while each individual variable may also undergo complex variations. Relying solely on local statistical features or modelling each variable independently is insufficient to fully characterize the coordinated variation patterns among variables under complex operating conditions. To address this issue, a multi-head self-attention mechanism is introduced to model the global dependencies among different features within the input window.

For the input representation $\mathbb{X}_{t,L} \in \mathbb{R}^{L\times N}$, the query $\mathbb{Q}_h \in \mathbb{R}^{L\times m}$, key $\mathbb{K}_h \in \mathbb{R}^{L\times m}$, and value $\mathbb{V}_h \in \mathbb{R}^{L\times m}$ matrices are defined as:

$$\mathbb{Q}_h = \mathbb{W}^Q\mathbb{X}_{t,L}, \mathbb{K}_h = \mathbb{W}^K\mathbb{X}_{t,L}, \mathbb{V}_h = \mathbb{W}^V\mathbb{X}_{t,L} \tag{9}$$

where m denotes the dimensionality of the high-dimensional features. $\mathbb{W}^Q, \mathbb{W}^K, \mathbb{W}^V$ are learnable parameter matrices. Single-head attention captures the importance of different features in high-dimensional sensor representations through dot-product attention, and adaptively reweights these features based on the obtained attention scores. It can be expressed as:

$$\mathbb{H}_h = \mathrm{softmax}\left(\frac{\mathbb{Q}_h\mathbb{K}_h^{\mathrm{T}}}{\sqrt{d_K}}\right)\cdot\mathbb{V}_h \tag{10}$$

where $d_K \in \mathbb{R}^L$ is the dimension of the key vector. Furthermore, the multi-head attention mechanism projects the input into multiple different subspaces to learn various correlation patterns in parallel, which patterns in parallel, which can be expressed as:

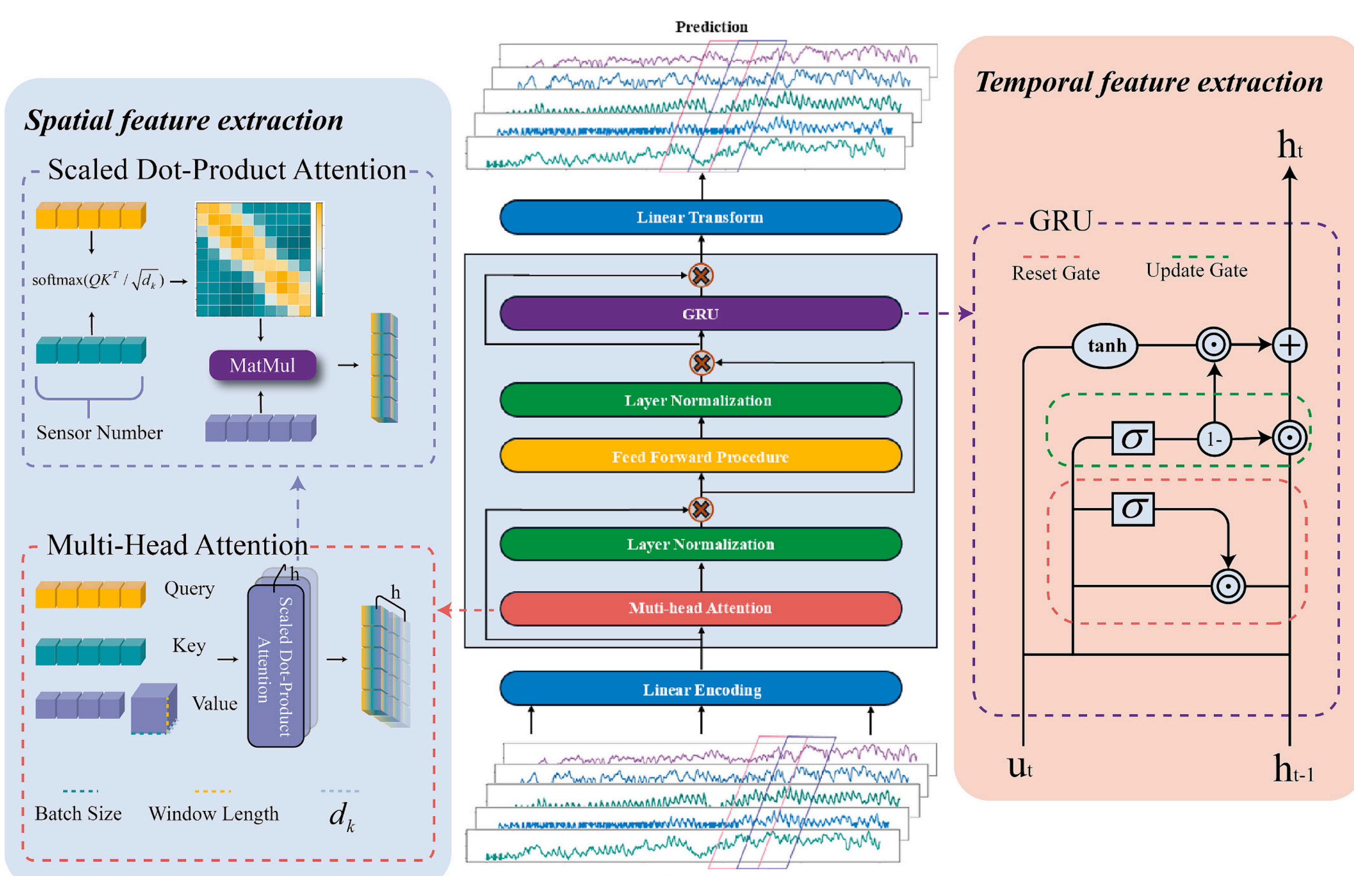


Fig. 2. The framework of the TG.

$$\widetilde{\mathbb{H}} = \mathrm{Concat}(\mathbb{H}_1, \ldots, \mathbb{H}_H) \cdot \mathbb{W}^O \tag{11}$$

where H is the attention heads, and $\mathbb{W}^O$ is the output projection matrix. $\widetilde{\mathbb{H}} \in \mathbb{R}^{L\times m}$ is the output of the Multi-head attention.

Through this mechanism, the model is able to learn the global dependencies among multivariate monitoring signals in different representation subspaces, thereby obtaining more discriminative spatial feature representations and further enhancing the separability between normal and abnormal behaviors.

The output of the spatial feature extraction module $\mathbb{u}_{t,L} \in \mathbb{R}^{L\times m}$ is obtained by passing through feedforward networks and layer normalization.

$$\mathbb{u}_{t,L} = \mathrm{LN}(\widetilde{\mathbb{H}} + \mathrm{FFN}(\widetilde{\mathbb{H}})) \tag{12}$$

3.1.2. Temporal feature extraction via GRU

On the basis of extracting the global correlations among variables, it is still necessary to further characterize the dynamic evolution of state variables over time. To this end, a GRU is employed to recursively model the temporal dynamics of the sequence representations generated by the self-attention module. By using the update gate and reset gate to regulate the fusion ratio between historical information and current input, the GRU can effectively capture long-sequence dependencies with relatively low computational complexity.

For the input representation $\mathbb{u}_{t,L} = [\mathbb{u}_{t-L+1}, \cdots, \mathbb{u}_t]$ at the future h-th step, the GRU computation can be expressed as:

$$\begin{bmatrix} \mathbb{q}_{t+h} \\ \mathbb{r}_{t+h} \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \end{bmatrix} \left(\begin{bmatrix} \mathbb{W}_q & \mathbb{U}_q \\ \mathbb{W}_r & \mathbb{U}_r \end{bmatrix} \begin{bmatrix} \mathbb{u}_{t,L} \\ \mathbb{h}_t \end{bmatrix} + \begin{bmatrix} \mathbb{b}_q \\ \mathbb{b}_r \end{bmatrix} \right) \tag{13}$$

$$\widetilde{\mathbb{h}}_{t+h} = \tanh(\mathbb{W}_h \mathbb{u}_{t+h} + \mathbb{U}_h(\mathbb{r}_{t+h} \odot \mathbb{h}_t) + \mathbb{b}_h) \tag{14}$$

$$\mathbb{h}_{t+h} = (1 - \mathbb{q}_{t+h}) \odot \mathbb{h}_t + \mathbb{q}_{t+h} \odot \widetilde{\mathbb{h}}_{t+h}) \tag{15}$$

where $\mathbb{h}_t$ and $\mathbb{h}_{t+h}$ denote the hidden states at the previous and current time steps, respectively, $\mathbb{q}_{t+h}$ is the update gate, $\mathbb{r}_{t+h}$ is the reset gate, $\widetilde{\mathbb{h}}_{t+h}$ is the candidate hidden state, $\sigma(\cdot)$ denotes the Sigmoid activation function, and $\odot$ represents the Hadamard product.

The current hidden state is transformed through a linear projection layer followed by layer normalization to obtain the predicted representation $\mathbb{u}_{t+h,L}$.

$$\mathbb{u}_{t+h,L} = \mathrm{LN}(\mathbb{W}_e \mathbb{h}_{t+h} + \mathbb{b}_e) \tag{16}$$

Through the recursive update of GRU, the model captures the temporal evolution patterns of monitoring variables and generates a temporal feature representation for future state prediction. Subsequently, the high-dimensional feature representations are transformed back into sensor-level features through a linear transformation.

$$\widehat{\mathbb{Y}}_{t+h,L} = g_\psi(\mathbb{u}_{t+h,L}) \tag{17}$$

$$g_\psi(\mathbb{u}_{t+h,L}) = \mathbb{W}_g \mathbb{u}_{t+h,L} + \mathbb{b}_g$$

where $\mathbb{W}_g$ and $\mathbb{b}_g$ represent the weight matrix and bias vector of the decoder, respectively, and the prediction output of the backbone network is obtained through this linear transformation.

3.2. Proposed physics-informed regularization module

Although the spatiotemporal predictive model is capable of learning complex mapping relationships from data, the latent representations obtained purely through a data driven manner often possess a high degree of freedom, lack an understanding of the underlying physical implications of temperature changes, and thus are insufficient to elucidate the intrinsic linkages between temperature anomalies and specific mechanical degradation processes. To address this issue, a physics-informed regularization module is introduced on top of the spatiotemporal predictive backbone network. By constraining the relationships among the predicted output features through a simplified heat balance equation, the module enhances the thermodynamic consistency and stability of the model predictions.

The output of the backbone network is denoted as $\widehat{\mathbb{Y}}_{t+h,L} = [\widehat{T}_{bear}(l), \widehat{T}_{NT}(l), \widehat{\omega}(l), \widehat{P}(l), \widehat{T}_{ET}(l)]^{\mathrm{T}} \in \mathbb{R}^{L\times 5}$ where the five predicted variables correspond to the bearing temperature, nacelle temperature, rotor speed, active power, and ambient temperature.

To constrain the relationships among these variables while preserving their local dynamic characteristics, the predicted sequence is first decomposed using a moving-average operator:

$$\widehat{\mathbb{Y}}_{t+h,L} = \widehat{\mathbb{Y}}^{sea}_{t+h,L} + \widehat{\mathbb{Y}}^{tr}_{t+h,L} \tag{17}$$

where, $\widehat{\mathbb{Y}}^{tr}_{t+h,L} = M_k(\widehat{\mathbb{Y}}_{t+h,L})$, $\widehat{\mathbb{Y}}^{sea}_{t+h,L} = \widehat{\mathbb{Y}}_{t+h,L} - \widehat{\mathbb{Y}}^{tr}_{t+h,L}$, $M_k(\bullet)$ denotes a moving-average operator with kernel size k.

Inspired by the Deep Hidden Physics Model proposed by Raissi et al. [41], a DeepHPM module is introduced as a physics-trend learner that restores $\widehat{\mathbb{Y}}^{sea}_{t+h,L}$ to $\widehat{\mathbb{Y}}_{t+h,L}$ by means of the predefined physics-informed formulation without resorting to purely data-driven reconstruction.

$$\widetilde{\mathbb{Y}}_{t+h,L} = \mathrm{DeepHPM}\left(\widehat{\mathbb{Y}}^{sea}_{t+h,L}; \theta_{physics}\right) \tag{18}$$

where, θ_{phsics} denotes the trainable parameters. The DeepHPM module adopts an autoencoder-inspired bottleneck structure that compresses the five dimensional input into a low-dimensional thermal state representation and subsequently maps it back to the original feature space:

$$\widetilde{\mathbb{Y}}_{t+h,L} = \mathbb{W}_2 \tanh\left(\mathbb{W}_1 \widehat{\mathbb{Y}}^{sea}_{t+h,L} + \mathbb{b}_1\right) + \mathbb{b}_2 \tag{19}$$

The bottleneck structure suppresses redundant fluctuations and measurement noise, thereby extracting the dominant coupled patterns among the bearing temperature, nacelle temperature, rotor speed, active power, and ambient temperature. The Tanh activation further captures nonlinear relationships and improves training stability.

A simplified discrete heat balance residual is constructed as [42]:

$$f = \widetilde{T}_{bear}(l) + \beta_1 \widetilde{T}_{bear}(l-1) + \beta_2 \widetilde{T}_{NT}(l) + \beta_3 \widetilde{\omega}^2(l) + \beta_4 \widetilde{P}(l) \tag{20}$$

$$loss_{physics} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} f^2} \tag{21}$$

where f quantifies the deviation of the predicted variables from the prescribed heat balance relationship. The discrepancy between the model output produced by data driven parameter updates and the predefined target labels is quantified using RMSE. By minimizing this metric, the model learns the optimal parameters θ_{phsics}.

$$loss_{data} = \sqrt{\frac{1}{L}\sum_{l=1}^{L} \left\| \mathbb{Y}_{t+h,L} - G_\theta(\mathbb{X}_{t,L}) \right\|^2} \tag{22}$$

where $G_\theta(\mathbb{X}_{t,L})$ represents the output from the data driven branch. Finally, the model is optimized end-to-end by minimizing a joint loss function:

$$loss = \lambda loss_{data} + (1-\lambda) loss_{physics} \tag{23}$$

where λ is a balancing coefficient used to regulate the relative contributions of different loss terms and to prevent the physical term from dominating the optimization process.

3.3. Multi-sensor anomaly fusion

In practical applications, SCADA data are susceptible to environmental disturbances, measurement noise, and the prediction stability varies across sensors. Sensors with larger prediction error variance are generally more strongly affected by disturbances, whereas those with smaller variance provide more stable condition information. Therefore, the inverse-variance weights are first calculated from the prediction errors within a predefined normal operating interval and are then fixed and applied to the entire test sequence for multi-sensor prediction error fusion.

The prediction output of the NBM can be expressed as:

$$\mathbb{Y}_{t+h,L} = G_{\theta^*}\left(\mathbb{X}_{t,L}\right) + \mathbb{e} \tag{24}$$

where:

$$\left(\mathbb{X}_{t,L}, \mathbb{Y}_{t+h,L}\right) \in \mathbb{D}_{normal} \tag{25}$$

$$e_i \sim \mathcal{N}\left(0, \sigma_i^2\right) \tag{26}$$

where e_i denotes the prediction error $\mathbb{e}$ of the i-th sensor, and σ_i^2 represents the variance of prediction errors, including measurement noise and data uncertainty. Due to differences in physical characteristics and monitoring environments among sensors, the uncertainty associated with each sensor varies:

$$\sigma_i^2 \neq \sigma_j^2, i \neq j \tag{27}$$

A larger σ_i^2 indicates greater prediction uncertainty under normal operating conditions, reflecting lower reliability of the corresponding sensor prediction. Therefore, this study utilizes prediction error variance within the normal operating interval to quantify sensor reliability and performs multi-sensor prediction error fusion.

For the i-th sensor, the prediction error is defined as:

$$e_i = \left(\begin{bmatrix} x_{t,i} \\ \vdots \\ x_{t+h,i} \end{bmatrix} \begin{bmatrix} \widehat{x}_{t,i} \\ \vdots \\ \widehat{x}_{t+h,i} \end{bmatrix}\right)^2 \tag{28}$$

Within the predefined normal operating interval $\mathbb{D}_{normal}$, the prediction error variance and normalized inverse-variance weight are calculated as:

$$\sigma_i^2 = \mathrm{Var}_{t \in \mathbb{D}_{normal}}(e_i), w_i = \frac{\left(\sigma_i^2 + \varepsilon\right)^{-1}}{\sum_{j=t}^{t+h}\left(\sigma_j^2 + \varepsilon\right)^{-1}} \tag{29}$$

where ε is a small positive constant introduced to avoid division by zero. The prediction error variance σ_i^2 characterizes the stability of sensor i under normal conditions. A sensor with a smaller prediction-error variance in the normal interval is assigned a larger fusion weight.

Once determined, the weights remain fixed and are used to fuse the prediction errors over the entire test sequence:

$$E_t^{\text{fused}} = \sum_{i=1}^{N} w_i e_{t,i}, t = 1, 2, \ldots, T \tag{30}$$

where E_t^{fused} represents the fused prediction error at time step t, which integrates the prediction deviations from all sensors according to their reliability weights.

4. Experimental setup

To validate the predictive capability, anomaly identification performance, and system level detection stability of the proposed PITG framework under complex operating conditions, this Section describes the experimental setup from three aspects: data description, data preprocessing and evaluation metrics, and comparative models with parameter settings.

Existing anomaly detection frameworks typically output binarized decision results, making it difficult to quantify the uncertainty associated with model predictions and thereby limiting their engineering reliability in practical applications. Fig. 3 illustrates the proposed PITG anomaly detection framework. The detailed implementation procedure is described as follows.

Step 1: Construction of the NBM

First, the raw SCADA data are preprocessed by removing outliers and missing values, followed by sample construction and feature extraction using a sliding-window strategy. To address gradient oscillation and reduced convergence efficiency caused by inconsistencies in variable scales under normal operating conditions, global normalization is applied to reduce the influence of scale differences and anomalous disturbances, while mini-batch processing is adopted to alleviate the computational burden. Finally, a predictive normal behavior model is constructed to learn low-dimensional representations of normal behavior data, thereby extracting key features and latent physical information.

Step 2: Physics-informed regularization

A physics-trend learning module is constructed based on the heat balance relationship among the predicted output variables, and the corresponding physical residual is used to characterize the coupling between bearing temperature and the related environmental and operational variables under normal operating conditions. The physical residual is then incorporated into the training objective as a soft constraint to improve the physical consistency and stability of the predictions, thereby enhancing the model's predictive and anomaly representation capabilities.

Step 3: Model training

Using the available normal dataset, the likelihood of the sensor observation data distribution is maximized to optimize the parameters of the NBM, thereby improving prediction accuracy.

Step 4: Prognostics

First, a normal behavior model is jointly constructed using multi-sensor data, and the prediction error between each sensor output and its corresponding observation is calculated to characterize the sensor specific operating condition. The inverse-variance weights are then estimated from the prediction-error variance within a predefined normal operating interval and are kept fixed for the entire test sequence. The weighted prediction errors from multiple sensors are fused to construct a system-level anomaly indicator. Finally, the 95th percentile of the fused errors in the normal interval is used as the detection threshold for system anomaly identification.

4.1. Data description and characterization

As shown in Table 1, two real world SCADA datasets collected from different wind farms were used in this study. Dataset 1 was obtained from a wind farm in East China and was primarily used to evaluate the performance of the normal behavior model. Dataset 2 [43] was collected from an onshore wind farm in Portugal and contains five operational subsets from different turbines, covering normal operation, a hydraulic system fault, and generator bearing failures. The sampling periods, event types, event intervals, and specific purposes of each subset are summarized in Table 1.

As shown in Table 2, five monitoring variables closely related to bearing thermal behavior were selected from each dataset. For Dataset 1, the main-shaft front-bearing temperature was taken as the target variable, while the main-shaft rear-bearing temperature, ambient temperature, nacelle temperature, and hub temperature were used as associated variables. For Dataset 2, the generator bearing temperature was selected as the target variable, together with nacelle temperature, rotor rotational speed, grid power, and ambient temperature.

Based on these variables, a corresponding heat balance relationship

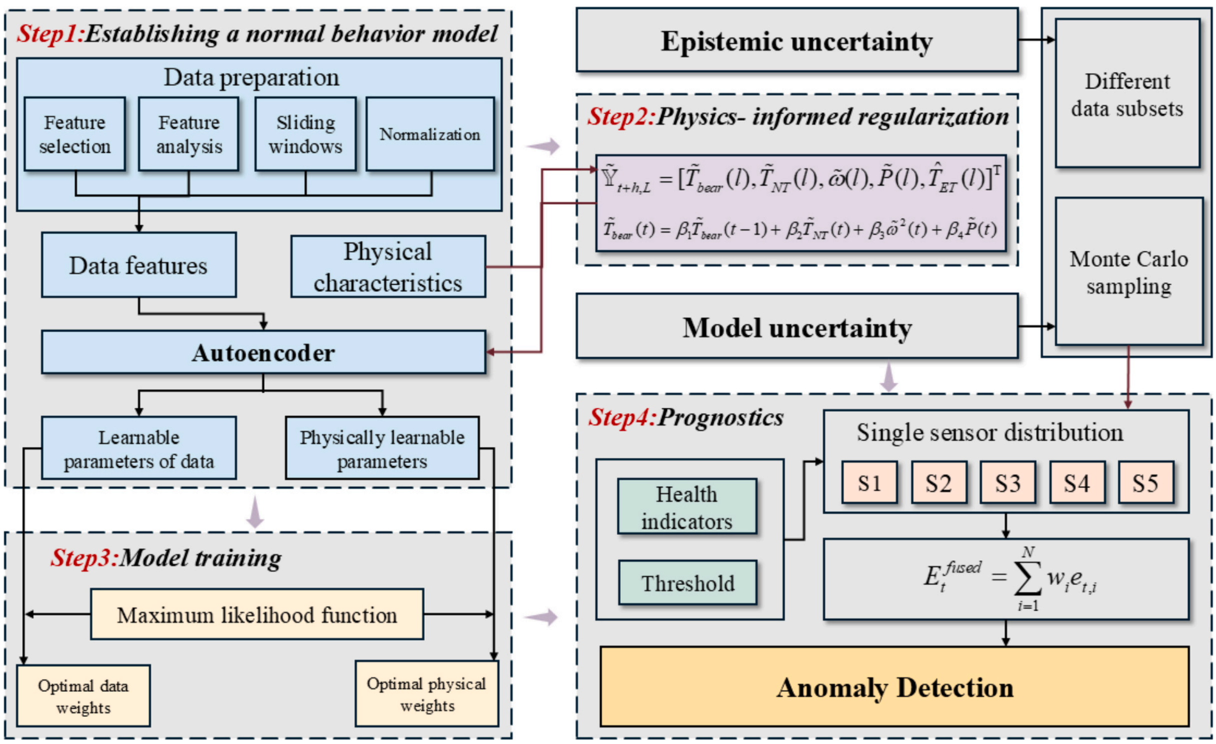


Fig. 3. Uncertainty-aware multi-sensor anomaly fusion framework for multi-sensor anomaly detection.

Table 1
Dataset overview. Dataset 2_a24 denotes Event 24 of Wind Farm A from Dataset 2; the same applies to the others.

Dataset	Turbine ID	Time span	Sampling interval	Number of valid samples	Operating condition	Event period	Main purpose
Dataset 1	1	2019.02–2020.03	1 min	470,000	Normal	–	NBM Evaluation
Dataset 2_a24	0	2022.04–2023.05	10 min	55,003	normal	–	NBM Evaluation, uncertainty analysis
Dataset 2_a71	0	2022.01–2023.01	10 min	54,744	normal	–	NBM Evaluation, uncertainty analysis
Dataset 2_a73	0	2022.06–2023.06	10 min	54,042	Hydraulic group	2023.06.10–2023.06.17	NBM Evaluation, uncertainty analysis
Dataset 2_a0	0	2022.08–2023.08	10 min	54,986	Generator bearing failure	2023.08.06–2023.08.20	Ablation study, uncertainty analysis, Anomaly Detection
Dataset 2_a40	10	2022.01–2023.01	10 min	56,158	Generator bearing failure	2022.12.26–2023.01.26	Uncertainty analysis, Anomaly Detection

Table 2
Monitoring variables selected for the experiments. Datasets 1 and 2 provide limited features; thus, different feature sets were used across experiments.

Dataset	Variable symbols	Variable name
Dataset 1	MSFBT	Main-shaft front-bearing temperature
	MSRBT	Main-shaft rear-bearing temperature
	ET	Ambient temperature
	NT	Nacelle temperature
	HT	Hub temperature
Dataset 2	GBT	Generator bearing temperature
	NT	Nacelle temperature
	RRS	Rotor rotational speed
	AP	Active power
	ET	Ambient temperature

was established for each dataset. The heat balance equation for Dataset 1 is formulated as follows:

$$\widehat{T}_{MSFBT}(l) = \beta_1 T_{MSFBT}(l-1) + \beta_2 T_{NT}(l) + \alpha_1 \omega^2(l) + \alpha_2 P(l) \tag{31}$$

$$\widehat{T}_{\text{MSRBT}}(l) = \beta_3 T_{\text{MSRBT}}(l-1) + \beta_4 T_{NT}(l) + \alpha_3 \omega^2(l) + \alpha_4 P(l) \tag{32}$$

Since the temperature difference between the front and rear main-shaft bearings is relatively small, and their primary distinction arises from their different positions within the nacelle, the temperature variations induced by heat generation are assumed to be approximately identical. Therefore, $\alpha_1 = \alpha_3$ and $\alpha_2 = \alpha_4$, (yielding)

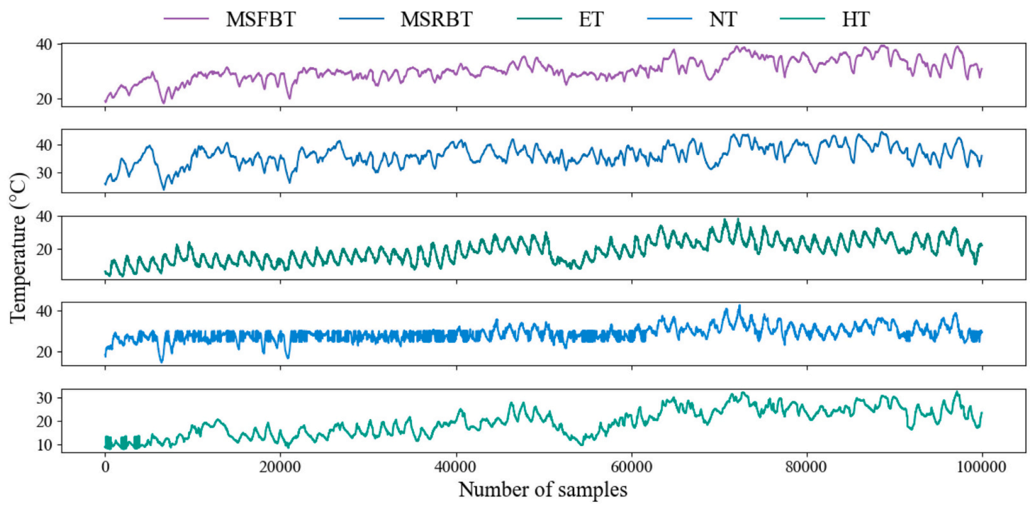
$$\widehat{T}_{MSFBT}(l) - \widehat{T}_{\text{MSRBT}}(l) \approx \beta_1 T_{MSFBT}(l-1) + \beta_3 T_{\text{MSRBT}}(l-1) + (\beta_2 - \beta_4) T_{NT}(l) \tag{33}$$

ET and HT are introduced as indirect heat transfer correction terms, and the model is reformulated in a parametric form.

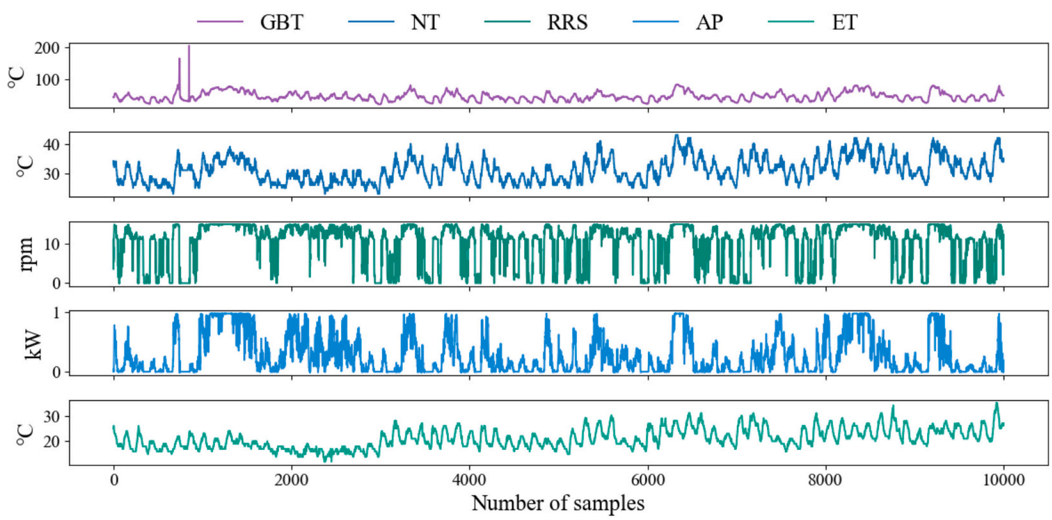
$$f = \widehat{T}_{MSFBT}(l) - \widehat{T}_{\text{MSRBT}}(l) + [\gamma_1 T_{MSFBT}(l-1) + \gamma_2 T_{\text{MSRBT}}(l-1) + \gamma_3 T_{ET}(l) + \gamma_4 T_{NT}(l) + \gamma_5 T_{HT}(l)] \tag{34}$$

Fig. 4 presents the time-domain variations of the five selected monitoring variables in three datasets.

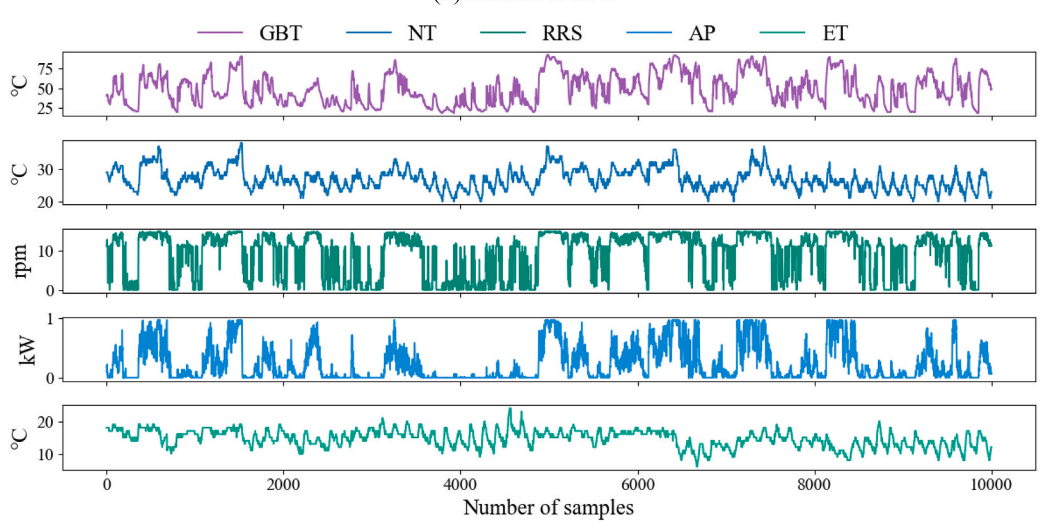
As shown in Fig. 4(a), the temperature variables in Dataset 1 remain generally stable, although several stage-dependent fluctuations can be observed. In Fig. 4(b) and Fig. 4(c), the rotor rotational speed and grid power in Dataset 2-a24 and Dataset 2-a40 exhibit more pronounced



(a) Dataset 1.



(b) Dataset 2-a24.



(c) Dataset 2-a40.

Fig. 4. Time-domain variations of the selected monitoring variables in different datasets.

variations, while the generator bearing temperature, nacelle temperature, and ambient temperature also change with the operating conditions. Overall, the monitoring variables show clear dynamic correlations, providing a data basis for multivariate modelling and the construction of heat balance relationships.

Fig. 5 further illustrates the correlations among the monitoring variables in the three datasets.

As shown in Fig. 5(a), the correlation coefficients of MSFBT with MSRBT and HT are 0.88 and 0.87, respectively, indicating strong relationships between the main-shaft bearing temperature and the other temperature related variables. In Fig. 5(b), the correlation coefficient between GBT and GP is 0.83, while that between NT and ET is 0.81. In Fig. 5(c), the correlation coefficients of GBT with GP and RRPM are 0.83 and 0.82, respectively. Although ET shows a relatively weak direct correlation with GBT, it is strongly correlated with NT, which is an important heat transfer variable in the heat balance relationship. Therefore, ET is retained as a correction term in the model. Overall, the temperature, rotational-speed, and power variables in the three datasets exhibit clear correlations, providing a basis for subsequent multivariate modelling and the construction of physics-based constraints.

4.2. Data preprocessing and evaluation indicators

Since SCADA monitoring variables differ significantly in physical dimensions and numerical scales, directly feeding them into the model may lead to imbalanced gradient updates, thereby affecting convergence stability and the reliability of threshold determination. Therefore, the data are first standardized before sample construction.

$$X = \{x_{i,j} | i = 1, 2, \ldots, n, j = 1, 2, \ldots, d\} \tag{35}$$

where i denotes the sequence length and j denotes the number of features.

Since each feature lies on a different numerical scale, the inter-feature differences can effectively reflect the characteristics of the data under normal operating conditions. To preserve these differences, normalization is performed along the temporal dimension.

$$\begin{cases} \mu_j = \dfrac{1}{n}\sum_{i=1}^{n} x_{i,j} \\ \sigma_j = \sqrt{\dfrac{1}{n}\sum_{i=1}^{n}\left(x_{i,j} - \mu_j\right)^2} \\ \hat{x}_{i,j} = \dfrac{x_{i,j} - \mu_j}{\sigma_j} \end{cases} \tag{36}$$

where μ_j and σ_j denote the mean and variance of the $x_{i,j}$ – th feature dimension, respectively.

To evaluate the model's ability to fit the state evolution patterns under normal operating conditions, the root mean square error (RMSE) and the coefficient of determination (R2) are adopted as predictive performance metrics. For a given target sensor, they are defined as follows:

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2} \tag{37}$$

$$\text{R}^2 = 1 - \frac{\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{m}(y_i - \bar{y}_i)^2} \tag{38}$$

where y_i and $\hat{y}_i$ denote the true value and the predicted value at the i – th time point, respectively, and m is the length of the prediction sequence. A smaller RMSE indicates a lower prediction error, whereas a value closer to 1 implies a stronger fitting capability of the model for the target variable. Here, the data have been normalized, and the reported accuracy is calculated as the average over different batches in each training epoch.

For each time step i, the ground-truth anomaly label and the predicted label are denoted by $y_i \in \{0, 1\}, \hat{y}_i \in \{0, 1\}$, respectively. Here, $y_i = 1$ indicates that the i-th time step is anomalous, whereas, $y_i = 0$ indicates a normal state. The predicted label $\hat{y}_i$ is obtained by comparing the anomaly score with the detection threshold. Precision, Recall, and F1-score are defined as:

$$\begin{cases} \text{Precision} = \dfrac{TP}{TP + FP} \\ \text{Recall} = \dfrac{TP}{TP + FN} \\ \text{F1} = \dfrac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \end{cases} \tag{39}$$

where TP, FP, and FN denote the numbers of true positives, false positives, and false negatives, respectively.

4.3. Compare models and parameter settings

To validate the effectiveness of the proposed model in normal-condition modelling, five representative time-series forecasting models were selected as benchmark methods: ModernTCN [44], which extracts multi-scale temporal features based on convolutional structures and can effectively characterize local dynamic variations; NHiTS [45], which employs a hierarchical architecture to capture variations at different temporal scales, thereby accounting for both global trends and local fluctuations; and TiDE [46], which utilizes a lightweight encoder–decoder structure to learn global temporal mapping features with high modelling efficiency. LSTM-AE [47], which employs an LSTM-based encoder–decoder architecture to reconstruct normal multivariate time-series patterns and detects anomalies based on reconstruction

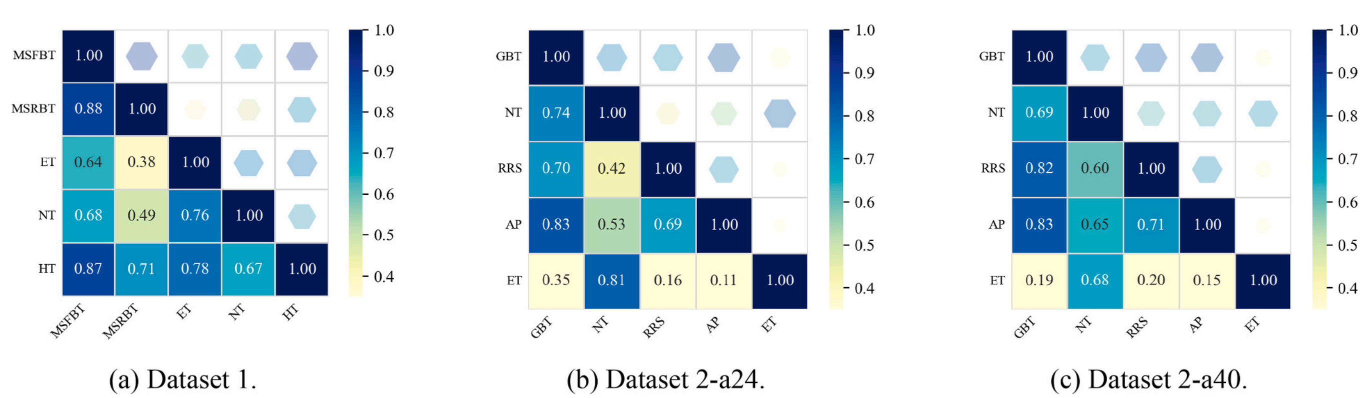


Fig. 5. Results of feature correlation calculation.

errors; VAE [48], which learns probabilistic latent representations through variational inference and identifies anomalies according to reconstruction deviations; DAGMM [49], which jointly learns low-dimensional representations and estimates sample density using a Gaussian mixture model for unsupervised anomaly detection; USAD [50], which adopts adversarially trained dual autoencoders to enhance the distinction between normal and abnormal multivariate time-series patterns; and OmniAnomaly [51], which combines stochastic recurrent modeling with probabilistic latent representations to capture temporal dependencies and detects anomalies based on reconstruction probabilities.

To ensure a fair comparison, all models use the same input window length and sliding step size, together with a unified data preprocessing procedure and identical training–testing partition strategies. Details are in Appendix A, Table 6. TG and PITG share the same spatiotemporal backbone network architecture, and the only difference is that PITG additionally introduces a physics-informed regularization term. Therefore, the performance gap between the two models can directly reflect the contribution of the physics-informed regularization module to model reliability.

5. Results and discussion

This chapter provides a systematic validation of the proposed PITG framework from four aspects: predictive performance, ablation analysis, uncertainty characterization, and anomaly detection performance.

5.1. Performance evaluation of the NBM

5.1.1. Accuracy analysis

To evaluate the performance of the PITG and TG model in constructing a NBM, three mainstream time-series forecasting models, namely ModernTCN, N-HiTS, and TiDE, together with five representative unsupervised anomaly detection models, namely LSTM-AE, VAE, DAGMM, USAD, and OmniAnomaly were selected as benchmark methods. Table 3 presents the RMSE and R2 results of the different models.

As shown in Table 3, after introducing the physics-informed module, ModernTCN, N-HiTS, TiDE, LSTM-AE, VAE, DAGMM, USAD, and OmniAnomaly all achieved lower average RMSE values across the four datasets, with average reductions ranging from 13.90% to 21.87%. In particular, the average RMSE reductions of ModernTCN, N-HiTS, and OmniAnomaly exceeded 20%. These results indicate that the physics-informed module provides an effective regularization effect for both forecasting-based and reconstruction-based models, thereby improving normal behavior modelling accuracy under most operating conditions. However, slight decreases in R^2 were observed for several model–dataset combinations, suggesting that the effectiveness of the physical constraint is influenced by both model architecture and operating condition.

In terms of overall performance, ModernTCN maintained relatively high prediction accuracy across the four datasets, while DAGMM exhibited comparatively stable results and can therefore be regarded as a representative baseline for forecasting-based and anomaly-detection models, respectively. By comparison, PITG achieved the lowest average RMSE of 0.1397 and the highest average R^2 of 97.21%, demonstrating the best overall predictive performance. Specifically, PITG obtained the lowest RMSE on Dataset 1 and Dataset 2-a24, and the highest R^2 on Dataset 1, Dataset 2-a24, and Dataset 2-a71. Although ModernTCN+PI achieved lower RMSE values than PITG on Dataset 2-a71 and Dataset 2-a73, PITG showed more balanced performance across the different datasets. These results suggest that the performance advantage of PITG arises not only from the physics-informed module, but also from the joint modelling of multivariate dependencies and temporal dynamics through the attention mechanism and GRU architecture, which further improves prediction accuracy and stability under varying operating conditions.

5.1.2. Ablation study

To verify the contribution of each component in PITG, an ablation study was conducted on the same test set by removing different modules from the model.

As shown in Table 4, the ablation study investigates the contributions of temporal feature extraction modules and the heat balance informed embedding module to normal behavior modelling. Transformer and GRU performances vary across different datasets, indicating that global feature correlation modelling and temporal dependency extraction capture different characteristics of wind turbine operational dynamics. Transformer + GRU achieves lower prediction errors on all datasets, with an average RMSE of 0.1727, which is reduced by approximately 38.53% and 39.94% compared with Transformer and GRU, respectively. This demonstrates the complementary capability of global dependency modelling and sequential dynamic representation.

After incorporating the heat balance informed embedding module, PITG achieves the best performance across all datasets. The average RMSE is further reduced from 0.1727 to 0.1397, corresponding to an

Table 3
Prediction Performance of the NBMs.

Model	Dataset							
	Dataset 1		Dataset 2_a24		Dataset 2_a71		Dataset 2_a73	
	RMSE	R2 (%)	RMSE	R2 (%)	RMSE	R2 (%)	RMSE	R2 (%)
ModernTCN	0.1788	96.64	0.1839	96.48	0.1797	96.55	0.2070	95.42
ModernTCN + PI	0.1547	96.89	0.1662	96.51	**0.1282**	96.46	**0.1364**	**97.56**
NHiTS	0.2031	95.42	0.3726	85.74	0.3248	88.91	0.3686	85.29
NHiTS + PI	0.1855	94.33	0.2621	85.80	0.2274	89.23	0.3334	85.39
TiDE	0.3479	86.58	0.4047	82.95	0.3726	85.36	0.3571	86.41
TiDE + PI	0.2497	89.97	0.3171	83.84	0.3429	85.15	0.3452	84.19
LSTM-AE	0.2949	91.00	0.5785	65.61	0.5138	72.57	0.5650	65.91
LSTM-AE + PI	0.2331	93.11	0.4761	63.38	0.4068	73.41	0.5039	66.33
VAE	0.3565	86.97	0.4777	76.56	0.4823	75.59	0.4434	78.83
VAE + PI	0.2967	88.71	0.3792	77.08	0.3507	80.17	0.3996	79.05
DAGMM	0.2349	94.29	0.3192	89.49	0.2926	91.12	0.3118	89.66
DAGMM + PI	0.2065	94.60	0.2552	89.65	0.2678	90.75	0.2488	89.58
USAD	0.2382	94.08	0.3506	87.33	0.3539	86.76	0.3374	87.80
USAD + PI	0.2073	94.55	0.2804	87.45	0.3038	88.20	0.3107	87.16
OmniAnomaly	0.2328	94.28	0.3420	87.80	0.3069	90.09	0.3460	87.13
OmniAnomaly + PI	0.2064	94.59	0.2580	89.23	0.2403	92.74	0.2690	90.62
TG	0.1774	96.74	0.1779	96.77	0.1640	97.24	0.1716	96.82
PITG	**0.1206**	**97.04**	**0.1483**	**97.22**	0.1421	**97.41**	0.1476	97.15

Table 4
Ablation results of PITG variants for normal behavior modeling.

Model	Dataset							
	Dataset 1		Dataset 2_a24		Dataset 2_a71		Dataset 2_a73	
	RMSE	R2 (%)	RMSE	R2 (%)	RMSE	R2 (%)	RMSE	R2 (%)
Transformer	0.1977	95.95	0.3195	89.53	0.2900	91.08	0.3168	89.30
Transformer + PI	0.1329	96.34	0.2695	90.89	0.2417	92.54	0.2700	90.46
GRU	0.2230	94.74	0.3138	89.93	0.2956	91.02	0.3179	89.16
GRU + PI	0.1619	94.40	0.2711	90.65	0.2548	91.81	0.2771	90.02
Transformer + GRU	0.1774	96.74	0.1779	96.77	0.1640	97.24	0.1716	96.82
Transformer + GRU + PI	**0.1206**	**97.04**	**0.1483**	**97.22**	**0.1421**	**97.41**	**0.1476**	**97.15**

improvement of approximately 19.15% over Transformer + GRU, while the average R^2increases from 96.89% to 97.21%. Moreover, introducing the physics-informed module into Transformer and GRU reduces their average RMSE by approximately 18.67% and 16.12%, respectively, demonstrating that the heat balance relationship provides additional physical constraints beyond purely data driven temporal representations.

Overall, the results indicate that Transformer-based global dependency modelling, GRU-based temporal evolution modelling, and heat balance informed physical constraints are complementary components. Their joint integration improves the accuracy of normal behavior representation and enhances the prediction stability under different operating conditions.

5.2. Uncertainty analysis

To evaluate the impact of physics-informed regularization on model reliability, this study takes TG as the baseline model and analyzes the improvements achieved by PITG from three perspectives: Epistemic uncertainty, Aleatoric uncertainty, and Model uncertainty.

Fig. 6 presents the distributions of the prediction losses for normal samples after aggregation across the feature dimension. The PITG losses are mainly concentrated in the low-RMSE region, with a higher probability-density peak, a narrower distribution, and a more rapidly increasing cumulative distribution function. By contrast, the TG losses exhibit a more pronounced right-skewed tail, indicating larger prediction errors. These results suggest that the physics-informed module constrains the model output at the feature level, reduces the dispersion of prediction errors, and improves the stability of normal behavior prediction.

Fig. 7 further compares the prediction loss distributions under healthy and faulty conditions. For PITG, the healthy condition losses are concentrated in the low error region, whereas the faulty condition losses shift toward higher values, resulting in clearer separation between the two operating states. In comparison, the healthy and faulty distributions of TG exhibit greater overlap, indicating weaker discriminability in the anomaly score space. Therefore, PITG not only reduces prediction fluctuations under healthy conditions but also enlarges the loss difference between healthy and faulty states, thereby reducing ambiguous samples near the detection threshold and improving anomaly discrimination.

Fig. 8 shows the prediction variances of different sensor outputs under Dropout rates ranging from 20% to 50%, based on 30 Monte Carlo forward passes. For both PITG and TG, the prediction variance increases as the Dropout rate increases, indicating that stronger stochastic perturbations lead to greater model output uncertainty. Nevertheless, PITG consistently exhibits lower prediction variance than TG at all Dropout rates. Taking the maximum sensor-wise variance as an example, the corresponding values for PITG at Dropout rates of 20%, 30%, 40%, and 50% are 0.02424, 0.04256, 0.06704, and 0.09667, respectively, compared with 0.03356, 0.05836, 0.09145, and 0.13326 for TG. These correspond to reductions of approximately 27.77%, 27.07%, 26.69%, and 27.45%, respectively. This indicates that the physics-informed module effectively suppresses stochastic prediction fluctuations and improves the consistency of Monte Carlo inference.

Overall, Figs. 6–8 demonstrates that PITG improves prediction stability under healthy conditions, enhances the separation between healthy and faulty states, and reduces prediction uncertainty under stochastic perturbations. These findings indicate that the physical constraint not only improves predictive accuracy but also regularizes model outputs, providing a more stable basis for subsequent anomaly detection and multi-sensor fusion.

As shown in Table 5, PITG and TG exhibit different prediction losses and fluctuation levels under different operating condition transfer scenarios. For State 02 (same turbine and same season), the mean loss of PITG increases from 0.4569 to 0.4673, while the standard deviation increases from 0.2807 to 0.3007. This indicates that when the training and testing distributions are highly consistent, the physics-informed embedding does not provide further improvement in prediction accuracy. This may be attributed to the high similarity of the data distribution, where the baseline model has already achieved relatively stable performance and the potential improvement introduced by additional

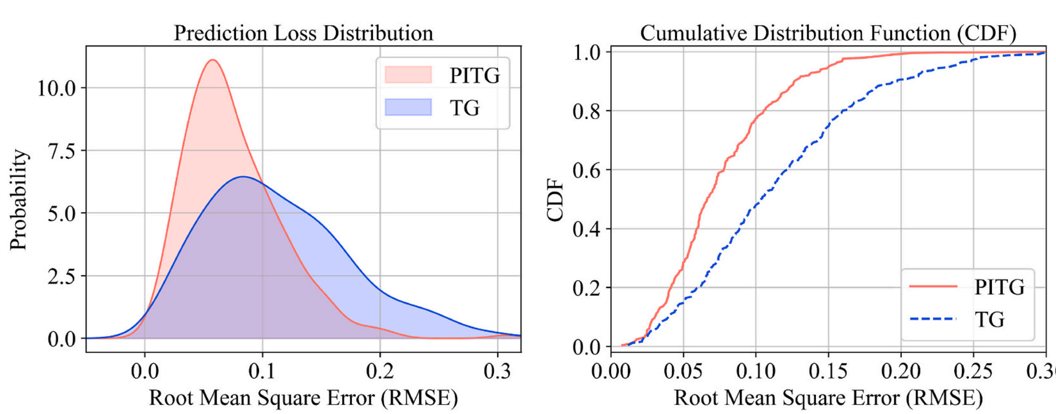


Fig. 6. Aleatoric uncertainty.

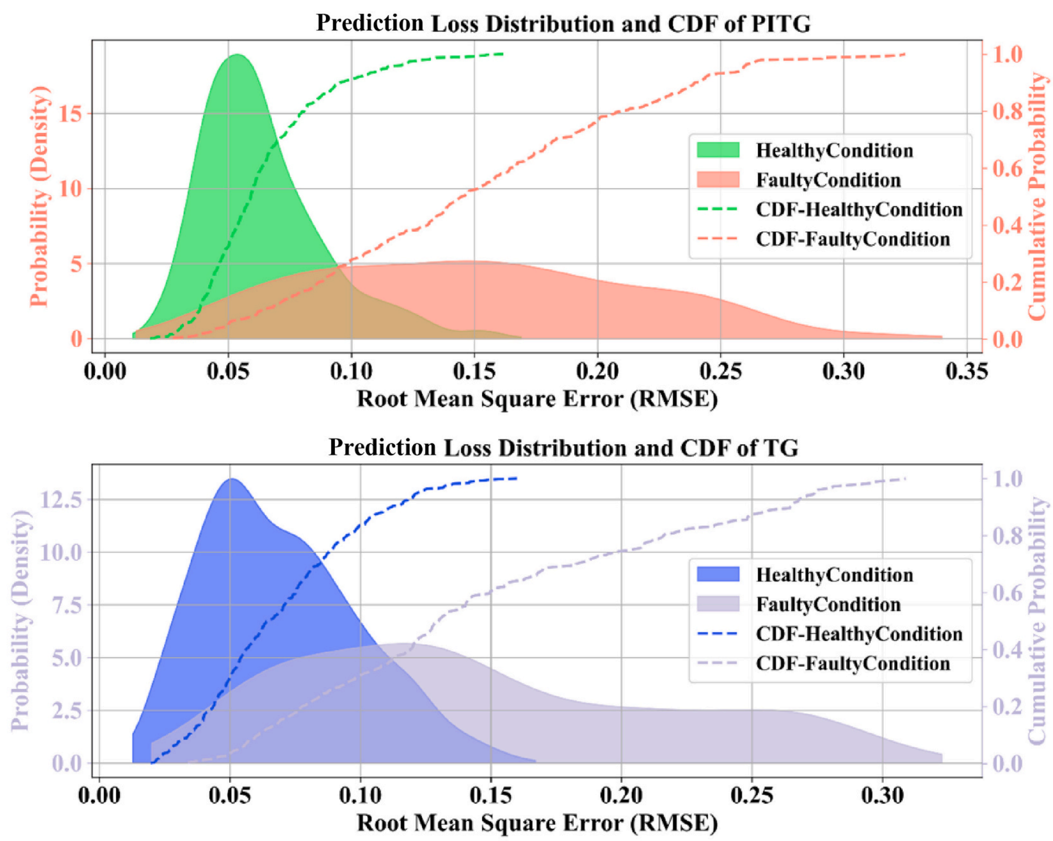


Fig. 7. Epistemic uncertainty.

physical constraints is limited.

For State 01 (same turbine and different seasons), PITG reduces the mean loss from 0.6146 to 0.5952 and the standard deviation from 0.3760 to 0.3607, corresponding to reductions of approximately 3.16% and 4.07%, respectively. This demonstrates that the physical constraint can partially alleviate the distribution shift caused by seasonal variations and improve prediction stability.

For State 40 (different turbines and different seasons), PITG also achieves lower prediction loss and fluctuation. The mean loss decreases from 0.5113 to 0.5054, while the standard deviation decreases from 0.4176 to 0.4013. Although the improvement is relatively limited, PITG maintains better stability under the more challenging transfer condition involving both turbine differences and seasonal variations.

Overall, PITG effectively reduces prediction error fluctuations under cross-season and cross-turbine conditions, indicating that the physics-informed embedding can utilize the thermal evolution relationships among temperature-related variables to improve the model's adaptability to distribution changes. However, under the same turbine and same season condition, where the data distribution difference is relatively small, the additional benefit of the physical constraint is less significant. This suggests that the proposed physical embedding is more effective in complex operating condition transfer scenarios.

5.3. Anomaly detection

As shown in the Fig. 9, different sensors exhibit distinct sensitivities to abnormal conditions. For both TG and PITG, the NT achieves relatively better overall detection performance, with F1-scores of 0.617 and 0.607, respectively. Although RRS achieves high Precision, reaching 0.897 under PITG, its Recall is only 0.196, indicating that the detection results are relatively conservative and may miss some abnormal samples. The GBT and AP exhibit a more balanced trade-off between Precision and Recall, whereas ET shows relatively limited anomaly detection capability. Overall, the single-sensor results indicate that different variables emphasize either detection accuracy or sensitivity, and relying on only one sensor is insufficient to robustly characterize the overall system condition.

From the perspective of model comparison, ModernTCN, as a representative high-performing time-series forecasting baseline, achieves relatively high Precision and F1-scores under most single-sensor and fusion scenarios, and generally outperforms DAGMM. As a representative anomaly detection model, DAGMM demonstrates certain anomaly recognition capability for some variables; however, its overall Recall is relatively low, resulting in limited F1 performance. In comparison, TG and PITG achieve better overall detection performance in most cases. Specifically, PITG improves the F1-score of generator bearing temperature from 0.480 for TG to 0.510, and achieves the highest Recall (0.438) and F1-score (0.566) among all methods using inverse-variance fusion. These results indicate that the physics-informed embedding contributes to improving anomaly identification capability and achieving a better balance between detection sensitivity and reliability.

Among the three fusion strategies, inverse-variance fusion achieves the highest Recall and F1-score for all four models. For example, the F1-score of PITG increases from 0.510 with equal-weight fusion and 0.524 with Dempster–Shafer (D—S) fusion to 0.566 with inverse-variance fusion. Similarly, the F1-score of ModernTCN improves from 0.369 and 0.465 to 0.506, respectively. This demonstrates that assigning sensor weights according to prediction reliability can reduce the interference of less reliable sensors and improve the overall anomaly detection performance. It should be noted that the fusion results do not always exceed the best-performing single sensor; however, they reduce the dependence on individual sensors and are more suitable for system-level condition assessment. Overall, the combination of PITG and

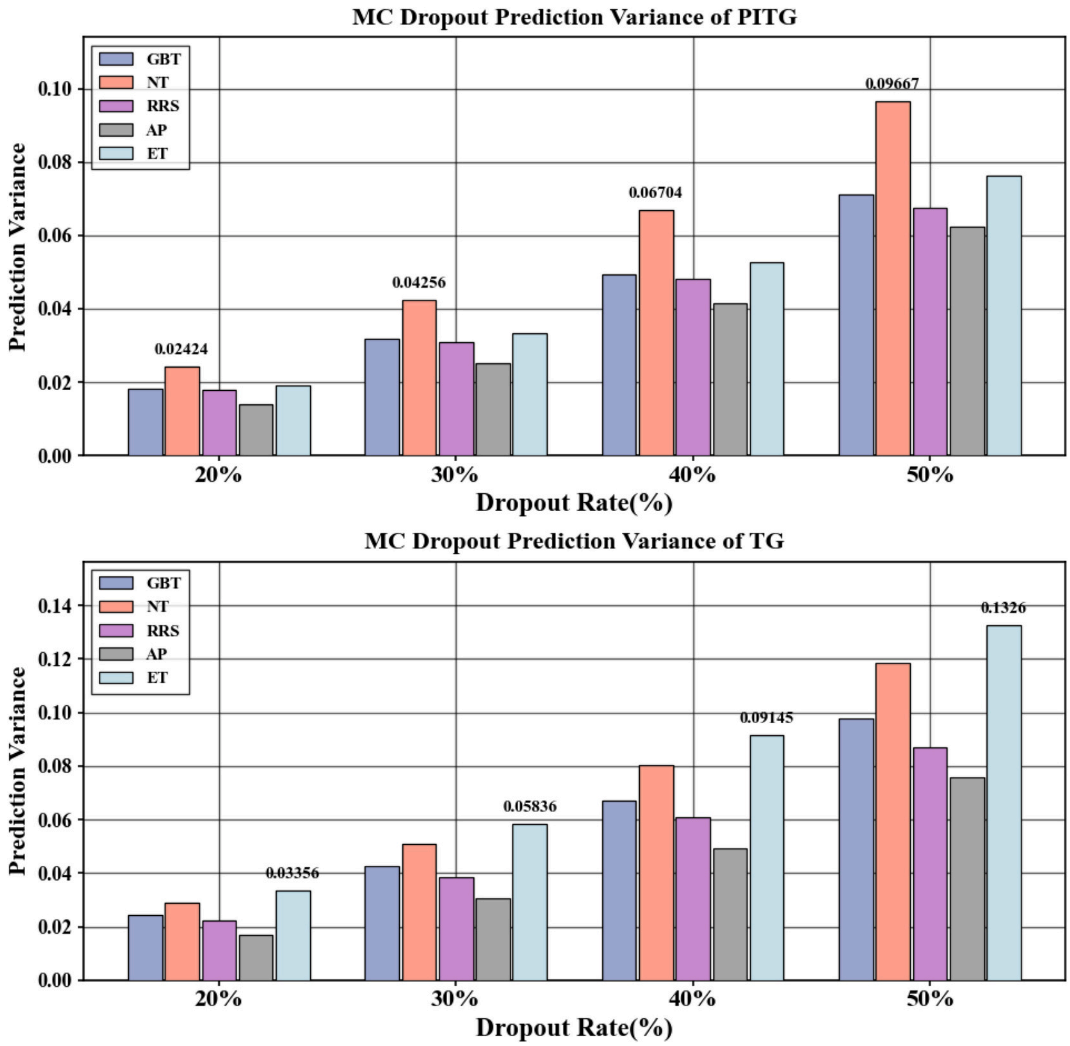


Fig. 8. Model uncertainty.

Table 5
Prediction uncertainty of PITG under different operating conditions.

State	Model	Mean loss	Standard deviation of loss
01	TG	0.6146	0.3760
	PITG	0.5952	0.3607
02	TG	0.4569	0.2807
	PITG	0.4673	0.3007
40	TG	0.5113	0.4176
	PITG	0.5054	0.4013

inverse-variance fusion achieves a better balance between Precision and Recall, providing a more stable solution for multi-sensor anomaly detection.

As shown in Fig. 10, under the same turbine but different seasonal conditions, both PITG and TG generate relatively continuous above-threshold responses within the fault interval. In particular, clear abnormal points can be observed near the fault initiation stage, the fluctuation stage during fault evolution, and the fault termination period. Due to the distribution shift in operating and environmental conditions caused by seasonal variations, several scattered false alarms still appear before and after the fault interval, indicating that seasonal differences have a certain impact on the stability of anomaly scores. Compared with TG, PITG produces a more concentrated anomaly response within the fault period; however, its lower detection threshold also results in several above-threshold points after fault recovery.

As shown in Fig. 11, under the same turbine and same seasonal conditions, the training and testing data exhibit higher distribution consistency. Both models maintain relatively low prediction losses before the fault and generate clear anomaly clusters after fault occurrence. PITG shows more continuous responses around the fault initiation and throughout the fault development process, enabling a clearer characterization of fault evolution. Compared with the other two operating conditions, the separation between normal and faulty periods is more evident in this scenario, indicating that similar data distributions contribute to improved detection stability.

As shown in Fig. 12, under different turbine and different seasonal conditions, the distribution discrepancy further increases. A larger number of above-threshold points appear before the fault occurrence, indicating that cross-turbine and cross-season transfer increases the risk of interpreting normal fluctuations as anomalies. Nevertheless, both PITG and TG are still able to detect multiple significant responses within the fault interval, forming abnormal point clusters around fault initiation, fault progression, and fault termination. Compared with the other two conditions, the overlap between normal and faulty loss distributions becomes more apparent, demonstrating that cross-turbine distribution shifts have a greater impact on detection reliability.

Overall, PITG and TG are both capable of identifying the major abnormal variations within the maintenance-recorded fault intervals under all three operating conditions. The detection results are most stable under the same turbine same season condition, while seasonal

Fig. 9. Performance comparison of different models and sensor fusion strategies for anomaly detection.

variations introduce additional fluctuations. When both turbine and seasonal conditions change, false alarms before fault occurrence become more pronounced. These results demonstrate that the proposed method has certain cross-condition anomaly response capability.

6. Conclusion

This study proposes an intelligent anomaly detection framework integrating a Physics-Informed Transformer-GRU deep model, a physics-informed embedding strategy, and an uncertainty-aware multi-sensor fusion approach for reliable anomaly detection of wind turbine drive-train systems. The proposed framework aims to improve the representation capability of NBMs for the dynamic evolution of monitoring variables, enhance the physical consistency of model predictions, and reduce the influence of uncertainty information during multi-sensor anomaly assessment. Furthermore, a modular anomaly detection pipeline is established, covering spatiotemporal feature modelling, physics-informed constraint learning, and uncertainty-aware fusion. The main conclusions of this study are summarized as follows:

1. A spatiotemporal normal behavior modelling framework was developed for multivariate SCADA data. By combining high-dimensional feature dependency modelling with temporal evolution learning, TG captures both inter-variable correlations and dynamic operating patterns. The experimental results demonstrate that TG improves prediction performance by effectively modelling spatiotemporal dependencies among monitoring variables.
2. A physics-informed embedding module was introduced by establishing physical relationships among predicted sensor features and constructing a thermal balance equation. The derived physical constraints were incorporated into the training process as regularization terms, thereby enhancing the physical consistency of model outputs. Experimental results demonstrate that this module improves the performance of different prediction models, confirming the effectiveness of physical constraints in enhancing state representation capability and prediction reliability.
3. An uncertainty-aware multi-sensor fusion strategy was introduced by estimating sensor-level data uncertainty through prediction errors and constructing inverse-variance weights to reduce the contribution of highly uncertain sensors in anomaly assessment. Experimental results demonstrate that the proposed inverse-variance fusion strategy effectively suppresses the influence of uncertainty in multi-sensor detection, reducing false alarms and missed detections while improving the stability and reliability of anomaly assessment.

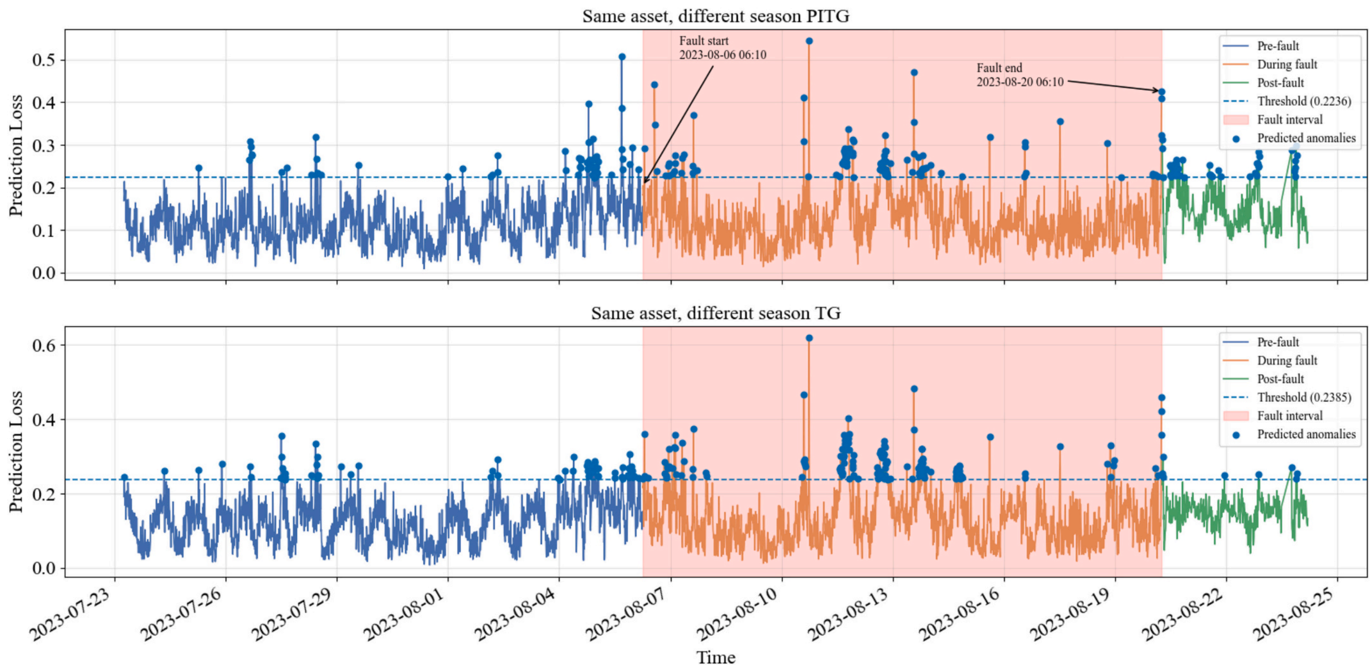


Fig. 10. Same turbine, different seasonal.

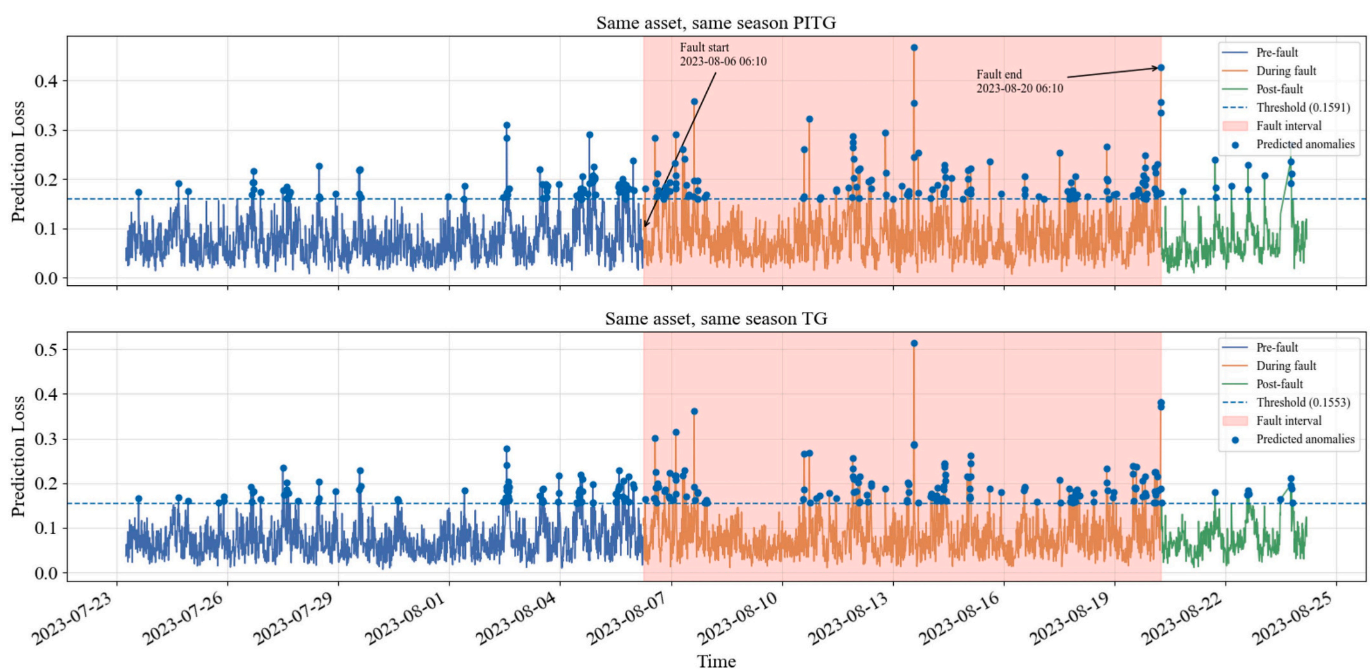


Fig. 11. Same turbine, same seasonal.

Overall, the proposed framework provides a structured solution for wind turbine drivetrain anomaly detection by jointly considering spatiotemporal feature evolution, physical consistency of prediction outputs, and uncertainty-aware multi-sensor fusion. The results demonstrate that the framework provides an effective approach toward more reliable and robust anomaly detection for complex wind turbine operating conditions.

CRediT authorship contribution statement

Hao Duan: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Qiang Zhang:** Writing – review & editing, Software, Methodology. **Huahao Ou:** Writing – review & editing, Software, Methodology. **Chun Li:** Writing – review & editing, Project administration, Investigation, Funding acquisition. **Wanfu Zhang:** Writing – review & editing, Funding acquisition. **Zifei Xu:** Writing – review & editing, Validation, Project administration, Methodology, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence

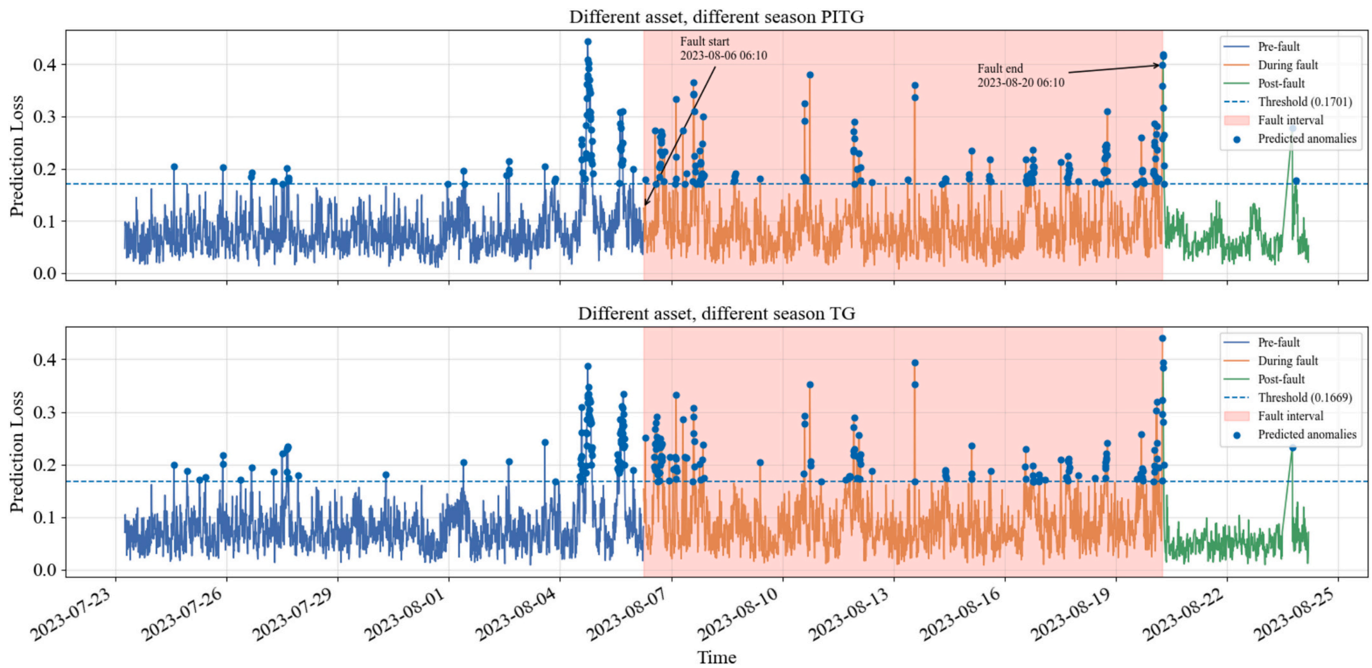


Fig. 12. Different turbine, different seasonal.

the work reported in this paper.

Acknowledgments

This study is funded by National Natural Science Foundation of China (Grant Nos. 52506259, 52375193), National Key Research and Development Program of China (2024YFA1012503), the Natural Science Fund of Shanghai Municipality (No. 24ZR1454800), the State Key Laboratory of Mechanical System and Vibration (Grant No. MSV202411).

Appendix A. Model hyperparameter settings and training–testing configuration

Table 6
Parameter settings of the proposed model.

Category	Parameter	Value
Data preprocessing	Sliding window length	64 samples
	Prediction horizon	10 min/12 min
	Batch size	32
	Normalization dimension	Temporal dimension
	Random seed	29
Modeling	Number of input features	5
	Hidden dimension	{64,100,128}
	Number of attention heads	{3,4,8}
	Feed-forward dimension	{32, 64, 256}
	Number of GRU layers	{1,2}
	Activation function	Relu
	Dropout rate	{0.2, 0.3, 0.4, 0.5}
Physics-informed module	Number of input features	5
	Kernel size	25
	Hidden-layer dimension	{1, 2}
	Activation function	Tanh
	Physical coefficients, $[\gamma_1,\gamma_2,\gamma_3,\gamma_4,\gamma_5]$	[−0.9983, 0.9986, −0.0003, −0.0005, −0.0006]
	Physical coefficients, $[\beta_1,\beta_2,\beta_3,\beta_4]$	[0.9982, 0.0006, 0.0007, 0.0021]
	λ	{0.9, 0.8, 0.75, 0.6}
Optimization	Optimizer	Adam
	Number of training epochs	80
	Learning rate	0.01
	Weight decay	0.00005
Uncertainty quantification	Number of MC samples	30

The principal hyperparameters used for data preprocessing, backbone-network construction, physics-informed embedding, model optimization, and uncertainty quantification are summarized in Table 6. The normalization parameters were calculated exclusively from the training set and subsequently applied to the test data to prevent information leakage and ensure consistency between the training and testing stages.

Among the evaluated fitting methods, least squares and Lasso were selected as the optimal fitting schemes for Dataset 1 and Dataset 2, respectively. Dataset 1 achieved an MSE of 3.8×10^{-5}, an RMSE of 0.006179, an MAE of 0.004605, and an R^2 of 0.997494, while Dataset 2 achieved an MSE of 3.3×10^{-5}, an RMSE of 0.005706, an MAE of 0.004135, and an $R^2$of 0.999776.

All experiments were conducted under the same hardware configuration, and all models were implemented within a unified training framework. The computational environment consisted of a 12th-generation Intel i5-12600KF processor and an NVIDIA RTX 4060 GPU. All AI-driven models and training frameworks were developed using PyTorch version 2.5.1.

Appendix B. Description of multiple operating conditions

Table 7
Dataset division under different operating conditions.

Label	Operating condition	Time period	Description
01	Same turbine, different season	2023.03.02–2023.05.10	Normal data; used for model training
02	Same turbine, same season	2022.08.04–2022.10.14	Normal data; used for model training
40	Different turbine, different season	2022.01.01–2022.03.11	Normal data; used for model training
0	Fault test dataset	2023.07.20–2023.8.24	Generator bearing failure; used for model validation

Models were trained separately using the normal datasets labeled 01, 02, and 40, which represent different asset and seasonal conditions. The trained models were then evaluated on Dataset 0 for anomaly detection. Dataset 0 contains the generator-bearing fault interval and was used exclusively for testing and validation.

Appendix C. Description of multiple operating conditions

Table 8
Dataset division under different operating conditions.

Stage	Total time (s)	Average cost per window (ms)	Throughput (windows/s)
Training	38.3168	5.5132	181.382
Inference	10.4663	44.7279	22.357

Table 8 summarizes the computational cost of the proposed method. The complete training process requires an average of 38.32 s, corresponding to 5.51 ms per window per epoch and a training throughput of approximately 181.38 windows/s. The complete testing pipeline requires 10.47 s, with an end-to-end inference cost of 44.73 ms per window and a throughput of approximately 22.36 windows/s. It should be noted that the reported testing time includes data loading, normalization, window construction, model loading, forward inference, prediction error calculation, and result export; therefore, it represents the end-to-end computational cost of the anomaly detection pipeline.

For SCADA data sampled at 1 min intervals, the inference time of 44.73 ms per window is substantially shorter than the data update interval. For data sampled at 10 min intervals, the computational cost is even more negligible relative to the sampling period. Therefore, the proposed method can support online anomaly detection for both 1 min and 10 min SCADA data, where online detection refers to timely inference after the arrival of each new data window rather than millisecond-level real-time control.

Data availability

Data will be made available on request.

References

[1] Huahao O, Weipao M, Chun L, et al. Research on multi-physical field coupling characteristics of three-dimensional variable cross-section blade H-type vertical axis wind turbine. Energy Convers Manag 2025;342:120105. https://doi.org/10.1016/j.enconman.2025.120105.
[2] Helsen Jan. Review of research on condition monitoring for improved O&M of offshore wind turbine drivetrains. Acoust Aust 2021;49(2):251–8. https://doi.org/10.1007/s40857-021-00237-2.
[3] Yongchao Z, Kun Y, Zihao L, et al. Integrated intelligent fault diagnosis approach of offshore wind turbine bearing based on information stream fusion and semi-supervised learning. Expert Syst Appl 2023;232. https://doi.org/10.1016/j.eswa.2023.120854.
[4] Ochella Sunday, Shafiee Mahmood, Dinmohammadi Fateme. Artificial intelligence in prognostics and health management of engineering systems. Eng Appl Artif Intell 2022;108.
[5] Zahra ZD, Geoffrey IW, Shirui P, et al. Deep learning for time series anomaly detection: a survey. ACM Comput Surv 2025;57(1). https://doi.org/10.1145/3691338.
[6] Peng Z, Chaozhe L, Huanyun X, et al. Bi-AAE: a binary adversarial autoencoder deep neural network model for anomaly detection in system-levels marine diesel engines. Ocean Eng 2024;302:117700.
[7] Shen L, Jinglong C, Zijun L, et al. Graph embedded patch-sense autoencoder with prior knowledge for multi-component system anomaly detection. Reliab Eng Syst Saf 2025;256. https://doi.org/10.1016/j.ress.2024.110784.
[8] Wadinger Marek, Kvasnica Michal. Adaptable and interpretable framework for anomaly detection in SCADA-based industrial systems. Expert Syst Appl 2024;246.
[9] Dao Phong B, Barszcz Tomasz, Staszewski Wieslaw J. Anomaly detection of wind turbines based on stationarity analysis of SCADA data. Renew Energy 2024;232.
[10] Jin X, Xu Z, Qiao W. Condition monitoring of wind turbine generators using SCADA data analysis. IEEE Trans Sustain Energy 2020;12(1):202–10.
[11] Moldovan SC, Iantovics LB. Review on information fusion-based data mining for improving complex anomaly detection. Wiley Interdiscip Rev: Data Min Knowl Discov 2025;15(2):e70017.
[12] Zhong Y, Baiyang C, Jia L, et al. Anomaly detection based on weighted fuzzy-rough density. Appl Soft Comput 2023;134:109995.
[13] Weiqiang W, Chunyue S, Jun Z, et al. Physics-informed gated recurrent graph attention unit network for anomaly detection in industrial cyber-physical systems. Inf Sci 2023;629:618–33. https://doi.org/10.1016/j.ins.2023.01.136.

[14] Yun H, Kim, et al. Autoencoder-based anomaly detection of industrial robot arm using stethoscope based internal sound sensor. J Intell Manuf 2023;34(3):1427–44. https://doi.org/10.1007/s10845-021-01862-4.
[15] Michau Gabriel, Frusque Gaetan, Fink Olga. Fully learnable deep wavelet transform for unsupervised monitoring of high-frequency time series. Proc Natl Acad Sci USA 2022;119(8).
[16] Q Junyu, Z Rui, L Chenyu, et al. Anomaly detection and multi-step estimation based remaining useful life prediction for rolling element bearings. Mech Syst Signal Process 2024;206.
[17] Kang Y, Zhenhan L, Zekun Y, et al. Improved PCA reconstruction-based unsupervised anomaly detection in uncontrolled structural health monitoring with correntropy. IEEE Trans Industr Inform 2025;21(10):8035–45.
[18] Z Song, L Yali, Z Xiaobin, et al. Coastline target detection based on UAV hyperspectral remote sensing images. Front Mar Sci 2024;11.
[19] Z Jun, W Chengkun, M Xiandong, et al. Abnormal vibration detection of wind turbine based on temporal convolution network and multivariate coefficient of variation. Mech Syst Signal Process 2022;174.
[20] D Shuyi, W Jiulong, W Meizhu, et al. A systematic data-driven approach for production forecasting of coalbed methane incorporating deep learning and ensemble learning adapted to complex production patterns. Energy 2023;263.
[21] S Giulia, Z Pamela, M Lucio, et al. Vehicle localization in an explainable dynamic Bayesian network framework for self-aware agents. Inf Fusion 2025;122.
[22] L Jiayang, W Xiaosun, X Fuqi, et al. Condition monitoring of wind turbines with the implementation of spatio-temporal graph neural network. Eng Appl Artif Intell 2023;121.
[23] W Anqi, P Yan, Q Zheng, et al. A two-stage anomaly decomposition scheme based on multi-variable correlation extraction for wind turbine fault detection and identification. Appl Energy 2022;321.
[24] Zhang Chen, Hu Di, Yang Tao. Anomaly detection and diagnosis for wind turbines using long short-term memory-based stacked denoising autoencoders and XGBoost. Reliab Eng Syst Saf 2022;222.
[25] Olayinka SO, Emerald UH, Olaniran JM, et al. Conditional monitoring and fault detection of wind turbines based on Kolmogorov-Smirnov non-parametric test. Energy Rep 2024;11:2577–91.
[26] S V, I O, S V, et al. Intelligent bearing fault diagnosis method combining mixed input and hybrid CNN-MLP model. Mech Syst Signal Process 2022;180.
[27] Ling X, Xin Y, Aijun H, et al. Condition monitoring and anomaly detection of wind turbine based on cascaded and bidirectional deep learning networks. Appl Energy 2022;305.
[28] Meng-Hui W, Shiue-Der L, Cheng-Che H, et al. Fault detection of wind turbine blades using multi-channel CNN. Sustainability 2022;14(3).
[29] Yongchao Z, Caichao Z, Jianjun T, et al. Anomaly detection and condition monitoring of wind turbine gearbox based on LSTM-FS and transfer learning. Renew Energy 2022;189:90–103.
[30] Younjeong L, Chanho P, Namji K, et al. LSTM-autoencoder based anomaly detection using vibration data of wind turbines. Sensors (Basel) 2024;24(9).
[31] Angel E, Luis M, Christian T, et al. Early fault detection in the main bearing of wind turbines based on gated recurrent unit (GRU) neural networks and SCADA data. IEEE/ASME Trans Mechatron 2022;27(6):5583–93.
[32] Cho Seongpil, Kim Sang-Woo, Kim Hyo-Jin. Fault detection in blade pitch systems of floating wind turbines utilizing transformer architecture. Struct Eng Mech 2024;92(2):121–31.
[33] Anqi W, Yan P, Yunyi Z, et al. Wind turbine fault detection and identification through self-attention-based mechanism embedded with a multivariable query pattern. Renew Energy 2023;211:918–37.
[34] Ashish V, Noam S, Niki P, et al. Attention is all you need. Adv Neural Inf Proces Syst 2017; 30.
[35] Yakupoglu Hidayet, Gozde Haluk, Taplamacioglu M Cengiz. Online noise-adaptive Kalman filter integrated novel autoencoder for multi-fault detection and early warning of wind turbines. Measurement 2025;256.
[36] Zifei Xu, Zhang Qiang. Predictive maintenance optimization for industrial equipment via reliable prognosis and risk-aware reinforcement learning. Complex Intell Syst 2025;12(1):1–20.
[37] X Jinjin, W Rongxi, L Zeming, et al. Physics-guided, data-refined fault root cause tracing framework for complex electromechanical system. Reliab Eng Syst Saf 2023;236.
[38] A Davide, F Benedetta, G Antonio, et al. An empirical evaluation of deep autoencoders for anomaly detection in the electricity consumption of buildings. Energ Buildings 2025;327.
[39] Yong Bang Xiang, Brintrup Alexandra. Bayesian autoencoders with uncertainty quantification: towards trustworthy anomaly detection. Comput Res Repos 2022;209.
[40] Gal Y, Ghahramani Z. Dropout as a Bayesian approximation: representing model uncertainty in deep learning. In: International conference on machine learning; 2015; abs/1506.02142.
[41] Raissi M. Deep hidden physics models: deep learning of nonlinear partial differential equations. J Mach Learn Res 2018;19(25):1–24.
[42] Cambron P, Tahan A, Masson C, et al. Bearing temperature monitoring of a wind turbine using physics-based model. J Qual Maint Eng 2017;23(4):479–88.
[43] Gueck C, Roelofs CMA, Faulstich S. CARE to compare: a real-world benchmark dataset for early fault detection in wind turbine data. Data 2024;9(12).
[44] Luo S, Wang X. ModernTCN: a modern pure convolution structure for general time series analysis. In: International conference on learning representations; 2024.
[45] Challu C, Olivares KG, Oreshkin BN, et al. N-HiTS: neural hierarchical interpolation for time series forecasting. In: Thirty-seventh AAAI conference on artificial intelligence. Vol. 37, No. 6; 2023. p. 6989–97.
[46] Das A, Kong W, Leach A, et al. Long-term forecasting with TiDE: time-series dense encoder. Tmlr 2024;2024.
[47] Malhotra P, Ramakrishnan A, Anand G, et al. LSTM-based encoder-decoder for multi-sensor anomaly detection. CoRR 2016. arXiv: 1607.00148.
[48] Kingma DP, Welling M. Auto-encoding variational bayes. ICLR 2014; 2014.
[49] Zong B, Song Q, Min MR, et al. Deep autoencoding gaussian mixture model for unsupervised anomaly detection. ICLR 2018; 2018.
[50] Audibert J, Michiardi P, Guyard F, et al. USAD: unsupervised anomaly detection on multivariate time series. Knowledge Discovery and Data Mining; 2020.
[51] Su Y, Zhao Y, Niu C, et al. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In: Kdd'19 Proceedings of the 25th ACM Sigkdd international conference on knowledge discovery and data mining; 2019. p. 2828–37.
[52] Tautz-Weinert J, Watson SJ. Using SCADA data for wind turbine condition monitoring – a review. IET Renew Power Gener 2017;11(4):382–94.
[53] Teng W, Ding X, Tang S, et al. Vibration analysis for fault detection of wind turbine drivetrains—a comprehensive investigation. Sensors (Basel, Switzerland) 2021;21(5).
[54] Tutiven C, Vidal Y, Insuasty A, et al. Early fault diagnosis strategy for WT main bearings based on SCADA data and one-class SVM. Energies 2022;15(12).
[55] Liu J, Yang G, Li X, et al. Wind turbine anomaly detection based on SCADA: a deep autoencoder enhanced by fault instances. ISA Trans 2023;139:586–605.
[56] Tchakoua P, Wamkeue R, Ouhrouche M, et al. Wind turbine condition monitoring: state-of-the-art review, new trends, and future challenges. Energies 2014;7(4):2595–630.
[57] Schroder L, Dimitrov NK, Verelst DR, et al. Using transfer learning to build physics-informed machine learning models for improved wind farm monitoring. Energies 2022;15(2).
[58] Lee C, Maceren EDC. Physics-informed anomaly and fault detection for wind energy systems using deep CNN and adaptive elite PSO-XGBoost. IET Gener Transm Distrib 2025;19(1).
[59] Gao Z, Xiao X, Mo H, et al. Physics-informed condition monitoring for wind turbines via change point detection under heteroscedasticity. Reliab Eng Syst Saf 2026;272.