Applied Artificial Intelligence Research

INNOVIAIR

Physics-informed Multi-task Tabular Transformer for Joint Fault Prediction and Remaining Useful Life Estimation of Industrial Equipment

Junling Hu[a,*]

[a]School of Intelligent Manufacturing Colleg, Chongqing Jianzhu College, Chongqing 400072, China

A R T I C L E I N F O

Keywords:
Predictive Maintenance
Fault Diagnosis
Remaining Useful Life
Physics-Informed Learning
Multi-Task Learning

A B S T R A C T

Accurate fault prediction and remaining useful life (RUL) estimation are essential for predictive maintenance of industrial equipment. However, existing methods mainly focus on time-series signals and often ignore the heterogeneous characteristics and degradation knowledge in tabular equipment data. This study proposes a Physics-informed Multi-task Tabular Transformer (PIMTT) for joint fault prediction and RUL estimation. The proposed model integrates numerical, categorical, and physics-informed tokens, where a type-aware physics gate is introduced to adaptively enhance degradation-related representations. A multi-task learning strategy is further developed to jointly optimize fault classification and RUL regression. Experiments on a synthetic industrial-equipment condition dataset with approximately 260,000 samples demonstrate that PIMTT achieves an F1-score of 0.8461 for fault prediction and an R^2 of 0.9899 for RUL estimation, outperforming several machine learning, Transformer-based, and multi-task baselines. The results verify the effectiveness of combining physical knowledge with tabular Transformer learning for intelligent equipment health assessment.

1.Introduction

The increasing complexity of modern industrial systems has created significant challenges for equipment reliability management. Unexpected failures in manufacturing equipment may lead to production interruptions, economic losses, and safety risks. Traditional maintenance strategies, such as corrective maintenance and fixed-interval preventive maintenance, cannot accurately reflect the actual health condition of equipment because they ignore the degradation process under different operating conditions[1,2]. Therefore, predictive maintenance (PdM), which aims to identify potential failures in advance and estimate equipment degradation trends, has become an important component of intelligent manufacturing systems.

Existing studies in prognostics and health management (PHM) have mainly focused on two related tasks: fault prediction and remaining useful life (RUL) estimation[3,4]. Fault prediction aims to determine whether equipment will enter an abnormal state within a future time window, while RUL estimation attempts to quantify the remaining operating time before failure. Although these two tasks describe different aspects of equipment health, they are inherently correlated because degradation accumulation usually leads to increasing failure probability and decreasing remaining lifetime.

However, many existing studies still treat fault classification and RUL estimation as independent problems, resulting in limited information sharing between short-term failure assessment and long-term degradation modeling.

Early research on equipment fault diagnosis mainly relied on handcrafted features extracted from vibration, acoustic, and electrical signals, followed by conventional machine learning classifiers such as support vector machines (SVMs), random forests, and boosting algorithms[5-7]. These approaches have achieved satisfactory performance in specific applications; however, their effectiveness strongly depends on expert-designed feature extraction and may degrade when dealing with heterogeneous operational information. In practical industrial environments, equipment monitoring data usually contain not only sensor measurements but also categorical attributes (e.g., equipment type), maintenance records, historical failure information, and operating conditions. Such heterogeneous tabular data contain valuable health-related information but are difficult to fully exploit using conventional feature engineering methods.

* Corresponding author.
E-mail address: jlinghu6@gmail.com.
https://doi.org/10.65455/s4kq0p29
Received 5 August 2026; Received in revised form 18 August 2026; Accepted 19 August 2026; Available 26 August 2026

Deep learning approaches have subsequently been introduced into PHM due to their capability of automatically learning complex feature representations[8]. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) have been widely applied to vibration-based fault diagnosis and sequential degradation modeling[9]. For example, CNN-based models can automatically extract discriminative patterns from raw sensor signals, while LSTM networks are capable of capturing temporal dependencies in degradation processes[10,11]. However, most existing deep learning methods are designed for time-series signals, where measurements are continuously collected over time. In contrast, many industrial scenarios provide equipment-level status records in a tabular format, where each sample represents a comprehensive health snapshot containing numerical variables, categorical information, and maintenance-related features. The characteristics of such data are fundamentally different from sequential sensor signals.

Recently, deep learning models for tabular data have received increasing attention. Neural networks with embedding layers have been widely adopted to handle heterogeneous categorical and numerical variables by mapping discrete attributes into continuous representations. Moreover, gradient boosting models, including XGBoost and CatBoost, remain strong baselines for structured data prediction due to their robustness and efficiency[6,12]. However, tree-based methods model feature interactions through hierarchical splitting rules, which may limit their ability to capture high-order relationships among multiple equipment factors. To address this limitation, Transformer-based tabular models have been proposed. FT-Transformer converts individual features into tokens and utilizes self-attention mechanisms to learn feature interactions, achieving competitive performance on various tabular benchmarks[13]. Nevertheless, existing tabular Transformer approaches are mainly data-driven and rarely incorporate domain-specific physical knowledge, which is important for industrial equipment health assessment.

In industrial systems, equipment degradation is governed by physical mechanisms rather than purely statistical correlations. For example, excessive temperature, abnormal vibration, insufficient lubrication, and poor cooling conditions are closely associated with mechanical degradation. However, these physical relationships are usually not explicitly represented in existing data-driven models. Incorporating physics-related information into deep learning models provides an opportunity to improve model interpretability and reliability, especially when equipment types have different failure mechanisms.

Another challenge in predictive maintenance is the joint modeling of failure risk and RUL. Multi-task learning provides an effective framework by sharing latent representations among related tasks while maintaining task-specific prediction branches. Previous studies have demonstrated that adaptive task weighting and mixture-of-experts (MoE) architectures can alleviate negative transfer between tasks and improve multi-objective optimization[14-16]. Nevertheless, existing multi-task PHM approaches are mainly developed for sequential sensor datasets, while joint learning on heterogeneous industrial tabular data with physical constraints remains insufficiently explored.

To address these limitations, this study proposes a PIMTT for joint equipment fault prediction and RUL estimation. Unlike conventional approaches that directly process raw tabular features, the proposed framework represents heterogeneous information using three types of tokens: numerical tokens, categorical tokens, and physics-informed tokens. Numerical and categorical tokens capture statistical characteristics of equipment states, while physics tokens explicitly encode domain knowledge derived from thermal, vibration, lubrication, cooling, and historical failure risks. These heterogeneous tokens are processed by a Transformer encoder to learn comprehensive equipment health representations. Furthermore, a multi-task learning framework is developed to simultaneously predict 7-day failure probability and RUL, where adaptive loss weights are employed to balance classification, regression, and physics consistency objectives.

The main contributions of this work are summarized as follows:

1. A physics-informed tabular Transformer framework is proposed for industrial equipment health assessment by integrating numerical features, categorical information, and physics-derived risk factors into unified token representations.
2. A joint fault prediction and RUL estimation architecture is developed, enabling the model to exploit the intrinsic relationship between short-term failure probability and long-term degradation trends.
3. An adaptive multi-task optimization strategy is introduced by dynamically balancing classification loss, RUL regression loss, and physics-guided regularization.

Extensive experiments are conducted on the synthetic equipment-condition dataset by comparing the proposed method with conventional machine-learning models (SVM, XGBoost, and CatBoost), sequence and physics-guided neural baselines, FT-Transformer, and MMoE.

2.Related work

2.1.Machine learning-based fault prediction

Early research on industrial fault prediction mainly relied on conventional machine learning methods combined with manually designed statistical features. Since industrial equipment monitoring data usually contain nonlinear relationships between operational variables and failure states, machine learning algorithms such as SVMs, random forests, and gradient boosting models have been extensively investigated[4,7].

SVMs have been widely adopted for fault classification due to their ability to construct nonlinear decision boundaries in high-dimensional feature spaces[5]. However, the performance of SVM-based approaches depends strongly on feature selection and kernel parameter optimization, which limits their scalability when dealing with increasingly heterogeneous industrial datasets.

Tree-based ensemble methods have become popular alternatives because of their robustness on structured data.

XGBoost, proposed by Chen and Guestrin[6], improves gradient boosting through an efficient regularized objective function and has achieved strong performance in many industrial prediction tasks. CatBoost further improves categorical feature handling by introducing ordered target encoding and permutation-based boosting strategies, making it suitable for datasets containing a large number of discrete variables[12].

Although these methods remain competitive for industrial tabular prediction, they mainly rely on manually designed splitting strategies to model feature interactions. For complex equipment health assessment, relationships among temperature, vibration, maintenance history, and operational conditions may involve high-order nonlinear dependencies that are difficult to explicitly capture through tree structures.

2.2.Deep learning for PHM: from time-series signals to equipment-level tabular data

With the development of deep learning, data-driven PHM approaches have gradually shifted from handcrafted feature extraction toward automatic representation learning. CNNs have been widely applied to mechanical fault diagnosis by extracting local patterns from vibration and acoustic signals[4]. Meanwhile, recurrent architectures, particularly LSTM, have demonstrated advantages in modeling degradation trajectories and RUL prediction due to their capability of capturing temporal dependencies[10].

For example, Zheng et al. proposed an LSTM-based framework for RUL prediction and demonstrated that recurrent models can effectively learn degradation trends from historical operating sequences[11]. Similarly, many studies have investigated CNN-LSTM hybrid architectures for bearing fault diagnosis and prognostics by combining local feature extraction and temporal dependency modeling[19-20].

Despite their success, most existing deep PHM methods assume that input data are collected as continuous time-series signals. In such scenarios, temporal ordering is the dominant information source. However, many industrial maintenance systems generate structured equipment records rather than raw signal streams. A typical equipment health record may include operating hours, equipment category, maintenance history, fault counts, environmental conditions, and sensor measurements. These variables describe the current health condition of equipment from multiple perspectives but do not necessarily contain explicit temporal sequences.

Therefore, directly applying sequence-based deep learning models to such datasets may not fully exploit the characteristics of tabular industrial data. A dedicated representation learning strategy is required to simultaneously model numerical measurements, categorical attributes, and domain-specific equipment information.

Table 1. Summary of representative related works

Study	Task	Dataset Type	Model
Chen & Guestrin[6]	Tabular prediction	Structured data	XGBoost
Prokhorenkova et al.[12]	Tabular prediction	Categorical-rich tabular data	CatBoost
Arik & Pfister[17]	Interpretable tabular learning	Tabular benchmark datasets	TabNet
Gorishniy et al.[13]	Tabular prediction	Tabular benchmark datasets	FT-Transformer
Zheng et al.[11]	Remaining useful life estimation	NASA C-MAPSS sensor data	LSTM-based RUL
Zhao et al.[9]	Machine health monitoring	Vibration/acoustic signals	Deep learning survey
Kendall et al.[14]	Multi-task learning	Computer vision datasets	Uncertainty-based weighting
Ma et al.[15]	Multi-task recommendation	Recommendation system data	MMoE
Raissi et al.[18]	Physics-informed learning	Physical systems governed by PDEs	PINN

2.3.Transformer-based tabular learning

Transformer architectures have achieved remarkable success in natural language processing by introducing self-attention mechanisms capable of capturing global dependencies among input elements[21]. Inspired by this idea, researchers have explored whether tabular features can also be represented as tokens and processed using Transformer architectures.

TabNet introduced sequential attention-based feature selection for interpretable tabular prediction, allowing the model to dynamically identify important features during decision making[17]. However, the sequential decision process may limit its ability to capture arbitrary feature interactions. TabTransformer further employs contextual embeddings and self-attention to model categorical feature interactions, demonstrating improved robustness over MLP-based approaches[22]. SAINT introduces row-wise attention and contrastive pre-training to enhance tabular representation learning by leveraging both feature-level and sample-level dependencies[23]. In addition, hybrid architectures such as neural oblivious decision ensembles combine the inductive bias of tree models with deep representation learning, achieving competitive performance on structured prediction tasks[24].

More recently, FT-Transformer proposed a feature tokenization strategy that converts numerical and categorical variables into token representations and applies Transformer encoders for tabular learning[13]. Compared with conventional multilayer perceptrons, FT-Transformer demonstrated improved capability in modeling complex feature interactions across various benchmark datasets.

Embedding-based neural networks have also been widely applied to structured prediction problems. By mapping categorical variables into dense latent representations, embedding models can capture semantic relationships among discrete attributes, which is particularly useful when dealing with high-cardinality categorical features such as equipment types.

However, existing tabular Transformer methods are primarily developed for general machine learning benchmarks and financial or commercial prediction tasks. In industrial equipment health assessment, feature relationships are constrained by physical degradation mechanisms. For example, increased vibration does not always indicate failure independently, but its combination with insufficient lubrication and excessive temperature may represent a strong degradation pattern. Therefore, incorporating physical knowledge into tabular Transformer architectures remains an open research problem.

2.4.Multi-task learning for fault prediction and RUL estimation

Fault prediction and RUL estimation represent two closely related aspects of equipment health management. Failure prediction focuses on identifying imminent abnormal conditions, while RUL estimation aims to quantify the remaining operating capability. Since both tasks rely on common degradation characteristics, multi-task learning provides a natural framework for integrating them[16,25].

The conventional multi-task learning paradigm usually adopts a shared-bottom architecture, where multiple tasks share a common feature extractor while maintaining task-specific output layers. Kendall et al. introduced uncertainty-based task weighting to automatically balance multiple objectives during training[14]. This approach avoids manually selecting fixed loss coefficients and has been widely adopted in multi-objective learning problems.

To improve task-specific representation learning, MoE architectures have been introduced. Ma et al. proposed the multi-gate mixture-of-experts (MMoE) framework, where different tasks utilize independent gating networks to select task-related expert representations[15]. This design reduces negative transfer between tasks and improves performance when task relationships are complex.

In PHM applications, several studies have explored joint fault diagnosis and RUL prediction. These methods generally employ shared feature extraction layers followed by task-specific prediction heads. However, existing approaches mainly focus on sequential sensor data and rarely consider heterogeneous tabular equipment information. Moreover, the relationship between physical degradation mechanisms and multi-task optimization remains insufficiently investigated.

2.5.Physics-informed learning for industrial prognostics

Purely data-driven models often require large amounts of labeled data and may generate predictions inconsistent with known physical principles. Physics-informed machine learning has emerged as a promising paradigm that integrates domain knowledge into neural network design, offering improved interpretability, generalization, and data efficiency[26,27]. In mechanical systems, degradation processes are governed by physical laws such as fatigue accumulation, thermal stress, and tribological wear, providing natural constraints for data-driven models.

Physics-informed neural networks (PINNs) represent a prominent line of research by embedding partial differential equations directly into the loss function[18]. While traditional PINNs mainly target forward and inverse problems in scientific computing, the underlying principle of incorporating physical knowledge has been increasingly extended to industrial prognostics[28].

For equipment prognostics, physics-guided features and constraints have been introduced to improve reliability under limited data conditions. However, most existing physics-informed PHM methods focus on time-series sensor signals or predefined degradation equations. Few studies investigate physics-informed representation learning for heterogeneous tabular equipment data.

Motivated by these observations (Table 1), this study introduces physics tokens into a Transformer-based tabular framework. Instead of directly concatenating manually designed physical features, the proposed method treats physical risk indicators as independent learnable tokens, allowing the Transformer to adaptively discover interactions between operational variables and degradation-related knowledge.

3.Methodology

3.1.Dataset

The dataset used in this study is a synthetic equipment condition dataset generated based on typical degradation patterns and failure characteristics of industrial equipment. Each sample represents the operational status and historical information of a single machine, encompassing heterogeneous features including machine type, operating hours, temperature, vibration, sound level, oil level, cooling level, power consumption, maintenance history, fault history, AI monitoring status, and recent error codes. The data generation process follows established physical degradation laws and maintenance logic observed in industrial practice, ensuring realistic feature distributions and failure patterns. The dataset (Table 2) contains approximately 260,000 samples. The problem is formulated as a joint prediction task consisting of two complementary objectives: binary classification of whether a machine will fail within the next seven days, with 93.96% normal and 6.04% failure samples indicating severe class imbalance, and continuous regression of RUL.

Table 2. Dataset statistics

Item	Value
Total samples	≈ 260000
Machine types	33
Numerical features	12
Categorical features	2
Physics-informed features	5
Fault samples	15184
Fault ratio	6.04%
RUL	0~1133

In practical industrial systems, equipment degradation is usually associated with several measurable physical phenomena, including thermal accumulation, mechanical vibration, lubrication deterioration, cooling performance degradation, and historical damage accumulation. However, these physical mechanisms are rarely directly provided as model inputs.Therefore, five physics-informed risk factors are constructed from the original monitoring variables. These factors do not represent strict physical equations but serve as domain-guided degradation indicators to introduce engineering knowledge into the deep learning model.

The five physical risk factors are described as follows:

(1) Thermal Risk Factor

$$R_{thermal}=\frac{T\cdot P}{C_l+\varepsilon} \quad (1)$$

where T is temperature, P is power consumption, C_l is coolant level or flow rate, and ε is a small constant to avoid division by zero.

(2) Vibration Impact Risk Factor

$$R_{vib}=V\cdot S \quad (2)$$

where V denotes vibration level and S denotes sound level.

(3) Lubrication Risk Factor

$$R_{lub}=\frac{log(1+H)\cdot log(1+D)}{O+\varepsilon} \quad (3)$$

where H is cumulative operating hours, D is days since last maintenance, and O is oil level percentage. Logarithmic transformation is applied to compress the dynamic range of time-related variables.

(4) Cooling Risk Factor

$$R_{cool}=\frac{T}{C+\varepsilon} \quad (4)$$

where C represents coolant level percentage.

(5) Historical Degradation Risk Factor

$$R_{hist}=\log(1+F_{\mathrm{h}})+\log(1+E_{30})+0.5\log(1+M_{\mathrm{h}}) \quad (5)$$

where F_h is the total number of historical faults, E_{30} is the number of error codes in the past 30 days, and M_h is the total number of maintenance records. The maintenance term is assigned a lower weight (0.5) since maintenance can be both a sign of aging and a corrective action.

The thermal risk factor is constructed by coupling temperature, power consumption and coolant level. Higher temperature and power consumption indicate increased heat generation, while lower coolant level reflects reduced heat dissipation capability. Vibration and acoustic signals jointly reflect mechanical impacts, imbalance and wear conditions. Their multiplicative coupling enhances the representation of abnormal mechanical states. The lubrication risk factor combines operating hours, maintenance interval and oil level. Logarithmic transformation is adopted to alleviate the influence of extremely large operating-hour values. Cooling system degradation directly threatens thermal stability. This factor quantifies the thermal load per unit of cooling resource. Historical failures and recent error codes are important indicators of equipment degradation. Maintenance history is incorporated with a smaller coefficient to reflect its indirect relationship with future failures.Collectively, these five risk factors form the physics-informed feature subset X_p.

3.2.Model framework

This study proposes a PIMTT for joint fault prediction and RUL estimation using heterogeneous equipment-level tabular data. The method is evaluated in this paper on synthetic records, where numerical measurements, categorical attributes, maintenance-related information, and physics-guided degradation indicators are simultaneously considered.

The overall framework consists of four major components (Fig 1.): (1) heterogeneous feature tokenization, (2) type-aware physics feature modulation, (3) Transformer-based health representation learning, and (4) multi-task prediction with adaptive optimization.

3.2.1.Heterogeneous feature tokenization

This study proposes a PIMTT for joint fault prediction and RUL estimation using heterogeneous equipment-level tabular data. The method is evaluated in this paper on synthetic records, where numerical measurements, categorical attributes, maintenance-related information, and physics-guided degradation indicators are simultaneously considered.

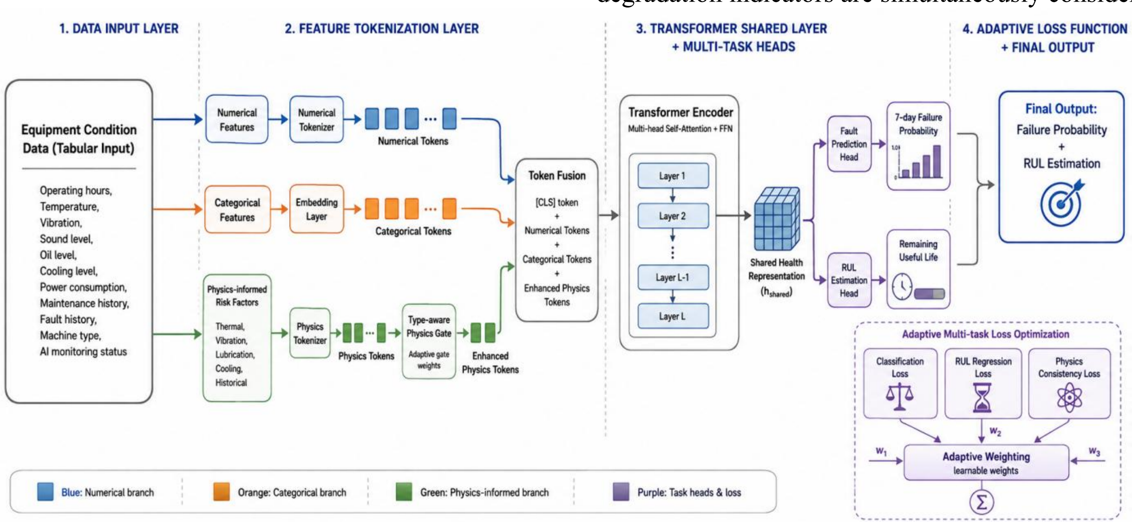

Fig 1. Model structure

Given an equipment condition sample, the input features are divided into three groups:

$$\boldsymbol{X}=\{X_n,X_c,X_p\} \quad (6)$$

where X_n denotes numerical operational features, X_c represents categorical equipment information, a X_p contains physics-informed risk factors. The numerical features include operating hours, temperature, vibration, sound level, oil level, cooling level, power consumption, maintenance-related variables, and historical failure information. The categorical features contain machine type and artificial intelligence monitoring status. Five physics-related variables are additionally constructed to represent thermal, vibration, lubrication, cooling, and historical degradation risks.

3.2.2.Type-aware physics enhancement

Different equipment types exhibit distinct degradation mechanisms: rotating machinery is more sensitive to vibration, whereas thermal critical equipment is primarily affected by thermal and cooling conditions. To capture this heterogeneity, a type-aware physics gate is introduced. The machine type embedding generates adaptive weights for the five physical tokens via a lightweight sigmoid-activated network, modulating the contribution strength of each physical factor. A residual enhancement strategy is adopted instead of discarding less relevant physical information:

$$Z_P' = Z_p(1+g) \tag{7}$$

where g represents the learned importance weights of physical factors.

3.2.3.Transformer-based shared representation learning

The proposed model transforms different feature types into unified token representations and feeds them into a Transformer encoder:

$$\boldsymbol{H} = \text{Transformer}([\text{CLS}], Z_n, Z_c, Z_p) \tag{8}$$

where Z_n, Z_c, and Z_p represent numerical tokens, categorical tokens, and physics tokens, respectively.

The output embedding of the classification token forms the shared equipment health representation $h = H_{\text{CLS}}$. This shared feature is fed into two task-specific prediction heads as follows:

$$\boldsymbol{h} = \boldsymbol{H}_{cls} \rightarrow \begin{cases} \hat{y}_f = f_f(\text{h}) \\ \hat{y}_r = f_r(\text{h}) \end{cases} \tag{9}$$

where $\hat{y}_f$ is the probability of failure within seven days, and $\hat{y}_r$ corresponds to the estimated RUL.

3.2.4.Adaptive multi-task loss optimization

The overall training objective integrates three components: fault classification loss, RUL regression loss and physics consistency loss:

$$L = \lambda_{cls}L_{cls} + \lambda_{rul}L_{rul} + \lambda_{\text{phy}}L_{\text{phy}} \tag{10}$$

Instead of manually assigning fixed coefficients, we adopt the uncertainty-based adaptive weighting strategy proposed by Kendall et al[14].

The loss is rewritten as:

$$L = \sum_{\text{i}} \left(\frac{1}{2\sigma_{\text{i}}^2} L_{\text{i}} + \log\sigma_{\text{i}} \right) \tag{11}$$

where σ_i represents the learned homoscedastic uncertainty of each task. This mechanism allows the model to automatically adjust the contribution of different objectives during training and avoids unstable optimization caused by manually selected loss weights.

3.3.Evaluation metrics

The proposed framework involves two prediction tasks: 7-day fault prediction and RUL estimation. Therefore, different evaluation metrics are adopted for classification and regression tasks and historical degradation risks.

3.3.1.Fault prediction task

Since the fault prediction task suffers from severe class imbalance, accuracy alone cannot fully reflect model performance. Therefore, Accuracy, Precision, Recall, and F1-score are adopted.

Accuracy measures the overall prediction correctness:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{12}$$

Precision evaluates the proportion of correctly identified failures among all predicted failures:

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

Recall measures the capability of detecting actual failures:

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

The F1-score provides a balanced measurement between Precision and Recall:

$$F1 = \frac{2Precision \times Recall}{Precision + Recall} \tag{15}$$

Considering that missed failures may cause severe economic losses in industrial applications, Recall and F1-score are considered more important indicators.

3.3.2.RUL task

For the RUL regression task, Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and coefficient of determination (R^2) are used.

MAE represents the average absolute deviation between predicted and actual RUL values:

$$MAE = \frac{1}{N}\sum_{\text{i}}^{N} |y_i - \hat{y}_i| \tag{16}$$

RMSE measures the prediction error with larger penalties for significant deviations:

$$RMSE = \sqrt{\frac{1}{N}\sum_{\text{i}}^{N} (y_i - \hat{y}_i)^2} \tag{17}$$

The coefficient of determination is calculated as:

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y}_i)^2} \tag{18}$$

Lower MAE and RMSE values indicate better RUL estimation accuracy, while a higher R^2 value indicates stronger explanatory capability.

4.Experiments

4.1.Experimental setup

To evaluate the effectiveness of the proposed PIMTT under controlled conditions, experiments were conducted on the synthetic industrial-equipment condition dataset introduced in Section 3.1.

The synthetic dataset was divided into training, validation, and testing subsets with a ratio of 6:2:2 using random seed 42 and stratification by the 7-day fault label. The training set was used for model optimization, the validation set was used for hyperparameter selection, early stopping, and classification-threshold selection, and the independent test set was used once for final performance evaluation. All preprocessing parameters, including missing-value statistics, numerical standardization, and categorical mappings, were fitted only on the training split and then applied unchanged to the validation and test splits.

All experiments were implemented in Python using PyTorch and scikit-learn-compatible libraries and executed on a workstation equipped with an NVIDIA V100 GPU. Random seeds for Python, NumPy, and PyTorch were fixed at 42;

deterministic CuDNN execution was enabled. Numerical features were standardized using training-set statistics, categorical variables were integer-encoded for embedding-based models and one-hot encoded for SVM, XGBoost, and the RUL target was standardized using the training-set mean and standard deviation for neural-network optimization before being transformed back to days for MAE, RMSE, and R^2 calculation. Fault imbalance was handled using class-balanced loss weights computed from the training split. The proposed model was compared with several representative baseline methods, including traditional machine learning models, deep learning models, Transformer-based tabular models, and multi-task learning models. The hyperparameter configurations of the proposed model are summarized in Table 3.

Table 3. Hyperparameters

Hyperparameter	Value
Token dimension	256
Num of Transformer layers	6
Hidden dimension	256
Number of attention heads	8
Feed-forward network dimension	1024
Dropout	0.15
Activation function	GELU
Output dimension	1
Loss function	Weight BCE/Huber
Optimizer	AdamW
Learning Rate	5×10^{-4}
Batch size	64
Epochs	100

4.2.Baseline implementation

The baseline models were trained using the same data split and random seed. SVM, XGBoost, CatBoost, LSTM, PINN and FT-Transformer were trained separately for fault classification and RUL prediction. MMoE and PIMTT were trained jointly for the two tasks. Data preprocessing was based only on the training set, while the validation set was used for early stopping and threshold selection.

SVM used C = 1.0 with balanced class weights. XGBoost used 300 trees, a learning rate of 0.06, and a maximum depth between 6 and 7. CatBoost used 320 iterations, a depth of 7, and a learning rate of 0.07. Since the dataset does not contain continuous time-series records, each sample was treated as a single-step input for LSTM. The LSTM had two layers with a hidden size of 128. The PINN baseline was implemented as a physics-guided multilayer perceptron with hidden sizes of 256, 128, and 64.

FT-Transformer contained 6 Transformer layers with 8 attention heads, a token dimension of 256, and a dropout rate of 0.15. MMoE used four shared experts and 2 task-specific gates. The neural models were generally trained using AdamW, a learning rate of 5×10^{-4}, a batch size of 64, and a maximum of 100 epochs. Early stopping was applied with a patience of 8 epochs.

4.3.Performance

The hyperparameters of PIMTT were determined through hyperparameter search based on the validation set. The selected configuration was used for all experiments. Table 4 presents the performance comparison between the proposed PIMTT and baseline methods on the testing set.

4.3.1.Fault prediction performance analysis

The classification results demonstrate that the proposed PIMTT achieves the best fault prediction performance among all compared methods.

Traditional machine learning models, including SVM, XGBoost, and CatBoost, achieve reasonable accuracy; however, their F1-scores remain relatively limited. Although XGBoost and CatBoost are strong baselines for structured data, they rely mainly on decision tree partitioning and cannot explicitly learn complex interactions among equipment operational variables.

The LSTM baseline improves fault prediction performance compared with traditional approaches, indicating that nonlinear representation learning can capture useful patterns in the synthetic equipment records. However, the input contains no temporal trajectory: each tabular sample is reshaped as a one-step pseudo-sequence. Consequently, this experiment should be interpreted as an architectural baseline rather than evidence of temporal modeling capability.

FT-Transformer achieves a higher F1-score than conventional models by representing the synthetic tabular features as tokens and learning their interactions through self-attention. This demonstrates the effectiveness of Transformer-based feature interaction modeling under the simulated data distribution.The MMoE model further improves the classification performance by introducing task-specific expert selection, achieving an F1-score of 0.8184. However, it does not explicitly incorporate equipment degradation knowledge.

The proposed PIMTT obtains the highest F1-score of 0.8461, improving by 3.38% compared with MMoE. This improvement indicates that introducing physics-informed tokens and equipment-aware feature modulation enables the model to better distinguish potential failure states under highly imbalanced conditions.

4.3.2.RUL prediction performance analysis

For RUL estimation, PIMTT achieves the lowest MAE and RMSE values and the highest R^2.

Compared with MMoE, the proposed model reduces RMSE from 34.72 to 30.39, corresponding to an improvement of approximately 12.5%. Meanwhile, R^2 increases from 0.9856 to 0.9899.

The improvement indicates that the integration of physics-informed tokens provides additional degradation-related information beyond conventional feature representations. By introducing thermal, vibration, lubrication, cooling, and historical degradation indicators, the model can better capture the relationship between equipment operating conditions and lifetime degradation.

4.4.Ablation study

To examine the contribution of the main components in PIMTT, ablation experiments were conducted by progressively introducing the physics-informed token representation, type-aware physics gate, and adaptive multi-task loss. The results are summarized in Table 5.

The base model achieves an F1-score of 0.7858 and an R^2 value of 0.9692. After adding the physics-informed tokens, the F1-score increases to 0.8133 and the R^2 value to 0.9757. The improvement suggests that the five constructed degradation factors provide useful information beyond the original equipment measurements.

Adding the type-aware physics gate further improves the F1-score to 0.8295 and the R^2 value to 0.9852. Compared with using physics-informed tokens alone, the gains are 0.0162 in F1 and 0.0095 in R^2. This result indicates that the relevance of physical degradation factors varies across equipment categories. Conditioning the physics representation on equipment type allows the model to emphasize more relevant factors while retaining the original physics information.

With the adaptive multi-task loss, the complete PIMTT reaches an F1 - score of 0.8461 and an R^2 value of 0.9899. Relative to the model without adaptive loss weighting, the F1-score increases by 0.0166 and the R^2 value by 0.0047. The result suggests that balancing fault prediction, RUL estimation, and physics-consistency objectives during training improves the shared health representation.

To further isolate the effect of joint learning, two single-task variants were evaluated using the same PIMTT backbone. PIMTT-ST-Fault is trained only for fault prediction and obtains an F1-score of 0.8214, whereas PIMTT-ST-RUL is trained only for RUL estimation and achieves an R^2 value of 0.9846. In comparison, the multi-task PIMTT improves the F1-score by 0.0247 and the R^2 value by 0.0053. Since the model architecture and physics representation are retained in the single-task variants, these differences provide more direct evidence that joint optimization of fault prediction and RUL estimation contributes to both tasks.

Overall, the ablation results show that the physics-informed tokens, type-aware physics gate, and adaptive multi-task learning each contribute to the final performance. The single-task comparison further indicates that the improvement of PIMTT is not solely due to its physical representation or Transformer backbone, but also benefits from the complementary supervision provided by fault prediction and RUL estimation.

Table 5. Ablation study of the proposed components

Model	Physics Token	Type-aware Gate	Fault Task	RUL Task	Multi-task Loss	F1	R^2
Base model	×	×	√	√	×	0.7858	0.9692
Physics Token	√	×	√	√	×	0.8133	0.9757
Type-aware Gate	√	√	√	√	×	0.8295	0.9852
PIMTT-Fault	√	√	√	×	×	0.8214	-
PIMTT-RUL	√	√	×	√	×	-	0.9846
PIMTT	√	√	√	√	√	0.8461	0.9899

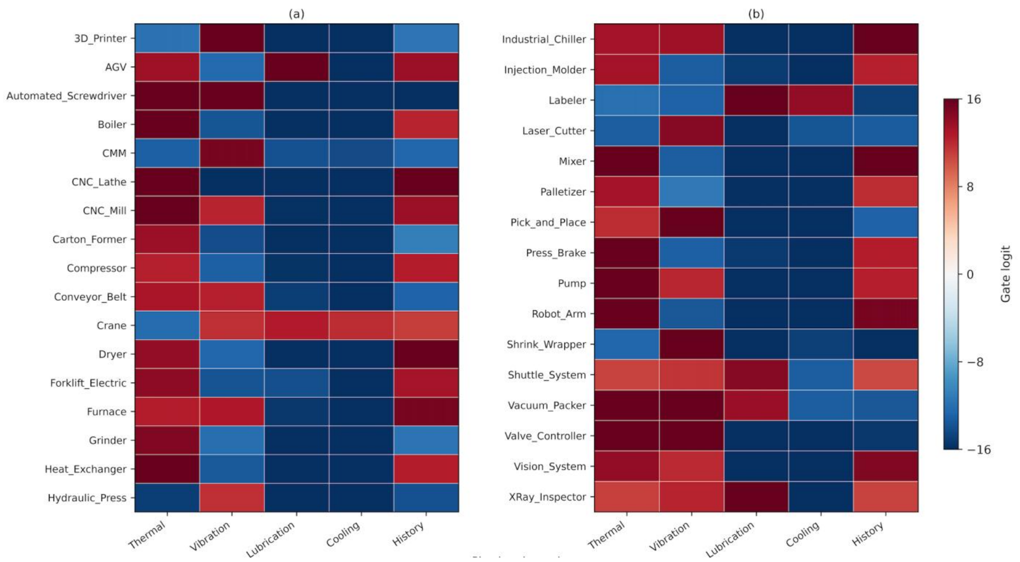

Fig 2. Visualization of type-aware physics gate strengths (logit scale)

5.Discussion

The analysis of model predictions shows that the five most influential features are operating hours, vibration impact risk factor, temperature, historical fault risk factor, and error codes within the last 30 days. Among these variables, operating hours has the largest contribution, indicating that accumulated operating conditions are closely related to equipment degradation. Long-term operation may cause component wear, fatigue accumulation, and gradual performance decline, leading to a higher probability of failure. The vibration impact risk factor also plays an important role, which is consistent with the fact that abnormal vibration is commonly associated with mechanical problems such as imbalance, looseness, and structural wear. Temperature reflects thermal stress during operation, while historical fault information and recent error codes provide additional information about previous abnormal conditions and maintenance status. These results indicate that both current operating states and historical degradation information are important for equipment health assessment.

The behavior of the type-aware physics gate was further investigated using the activation statistics of physics channels and the gate strength visualization. As shown in Table 6, thermal information has the highest activation ratio among the five physical factors, with 25 out of 33 equipment types selecting this channel. This suggests that thermal-related degradation is a common phenomenon across different industrial equipment. Historical degradation and vibration factors also show relatively high activation ratios, indicating that past operating conditions and mechanical behavior are important degradation indicators.

The lubrication and cooling factors show lower activation ratios, which is mainly related to their equipment-dependent characteristics. Lubrication degradation is more relevant to equipment with rotating parts, transmission structures, or friction-related mechanisms, while it has limited influence on equipment without these components. Similarly, cooling information is mainly required for equipment with specific thermal management requirements. For many devices, temperature-related degradation can already be partially represented by thermal information, reducing the necessity of separately emphasizing cooling conditions. Therefore, the lower activation frequency of these factors reflects the selective behavior of the gate rather than their lack of usefulness.

Table 6. Statistics of physics-channel activation

Channel	Activated machine types	Activited ratio
Thermal	25/33	75.76%
History	19/33	57.58%
Vibration	17/33	51.52%
Lubrication	6/33	18.18%
Cooling	2/33	6.06%

The visualization of type-aware physics gate strengths (Fig 2.) further shows that different equipment categories have different physical preferences. The gate does not assign identical physical representations to all equipment types but adjusts the contribution of each factor according to equipment characteristics. For example, machining-related equipment tends to place more emphasis on thermal and vibration information, which is consistent with the influence of machining processes, mechanical stress, and component wear. Equipment with transmission mechanisms shows stronger responses to lubrication and historical degradation information. These patterns indicate that the proposed gate can capture differences in degradation mechanisms among heterogeneous industrial equipment.

Although the proposed PIMTT model achieves promising results, several limitations remain. First, the current study focuses on static tabular equipment records and does not consider continuous sensor sequences. In practical applications, equipment degradation is usually a progressive process, and temporal changes may provide additional information. Second, the physics-informed factors used in this work are constructed based on general degradation knowledge. Although these factors improve model performance, more automatic methods for extracting physical representations could be explored in the future. Third, the validation is currently performed on one equipment dataset, and further evaluation on data from different factories and operating environments is required to verify the generalization ability of the model.

6.Conclusion

In this paper, we propose a PIMTT for joint fault prediction and RUL estimation from heterogeneous equipment-level tabular records. The method is evaluated on a synthetic dataset designed to emulate operational conditions, equipment attributes, maintenance information, and degradation-related measurements; no real factory observations are used in the present experiments.

The proposed method characterizes equipment status with three types of tokens: numerical tokens, categorical tokens and physics-informed tokens. From raw monitoring variables, we construct five degradation-related risk factors to reflect thermal, vibration, lubrication, cooling and historical degradation features. We also design a type-aware physics gate to adjust the contribution of physical information for different equipment categories. A multi-task learning strategy is adopted to jointly optimize seven-day fault prediction and RUL estimation, so that the model can learn shared equipment health representations from both classification and regression tasks.

We carry out experiments on a synthetic industrial-equipment condition dataset. Under this controlled simulated setting, PIMTT performs better than the evaluated traditional machine-learning models, general deep-learning methods, tabular Transformer models, and multi-task baselines. The model achieves an F1-score of 0.8461 for fault prediction and an R^2 value of 0.9899 for RUL estimation. Ablation studies indicate that physics-informed representations, type-aware physical modulation, and multi-task optimization contribute to performance on this synthetic benchmark. These results should not be interpreted as validated real-world industrial performance.

Further analysis of the learned physics gate indicates that different equipment categories show different degradation preferences. This suggests our method can capture category-specific degradation features, rather than applying identical physical assumptions to all devices.

For future work, we will validate the proposed method on independently collected real-world industrial datasets with time-stamped sensor streams, diverse operating regimes, and documented failure events. We will also investigate temporal degradation patterns, equipment interdependencies, and cross-domain generalization to quantify the gap between the present synthetic benchmark and operational deployment.

References

[1] MOBLEY R K. An introduction to predictive maintenance. Oxford, UK: Butterworth-Heinemann, 2002.

[2] PENG Y, DONG M, ZUO M J. Current status of machine prognostics in condition-based maintenance: a review. The International Journal of Advanced Manufacturing Technology, 2010, 50(1): 297-313.

[3] LEI Y, LI N, GUO L, et al. Machinery health prognostics: A systematic review from data acquisition to RUL prediction. Mechanical Systems and Signal Processing, 2018, 104: 799-834.

[4] JARDINE A K S, LIN D, BANJEVIC D. A review on machinery diagnostics and prognostics implementing condition-based maintenance. Mechanical Systems and Signal Processing, 2006, 20(7): 1483-1510.

[5] VAPNIK V N. The Nature of Statistical Learning Theory. New York: Springer, 1995.

[6] CHEN T, GUESTRIN C. XGBoost: A scalable tree boosting system//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY: ACM, 2016: 785-794.

[7] CARVALHO T P, SOARES F A, VITA R, et al. A systematic literature review of machine learning methods applied to predictive maintenance. Computers & Industrial Engineering, 2019, 137: 106024.

[8] LECUN Y, BENGIO Y, HINTON G. Deep learning. Nature, 2015, 521(7553): 436-444.

[9] ZHAO R, YAN R Q, CHEN Z H, et al. Deep learning and its applications to machine health monitoring. Mechanical Systems and Signal Processing, 2019, 115: 213-237.

[10] HOCHREITER S, SCHMIDHUBER J. Long short-term memory. Neural Computation, 1997, 9(8): 1735-1780.

[11] ZHENG S, RISTOVSKI K, FARAHAT A, et al. Long short-term memory network for remaining useful life estimation//2017 IEEE International Conference on Prognostics and Health Management. Piscataway, NJ: IEEE, 2017: 88-94.

[12] PROKHORENKOVA L, GUSEV G, VOROBEV A, et al. CatBoost: unbiased boosting with categorical features. Advances in Neural Information Processing Systems, 2018, 31: 6638-6648.

[13] GORISHNIY Y, RUBACHEV I, KHRULKOV V, et al. Revisiting deep learning models for tabular data. Advances in Neural Information Processing Systems, 2021, 34: 18932-18943.

[14] KENDALL A, GAL Y, CIPOLLA R. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics//IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 7482-7491.

[15] MA J, ZHAO Z, YI X, et al. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts//Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, NY: ACM, 2018: 1930-1939.

[16] ZHANG C, ZHANG X, WANG Y, et al. Multi-task learning for machinery fault diagnosis and remaining useful life prediction: A review. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 1-15.

[17] ARIK S O, PFISTER T. TabNet: Attentive interpretable tabular learning//AAAI Conference on Artificial Intelligence. Menlo Park, CA: AAAI Press, 2021: 6679-6687.

[18] RAISSI M, PERDIKARIS P, KARNIADAKIS G E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics, 2019, 378: 686-707.

[19] LI X, ZHANG W, MA H, et al. Transformer-based fault diagnosis and remaining useful life prediction for rotating machinery: A review. Knowledge-Based Systems, 2024, 292: 111612.

[20] MO H, ZUO M J, WANG D. Remaining useful life estimation using a temporal convolutional network with attention mechanism. IEEE Transactions on Industrial Electronics, 2023, 70(10): 10458-10467.

[21] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need. Advances in Neural Information Processing Systems, 2017, 30: 5998-6008.

[22] HUANG X, KHETAN A, CVITKOVIC M, et al. TabTransformer: Tabular data modeling using contextual embeddings[Z/OL]. arXiv:2012.06678, 2020.

[23] SOMEPALLI G, GOLDBLUM M, SCHWARZSCHILD A, et al. Saint: Improved neural networks for tabular data via row attention and contrastive pre-training[Z/OL]. arXiv:2106.01342, 2021.

[24] POPOV S, MOROZOV S, BABENKO A. Neural oblivious decision ensembles for deep learning on tabular data//International Conference on Learning Representations, 2019.

[25] FINK O, ZIO E, WEIDMANN U. A framework for prognostic uncertainty assessment based on multiple degradation process characteristics. IEEE Transactions on Reliability, 2014, 63(2): 439-448.

[26] KARNIADAKIS G E, KEVREKIDIS I G, LU L, et al. Physics-informed machine learning. Nature Reviews Physics, 2021, 3(6): 422-440.

[27] WILLARD J, JIA X, XU S, et al. Integrating physics-based modeling with machine learning: A survey[Z/OL]. arXiv:2003.04919, 2020.

[28] WANG H, XU C, LIU Y, et al. Physics-informed neural networks for machine fault diagnosis and remaining useful life prediction: A review. Reliability Engineering & System Safety, 2024, 245: 109965.