



Article

Tool Wear Prediction in Complex Machining Processes: A Hybrid Residual-Compensated Deep Learning Framework

Fucong Liu ^{1,2,*} , Faqiang Wen ^{1,*}, Baokaidi Tian ¹, Lei Yu ¹, Tianxiang Yu ¹, Min Li ¹, Yixin Geng ¹ and Sai Lou ²

¹ Tianjin High-End Intelligent Machine Tools Engineering Research Center, Tianjin University of Technology and Education, No 1310, Dagou South Road, Tianjin 300222, China

² Tianjin BestBond (Baishibingde) Intelligent Technology Co., Ltd., No. 42-B10, Baidi Road, Nankai District, Tianjin 300300, China

* Correspondence: tjuchong@126.com (F.L.); andsen304218@163.com (F.W.)

Abstract

Accurate tool wear prediction is essential for predictive maintenance in complex machining processes, but non-stationary sensor signals make it difficult for a single model to capture both long-term degradation trends and local transient disturbances. This study introduces a residual-compensated Hybrid CNN-Informer + LightGBM framework for tool wear prediction. The workflow first preprocesses multi-source sensor signals and selects wear-sensitive statistical descriptors to guide FiLM-based deep feature modulation. A CNN-Informer backbone then estimates the main wear trend by combining local feature extraction with long-range temporal modeling, and a LightGBM module performs secondary compensation on the remaining prediction residuals. On the PHM 2010 milling benchmark, the proposed framework achieved an RMSE of 4.0016, MAE of 2.8271, and R^2 of 0.9870, reducing RMSE and MAE by 47.9% and 48.5% compared with a standard Transformer. Ablation results showed that both the CNN branch and residual compensation contributed to the final accuracy. External validation on the HMoTP dataset using an independent held-out tool further yielded an RMSE of 9.1111, MAE of 7.3522, and R^2 of 0.9729. These results indicate that separating main-trend learning from residual correction provides a practical strategy for robust tool wear prediction under the tested machining conditions.

Keywords: tool wear prediction; residual compensation; hybrid deep learning; non-stationary sensor signals; predictive maintenance

1. Introduction

In modern advanced manufacturing systems, the surface finish quality achieved through precision machining directly determines the fatigue life and in-service performance of high-end equipment (such as aircraft engine blades and new energy vehicle transmission systems) [1]. However, machine tools operate under harsh conditions characterized by high loads and intense friction over extended periods, making severe wear of cutting tools inevitable. Severe tool wear not only leads to deterioration of surface integrity (such as chatter marks and thermal burns) but also causes significant unplanned downtime [2,3]. Traditional reactive maintenance or periodic preventive maintenance strategies often result in either excessive maintenance or sudden equipment failure. Therefore, the development of data-driven predictive maintenance (PdM) technologies—which utilize multi-source, heterogeneous sensor signals to assess tool health in real time and accurately predict tool wear—has become a key research focus in the field of smart manufacturing [4].



Received: 28 May 2026

Revised: 9 July 2026

Accepted: 10 July 2026

Published: 12 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

In recent years, significant progress has been made in data-driven methods for predicting tool wear. Early research primarily relied on traditional machine learning algorithms, such as Support Vector Regression (SVR) and Random Forest (RF). These methods are highly dependent on manual feature engineering by domain experts and often encounter performance bottlenecks due to limited feature extraction capabilities when dealing with massive, high-dimensional industrial non-stationary monitoring signals [5]. With the advancement of deep learning, deep architectures represented by Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) Networks have been widely applied to tool condition monitoring, enabling end-to-end feature learning [6]. CNNs excel at capturing local spatial features of signals but have limited scope when processing long sequences [7], whereas the serial computation mechanism of LSTMs, despite their temporal memory capabilities, limits the efficiency of modeling long-range dependencies, and they remain prone to the vanishing gradient problem during backpropagation in ultra-long sequences [8].

To overcome the bottlenecks in long-sequence modeling, Transformer-based models utilizing self-attention mechanisms have gradually been introduced into the field of fault diagnosis and prediction, demonstrating outstanding performance in capturing global degradation trends across the entire lifecycle [9,10]. Nevertheless, standard Transformer models still face two major challenges in practical machining applications: first, the computational complexity of the standard attention mechanism grows quadratically with sequence length ($O(L^2)$), resulting in high computational costs when processing high-frequency sensor signals [11]; second, global attention mechanisms often overemphasize macro-level features while being insufficiently sensitive to local high-frequency transient shocks caused by micro-chipping, micro-cracks, or chatter during milling processes, which can easily lead to model failure in detecting early-stage, subtle faults or result in prediction lag [12].

However, most existing deep learning-based prediction models rely on an idealized assumption: that the tool degradation process is relatively stationary. In real-world machine tool cutting operations, sensor signals are highly non-stationary. These signals not only contain a “stationary main trend” reflecting gradual tool wear but are also interspersed with a large number of nonlinear random disturbances caused by sudden changes in the depth of cut, material inhomogeneities, or electrical interference [13,14].

When dealing with complex, non-stationary signals from machine tool cutting processes, single-architecture deep neural networks face significant trade-offs in multi-scale feature extraction and modeling. When attempting to use a single loss function to simultaneously fit both the smooth, main degradation trend and high-frequency abrupt disturbances, the model is highly prone to parameter explosion and “black-box” feature confusion. If the network tends to fit the smooth main trend, it will generate substantial prediction errors at local abrupt changes; conversely, if it focuses excessively on local disturbances, it is highly prone to overfitting. How to decouple the architecture to enhance the model’s predictive accuracy while improving its interpretability of wear-related signal features has become a major challenge in current research [15].

To overcome the limitations of single-model architectures in capturing both stationary and non-stationary features, this paper proposes a two-stage hybrid deep learning framework for tool wear prediction based on residual compensation. Instead of attempting to fit all signal components using a single end-to-end architecture or direct feature concatenation, the proposed framework separates main-trend modeling from residual-error compensation and further incorporates wear-sensitive statistical descriptors to guide deep feature learning. The main contributions of this paper are summarized as follows:

- We propose a dual-stream parallel hybrid CNN-Informer deep prediction architecture for main degradation trend feature extraction, cascaded with LightGBM for

residual compensation. By utilizing a 1D-CNN branch to efficiently extract local time-frequency features, while introducing an Informer network (ProbSparse Attention) with $\mathcal{O}(L \log L)$ complexity to capture long-range temporal dependencies across the entire lifecycle, we achieve effective baseline prediction of macro-level trends.

- We propose a statistical-feature-guided data-driven dual modeling strategy for decoupling the main degradation trend from the residual. During the deep feature extraction stage, we innovatively introduce Feature-wise Linear Modulation (FiLM) to utilize wear-sensitive time–frequency statistical features to guide the global main trend prediction by deep networks; simultaneously, treating the prediction residuals of the deep model as under-exploited nonlinear residual information, we employ LightGBM gradient-boosted trees to perform a secondary fine compensation. This effectively decouples the stationary degradation patterns from non-stationary random disturbances, providing a consistent improvement to the model’s robustness under complex and variable operating conditions.
- The model’s robust performance, statistical interpretability and wear-related signal analysis were validated using a benchmark dataset of complex machining operations. Systematic experiments based on the PHM 2010 milling dataset demonstrate that the predictive accuracy of this hybrid framework significantly outperforms existing mainstream methods. More importantly, statistical analysis shows that this framework has effectively extracted the wear-related signal features from the monitoring signals.

2. Theoretical Background

To effectively extract complex features from sensor signals in machining processes and decouple their stationary and non-stationary components, this paper proposes a hybrid prediction framework that integrates local perception, global long-range modeling, and nonlinear residual approximation. This section briefly outlines the core algorithmic foundations underlying this framework and its mathematical rationale as applied to the processing of time-series signals in machine tool cutting operations.

2.1. One-Dimensional Convolutional Neural Network (1D-CNN)

Due to its weight sharing and local receptive field characteristics, the 1D-CNN is highly effective at extracting features when processing high-frequency time series [16]. During machining operations, minute tool damage often manifests as localized, transient high-frequency spikes in cutting force or vibration signals. By sliding the convolutional kernels along the time axis, 1D-CNNs can effectively capture these micro-scale signal distortions. Its forward propagation process can be expressed as:

$$y_i^l = f \left(\sum_{k=1}^K w_k^l \cdot x_{i+k-1}^{l-1} + b^l \right) \quad (1)$$

Here, x^{l-1} represents the input feature map of layer $l - 1$, y_i^l represents the local features output by layer l , w_k^l represents the weights of a convolution kernel of size K , b^l represents the bias term, and $f(\cdot)$ denotes the nonlinear activation function (ReLU). By stacking multiple layers of 1D-CNNs and pooling layers, the network is able to extract high-level local representations with translation invariance from the raw noisy signal.

2.2. Informer and ProbSparse Self-Attention

Although 1D-CNN excels at local feature extraction, its ability to capture long-range degradation dependencies across the full tool life cycle is hindered by its limited receptive field when processing ultra-long sequences ($L = 4096$). While standard Transformer models address long-range dependency issues through self-attention mechanisms, their

standard dot-product attention matrices entail a computational complexity of $\mathcal{O}(L^2)$. This leads to significant memory overhead and computational bottlenecks during long-sequence prediction, as illustrated in Figure 1 [17].

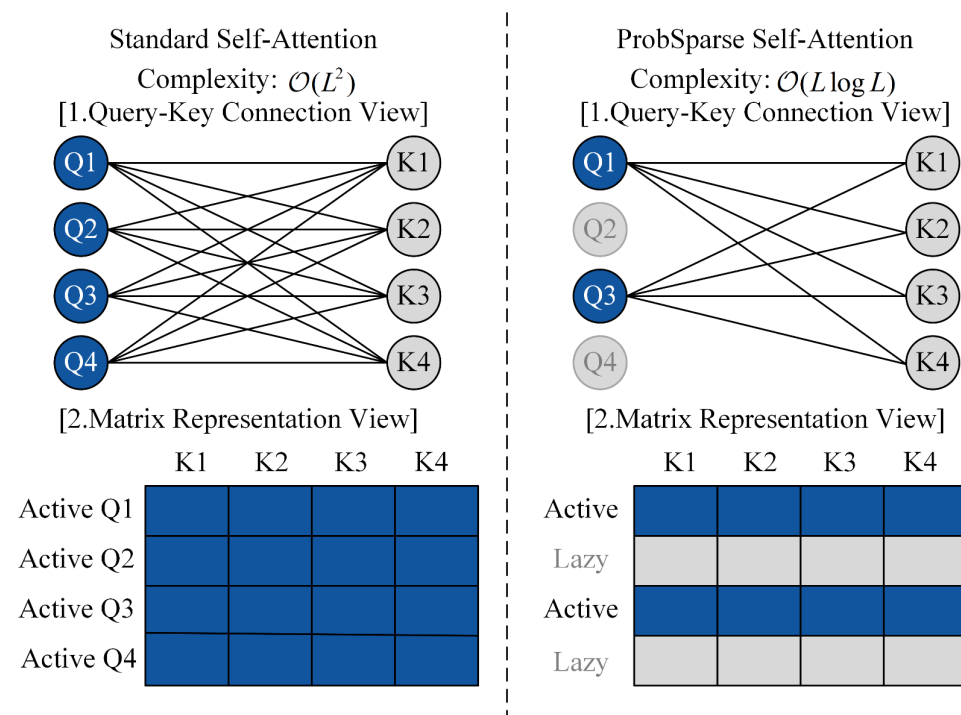


Figure 1. Comparison between Standard Self-Attention and ProbSparse Self-Attention mechanisms (blue indicates active queries; gray indicates inactive or lazy queries).

To overcome these limitations, the Informer model is introduced [18]. The core innovation of Informer lies in the ProbSparse Self-Attention mechanism. Its mathematical foundation is based on the observation that in self-attention calculations, only a small fraction of Queries significantly contribute to the final attention weight distribution, exhibiting a long-tail distribution. Informer assesses Query “activeness” by calculating the approximate Kullback–Leibler (KL) divergence between the Query and Key distributions. Only the top- u active queries are selected for dot-product operations:

$$\mathcal{A}(Q, K, V) = \text{Softmax}\left(\frac{\overline{Q}K^T}{\sqrt{d_k}}\right)V \quad (2)$$

where $\overline{Q}$ is a sparse matrix containing only the selected active queries, with its dimension reduced from $L \times d_k$ to $u \times d_k$ (where $u = c \cdot \ln L$, and c is a constant factor). This mechanism successfully reduces the time and space complexity of long-sequence modeling from $\mathcal{O}(L^2)$ to $\mathcal{O}(L \log L)$, enabling the model to efficiently capture the macroscopic main degradation trend of the tool within a large observation window.

2.3. Gradient Boosting Decision Trees and Residual Approximation via LightGBM

While deep neural networks (such as CNN and Informer) excel at approximating continuous and smooth functional mappings, they often produce structural prediction errors (residuals) when fitting highly non-stationary machining signals that contain abrupt changes and significant random noise. To perform secondary modeling on these nonlinear perturbations that are difficult for deep networks to absorb, this paper introduces LightGBM as a residual compensator [19].

LightGBM is a distributed gradient boosting framework based on decision tree algorithms. Compared to traditional GBDT, it introduces Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB), significantly enhancing training speed and generalization capability for high-dimensional features. In the residual compensation mechanism proposed in this paper, the goal of LightGBM is to fit the pseudo-residuals generated by the preceding deep model. Its additive modeling and forward distribution algorithm can be expressed as:

$$F_M(x) = \sum_{m=1}^M \gamma_m h_m(x) \quad (3)$$

where $h_m(x)$ denotes the m -th regression tree and γ_m is its weight (often set to 1, with the leaf weights absorbing the magnitude). In the m -th iteration, the model first computes the negative gradient (pseudo-residual) of the loss function with respect to the previous prediction:

$$r_{i,m} = - \left[\frac{\partial \mathcal{L}(y_i, F(x_i))}{\partial F(x_i)} \right]_{F(x)=F_{m-1}(x)} \quad (4)$$

Then, a new regression tree h_m is grown to fit these pseudo-residuals by minimizing the squared error:

$$h_m = \operatorname{argmin}_h \sum_{i=1}^n (r_{im} - h(x_i))^2 \quad (5)$$

This tree-based ensemble learning method exhibits exceptional robustness against outliers and non-stationary fluctuations, effectively compensating for the limitations of deep learning models when addressing transient shocks. The residual-fitting process is summarized in Figure 2.

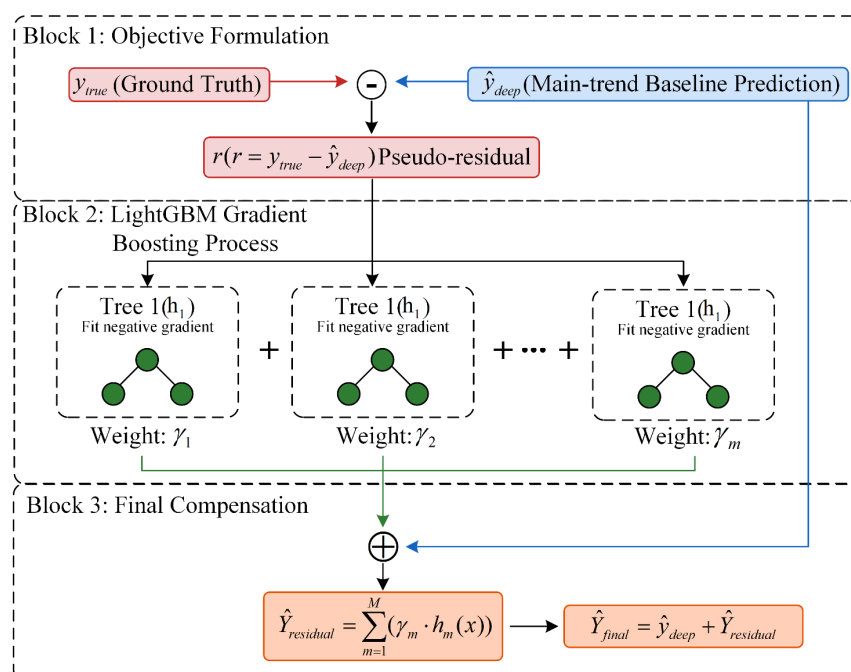


Figure 2. Schematic diagram of residual fitting via LightGBM decision trees (additive model).

3. Proposed Methodology

To address the highly non-stationary nature of sensor signals in machining processes and the strong coupling between local transient shocks and global degradation features, this paper proposes a two-stage (Baseline-to-Fine Compensation) hybrid tool wear prediction

framework based on residual compensation. This section details the overall architecture of the framework, the multi-stage data preprocessing pipeline, the statistical-feature-guided FiLM fusion mechanism, and the specific implementations of residual compensation and automated hyperparameter optimization. Figure 3 provides an overview of the complete framework.

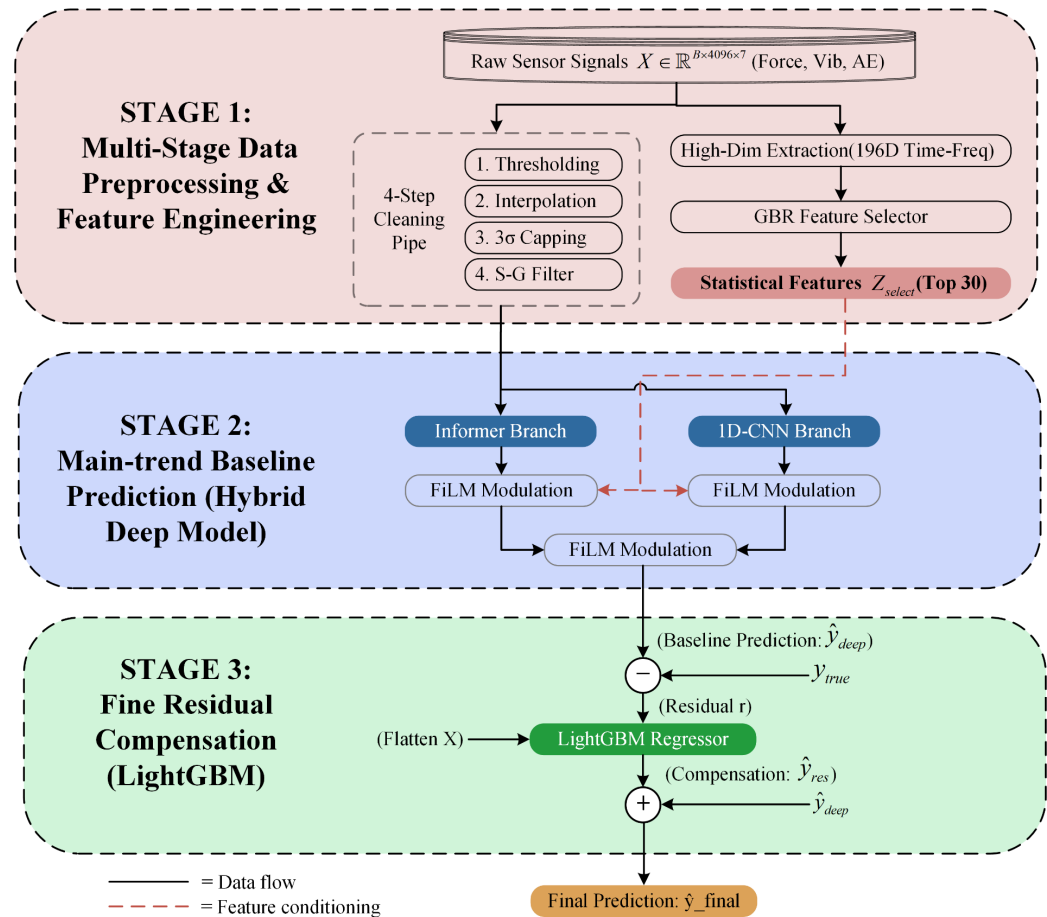


Figure 3. The overall architecture of the proposed hybrid residual-compensated framework.

3.1. Overall Architecture

The proposed two-stage prediction architecture consists of the following three progressive execution stages:

Stage 1: Multi-Stage Data Preprocessing and Feature Engineering. The collected multi-source heterogeneous raw sensor signals ($X \in \mathbb{R}^{B \times 4096 \times 7}$) undergo a four-step fidelity denoising and standardization pipeline. Simultaneously, high-dimensional time–frequency statistical features are extracted, and a Gradient Boosting Regressor (GBR) is utilized to select the most representative wear-sensitive statistical features.

Stage 2: Main-trend Baseline Prediction via Conditional Modulation. A parallel Hybrid CNN-Informer backbone network is constructed. Within this backbone, the 1D-CNN branch is responsible for extracting transient shock patterns, while the Informer branch handles long-range sequence modeling. Feature-wise Linear Modulation (FiLM) layers are innovatively introduced to dynamically calibrate the deep feature maps using the wear-sensitive statistical features extracted in Stage 1, outputting the main degradation trend baseline prediction $\hat{y}_{deep}$ for the tool wear state.

Stage 3: Fine Residual Compensation. The pseudo-residual $r = y - \hat{y}_{deep}$ between the true degradation label and the deep model’s prediction is calculated. Leveraging LightGBM’s robust nonlinear fitting and anti-noise capabilities, secondary modeling is

performed on this residual. The final tool wear prediction is obtained through the linear addition of the deep network's prediction and the LightGBM's compensation value: $\hat{y}_{final} = \hat{y}_{deep} + \hat{y}_{res}$.

3.2. Multi-Stage Data Preprocessing and Feature Engineering

Industrial sensors (e.g., cutting force, vibration, and AE) are susceptible to environmental interferences such as electrical discharge and mechanical resonance. To prevent non-physical extreme values and high-frequency background noise from disrupting the gradient updates of the deep neural network, this paper designs a rigorous four-stage data cleaning and feature extraction pipeline.

3.2.1. Four-Stage Fidelity Signal Cleaning Pipeline

Physical Range Validation: Hard threshold boundaries are established based on sensor measurement ranges and common knowledge of machining operations (e.g., cutting force $[-2000, 2000]$ N). Abrupt data points exceeding these physical boundaries are forced to NaN (not a number) values to eliminate severe sensor outliers.

Linear Interpolation: A linear interpolation method is applied to fill the missing values using valid data from preceding and succeeding timestamps, ensuring the continuity of the sequence.

Statistical Capping (3σ): To eliminate occasional high-amplitude electromagnetic interference (spikes), all data points falling outside the range of $\mu \pm 3\sigma$ are smoothly truncated to the boundary values $x_{clip} = \text{clip}(x, \mu - 3\sigma, \mu + 3\sigma)$ in accordance with Pauta's Criterion (3σ rule).

Savitzky–Golay Filter (S-G Filter): Traditional moving average filters severely smooth out signal peaks that reflect local tool chipping. Therefore, this paper adopts the Savitzky–Golay filter. By utilizing local polynomial least-squares fitting, the S-G filter effectively preserves the width and amplitude of the local transient shocks in the original signal while smoothing out high-frequency white noise. Figure 4 compares the raw and filtered signals [20].

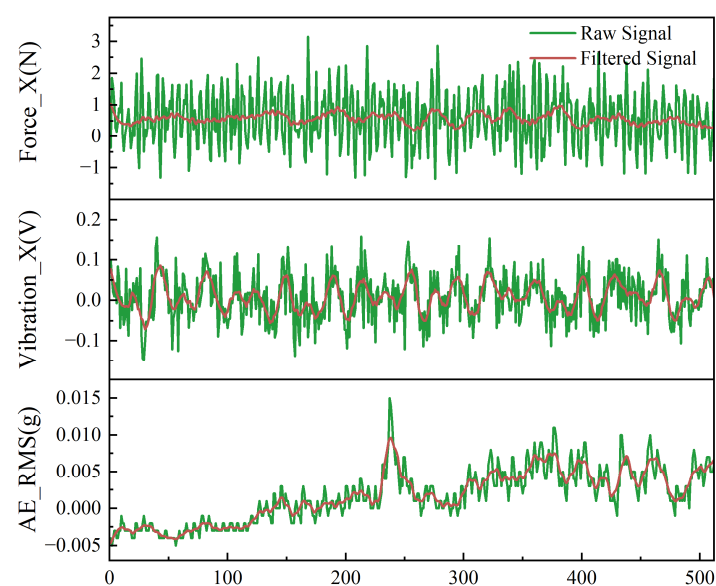


Figure 4. Comparison between raw sensor signals and filtered signals after the four-stage fidelity cleaning pipeline.

3.2.2. High-Dimensional Statistical Feature Extraction and GBR Selection

Directly inputting cleaned time series (length $L = 4096$) into deep models preserves raw information but often leads to slow convergence. To inject wear-sensitive statistical descriptors of machining into the model, 196-dimensional high-level statistical features are extracted from both the time and frequency domains. Time-domain features include root mean square (RMS), crest factor, skewness, and kurtosis, which are highly sensitive to early mechanical shocks and friction [21]. Frequency-domain features are extracted via fast Fourier transform (FFT), including Spectral Centroid and frequency-band energy ratios.

To eliminate feature collinearity and redundancy, a Gradient Boosting Regressor (GBR) model is employed for importance evaluation and feature selection. The top- K (e.g., $K = 30$) core statistical features $Z_{select} \in \mathbb{R}^{B \times 30}$ with the highest cumulative contribution rates are selected to serve as the input for subsequent conditional modulation of the deep network. To avoid information leakage, the GBR feature selector was trained only on the training subset after the training/validation/testing split, and the resulting top-30 feature set was then fixed and applied unchanged to the validation and testing subsets. Figure 5 shows the resulting feature-importance ranking.

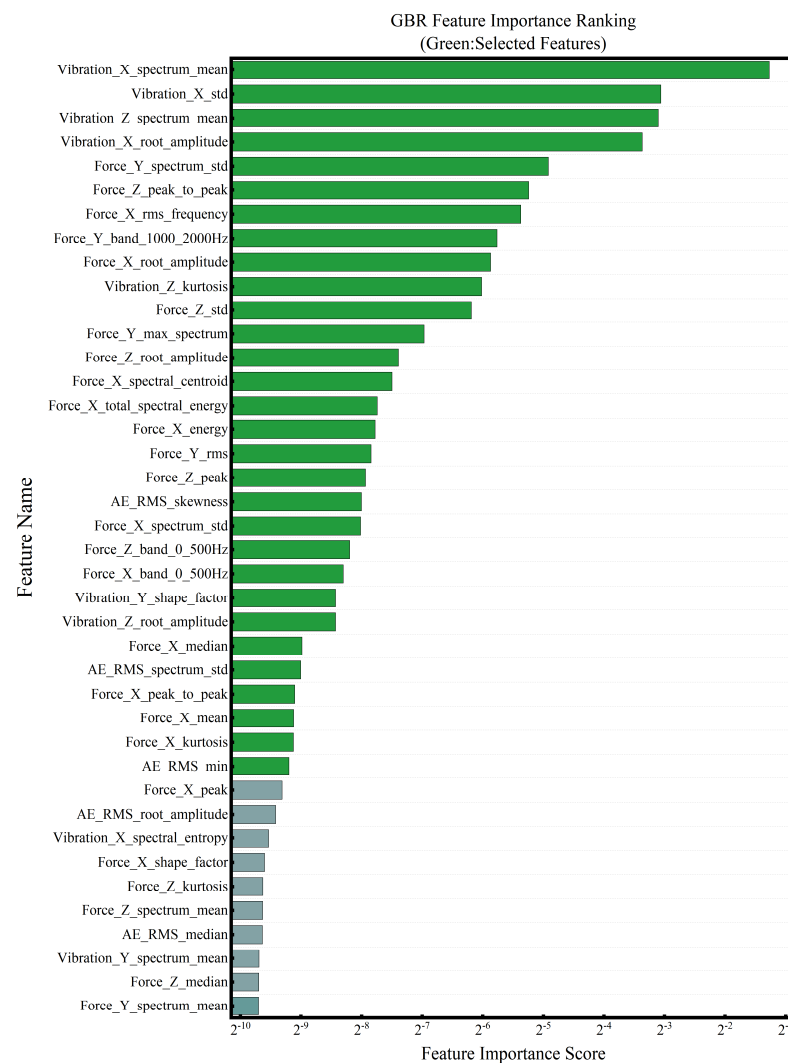


Figure 5. Importance evaluation of the Top-30 core statistical features based on GBR (green indicates selected features).

3.3. Statistical-Feature-Guided FiLM Modulation

Most existing studies employ a “late fusion” strategy by simply concatenating hand-crafted statistical features with deep features. This approach may weaken the contribution of wear-sensitive time- and frequency-domain descriptors within high-dimensional deep representations. In this study, these descriptors are not treated as physical laws, mechanistic constraints, or tribological wear equations; rather, they are data-derived, process-relevant indicators extracted from cutting force, vibration, and acoustic emission signals. To make better use of these descriptors, this paper introduces a Feature-wise Linear Modulation (FiLM) mechanism into the Hybrid CNN-Informer backbone.

As an early fusion strategy, the FiLM module allows manual statistical features Z_{select} to dynamically intervene in the activation response of feature maps during the forward propagation of the deep network. Specifically, for the local features $F_{cnn} \in \mathbb{R}^{B \times C_{cnn}}$ extracted by the CNN branch and the global features $F_{inf} \in \mathbb{R}^{B \times C_{inf}}$ extracted by the Informer branch, the FiLM layer utilizes two independent multi-layer perceptron (MLP) networks, f_γ and f_β , to map the statistical features into scaling factors γ and shifting factors β for an affine transformation:

$$\gamma_c = \text{Sigmoid}(f_\gamma(Z_{select})), \quad \beta_c = f_\beta(Z_{select}) \quad (6)$$

Subsequently, a channel-wise affine calibration is performed on the deep feature maps:

$$F'_{mod} = \gamma \odot F_{original} + \beta \quad (7)$$

where the scaling factor $\gamma \in (0, 1)$ is constrained by a sigmoid function to prevent gradient explosion. From a process-signal perspective, the scaling factor adaptively amplifies or suppresses channel responses associated with specific frequency bands according to global vibration or energy levels [22], while the shifting factor helps correct signal baseline drift caused by different machining operations or idling states. By introducing a residual connection $F_{out} = \text{LayerNorm}(F_{original} + \text{Dropout}(F'_{mod}))$, the stability of gradient backpropagation is ensured. Figure 6 summarizes the FiLM-based conditional modulation.

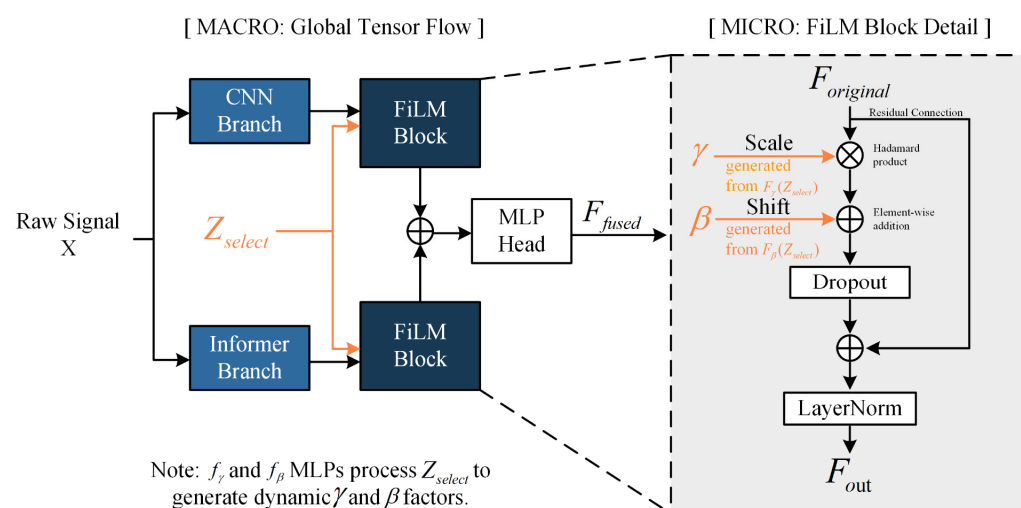


Figure 6. The architecture of FiLM conditional modulation integrating time–frequency statistical features with deep representations.

3.4. Hybrid Informer-CNN Backbone

To fully decouple the time–frequency characteristics of machining signals, the feature extraction backbone adopts a dual-stream parallel deep architecture, as illustrated in Figure 7.

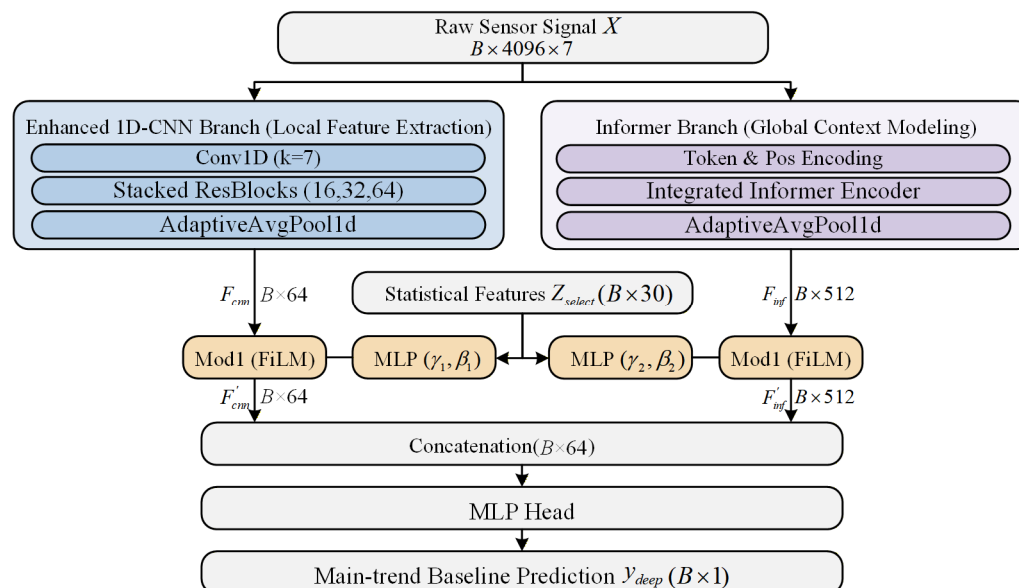


Figure 7. The detailed internal architecture and tensor flow of the Hybrid CNN-Informer backbone with FiLM early fusion.

Enhanced 1D-CNN Local Perception Branch: A one-dimensional residual convolutional branch is constructed to extract local transient patterns from high-frequency machining signals. The CNN branch is defined by a base channel number and a convolution kernel size. In the final implementation selected through Bayesian hyperparameter optimization, (*cnn_base_channel* = 16) and (*k* = 7). Accordingly, the residual convolutional stages adopt a pyramidal channel expansion strategy of 16 → 32 → 64. Identity-mapping skip connections are embedded in each stage to alleviate gradient degradation, and Adaptive Average Pooling is finally used to obtain a compact local feature representation.

Informer Global Long-range Dependency Branch: Targeting ultra-long machining sequences (*L* = 4096), the Informer branch selects a small number of active queries carrying rich semantics for dot-product operations by calculating the sparsity measurement between the Query and Key distributions. This branch not only significantly reduces memory consumption but also provides a macroscopic view of the full-lifecycle global degradation trend, ranging from “initial run-in” to “severe failure.”

After the parallel feature extraction, two independent FiLM modules are introduced to perform conditional modulation on the local features from the CNN and the global features from the Informer, respectively, utilizing the same set of manual statistical features. Subsequently, these two channels, calibrated by statistical descriptors, are merged at the end of the network through concatenation fusion. Finally, a multi-layer perceptron (MLP Head) performs dimensionality mapping to output the main degradation trend baseline prediction $\hat{y}_{deep}$ for the maximum flank wear at the corresponding time step.

3.5. Fine Residual Compensation via LightGBM

Although the Hybrid CNN-Informer can accurately fit the smooth main degradation trend of tool wear, deep learning models often exhibit structural prediction lag or

underestimation when encountering random nonlinear mutations caused by material inhomogeneities or spindle chatter in real-world industrial settings [23].

To address this, the training strategy for the second stage of this research focuses on the “secondary mining” of deep model errors. Using the flattened raw signals (X_{flat}) as the input space, the difference between the ground-truth labels (y_{true}) and the predictions of the deep model ($\hat{y}_{deep}$) is defined as the residual label r :

$$r = y_{true} - \hat{y}_{deep} \quad (8)$$

The sequence pair (X_{flat}, r) is then input into the LightGBM regressor for fitting. Since decision trees are inherently piecewise-constant functions in terms of space partitioning, they possess a superior capability to capture high-frequency oscillations and outliers that are challenging for deep learning models to approximate. During the inference phase, the final prediction $\hat{y}_{final}$ is obtained by precisely summing the main degradation trend and the residual correction term:

$$\hat{y}_{final} = \hat{y}_{deep} + \hat{y}_{res} \quad (9)$$

3.6. Bayesian Hyperparameter Optimization

To maximize the synergistic performance of the hybrid model and eliminate the suboptimal bias introduced by manual parameter tuning, this paper introduces a Bayesian automated optimization system based on the Optuna framework.

The optimization process employs the Tree-structured Parzen Estimator (TPE) algorithm. By constructing a probabilistic surrogate model of the parameter space to guide the search direction, the TPE algorithm focuses on exploring promising regions that minimize the validation set RMSE. More importantly, given the extremely high computational cost of training deep networks, a Hyperband dynamic -pruning strategy based on multi-armed bandit theory is simultaneously introduced. The system monitors the validation loss during the early training epochs of each trial in real time. If the convergence trend of a specific hyperparameter configuration is found to be significantly inferior to historical records, early termination (Early Pruning) is immediately triggered. This dynamic evaluation-based early pruning strategy and computational resource-allocation mechanism enable this study to accomplish hundreds of global joint optimizations for both the model architecture and hyperparameters within a limited computational budget [24].

4. Experimental Setup and Results

To comprehensively validate the effectiveness, robustness, and generalization capability of the proposed prediction framework, this section conducts systematic experimental evaluations on the PHM 2010 benchmark dataset and an external HMoTP validation dataset for tool wear prediction in machining operations. All experiments were implemented in Python 3.11.9 using PyTorch 2.9.1 with CUDA 13.0 on a Windows 11 operating system. The hardware environment was equipped with an NVIDIA GeForce RTX 5070 Ti GPU (16 GB GDDR7 VRAM, NVIDIA Corporation, Santa Clara, CA, USA) and 32 GB of DDR5 memory to accelerate the training and inference processes.

4.1. Dataset Description and Preprocessing Details

This paper first adopts the PHM 2010 High-Speed CNC Milling Machine Tool Wear Challenge dataset (PHM 2010 Data Challenge Dataset) as the main benchmark platform for algorithm validation [25]. This dataset provides a highly representative physical scenario of machining processes, characterizing severe transient high-frequency shocks and nonlinear degradation processes.

To provide a comprehensive physical context and clarify the tribological details underlying this computational study, the specific experimental equipment and machining parameters are summarized in Table 1. The machining proceeded line by line along the X-axis of the workpiece.

Table 1. Experimental conditions and machining parameters for the PHM 2010 dataset.

Parameter Category	Specification/Value
Machine Tool	Röders Tech RFM760 high-speed CNC milling machine
Workpiece Material	Stainless steel (Hardness: HRC 52)
Cutting Tool	6 mm diameter, three-flute ball-nose tungsten carbide cutter
Cutting Condition	Dry milling
Spindle Speed	10,400 RPM
Feed Rate	1555 mm/min
Axial Depth of Cut (Z)	0.2 mm
Radial Depth of Cut (Y)	0.125 mm

During the experiment, the three-axis cutting forces were recorded using a Kistler 9265B three-component dynamometer, the cutting vibration signals were captured using Kistler 8636C piezoelectric accelerometers, and the AE-RMS signal was acquired using a Kistler 8152-series acoustic emission sensor. All multi-source sensor signals were sampled simultaneously at 50 kHz. This 7-channel multi-modal monitoring system provides a minimal sufficient set of descriptors for tool wear progression: the 3-axis cutting forces (F_x, F_y, F_z) directly reflect the macroscopic mechanical load and severe friction changes at the tool–workpiece interface; the 3-axis vibration signals (V_x, V_y, V_z) capture the dynamic response and stability of the machine–tool fixture system; and the acoustic emission (AE) signal is highly sensitive to micro-scale plastic deformation and micro-chipping events.

In milling operations, tool wear is commonly characterized by flank wear (VB) and crater wear (KB) [26]. Among these, flank wear VB is widely adopted in tool wear monitoring and prediction because it directly affects the surface quality, process reliability, and dimensional accuracy of the machined workpiece [27]. Following the annotation scheme of the PHM 2010 dataset, this study adopts flank wear VB as the target indicator of tool degradation. Here, VB denotes the distance between the cutting edge and the end of the wear land on the tool flank face. After each complete cutting pass, the tool was removed, and the flank wear of each flute was measured using a LEICA MZ12 microscope. Since tool failure is generally governed by the most severely worn cutting edge, the maximum flank wear among the three flutes was used as the ground-truth wear label for the corresponding time step.

To ensure long-range sequence dependencies and the integrity of process-related signal features, the multi-source signal from a single cutting pass was down-sampled and truncated into a sequence of fixed length $L = 4096$. Consequently, the input tensor dimension is constructed as $X \in \mathbb{R}^{B \times 4096 \times 7}$.

Regarding the dataset division strategy, six identical cutters were initially tested in the challenge. However, due to structural limitations in the official data release where cutters C2, C3, and C5 lack complete and continuous ground-truth wear labels, this study strictly utilizes cutters C1, C4, and C6 to prevent the injection of artificial bias through label interpolation. The samples from tools C1, C4, and C6 were divided into training, validation, and testing sets at a ratio of 7:2:1 using a sample-level random split.

It should be noted that the PHM 2010 dataset was divided at the sample level rather than the tool level; therefore, samples from the same tool-life trajectory may appear in different subsets. This setting may introduce trajectory-level information overlap among

the training, validation, and testing data, which represents a potential source of information leakage. Thus, the PHM 2010 results mainly reflect model performance under this benchmark split. Section 4.9 further provides a complementary held-out-tool validation on the HMoTP dataset.

4.2. Training Strategy and Hyperparameters

The training of the deep model in this study is based on the Adam optimizer, and the objective function employs the Mean Squared Error (MSE) loss to penalize outlier predictions that deviate from the true wear values. During the network initialization phase, the parameters of the deep backbone network are set according to the final hyperparameter optimization results. Specifically, the CNN branch uses a base channel number of (*cnn_base_channel* = 16) with a convolution kernel size of (*kernel_size* = 7), corresponding to the pyramidal channel expansion described in Section 3.4.

During the training process, the batch size is set to 16, and the initial learning rate is set to 5×10^{-5} , accompanied by an early stopping mechanism. The maximum limit for training epochs is set to 150. After the training of the deep model converges and the main degradation trend is extracted, the weights of the deep model are frozen. Subsequently, the extracted prediction residuals are utilized to train the LightGBM compensator. After optimization via Optuna, the core hyperparameter of LightGBM—the learning rate—is adaptively locked at approximately 0.0396. This is combined with a dynamic limit on the number of leaf nodes to achieve optimal nonlinear approximation of high-frequency residuals.

Unless otherwise specified, all PHM 2010 results reported in Sections 4.4–4.8 were obtained using the final configuration summarized in Table 2.

Table 2. Final configuration used for the reported PHM 2010 results.

Configuration Item	Final Setting
Statistical features for FiLM CNN branch	Top-30 GBR-selected statistical features; manual_feature_dim = 30 base channel = 16; kernel size = 7; channel expansion = 16, 32, 64
Informer branch	hidden dimension = 256; encoder layers = 4; attention heads = 4; dropout = 0.22276
Deep-model training	batch size = 16; initial learning rate = 5×10^{-5} ; maximum epochs = 150; early-stopping patience = 30
LightGBM residual compensator	learning rate = 0.03965; num_leaves = 18; feature_fraction = 0.8697; num_boost_round = 100
Residual compensation ensemble	5 LightGBM models averaged; seeds = 42, 123, 456, 789, and 2024
Data split and random seed	C1/C4/C6 samples randomly split 70%/20%/10% for training/validation/testing; split seed = 42
Hyperparameter-selection criterion	Best Optuna/TPE trial selected by minimum validation RMSE; best validation RMSE = 4.6691

4.3. Evaluation Metrics

To quantitatively evaluate the accuracy and goodness-of-fit of the model in predicting tool wear, this paper adopts four standard statistical metrics: Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2). Their specific mathematical definitions are as follows:

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 \quad (10)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2} \quad (11)$$

$$MAE = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i| \quad (12)$$

$$R^2 = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (13)$$

where N is the total number of samples in the testing set, y_i is the actual measured value, $\hat{y}_i$ is the model's predicted value, and $\bar{y}$ is the mean of the actual measured values. RMSE is more sensitive to prediction outliers, while an R^2 value closer to 1 indicates that the model has a stronger explanatory capability for the overall degradation trend.

4.4. Comparative Analysis

To validate the predictive performance of the proposed model on the PHM 2010 benchmark dataset, this section constructs multiple baseline models for a comprehensive comparison, covering traditional machine learning methods (SVR, LightGBM), classic time-series deep learning architectures (CNN-ResNet, TCN, BiLSTM+Attn), and the standard self-attention model (Transformer). The quantitative comparison of the prediction results is presented in Table 3.

Table 3. Performance comparison of different prediction models.

Model	RMSE	MAE	R^2	Training Cost (s)
SVR	15.9605	8.0212	0.7932	48
LightGBM (Raw Input)	8.2279	5.3726	0.9450	85
Transformer	7.6882	5.4927	0.9520	1826
TCN	5.4445	3.9157	0.9759	144
CNN-ResNet	6.7153	4.6203	0.9634	46
BiLSTM+Attn	5.1514	3.0063	0.9785	4176
Ours (Hybrid+LightGBM)	4.0016	2.8271	0.9870	1027

Result Analysis: As shown in Table 3 and Figure 8, the traditional shallow model (SVR) exhibits severe performance bottlenecks when dealing with high-dimensional long sequences (with an RMSE of up to 15.9605). Although the standard Transformer theoretically possesses a global receptive field, its ability to extract local transient high-frequency shocks is weak due to computational resource constraints (RMSE of 7.6882), resulting in persistently high errors. In contrast, the model proposed in this paper achieves significant advantages, with its RMSE (4.0016) and MAE (2.8271) being the lowest among all models, and the coefficient of determination (R^2) increasing to the highest value of 0.9870. Compared to the standard Transformer, the proposed model significantly reduces the RMSE and MAE by 47.9% and 48.5%, respectively. Although the prediction accuracy of the BiLSTM+Attn combination is close to that of the proposed model, its training time reaches 4176 s—more than four times that of our model—indicating excessive computational overhead. This indicates that the cascaded architecture of “hybrid deep network capturing the main degradation trend” combined with “LightGBM correcting local residuals” can effectively adapt to the complex dynamic process of nonlinear tool degradation, achieving an excellent balance between accuracy and efficiency.

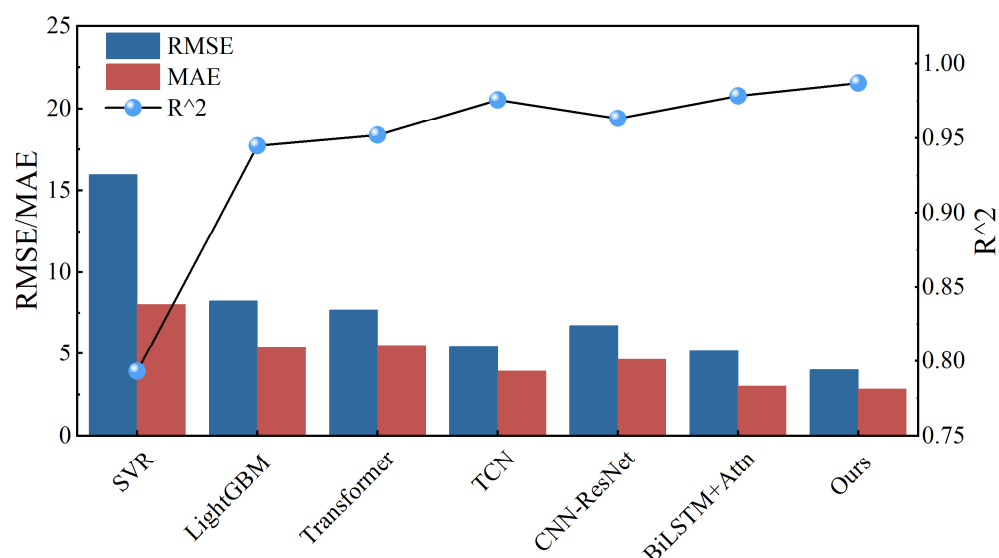


Figure 8. Histogram comparison of performance metrics across different baseline models.

4.5. Ablation Study

To thoroughly investigate the independent contributions of the core modules within the proposed framework (i.e., the CNN local perception module and the LightGBM residual compensation module), an ablation study is designed in this section. Two model variants are constructed: Model A (removing LightGBM, retaining only the deep model) and Model B (removing the CNN branch, retaining only the Informer + LightGBM architecture).

Result Analysis: As shown in Table 4, when the LightGBM residual compensation mechanism is removed (Model A), the model's RMSE deteriorates from 4.0016 to 4.3238. This indicates that relying solely on deep learning for end-to-end fitting will still miss a portion of the high-frequency residual components caused by transient operating conditions. Similarly, when the CNN branch is stripped away (Model B), the model also experiences a performance drop (RMSE of 4.7591), proving that the 1D-CNN plays a crucial role in extracting local transient high-frequency shocks when dealing with machining vibration signals. The ablation study demonstrates that while the deep learning backbone captures the primary degradation trend, the integration of decision-tree residual correction provides an effective complementary mechanism to further refine local transient errors.

Table 4. Ablation study on the contributions of different modules.

Model Variant	MSE	RMSE	MAE	R ²
Model A (CNN-Informer without Residual Comp)	18.6948	4.3238	2.9423	0.9848
Model B (Informer-LightGBM without CNN)	22.6489	4.7591	3.4929	0.9816
Ours (Complete Framework)	16.0125	4.0016	2.8271	0.9870

4.6. Visualization of Prediction Performance on the Test Set

To intuitively demonstrate the dynamic tracking capability and overall goodness-of-fit of the proposed hybrid framework under highly non-stationary conditions, a multi-dimensional visualization analysis was performed on the tool wear predictions for the test set. This includes the prediction trajectory on the test set, zoomed-in views of local fluctuating segments, and a scatter correlation plot between predicted and ground-truth values.

Visual Analysis: As illustrated in Figure 9a, the ground-truth wear values in the test set do not follow a perfectly smooth trend. Instead, the macroscopic upward trend is interspersed with numerous local nonlinear steps and high-frequency oscillations caused by

chip entanglements or material hard spots. Under these complex conditions, the prediction trajectory of the proposed model closely follows the ground-truth labels across the test samples without significant deviation or divergence, validating the model's global tracking performance over long sequences.

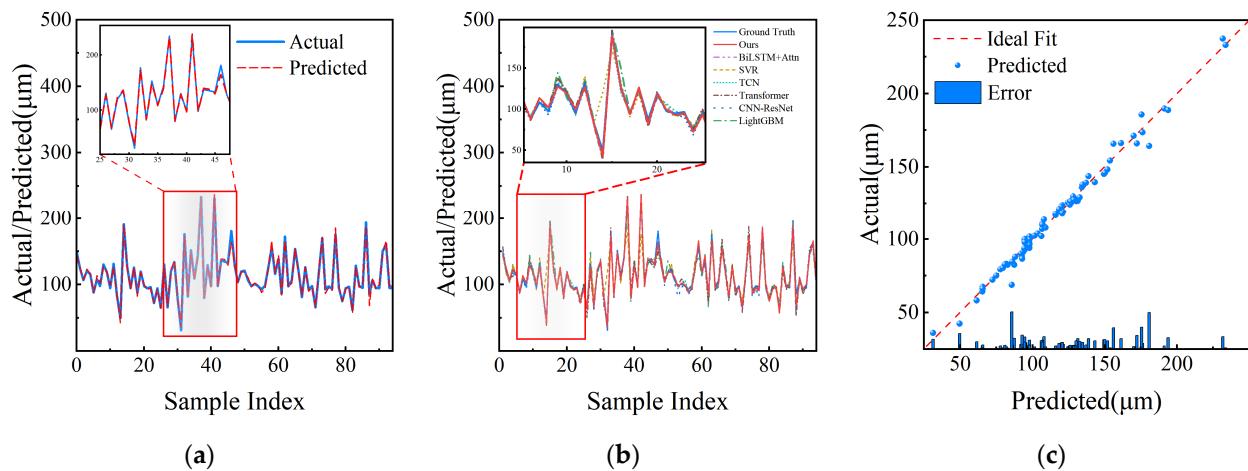


Figure 9. Visualization of the prediction performance: (a) Prediction trajectory on the test set, (b) Zoomed-in view of a highly fluctuating segment, (c) Scatter plot of predicted versus true flank wear.

The zoomed-in view in Figure 9b clearly shows that conventional models, such as LSTM, often exhibit a significant prediction lag (Lagging Effect) when facing transient high-frequency mutations in wear volume, resulting in over-smoothed and delayed responses. In contrast, the proposed model not only preserves high-frequency details but also better follows local fluctuations of the ground truth.

Furthermore, the scatter correlation plot in Figure 9c confirms the superior performance of the model from a global statistical perspective. The prediction points are tightly clustered along the $y = x$ diagonal, showing no significant systematic bias or outliers. This high degree of linear correlation aligns with the high R^2 of 0.9870 reported in Table 3, further demonstrating the good dynamic tracking capability of the proposed framework in complex, non-stationary machining scenarios.

4.7. Optimization Process Analysis

The performance of deep learning frameworks is highly dependent on the configuration of network architectures and hyperparameters. The Optuna TPE Bayesian optimization algorithm, combined with the Hyperband pruning strategy employed in this study, helps improve the search process.

As illustrated in Figure 10, the optimization algorithm extensively evaluated the scattered parameter space during the initial exploration phase, resulting in significant fluctuations in the objective function values (validation set RMSE). As the number of iterations increased, the TPE algorithm converged rapidly, utilizing the constructed probabilistic surrogate model to concentrate sampling on promising, high-potential regions with low errors (exploitation). Simultaneously, a large number of poorly performing trials were successfully pruned by the Hyperband strategy during early epochs. While ensuring convergence toward a low-validation-error solution (represented by the red convergence baseline), this dynamic evaluation strategy effectively avoided prolonged evaluation of sub-optimal configurations.

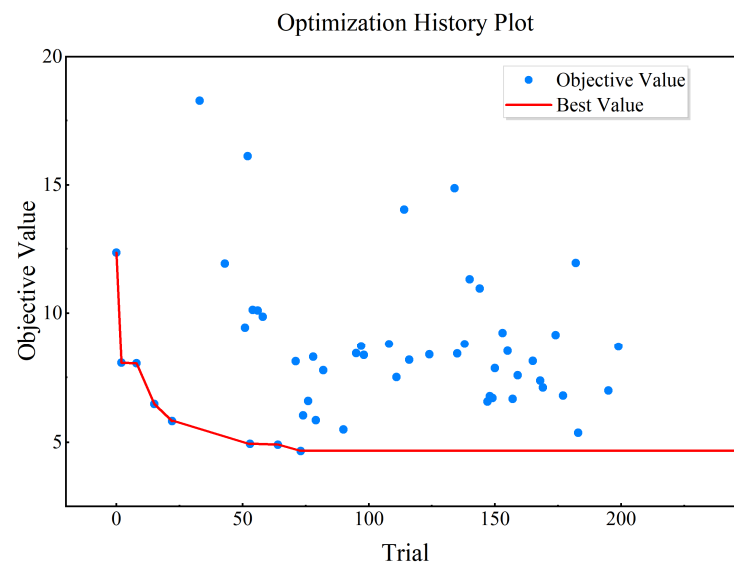


Figure 10. Convergence history of the Bayesian hyperparameter optimization process using TPE and Hyperband pruning.

4.8. Statistical and Visual Analysis of Residuals

In this study, residual analysis serves not only as a means to evaluate model accuracy but also as robust statistical evidence confirming whether the model has sufficiently extracted the wear-related signal information from the data.

Figure 11 presents a comprehensive visual diagnosis of the final prediction residuals. The histogram (Figure 11a) exhibits an effective “bell-shaped” distribution with its statistical mean extremely close to zero (0.4318). The scatter plot of residuals versus predicted values (Figure 11b) shows that the errors are randomly and uniformly distributed around the zero-mean reference line without any “funnel-shaped” divergence, visually suggesting constant variance across all tool degradation stages. Furthermore, in the Q-Q plot (Figure 11c), the sample quantiles adhere closely to the normal reference diagonal in the central region, preliminarily indicating a near-Gaussian distribution. Finally, the residual autocorrelation function (ACF) in Figure 11d shows that the autocorrelations across various lags remain predominantly within the 95% Bartlett confidence bounds, suggesting the absence of strong time-dependent structures.

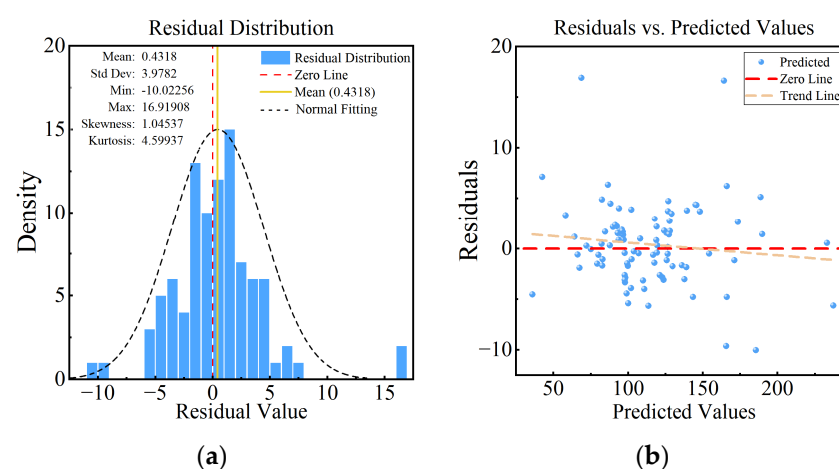


Figure 11. Cont.

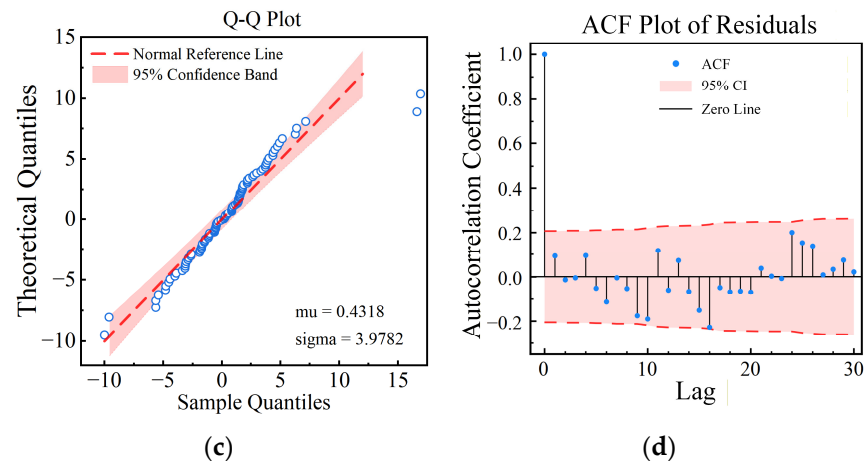


Figure 11. Statistical visualization of the final prediction residuals: (a) Histogram with normal distribution fit, (b) Residuals vs. Predicted Values, (c) Q-Q plot, (d) residual autocorrelation function (ACF) with 95% Bartlett confidence bounds.

To quantitatively validate these visual interpretations, rigorous statistical diagnostic tests were conducted at a significance level of 0.05 (Table 5). It is crucial to note that in practical high-frequency machining operations, monitoring signals naturally exhibit slightly heavy-tailed behavior due to intermittent cutting shocks, material hard spots, and measurement uncertainty. Consequently, applying strict normality tests directly to the raw absolute residuals can be overly sensitive to these physically inevitable outliers. To objectively evaluate the model's fit to the central degradation trend [28], a pre-defined two-sided 1.5% trimming strategy was introduced as a sensitivity analysis prior to the normality testing.

Table 5. Statistical diagnostic tests of final prediction residuals.

Residual Set	n	Shapiro–Wilk p	Anderson–Darling	Durbin–Watson	Breusch–Pagan p
Original residuals	95	4.86×10^{-5}	1.392	1.812	0.263
Trimmed residuals	91	0.0825	0.377	1.785	0.183

As detailed in Table 5, the diagnostic results provide quantitative support for the robustness of the proposed hybrid framework across three statistical aspects. First, after excluding the extreme 1.5% tails, the Shapiro–Wilk test yields a p -value of 0.0825, failing to reject the null hypothesis of normality. This near-Gaussian central distribution indicates that the framework has successfully decoupled the deterministic wear patterns from the non-stationary signals, leaving primarily unpredictable random noise. Second, regarding temporal independence, the Durbin–Watson statistics for both the original (1.812) and trimmed residuals (1.785) align closely with the ideal value of 2. This confirms the ACF visual findings, suggesting no evident systematic “memory effect” or prediction lag against dynamic cutting shocks. Finally, the Breusch–Pagan test returns a p -value of 0.183, confirming homoscedasticity. In real-world machining, signal variance typically explodes during the severe wear stage just before tool failure. The non-significant Breusch–Pagan result indicates that the residual variance does not show an obvious systematic dependence on the predicted wear level, suggesting stable error dispersion across the prediction range.

4.9. External Generalization Validation on the HMoTP Dataset

To further evaluate the generalization capability of the proposed framework beyond the PHM 2010 benchmark dataset, an external validation experiment was conducted on the

HMoTP dataset. Different from the sample-level random 7:2:1 split used for PHM 2010, the HMoTP experiment adopted an independent tool-based validation strategy, in which one complete tool life was held out for testing and the remaining tools were used for model development. This protocol provides a stricter cross-tool evaluation and helps prevent data leakage from the same tool life cycle.

4.9.1. HMoTP Dataset Description and Tool-Based Data Division

This study further adopts the High-speed Milling of Thin-walled Parts (HMoTP) dataset as an external validation dataset [29]. The HMoTP dataset was released by Prof. Qinghua Song's research group at Shandong University and was originally developed for tool wear monitoring in high-performance milling of titanium alloy thin-walled parts using domain knowledge [30]. Compared with the PHM 2010 benchmark dataset, HMoTP represents a different machining scenario in terms of workpiece structure, material properties, sensing configuration, and signal evolution characteristics.

Thin-walled titanium alloy parts are widely used in aerospace manufacturing because of their lightweight structure and high specific strength. However, their low structural stiffness makes the milling process more susceptible to workpiece deformation, chatter, and force–vibration coupling. Under high-speed cutting conditions, the monitoring signals may contain not only the gradual degradation trend caused by progressive tool wear, but also local transient fluctuations induced by structural flexibility and dynamic instability. Therefore, this dataset provides a suitable scenario for examining whether the proposed hybrid framework can capture both long-range wear evolution and local non-stationary signal disturbances.

To provide a clear physical context for this external validation, the main experimental platform and data characteristics of the dataset are summarized in Table 6.

Table 6. Experimental platform and data characteristics of the HMoTP dataset.

Parameter Category	Specification/Value
Machine tool	Five-axis boring and milling machining center with a HEIDENHAIN numerical control system
Workpiece material and size	Ti6Al4V titanium alloy thin-walled part, 150 mm × 100 mm × 5 mm
Cutting tool	14 mm TiAlN-coated detachable carbide end mill
Force/moment measurement	Kistler 9170B rotary dynamometer
Vibration measurement	Dytran 3263A1 three-axis acceleration sensor
Data acquisition system	NI PXIe-1073 acquisition system
Sampling frequency	5000 Hz
Monitoring signals	Three-axis milling force, axial moment, and three-axis workpiece vibration
Signal channels	Fx, Fy, Fz, Mz, Ax, Ay, and Az
Tool-life tests	Three tools: T01, T02, and T03
Raw cutting operations	100 cutting operations per tool
Wear measurement	Portable digital microscope, E3ISPM18000KPA
Wear label	Measured tool wear value corresponding to each cutting operation

For each cutting operation, the monitoring signals were stored as an independent CSV file, and the corresponding measured mean wear value was used as the prediction label. To maintain consistency with the PHM 2010 input setting, the cleaned signals were segmented using a non-overlapping sliding-window strategy with both the input sequence length and sliding step set to 4096. Since each cutting record contains 32,001 sampling points, it was segmented into seven complete non-overlapping windows, and the remaining points were discarded. Accordingly, each tool yielded 700 window-level samples, with all windows from the same cutting operation assigned the corresponding measured mean wear label.

Tool T03 was completely held out as the unseen test tool, while T01 and T02 were used only for model development. The window-level samples from T01 and T02 were randomly divided into training and validation subsets at an 8:2 ratio, whereas all samples from T03 were used only for independent testing.

4.9.2. Generalization Results and Discussion

To verify the generalization capability of different prediction methods on the external validation dataset, the proposed model was compared with several representative baseline methods. The prediction performance was evaluated using RMSE, MAE, and R^2 . Figure 12 presents the tool wear prediction results for the testing tool T03, from which it can be observed that the proposed method provides more accurate predictions than the compared methods.

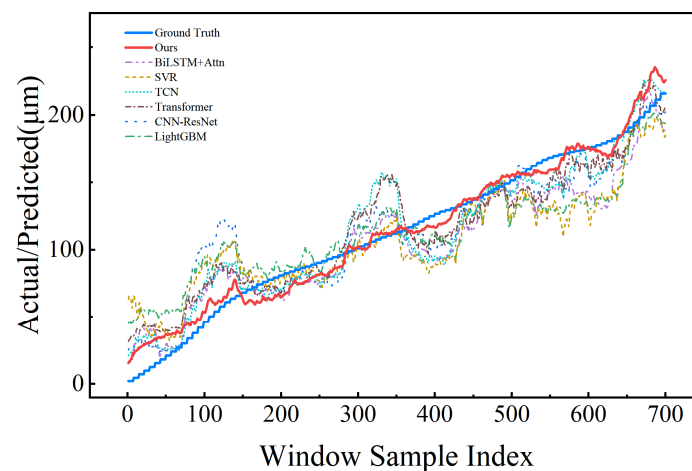


Figure 12. Prediction curves of different models on the held-out T03 tool.

As shown in Table 7, the proposed Hybrid CNN-Informer + LightGBM framework achieves the best overall performance on the held-out T03 tool. These results indicate that the proposed method achieves lower prediction errors and a higher coefficient of determination than the baseline models.

Table 7. Cross-tool generalization performance on the held-out T03 tool.

Model	RMSE	MAE	R^2
SVR	27.4631	21.9832	0.7539
LightGBM	26.9491	22.0552	0.7630
Transformer	19.2135	16.3043	0.8795
TCN	19.6990	16.0195	0.8734
CNN-ResNet	22.0326	16.5738	0.8416
BiLSTM+Attn	20.7505	17.6576	0.8595
Ours (Hybrid+LightGBM)	9.1111	7.3522	0.9729

As shown in Table 8, the model with residual compensation achieves lower RMSE and MAE than the uncompensated Hybrid CNN-Informer baseline on the held-out T03 tool, while also obtaining a higher R^2 value. This indicates that the LightGBM module mainly corrects the remaining local errors after the deep backbone has captured the main wear degradation trend. The improvement observed under the cross-tool validation setting further suggests that the residual correction provides a modest supplementary refinement for unseen tool-wear data.

Table 8. Effect of residual compensation on the held-out T03 tool.

Model	RMSE	MAE	R^2
Hybrid CNN-Informer	9.8183	8.1016	0.9685
Hybrid CNN-Informer + LightGBM	9.1111	7.3522	0.9729

The prediction curves in Figure 12 further demonstrate the effectiveness of the proposed method. Compared with the baseline models, the predicted results of the proposed framework are closer to the measured tool wear values, showing more accurate prediction performance on the unseen T03 tool.

It should be noted that the leave-one-tool-out validation used in this section is stricter than a random sample-level split. In a random split, training and testing samples may come from the same tool life trajectory, whereas in this experiment, the T03 tool is completely excluded from model training and validation. Therefore, the results on T03 demonstrate that the proposed framework has good generalization capability for predicting the wear trend of unseen cutting tools.

5. Conclusions and Future Work

To address the dual challenges in tool wear monitoring during modern precision machining operations—specifically, the difficulty in capturing long-term dependencies and the tendency to overlook nonlinear transient features—this paper proposes a two-stage hybrid tool wear prediction framework based on residual compensation, featuring a “main-trend baseline extraction and fine local compensation” strategy. This framework effectively integrates the deep feature extraction capabilities of the Hybrid CNN-Informer with the robust nonlinear fitting advantages of LightGBM. Through systematic validation on the PHM 2010 benchmark dataset and additional external validation on the HMoTP dataset, the following conclusions are drawn:

Methodological Contribution: The statistical-feature-guided early fusion strategy based on the FiLM conditional modulation mechanism successfully prevents time-frequency statistical features from being submerged within deep networks. Furthermore, decoupling the prediction task into “main degradation trend extraction” and “secondary residual compensation” provides a practical way to alleviate the limitation of single-architecture deep networks in balancing stationary and non-stationary signal components.

Quantitative Performance: Comparative experiments demonstrate that the error metrics of the proposed framework significantly outperform existing mainstream methods. The model achieves an RMSE as low as 4.0016, an MAE of 2.8271, and a coefficient of determination (R^2) of up to 0.9870. Compared to the standard Transformer, the prediction accuracy shows clear improvement, with improvements exceeding 30%. In the external HMoTP validation, the proposed framework also achieved the best performance on the held-out T03 tool, with an RMSE of 9.1111, MAE of 7.3522, and R^2 of 0.9729, indicating good predictive capability under an independent tool-based testing protocol.

Theoretical Validation: Statistical visualization analysis confirms that the final system prediction errors, after residual compensation, effectively approach a robust, near-Gaussian distribution. This statistically corroborates that the model can sufficiently extract the deterministic wear-related signal components from sensor signals, demonstrating good robustness under the tested conditions and freedom from systematic bias.

Future Work: Although this study demonstrates the effectiveness and predictive capability under the tested datasets of the hybrid framework on the PHM 2010 benchmark dataset and the external HMoTP dataset, future research needs to advance toward more extreme machining scenarios. Future work will focus on exploring transfer learning and

domain adaptation technologies, aiming to migrate this framework from the current milling environment to high-precision gear grinding scenarios characterized by data scarcity and harsh operating conditions. Simultaneously, research will be conducted to compress the deep model via pruning or quantization techniques, facilitating low-latency, online closed-loop predictive maintenance (PdM) on machine tool edge devices.

Author Contributions: Conceptualization, F.L. and F.W.; methodology, F.L. and F.W.; software, F.W.; validation, F.W., L.Y. and B.T.; formal analysis, F.W. and T.Y.; investigation, F.W.; resources, F.L.; data curation, F.W., T.Y. and M.L.; writing—original draft preparation, F.W.; writing—review and editing, F.L.; visualization, F.W. and Y.G.; supervision, F.L. and S.L.; project administration, F.L.; funding acquisition, F.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Tianjin Science and Technology Program Project (Grant no. 24PTGTQY00100), in part by the Tianjin Metrology Science and Technology Project (Grant no. 2024TJMT073), and in part by the Tianjin Science and Technology Plan Project (Grant no. 24PTLYHZ00030).

Data Availability Statement: The data that support the findings of this study are openly available from the PHM Society Data Repository for the PHM 2010 Data Challenge dataset at <https://www.phmsociety.org/competition/phm/10> (accessed on 9 July 2026). The HMoTP dataset used for external validation is openly available from Mendeley Data, as described in Ref. [29].

Conflicts of Interest: Authors Fucong Liu and Sai Lou are employed by the company “Tianjin BestBond (Baishibingde) Intelligent Technology Co., Ltd.” The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Jouini, N.; Yaqoob, S.; Ghani, J.A.; Mehrez, S. Tool Wear Effect on Machinability and Surface Integrity in MQL and Cryogenic Hard Turning of AISI 4340. *Materials* **2025**, *18*, 5423. [\[CrossRef\]](#) [\[PubMed\]](#)
2. Laghari, R.A.; Dahri, F.H.; Wang, S.; Padhiar, M.A.; Laghari, A.A.; Saeed, U.; Waqas, M. Machine learning-driven framework for tool wear prediction and optimization in machinability of SiCp/Al metal matrix composite materials. In *Proceedings of the Fourth International Conference on New Materials, Machinery, and Vehicle Engineering (NMMVE 2025)*; SPIE: Pune, India, 2025; Volume 13953, pp. 403–411. [\[CrossRef\]](#)
3. Kuntoğlu, M.; Sağlam, H. Investigation of progressive tool wear for determining of optimized machining parameters in turning. *Measurement* **2019**, *140*, 427–436. [\[CrossRef\]](#)
4. Yedurkar, D.P.; Schlech, T.; Sause, M.G. A systematic review on smart and predictive maintenance in tool condition monitoring. *IEEE Access* **2025**, *13*, 106246–106286. [\[CrossRef\]](#)
5. Kong, D.; Chen, Y.; Li, N.; Tan, S. Tool wear monitoring based on kernel principal component analysis and v-support vector regression. *Int. J. Adv. Manuf. Technol.* **2017**, *89*, 175–190. [\[CrossRef\]](#)
6. Saatci, B.T.; Ulas, M.; Gurgenc, T. Hybrid AI systems for tool wear monitoring in manufacturing: A systematic review. *Appl. Sci.* **2026**, *16*, 208. [\[CrossRef\]](#)
7. Bhatt, D.; Patel, C.; Talsania, H.; Patel, J.; Vaghela, R.; Pandya, S.; Modi, K.; Ghayvat, H. CNN variants for computer vision: History, architecture, application, challenges and future scope. *Electronics* **2021**, *10*, 2470. [\[CrossRef\]](#)
8. Al-Selwi, S.M.; Hassan, M.F.; Abdulkadir, S.J.; Muneer, A. LSTM inefficiency in long-term dependencies regression problems. *J. Adv. Res. Appl. Sci. Eng. Technol.* **2023**, *30*, 16–31. [\[CrossRef\]](#)
9. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation, Inc.: San Diego, CA, USA, 2017; Volume 30. [\[CrossRef\]](#)
10. Guo, H.; Zhu, K.P. Attention-based dual-scale hierarchical LSTM for tool wear monitoring. *Manuf. Lett.* **2021**, *29*, 99–103. [\[CrossRef\]](#)
11. Keles, F.D.; Wijewardena, P.M.; Hegde, C. On the computational complexity of self-attention. In *Proceedings of the International Conference on Algorithmic Learning Theory*; PMLR; JMLR, Inc.: New York, NY, USA, 2023; pp. 597–619. [\[CrossRef\]](#)
12. Wu, Q.; Zhou, X.; Pan, X. Cutting tool wear monitoring in milling processes by integrating deep residual convolution network and gated recurrent unit with an attention mechanism. *Proc. Inst. Mech. Eng. Part B J. Eng. Manuf.* **2023**, *237*, 1171–1181. [\[CrossRef\]](#)

13. Zhang, X.; Wan, L.; Ran, X. Research progress on the chatter stability in machining systems. *Int. J. Adv. Manuf. Technol.* **2024**, *131*, 29–62. [\[CrossRef\]](#)
14. Pimenov, D.Y.; Bustillo, A.; Wojciechowski, S.; Sharma, V.S.; Gupta, M.K.; Kuntoğlu, M. Artificial intelligence systems for tool condition monitoring in machining: Analysis and critical review. *J. Intell. Manuf.* **2023**, *34*, 2079–2121. [\[CrossRef\]](#)
15. Hu, Q.; Su, B.; Kong, J.; Liu, W. Adaptive wavelet convolution stochastic analysis graph neural network for remaining useful life prediction. *IEEE Sens. J.* **2025**, *25*, 19428–19441. [\[CrossRef\]](#)
16. El Zaar, A.; Mansouri, A.; Benaya, N.; Bakir, T.; El Allati, A. Hybrid Transformer-CNN architecture for multivariate time series forecasting: Integrating attention mechanisms with convolutional feature extraction. *J. Intell. Inf. Syst.* **2025**, *63*, 1233–1264. [\[CrossRef\]](#)
17. Li, S.; Jin, X.; Xuan, Y.; Zhou, X.; Chen, W.; Wang, Y.X.; Yan, X. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation, Inc.: San Diego, CA, USA, 2019; Volume 32. [\[CrossRef\]](#)
18. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 2–9 February 2021; Volume 35, pp. 11106–11115. [\[CrossRef\]](#)
19. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Liu, T.Y. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*; Neural Information Processing Systems Foundation, Inc.: San Diego, CA, USA, 2017; Volume 30. Available online: https://proceedings.neurips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html (accessed on 9 July 2026).
20. Savitzky, A.; Golay, M.J. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* **1964**, *36*, 1627–1639. [\[CrossRef\]](#)
21. Kossakowska, J.; Bombiński, S.; Ejsmont, K. Analysis of the Suitability of Signal Features for Individual Sensor Types in the Diagnosis of Gradual Tool Wear in Turning. *Energies* **2021**, *14*, 6489. [\[CrossRef\]](#)
22. Perez, E.; Strub, F.; De Vries, H.; Dumoulin, V.; Courville, A. Film: Visual reasoning with a general conditioning layer. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32. [\[CrossRef\]](#)
23. Li, Z.; Lei, X.; You, Z.; Huang, T.; Guo, K.; Li, D.; Liu, H. Tool Wear Prediction Based on Residual Connection and Temporal Networks. *Machines* **2024**, *12*, 306. [\[CrossRef\]](#)
24. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*; Association for Computing Machinery: New York, NY, USA, 2019; pp. 2623–2631. [\[CrossRef\]](#)
25. PHM Society. 2010 PHM Society Conference Data Challenge. Available online: https://phmsociety.org/phm_competition/2010-phm-society-conference-data-challenge/ (accessed on 9 July 2026).
26. Sayyad, S.; Kumar, S.; Bongale, A.; Kamat, P.; Patil, S.; Kotecha, K. Data-driven remaining useful life estimation for milling process: Sensors, algorithms, datasets, and future directions. *IEEE Access* **2021**, *9*, 110255–110286. [\[CrossRef\]](#)
27. Palanisamy, P.; Rajendran, I.; Shanmugasundaram, S. Prediction of tool wear using regression and ANN models in end-milling operation. *Int. J. Adv. Manuf. Technol.* **2008**, *37*, 29–41. [\[CrossRef\]](#)
28. Shapiro, S.S.; Wilk, M.B. An analysis of variance test for normality (complete samples). *Biometrika* **1965**, *52*, 591–611. [\[CrossRef\]](#)
29. Wang, R.; Song, Q.; Peng, Y.; Qin, J.; Liu, Z.; Liu, Z. *High-Speed Milling of Thin-Walled Parts (HMoTP) Dataset; V1*; Mendeley Data: Amsterdam, The Netherlands, 2024. [\[CrossRef\]](#)
30. Wang, R.; Song, Q.; Peng, Y.; Qin, J.; Liu, Z.; Liu, Z. Toward digital twins for high-performance manufacturing: Tool wear monitoring in high-speed milling of thin-walled parts using domain knowledge. *Robot. Comput.-Integr. Manuf.* **2024**, *88*, 102723. [\[CrossRef\]](#)

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