ENGINEERING Information Technology & Electronic Engineering
https://www.academax.com/EITEE; https://www.engineering.org.cn
https://ieeexplore.ieee.org/xpl/RecentIssue.jsp?punumber=11344065
E-mail: eng.itee@zju.edu.cn

Engineering Journals
Available on IEEE Xplore®

Research Article

https://doi.org/10.1631/ENG.ITEE.2025.0111

CdualTAL: multi-domain tool wear prediction using a dual-channel Transformer and cross-attention network

Na LI[1,2], Zhendong LIU[3]✉, Xiao WANG[1,2], Jiamin JIANG[3], Yanjie WEI[4]

[1]*School of Intelligent Manufacturing and Control Engineering, Qilu Institute of Technology, Jinan 250200, China*
[2]*Shandong Provincial Key Laboratory of Industrial Big Data and Intelligent Manufacturing, Jinan 250200, China*
[3]*School of Computer and Information Engineering, Shanghai Polytechnic University, Shanghai 201209, China*
[4]*Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China*

Abstract: Accurate tool wear prediction is crucial for manufacturing efficiency, yet effectively using multi-domain sensor features is difficult due to redundant noise. There is a critical need to strategically leverage highly predictive strong features and potentially informative weak features. To address this issue, we propose CdualTAL, an improved Transformer-based encoder–attention–decoder algorithm. Its name represents the model's key components: a correlation-adaptive feature selection algorithm module, a dual-channel Transformer encoder, an attention mechanism, and a long short-term memory (LSTM) decoder. CdualTAL employs a dual-channel encoder to independently process the full set of multi-domain features, along with a subset of strong features selected using a designed correlation-adaptive feature selection algorithm. A custom cross-attention mechanism is then used to fuse these representations, sharpening focus on strong features while judiciously integrating information from weak ones. Finally, a hierarchical LSTM decoder captures deep temporal dependencies. Validated on tool wear datasets, CdualTAL outperforms 11 state-of-the-art methods, achieving superior prediction stability and accuracy with an average R^2 of 0.983 and a root mean square error (RMSE) of 4.373.

Key words: Multi-domain features; Dual-channel; Feature fusion; Tool wear; Attention mechanism; Feature enhancement

1 Introduction

As information technology advances rapidly, traditional manufacturing industries are undergoing digital transformation toward greater intelligence. A critical component of intelligent manufacturing is tool wear monitoring, which can significantly enhance production efficiency and automation (Chehrehzad et al., 2024). During the operation of computer numerical control (CNC) machines, cutting tools inevitably experience friction with workpieces during machining. Tool wear worsens continuously with processing time until tool failure (Hou et al., 2025); therefore, real-time monitoring of tool status and accurate wear prediction are crucial. Such measures reduce cutting downtime, enable the more efficient allocation of resources and lower production costs, and improve product quality.

Tool prediction and replacement have traditionally relied heavily on operator experience. Premature replacement often leads to unused tool life, while delayed replacement results in tool failure and workpiece damage (Kumar et al., 2025). Subsequent to the widespread adoption of machine learning, researchers have made significant progress in tool wear assessment by integrating multi-source signals, such as cutting force, vibration, and acoustic emission, using various machine learning methodologies (Shi et al., 2020; Marani et al., 2021; Ou et al., 2021; Duan et al., 2022). Analyses conducted across time, frequency, and time-frequency domains have proven effective in characterizing tool wear status (Huang et al., 2020, 2024; Yan et al., 2021; Guo et al., 2022; He et al., 2022), enriching datasets and enhancing recognition accuracy and reliability. However, while multi-domain features offer comprehensive insights into the complex dynamics of tool wear, they inevitably introduce numerous irrelevant or noisy components. These so-called redundant features pose a significant challenge, as they dilute the model's focus on the most predictive attributes (Gao et al., 2022; He et al., 2022).

Deep learning methods have emerged as pivotal technologies for tool wear monitoring, demonstrating remarkable capabilities for establishing complex nonlinear mappings between multi-domain features and wear states, thereby reducing dependence on expert knowledge and manual feature engineering (Duan et al., 2023). Convolutional

✉ Zhendong LIU, liuzd2000@126.com
Na LI, https://orcid.org/0000-0002-6127-182X
Zhendong LIU, https://orcid.org/0000-0002-4131-313X
CLC number: TP391.41; TP274

Received: Nov. 1, 2025; Revision accepted: Jan. 31, 2026;
Crosschecked: Jan. 31, 2026

neural networks (CNNs) and recurrent neural networks (RNNs) are widely employed in this domain. Huang et al. (2020) designed a deep convolutional neural network (DCNN) to adaptively extract sensitive features from multi-domain features derived from multi-sensor signals. Building on this approach, Huang et al. (2024) incorporated deep adversarial domain confusion to improve prediction accuracy across different operational domains. Cai et al. (2020) proposed a hybrid model that incorporates long short-term memory (LSTM) to extract abstract deep features from sequential signals. Shah et al. (2024) explored generative adversarial networks (GANs) to synthesize spectral wear data, addressing data scarcity and improving prediction using LSTM, gated recurrent unit (GRU), and CNN models. Duan et al. (2023) introduced a hybrid attention-based parallel deep learning (HABPDL) network that uses stacked ResNet and BiLSTM blocks with attention to capture both spatial and temporal dependencies. Wang S et al. (2022) presented a multi-channel feature fusion CNN-LSTM model optimized with particle swarm optimization (PSO) for spatiotemporal feature learning and prediction enhancement.

Additionally, recent studies have explored advanced architectures to address specific challenges. Shah et al. (2022) used singular generative adversarial networks (SinGANs) combined with LSTM to synthesize data and predict tool wear during stainless steel face milling, effectively mitigating data scarcity issues. Furthermore, Lu et al. (2022) proposed a prediction model based on the attention mechanism and independent recurrent neural network (IndRNN), which improves the capture of long-term dependencies while addressing the gradient vanishing problem common in traditional RNNs.

Despite these advances, significant challenges persist. First, the highly complex and nonlinear relationships in tool wear data make it difficult for single models to fully capture the intricate interactions among multi-source sensors and multi-dimensional features, leading to inadequate representation of dynamic wear processes. Second, models that rely solely on single-domain features (e.g., time or frequency domain) often overlook crucial cross-domain interactions (Khoshouei et al., 2024). Third, while hybrid models incorporating multi-domain features provide richer information, the practice of extracting only strong features can result in neglecting those that exhibit weak direct correlation with wear (weak features). These weak features may contain critical complementary information or subtle patterns essential for a holistic understanding; overlooking them can compromise the comprehensiveness and accuracy of predictions (Kejriwal et al., 2024). This limitation hinders the model's ability to achieve both high accuracy and comprehensive representation. Additionally, effectively capturing long-range dependencies within the sequence data remains a challenge for many architectures.

To address these challenges, particularly the effective utilization of both strong and weak multi-domain features and the modeling of long-range dependencies, in this work we propose a novel tool wear prediction algorithm, CdualTAL. This algorithm reconstructs the multi-sensor signal dataset across the time, frequency, and time–frequency domains. It analyzes and selects strong feature data using a correlation-adaptive feature selection module and constructs a dual-channel Transformer encoder to extract key strong features from these domains. The encoder independently processes strong features and multi-domain key features in parallel, effectively isolating feature interference while enabling targeted learning of discriminative patterns. Subsequently, a custom cross-attention mechanism with feature-type gating is designed to adaptively balance strong and weak features. The mechanism incorporates dynamic weighting, which adjusts feature contributions based on learned importance scores. Finally, a deep LSTM decoder is used to mine deep features for global feature extraction. The cross-validation experiments on three tool wear datasets reveal that CdualTAL exhibits stable, high prediction accuracy, achieving an average R^2 of 0.983, thereby providing a practical solution for multi-domain tool wear monitoring. In conclusion, CdualTAL algorithm establishes a new heterogeneous paradigm for industrial monitoring. By synergistically combining feature isolation, adaptive fusion, and hierarchical decoding, it achieves cutting-edge performance while addressing the feature interference problem in multi-sensor systems.

2 CdualTAL algorithm

We propose the novel algorithm, CdualTAL, to enhance the prediction model's ability to extract key features from multi-domain features. The main structure of CdualTAL is depicted in Fig. 1, comprising four primary components: a correlation-adaptive feature selection module, a dual-channel Transformer encoder, an attention mechanism, and an LSTM decoder. Each of these components plays a pivotal role in improving prediction accuracy by addressing challenges such as feature redundancy and noisy data in multi-sensor systems.

To achieve high-accuracy predictions, CdualTAL processes input data across multiple stages. First, the extracted multi-sensor tool signals are transformed into multi-domain interpretable features, including various time- and frequency-domain characteristics. From these features, the correlation-adaptive feature selection module captures the time-varying characteristics of feature correlations, thereby reducing noise interference. The remaining subset of features, along with the original multi-domain feature set, is inputted into a one-dimensional (1D) CNN (Conv1D) module for feature extraction (Mo et al., 2024). While optimizing feature representation, the flattening layer provides a consistent data format for model training (Max et al., 2024). The processed data are divided into two streams and pass through a dual-channel encoder. The dual-channel encoder adopts a parallel, independent structure of dual-channel Transformer encoders, preventing correlation interference between features and enabling the independent extraction of strong features across multiple domains. Each stream is processed by an independent Transformer encoder to enable the separate handling of strong and weak feature sets, which enhances the model's ability to capture complex patterns in the data. The outputs from the two encoders are fused via a cross-attention module, using scaled dot-product weighting to perform a weighted fusion of multi-domain and strong features. This effectively emphasizes the most relevant features while reasonably balancing the use of potentially key information from weaker features, thereby improving overall prediction performance. The fused features enter a three-layer LSTM decoder with residual connections, which cumulatively connects different

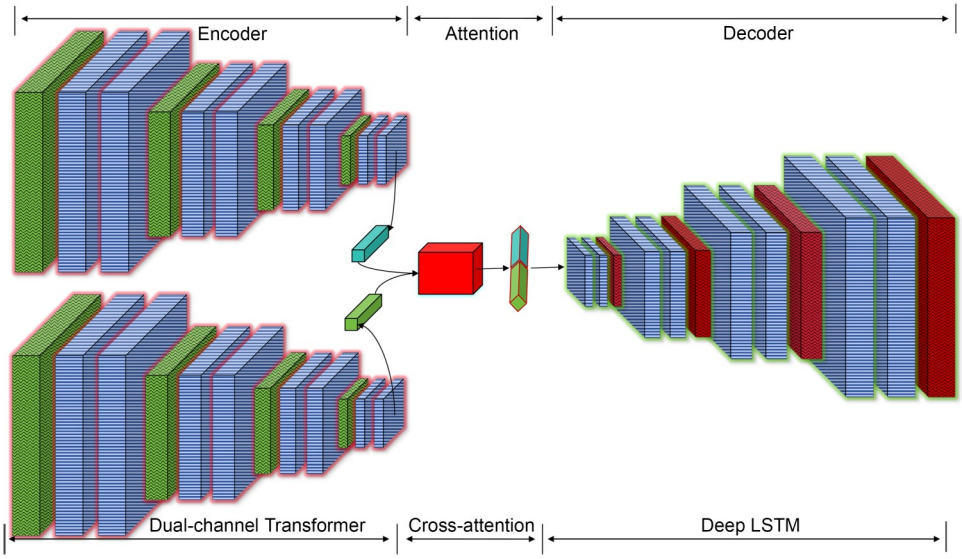

Fig. 1 Main structure of CdualTAL

output feature layers through residual connections, mining deep features inherent in the multi-domain tool wear sequences and enabling the model to capture deeper temporal dependencies and long-range patterns. Finally, the decoder output is processed through pooling and dense layers to generate the final wear predictions. The pseudocode for the overall data processing flow of CdualTAL is shown in Algorithm 1.

2.1 Correlation-adaptive feature selection

We propose a correlation-adaptive feature selection algorithm based on a sliding window and the maximal information coefficient, aiming to achieve dynamic optimization of multi-domain sensor features. At the algorithm design level, this module uses the maximal information coefficient as the feature measurement method, capturing complex nonlinear relationships between features and wear levels through mutual information calculation and adaptive binning. The main calculation formula is as follows:

$$\mathcal{F}^{t}_{\text{strong}}=\left\{ \boldsymbol{f}_i \left| \max_{w\in[t-k,\,t]} \frac{I_w(\boldsymbol{f}_i,\boldsymbol{Y})}{\log_2\left(\min\left(|B_X|,|B_Y|\right)\right)}>\tau \right. \right\}, \tag{1}$$

where $\mathcal{F}^{t}_{\text{strong}}$ is the current time step strong feature set, I_w represents the mutual information within window w, calculating the nonlinear correlation between feature $\boldsymbol{f}_i$ and wear amount Y. $|B_X|$ and $|B_Y|$ represent the number of grid bins for features and wear amount, respectively. The sliding window size k is set to 30 based on empirical performance validation. To ensure robustness against parameter sensitivity and signal noise, we employ an adaptive threshold τ equal to the median of the wear-related MIC values. This data-driven mechanism ensures that the feature selection process relies on relative feature importance rather than absolute magnitudes. The pseudocode for the complete data processing flow of this correlation-adaptive feature selection algorithm is shown in Algorithm 2.

2.2 Dual-channel Transformer encoder

When constructing a multi-domain feature-driven tool wear prediction model, traditional single-channel or simple feature concatenation methods often struggle to avoid mutual interference between strong and weak features, leading to key signals being drowned out by noise and limiting further improvement in model performance. To fundamentally address the issues of feature redundancy and interaction interference in multi-source sensor information, we

Algorithm 1 CdualTAL

Input: $\textbf{tool}_{\text{data}}$=[C1, C4, C6] which represents a 315×7 sensor data matrix for each cutting tool and $\textbf{wear}_{\text{labels}}$=[C1, C4, C6] which represents the true wear value vector for cutting tool.

Output: predicted tool wear values $\boldsymbol{Y}_{\text{pred}}$.

Begin

For tool in $\textbf{tool}_{\text{data}}$ **do**
  For t←1 to 315 **do** // Per machining pass of each tool
    For s←0 to 5 **do** // Use the first six sensors
      features[t][s]←Compute features with the formulas in Table 1
    End for
    fm[t]←Construct feature vectors
  End for
End for
af ←Concatenate **fm** from all tools
aw←Apply identical operations to $\textbf{tool}_{\text{data}}$
$\boldsymbol{X}_{\text{strong}}$←Use dynamic feature selection to select features
For tool in [C1, C4, C6] **do**
  $\boldsymbol{X}_{\text{full}}$←**fm**[**tool**] // Full feature set
  $\textbf{Conv}_{\text{A}}$, $\textbf{Conv}_{\text{B}}$←Apply 1D CNN to $\boldsymbol{X}_{\text{full}}$ and $\boldsymbol{X}_{\text{strong}}$
  $\boldsymbol{E}_{\text{A}}$←Encode multi-domain features using the $\textbf{Conv}_{\text{A}}$ channel
  $\boldsymbol{E}_{\text{B}}$←Encode strong features using the $\textbf{Conv}_{\text{B}}$ channel
  Att ←Compute cross-attention between $\boldsymbol{E}_{\text{A}}$ and $\boldsymbol{E}_{\text{B}}$
  Fused ←$\alpha\cdot$**Att** + $\beta\cdot\boldsymbol{E}_{\text{B}}$ // Perform weighted fusion
  $\boldsymbol{H}_1$ ←LSTM_layer1(**Fused**) // LSTM decoding
  $\boldsymbol{H}_2$ ←LSTM_layer2($\boldsymbol{H}_1$)
  $\boldsymbol{H}_3$ ←LSTM_layer3(Concat($\boldsymbol{H}_1$, $\boldsymbol{H}_2$)) // Residual connections
  $\boldsymbol{P}$←Apply pooling operations to $\boldsymbol{H}_3$ outputs
  $\boldsymbol{Y}_{\text{pred}}$[**tool**] ←Generate tool-specific predictions from $\boldsymbol{P}$
End for
Return $\boldsymbol{Y}_{\text{pred}}$

End

Algorithm 2 Correlation-adaptive feature selection

Input: original multi-domain feature set $\boldsymbol{F}=\{\boldsymbol{f}_1,\boldsymbol{f}_2,\cdots,\boldsymbol{f}_{144}\}$; wear quantity sequence $\boldsymbol{Y}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_T\}$ (T=315); sliding window size k; adaptive threshold τ.

Output: strong feature subset $\boldsymbol{X}_{\text{strong}}$.

Begin

For $t = k$ to T **do** // Traverse each time window
  $\textbf{window}_{\text{features}}=\boldsymbol{F}[t-k:t]$
  $\textbf{window}_{\text{wear}}=\boldsymbol{Y}[t-k:t]$
  For i=1 to 144 **do** // Traverse each feature
    $I_w=\sum_{x\in f_i}\sum_{y\in \boldsymbol{Y}} p(x,y)\log_2\frac{p(x,y)}{p(x)p(y)}$ // Mutual information
    $B_X=|\textbf{window}_{\text{features}}|^{0.6}$
    $B_Y=|\textbf{window}_{\text{wear}}|^{0.6}$
    $\text{MIC}=\frac{I_w(\boldsymbol{f}_i,\boldsymbol{Y})}{\log_2\left(\min\left(|B_X|,|B_Y|\right)\right)}$
    If MIC>τ **then** // τ is taken as the median of the wear
      $\boldsymbol{X}_{\text{strong}}[t]$.append($i$) // Add as strong features
    End if
  End for
  For (i, j) in ($\boldsymbol{X}_{\text{strong}}[t]$, 2) **do**
    $\gamma_{i,j}=\frac{\text{Cov}(\boldsymbol{f}_i,\boldsymbol{f}_j)}{\sigma_{f_i}\sigma_{f_j}}$ // Inter-feature synergy coefficient
    If $|\gamma_{i,j}|$<0.3 **then**
      $\boldsymbol{X}_{\text{strong}}[t]$.remove($i$ or j) // Remove weak features
    End if
  End for
End for
Return $\boldsymbol{X}_{\text{strong}}[k:T]$

End

propose the innovative concept of feature isolation encoding and design a dual-channel Transformer encoder structure based on this idea. This module employs independent, parallel feature-processing channels that focus separately on global multi-domain features and dynamically select strongly correlated features, thereby preventing the contamination of key features by irrelevant or weakly correlated features at the source. This approach provides a cleaner and more discriminative feature representation for subsequent deep feature fusion and dependency modeling.

The encoder part of CdualTAL is designed based on a dual-channel Transformer model; its encoder structure is shown in Fig. 2, primarily comprising layer normalization, multi-head attention, and feed forward modules. After Conv1D processing, the multi-domain key features and strong features independently pass through separate Transformer channels to extract long-term sequence features. With this approach, the model can capture complex temporal dependencies more accurately. The independent Transformer channels enable CdualTAL to focus more on learning the strong and key multi-domain features of tool wear. This design not only significantly enhances the model's capability to discern and leverage discriminative patterns from strong features across multiple domains, but also effectively mitigates interference and redundancy among heterogeneous features. For an input sequence with a length of 315 and a feature dimension of 14, the specific calculation process is as follows.

1. Each channel performs input normalization through layer normalization to ensure the stability of gradients during the training process:

$$\text{Norm}(\boldsymbol{X})=\hat{\boldsymbol{X}}=\frac{\boldsymbol{X}-\frac{1}{d}\sum_{j=1}^{d}\boldsymbol{X}_j}{\sqrt{\frac{1}{d}\sum_{j=1}^{d}\left(\boldsymbol{X}_j-\frac{1}{d}\sum_{j=1}^{d}\boldsymbol{X}_j\right)^2}}, \tag{2}$$

where $\boldsymbol{X}$ is a strong feature matrix and d is the dimensionality of the matrix.

2. The multi-head attention mechanism is used for the dynamic weighting of features to capture dependencies in long time series. The model's capacity to capture long-range dependencies is significantly enhanced by the attention mechanism, which facilitates parallel processing of the entire input sequence. Assuming that the weight matrices of the queries, keys, and values are $\boldsymbol{W}_{\text{Q}}$, $\boldsymbol{W}_{\text{K}}$, and $\boldsymbol{W}_{\text{V}}$, respectively, the multi-head attention computation following the linear transformation is shown below:

$$\textbf{Att}=\boldsymbol{W}_{\text{o}}\cdot\text{Concat}\begin{bmatrix}\text{Softmax}\left(\frac{\hat{\boldsymbol{X}}\boldsymbol{W}_{\text{Q}_1}(\hat{\boldsymbol{X}}\boldsymbol{W}_{\text{K}_1})^{\text{T}}}{\sqrt{d_k}}\right)\hat{\boldsymbol{X}}\boldsymbol{W}_{\text{V}_1}\\ \vdots\\ \text{Softmax}\left(\frac{\hat{\boldsymbol{X}}\boldsymbol{W}_{\text{Q}_h}(\hat{\boldsymbol{X}}\boldsymbol{W}_{\text{K}_h})^{\text{T}}}{\sqrt{d_k}}\right)\hat{\boldsymbol{X}}\boldsymbol{W}_{\text{V}_h}\end{bmatrix}, \tag{3}$$

where h represents the total number of attention heads, h=4. $\sqrt{d_k}$ is the scaling factor for calculating the attention score, and $\boldsymbol{W}_{\text{o}}$ is the

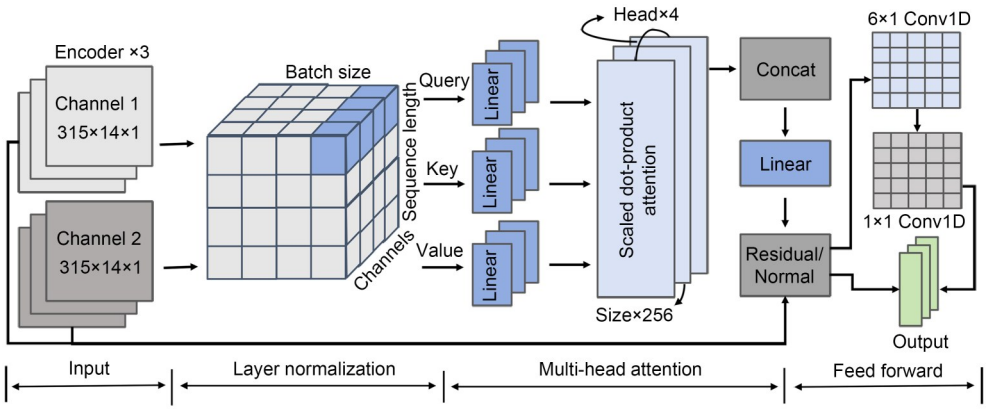

Fig. 2 Structure of the dual-channel encoder

output weight matrix. Adaptive focus on the most relevant segments of the input sequence is facilitated by this mechanism.

3. The output of the multi-head attention mechanism is subjected to ADD & Norm operations to obtain $\hat{\boldsymbol{X}}$. Here, ADD refers to the element-wise addition. Subsequently, the feature representation is further enhanced in the feed forward module using the nonlinear transformations of the two convolutional layers, which are computed as shown below:

$$\text{Encoder}=\textbf{Conv}\left[\text{ReLU}\left(\textbf{Conv}\left[\hat{\boldsymbol{X}},\boldsymbol{W}_{f_1},\boldsymbol{b}_{f_1}\right]\right),\boldsymbol{W}_{f_2},\boldsymbol{b}_{f_2}\right]+\hat{\boldsymbol{X}}, \quad (4)$$

where ReLU() is the activation function, $\boldsymbol{W}$ is the weight matrix, and $\boldsymbol{b}$ is the bias. The subscripts f_1 and f_2 denote the different layers of the filter.

2.3 Custom cross-attention

The dual-channel encoder achieves the isolated encoding of strong features and global multi-domain features, effectively avoiding interference from irrelevant features. However, efficiently fusing these two heterogeneous feature streams (the highly discriminative information from the strong feature channel and the potentially complementary information from the multi-domain channel) while endowing the model with the ability to dynamically focus on key information remains a critical challenge. To this end, we innovatively design a gated feature weighting fusion mechanism, namely the custom cross-attention module. This module transcends the limitations of traditional attention mechanisms that focus on homogeneous features, instead emphasizing cross-channel interactions. It introduces feature-type-aware dynamic weighting gating to ensure that, while strengthening the dominant role of strong features, it prudently excavates and integrates the potential value of weak features, thereby maximizing the synergistic benefits of heterogeneous feature streams.

In CdualTAL, the cross-attention mechanism uses scaled dot-product attention to enhance the selection capability of multi-domain features. The cross-attention structure is shown in Fig. 3, comprising format conversion, scaled dot-product attention, fusion, and feed forward modules. By applying the dot-product attention matrix, cross-attention attempts to balance global and local attention, capturing the complex dependencies between the dual input streams. The weighted fusion component employs a feature-type-aware gating mechanism to adaptively merge the outputs from the dual-channel encoder. This dynamic weighting strategy assigns probabilistic weights based on the learned importance of both strong and weak features. It meticulously preserves critical information embedded within the key strong features while judiciously integrating potentially valuable signals from the complementary weak features. Crucially, this process intensifies the model's focus on the most predictive strong features distributed across the multi-domain feature space, thereby optimizing the feature representation for subsequent decoding. This dynamic fusion ensures that the model places greater emphasis on relevant features while mitigating the impact of weaker ones. The 4×1 Conv1D and 1×1 Conv1D layers provide additional transformation layers for the feed forward module, ensuring that the model captures both short- and long-term dependencies in the sequence data. The cross-attention mechanism enables CdualTAL to focus on strong features while preserving weak but important features, leading to more robust and accurate tool wear prediction.

2.4 Hierarchical LSTM decoder

The cross-attention module achieves dynamic balancing and deep information fusion of strong and weak features, providing the model with rich and focused feature representations. However, tool wear is essentially a continuous degradation process characterized by strong temporal dependence and complex evolutionary patterns. This requires the decoder to possess powerful long-term memory capabilities and deep feature extraction abilities to fully exploit the deep dynamic patterns embedded in the fused feature sequences. To meet this need and overcome the limitations of shallow LSTM models in modeling long sequence dependencies, we have discarded the simple feature stacking strategy and innovatively designed a hierarchical LSTM decoder.

Deep LSTM models significantly outperform shallow LSTM models on prediction tasks. However, stacking multiple LSTM layers proves effective only within a certain range of layers (Yu et al., 2024). Excessive stacking can lead to overfitting or increased computational complexity. Therefore, this proposed method employs a deep decoder comprising three stacked LSTM layers to extract deep features of tool wear, combined with pooling operations to enhance their representation.

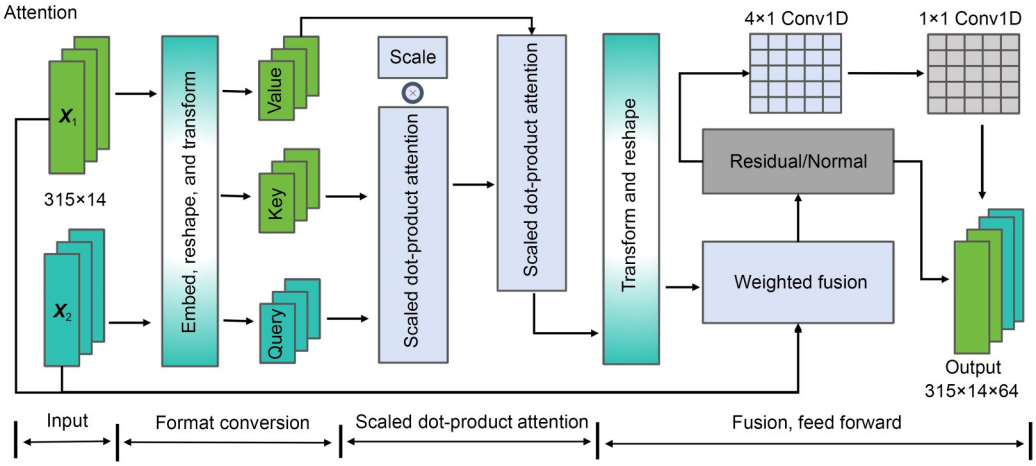

Fig. 3 Structure of cross-attention

The structure of the deep decoder (Fig. 4) consists of three primary components: three-layer LSTM, pooling, and dense. The three-layer LSTM part extracts deep features through three stacked LSTM units, with each layer's output serving as the input for the next. A residual connection between the outputs of the first and second LSTM layers serves as input to the third layer, which integrates features across different time steps and levels to capture long-term and deep dependencies in the data. This design ensures that the model can capture the full complexity of the tool-wear process over time. Additionally, the introduced hybrid pooling operation further strengthens the extraction and compression of key temporal features, providing a more discriminative high-level representation for predicting the final wear amount. The pooling part further compresses and extracts relevant features by combining average and max pooling, preserving the overall trend while capturing local extremum features. This helps the model generalize better to unseen data while maintaining high prediction accuracy. The computational process of the deep decoder for the output sequence of the cross-attention module is described as follows:

1. The LSTM of the L^{th} layer can be represented as

$$\boldsymbol{H}_t^L=\mathrm{LSTM}^L\left(\boldsymbol{X}_t,\boldsymbol{H}_{t-1}^L,\boldsymbol{C}_{t-1}^L\right), \tag{5}$$

where $\boldsymbol{H}_{t-1}^L$ is the hidden state of the L^{th} layer LSTM at time step t, $\boldsymbol{C}_{t-1}^L$ is the cell state, and the gating mechanism is calculated as follows:

$$\begin{cases} \boldsymbol{f}_t=\sigma\left(\boldsymbol{W}_{\mathrm{f}}^L\begin{bmatrix}\boldsymbol{H}_{t-1}^L\\ \boldsymbol{x}_t\end{bmatrix}+\boldsymbol{b}_{\mathrm{f}}^L\right),\\ \boldsymbol{i}_t=\sigma\left(\boldsymbol{W}_{\mathrm{i}}^L\begin{bmatrix}\boldsymbol{H}_{t-1}^L\\ \boldsymbol{x}_t\end{bmatrix}+\boldsymbol{b}_{\mathrm{i}}^L\right),\\ \tilde{\boldsymbol{c}}_t=\tanh\left(\boldsymbol{W}_{\mathrm{c}}^L\begin{bmatrix}\boldsymbol{H}_{t-1}^L\\ \boldsymbol{x}_t\end{bmatrix}+\boldsymbol{b}_{\mathrm{c}}^L\right),\\ \boldsymbol{C}_t^L=\boldsymbol{f}_t^L\boldsymbol{C}_{t-1}^L+\boldsymbol{i}_t^L\tilde{\boldsymbol{c}}_t^L,\\ \boldsymbol{o}_t=\sigma\left(\boldsymbol{W}_{\mathrm{o}}^L\begin{bmatrix}\boldsymbol{H}_{t-1}^L\\ \boldsymbol{x}_t\end{bmatrix}+\boldsymbol{b}_{\mathrm{o}}^L\right),\\ \boldsymbol{H}_t^L=\boldsymbol{o}_t^L\tanh\left(\boldsymbol{C}_t^L\right), \end{cases} \tag{6}$$

where $\tilde{\boldsymbol{c}}_t$ represents the cell state update.

The three-layer LSTM is denoted as

$$\boldsymbol{H}_t^3=\mathrm{LSTM}^3\left(\mathrm{Concat}\left[\boldsymbol{H}_t^1,\boldsymbol{H}_t^2\right],\boldsymbol{H}_{t-1}^3,\boldsymbol{C}_{t-1}^3\right), \tag{7}$$

where $\boldsymbol{f}_t$ is the forgetting gate, $\boldsymbol{i}_t$ is the input gate, $\boldsymbol{o}_t$ is the output gate, $\boldsymbol{C}_t^L$ is the memory cell, σ is the sigmoid function, and tanh is the hyperbolic tangent function. Each LSTM layer has its own weight $\boldsymbol{W}$ and bias $\boldsymbol{b}$.

2. The pooling and fully connected layers further compress and process the feature representation of the LSTM output, defined as

$$\begin{aligned}\boldsymbol{C}_{\mathrm{Pred}}=\mathrm{Sigmod}\Bigg(\boldsymbol{W}_{\mathrm{dense2}}\cdot\mathrm{ReLU}\Bigg(\boldsymbol{W}_{\mathrm{dense1}}\\ \cdot\mathrm{Concat}\left[\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{H}_t^3,\max_{t\in\{1,2,\cdots,T\}}\boldsymbol{H}_t^3\right]\Bigg)\Bigg),\end{aligned} \tag{8}$$

where $\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{H}_t^3$ represents the global average pooling operation, $\max_{t\in\{1,2,\cdots,T\}}\boldsymbol{H}_t^3$ represents the global maximum pooling operation, $\boldsymbol{W}_{\mathrm{dense}_i}$ (i=1, 2) is the weight matrix of dense, and T represents the number of time steps.

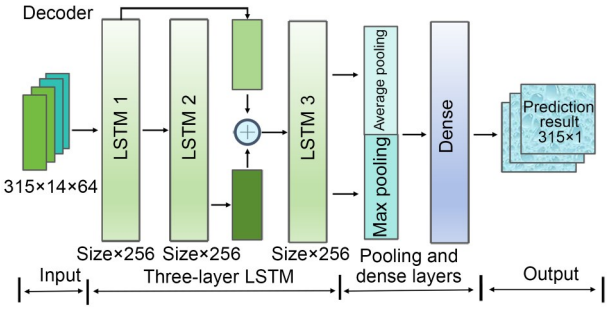

Fig. 4 Structure of the hierarchical decoder

3 Experiment setup

3.1 Experimental conditions

To validate the effectiveness of CdualTAL tool wear prediction method and its closest competitors, we test them on publicly available tool wear datasets hosted by the American Society of Prognostics and Health Management (PHM) at phmsociety.org. Specifically, the experimental data are derived from the PHM 2010 Challenge dataset. Regarding the wear labeling criteria, the ground truth is defined as the average flank wear width (VB) of the cutter. Since continuous online measurement of VB is not feasible, the ground truth is obtained through offline measurement. After each cutting pass, the tool is dismounted, and the wear value is precisely measured using a LEICA MZ12 stereo microscope. Six full-life-cycle tests are conducted. The end-milling material is a rectangular workpiece, with a consistent cutting time used for each operation and a milling length of 108 mm. Each test consists of 315 passes, during which tool flank wear is measured. The collected data include cutting force signals in the X, Y, and Z directions, vibration signals, and the root mean square (RMS) values of the acoustic emission signals during milling. The tool condition monitoring system for high-speed milling is shown in Fig. 5.

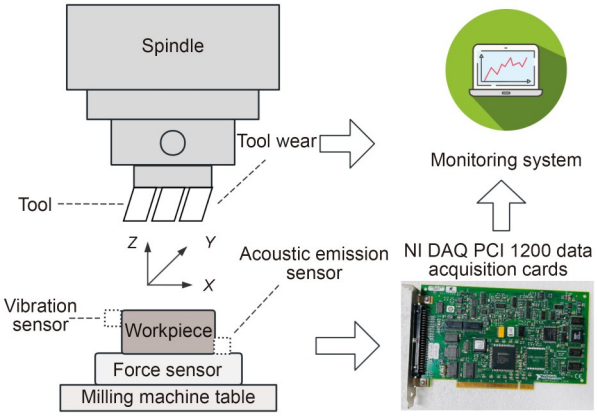

Fig. 5 Tool condition monitoring for the milling process

The tool wear prediction experiments are conducted in a Python 3.12.4 and Keras 3.4.1 environment, running on the NVIDIA GeForce RTX 4060Ti. Keras is backed by TensorFlow 2.16.1, and

the server runs Ubuntu 22.04.4 LTS. Under this experimental setup, the total number of parameters for CdualTAL model is approximately 2.0×10^6. The average training time per epoch is 14 s, and the inference latency is 190 ms per step, satisfying the requirements for near-real-time industrial monitoring.

3.2 Experimental datasets and preprocessing method

The tool wear dataset comprises six parts, where C1, C4, and C6 are labeled tool wear datasets for three different tools, and C2, C3, and C5 are unlabeled datasets for the same three tools. The labeled C1, C4, and C6 datasets are selected as experimental data. Each dataset has a dimension of 315×7, containing milling force signals in the *Z*, *Y*, and *X* directions, vibration signals, and acoustic emission signals collected over 315 tool passes. The C6 milling test is chosen as the focus; Fig. 6 illustrates the wear evolution curves of the three cutting edges of the tool.

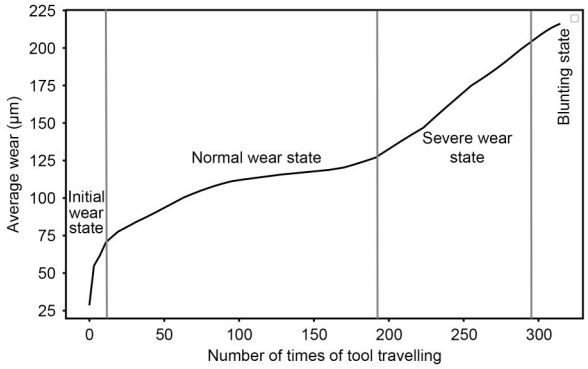

Fig. 6 Tool wear curve of the C6 tool

Fig. 6 shows that during the initial wear stage, the C6 tool's wear rate increases rapidly, resulting in a steep tool-pass curve, indicating significant wear in this phase. This is primarily due to the rapid smoothing of the tool's microscopic asperities and the running-in process between the tool and the workpiece. At the normal wear stage, the wear rate slows down, resulting in a smooth upward trend as the contact area stabilizes and wear proceeds more uniformly. At the severe wear stage, the wear rate increases significantly, indicating an exacerbation of the wear phenomenon, often caused by fatigue damage accumulation and thermal softening. Eventually, when the wear amount exceeds the tool's critical threshold, it enters the blunting failure stage, where cutting performance deteriorates abruptly. Throughout this process, the initial and dulling phases exhibit significant noise in the wear features, which has little effect on tool wear prediction because of their transient and unstable nature. Therefore, the model construction for tool wear prediction focuses on extracting signal features during the normal and rapid wear stages, where trends are more consistent and better reflect the actual degradation process.

Given the zero-drift phenomenon in the C1, C4, and C6 datasets (He et al., 2022), the tool exhibits continuous wear during actual use. However, the sensor fails to accurately reflect this change. Therefore, the *X*, *Y*, and *Z* sensor signals from all three datasets are preprocessed to mitigate the effects of this issue. The zero-drift noise and redundant information are removed, and more stable and valid features are extracted (Oppliger et al., 2024). The raw multi-sensor signals (cutting forces in *X*/*Y*/*Z* directions, vibration, and acoustic emission RMS) undergo essential preprocessing to address zero-drift noise and redundant information. The force and vibration signals are acquired at a sampling frequency of 50 kHz, consistent with the standard PHM 2010 dataset configuration. For each cutting pass, we extract 24 representative features, as listed in Table 1. In Table 1, N denotes the total number of samples, x_i denotes the i^{th} data value, z_i denotes the i^{th} data value after standardization, v_i denotes the i^{th} frequency point, and X represents the frequency domain. Specifically, wavelet packet decomposition (WPD) is employed to extract time–frequency energy features (E1–E8). Prior to model input, all feature sequences are standardized using *Z*-score normalization to ensure stable training and convergence. These manually engineered features provide physically meaningful representations of tool wear states, complementing the deep learning model's ability to learn hierarchical representations (Fawzi et al., 2022).

From the six different sensor signals of the three milling tools, a total of 3×144 signal features are extracted (i.e., 3×6×24). The signal features with correlation coefficients greater than 0.9 are selected from the three milling tools as the key strong-feature data. These strong-feature data and all signal-feature data are used as inputs to the dual-channel encoder model (Chen et al., 2024). It is worth noting that while cutting conditions (speed, feed rate, and depth of cut) vary across different machining tasks, they are not modeled as explicit input variables in this study. Instead, they are treated as latent factors intrinsically encoded in the variations of the multi-sensor signals (cutting force, vibration, and acoustic emission). The proposed deep learning architecture captures the physical effects of these conditions by directly learning complex nonlinear representations from the extracted signal features. Since the tool wear dataset includes three labeled tool wear datasets (C1, C4, and C6) with consistent data sizes (315 data points each), this work uses a three-fold cross-validation method to validate the generalization ability of CdualTAL.

The cross-validation method alternates between training and validation sets. This process enhances the comprehensiveness and reliability of CdualTAL's performance evaluation, and reduces the bias and instability introduced by a single data split (Turbé et al., 2023). The dataset partitioning method is shown in Table 2, where the training and validation sets are split in an 8:2 ratio.

3.3 Evaluation metrics

The evaluation metrics commonly used in existing tool wear prediction studies include the root mean square error (RMSE) and mean absolute error (MAE), with little emphasis on R^2. R^2 provides a more straightforward and intuitive measure of the model's fit to the data. The higher the R^2 value, the stronger the model's ability to capture the data trends and structure. Therefore, we adopt three commonly used evaluation metrics for regression prediction—R^2, RMSE, and MAE—as the evaluation criteria for CdualTAL. The formulas for the three evaluation metrics are as follows:

$$R^2=1-\frac{\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i-\overline{Y}_i\right)^2}, \tag{9}$$

Table 1 Multi-sensor signal extraction expressions

No.	Expression
1	$\text{Absolute}=\frac{1}{N}\sum_{i=1}^{N}\lvert x_i\rvert$
2	$\text{Max}=\max(x_i)$
3	$\text{RootMeanSquare}=\sqrt{\frac{1}{N}\sum_{i=1}^{N}X_i^2}$
4	$\text{SquareRootAmplitude}=\left(\frac{1}{N}\sum_{i=1}^{N}\sqrt{\lvert x_i\rvert}\right)^2$
5	$\text{Skewness}=\frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\text{AbsoluteMean})^3}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i-\text{AbsoluteMean})^2\right)^{3/2}}$
6	$\text{Kurtosis}=\frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\text{AbsoluteMean})^4}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i-\text{AbsoluteMean})^2\right)^2}$
7	$\text{ShapeFactor}=\frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}x_i^2}}{\frac{1}{N}\sum_{i=1}^{N}\lvert x_i\rvert}$
8	$\text{PulseFactor}=\frac{\max\left(\lvert x_i\rvert\right)}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}x_i^2}}$
9	$\text{SkewnessFactor}=\frac{\frac{1}{N}\sum_{i=1}^{N}(z_i-\bar{z})^3}{\left(\frac{1}{N}\sum_{i=1}^{N}(z_i-\bar{z})^2\right)^{3/2}}$
10	$\text{CrestFactor}=\frac{\max\left(\lvert x_i\rvert\right)}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}x_i^2}}$
11	$\text{ClearanceFactor}=\frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}x_i^2}}{\frac{1}{N}\sum_{i=1}^{N}x_i}$
12	$\text{KurtosisFactor}=\frac{\text{Kurtosis}}{(\text{RootMeanSquare})^4}$
13	$\text{CenterFrequency}=\frac{\sum_{i=1}^{N}v_i\cdot\lvert X(v_i)\rvert^2}{\sum_{i=1}^{N}\lvert X(v_i)\rvert^2}$
14	$\text{MeanSquareFrequency}=\frac{\sum_{i=1}^{N}v_i^2\cdot\lvert X(v_i)\rvert^2}{\sum_{i=1}^{N}\lvert X(v_i)\rvert^2}$
15	$\text{RootMeanSquareFrequency}=\sqrt{\frac{\sum_{i=1}^{N}v_i^2\cdot\lvert X(v_i)\rvert^2}{\sum_{i=1}^{N}\lvert X(v_i)\rvert^2}}$
16	$\text{FrequencyVariance}=\frac{\sum_{i=1}^{N}\lvert X(v_i)\rvert^2(v_i-\bar{v})^2}{\sum_{i=1}^{N}\lvert X(v_i)\rvert^2}$
17–24	E1–E8

Table 2 Dataset partitioning

Training & validation datasets	Test dataset
C1+C4	C6
C1+C6	C4
C4+C6	C1

$$\text{RMSE}=\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Y_i-\hat{Y}_i\right)^2}, \tag{10}$$

$$\text{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|Y_i-\hat{Y}_i\right|, \tag{11}$$

where n is the workload. The real observed value is the corresponding predicted value $\hat{Y}_i$, and the average of the actual observed values $\overline{Y}_i$ is represented.

4 Experiments

Fig. 7 presents the prediction results of CdualTAL. Fig. 7a shows the wear prediction for the C1 tool, Fig. 7b shows the wear prediction for the C4 tool, and Fig. 7c shows the wear prediction for the C6 tool. The red line represents the values predicted by CdualTAL, while the black line represents the actual wear values. The prediction errors between the predicted and actual values are represented by the red histogram at the bottom, and the prediction error range is highlighted by the green shaded area in the figure.

From Fig. 7, it is evident that the wear curves predicted by CdualTAL closely match the actual wear curves, with most prediction errors falling within a narrow range. This demonstrates that CdualTAL excels at capturing the key features of and deep information from the tool wear data, enabling precise modeling of the wear trend. Additionally, CdualTAL exhibits outstanding stability in wear prediction at different stages, fully validating its reliability and applicability for tool wear prediction under complex working conditions.

However, higher prediction errors are concentrated mainly at the blunting failure stage of the wear curve. During this stage, the sensor-acquired signals contain more noise, which negatively affects the model's prediction performance, leading to larger prediction errors. Furthermore, in the initial wear phase, some deviation between the three predicted curves and the actual wear curve can be observed. This is mainly because, at the early stages of wear, the wear on the tool is small, and there are fewer wear features for the model to learn, which affects the accuracy of the predictions.

4.1 Model comparison

This work systematically compares the predictive performance of CdualTAL with those of several advanced tool wear prediction methods. To ensure a fair and robust comparison, the network architectures and specific hyperparameter configurations for all baseline models are set strictly in accordance with the optimal parameters recommended in their respective original literature. This approach guarantees that each baseline model operates at its intended performance level. All models are evaluated using the

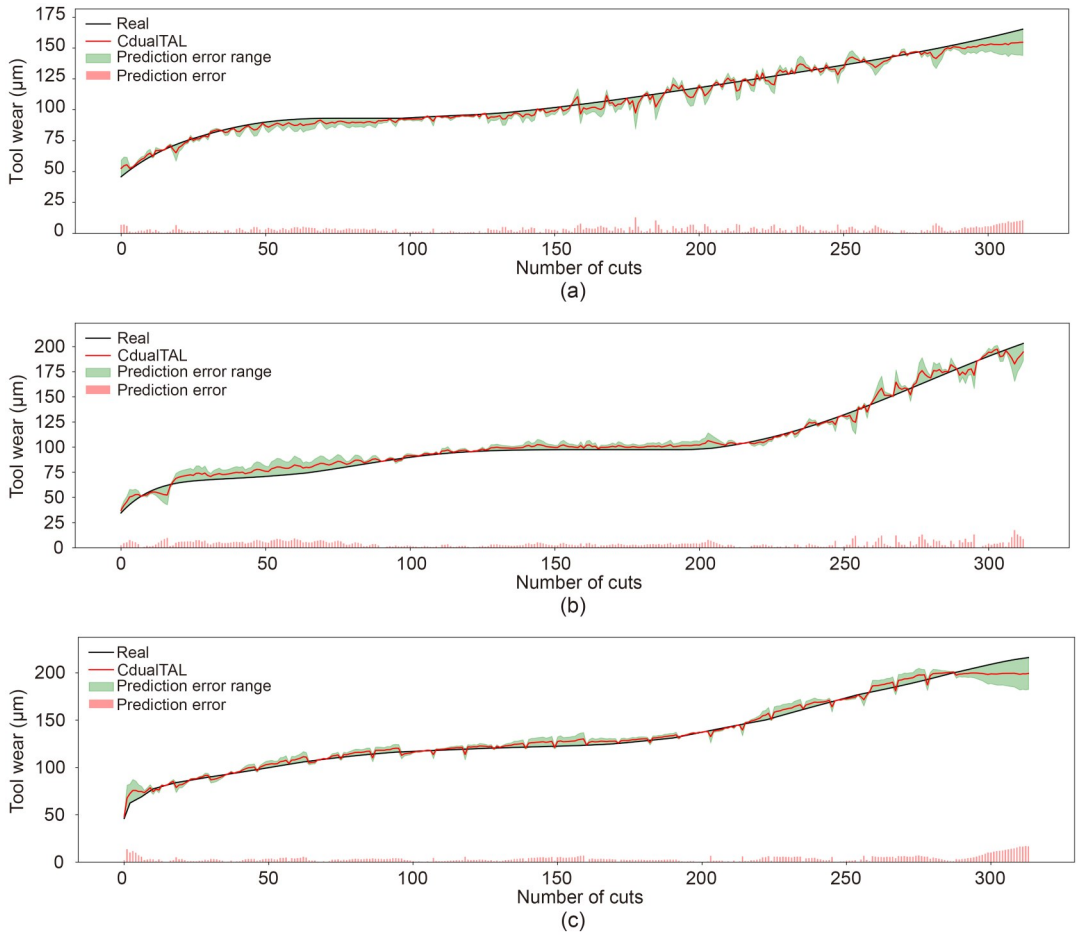

Fig. 7 Prediction results with error bar of CdualTAL: (a) C1 tool; (b) C4 tool; (c) C6 tool

Table 3 Comparison of prediction metrics between the CdualTAL and baseline models

Model	C1			C4			C6			Average		
	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE	R^2	RMSE	MAE
CdualTAL	0.979	3.780	3.051	0.984	4.680	3.793	0.985	4.660	3.514	0.983	4.373	3.453
CNN–XGBoost	0.910	8.110	6.187	0.900	11.640	8.697	0.930	10.660	8.722	0.913	10.137	7.869
DCNN–SLSTM	–	8.340	6.940	–	9.930	6.590	–	10.020	7.740	–	9.430	7.090
CNN–LSTM (Qiao et al., 2018)	–	13.770	11.180	–	11.850	9.390	–	14.330	11.340	–	13.317	10.637
Deep–LSTM (Zhao et al., 2016)	–	8.300	12.100	–	10.200	8.700	–	18.900	15.200	–	12.467	12.000
BiLPRes (Si et al., 2024)	–	4.341	3.303	–	5.483	4.520	–	7.395	6.797	–	5.740	4.873
DH–GRU (Wang JJ et al., 2019)	–	4.660	3.700	–	8.730	7.070	–	6.940	5.080	–	6.777	5.283
CNN–BiLSTM (Wang CG et al., 2024)	–	6.930	5.530	–	10.100	7.700	–	11.840	8.660	–	9.623	7.300
HLLSTM (Chan et al., 2022)	–	8.000	6.600	–	7.500	6.000	–	8.800	7.100	–	8.100	6.567
IRM–CFAM (Wu et al., 2021)	–	6.018	4.712	–	7.395	5.084	–	6.667	5.613	–	6.693	5.136
SSAE–BPNN (He et al., 2022)	0.885	9.258	–	0.864	13.988	–	0.864	14.767	–	0.871	12.671	–
IE–SBiGRU (Li et al., 2022)	–	5.056	3.694	–	6.884	5.189	–	4.527	3.398	–	5.489	4.094

"–" indicates that specific experimental results are not reported in the original paper

same dataset to further verify the effectiveness, stability, and applicability of CdualTAL. Table 3 presents the prediction metrics (R^2, RMSE, and MAE) for CdualTAL and 11 baseline models on the three datasets (C1, C4, C6). It is important to note that these results are derived from the three-fold cross-validation experiments described in Section 3.2. The average column represents the consolidated performance across the three experimental runs, demonstrating the model's stability and repeatability without requiring additional single-seed repetitions. CdualTAL performs the best on the C1 and C4 datasets. Its predicted RMSE and MAE values are significantly lower than those of the other models, indicating excellent predictive performance.

However, the prediction results of CdualTAL are slightly inferior to those of the IE-SBiGRU model on the C6 dataset. In-depth analysis, combined with the experiments shown in Fig. 8, indicates that the C6 tool exhibits more severe wear fluctuations and sensor

signal noise during the passivation failure stage, posing specific challenges for CdualTAL, which relies on global feature representation and long sequence modeling. Although CdualTAL's dual-channel isolated encoding and cross-attention mechanism can effectively suppress noise interference and fuse key information under most operating conditions, during the extreme passivation stage of C6, high-frequency noise may partially obscure strong feature signals, resulting in a slight decrease in the model's accuracy in capturing local abrupt changes. This will inform future model optimization.

To further assess the models' overall performance, we calculate the average prediction metrics across three datasets. CdualTAL achieves an average R^2 of 0.983, an RMSE of 4.373, and an MAE below 3.5, significantly surpassing all 11 benchmark models. This superior performance is due to the model's innovative dual-channel encoder, which isolates feature interference, and the custom cross-attention mechanism, which adaptively balances strong and weak features, collectively enhancing the model's ability to distill discriminative patterns from noisy multi-domain data. While the IESBiGRU model demonstrates slight advantages on the C6 dataset because its bidirectional gated units excel at local temporal variations, the consistent superiority of CdualTAL in terms of average metrics underscores its exceptional robustness and generalizability, which other models cannot achieve. The high R^2 value (close to 1.0) confirms the model's near-perfect fit to wear trends, while the low RMSE and MAE highlight its precision in minimizing prediction errors, validating CdualTAL as a comprehensive solution for diverse tool conditions. CdualTAL offers higher prediction accuracy than other prediction methods, and demonstrates good adaptability and prediction stability across different tool conditions, fully validating its effectiveness and reliability for tool wear prediction.

4.2 Ablation study

CdualTAL comprises four primary modules: feature selection, encoder, attention, and decoder. To comprehensively analyze each module's contribution to the performance metrics, we design four experiments on the C6 dataset, each validating a specific module. This targeted ablation study focuses on the C6 tool for two key reasons: first, as shown in Fig. 6, the C6 wear curve exhibits the most comprehensive degradation characteristics, including distinct initial, normal, rapid, and failure stages, providing a holistic benchmark for evaluating module robustness across all wear phases. Furthermore, while the three-fold cross-validation validates the model's generalizability, isolating experiments to C6 eliminates tool-specific variability, allowing us to accurately attribute performance changes to architectural modifications. The detailed experiment descriptions are presented in Table 4.

Table 4 Description of the setup for ablation study

Model	Setup of each model
1	Replace the feature selection module with the static method for selecting strong features based on correlation coefficients
2	Control the number of layers in the dual-channel Transformer encoder module
3	Replace the custom cross-attention module with the common cross-attention module
4	Control the number of layers in the hierarchical LSTM decoder module

Extensive studies have shown that deeper neural network architectures typically yield superior predictive capabilities compared to their shallower counterparts, primarily due to their enhanced capacity for hierarchical feature abstraction and complex pattern recognition. However, this advantage is constrained by a critical trade-off: excessive layering leads to diminishing marginal returns in performance gains, while exponentially increasing computational complexity and the risk of overfitting. Therefore, this research limits the maximum depth of the encoder and decoder modules to three layers.

To evaluate the necessity and superiority of the proposed correlation-adaptive feature selection algorithm, we construct a contrast model (model 1). This model replaces the correlation-adaptive feature selection module with a conventional static feature selection method. Specifically, we calculate the Pearson correlation coefficients between each feature and the tool wear value across the entire training dataset. The top-k features with the highest absolute correlation values are then selected as the "strong features" for the entire lifecycle of the test tool, thereby implementing a static, global feature-ranking strategy. This contrast model retains the dual-channel encoder and cross-attention architecture, but lacks the dynamic and adaptive feature evaluation capabilities inherent in our proposed module.

We conduct experiments across four module configurations, and the resulting prediction metrics are presented in Table 5. Concurrently, we record the experimental results of completely removing these innovative modules, as shown in Fig. 8.

Table 5 Ablation study of different modules in CdualTAL model

Model	Configuration	R^2	RMSE	MAE	Time per step (ms)
1	Static selection	0.941	9.857	8.932	113
	Correlation-adaptive selection	0.985	4.660	3.514	190
2	Dual-channel Transformer-0	0.935	10.08	8.068	124
	Dual-channel Transformer-1	0.941	9.62	7.717	154
	Dual-channel Transformer-2	0.967	7.26	5.730	176
	Dual-channel Transformer-3	0.985	4.66	3.514	190
3	Multi-head attention	0.927	10.73	8.509	213
	Cross-attention	0.985	4.66	3.514	190
4	Hierarchical LSTM-0	0.893	13.00	10.821	80
	Hierarchical LSTM-1	0.914	11.67	9.310	109
	Hierarchical LSTM-2	0.940	9.72	7.683	149
	Hierarchical LSTM-3	0.985	4.66	3.514	190

The critical role of the correlation-adaptive feature selection module is immediately apparent in Table 5 and Figs. 8a–8b. Model 1 exhibits poorer performance, with an R^2 of 0.941 and significantly elevated RMSE (9.857) and MAE (8.932). While its computational cost is lower due to the simplicity of static correlation calculation, this comes at the great expense of predictive accuracy.

The performance gap between model 1 and the complete CdualTAL underscores a key insight: the correlation between sensor features and tool wear is not static but evolves throughout the tool's lifecycle. A feature that is highly correlated with wear at the initial or normal stage may become less informative or even noisy

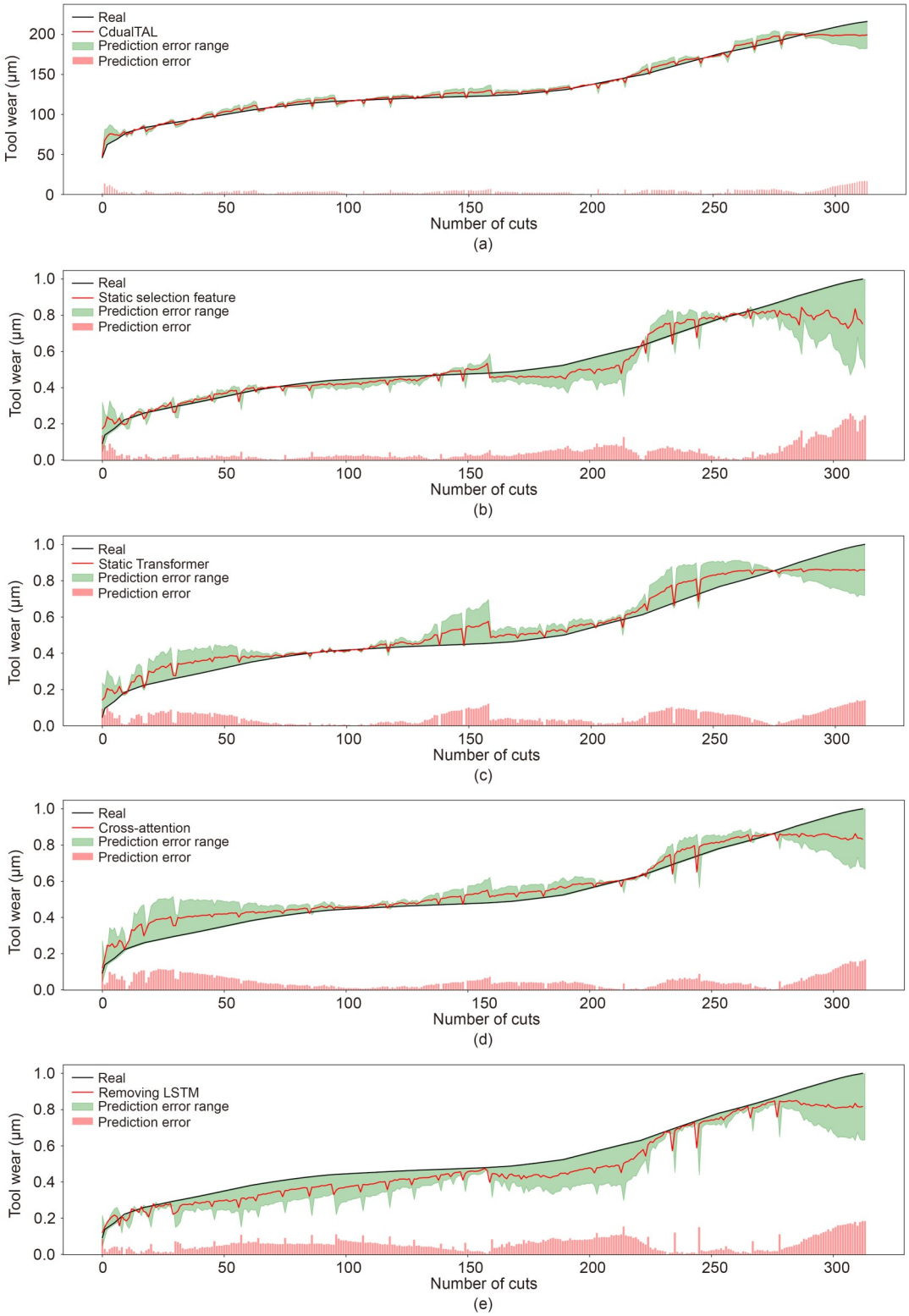

Fig. 8 Prediction results with error bars of different models: (a) CdualTAL model; (b) static selection; (c) removing Transformer; (d) cross-attention; (e) removing LSTM

at the rapid wear or failure stages, and vice versa. The static method fails to capture this time-varying characteristic, forcing the model to rely on a fixed, potentially suboptimal set of features for the entire sequence, leading to significant prediction errors, particularly during phase transitions. Conversely, by continuously re-evaluating feature relevance within a sliding window, the proposed module dynamically adapts the set of strong features, ensuring that the model consistently prioritizes the most discriminative signals for the current wear state, which is paramount for achieving high accuracy throughout the degradation process. This ablation study conclusively proves that the dynamic nature of feature selection is a major contributor to the model's state-of-the-art performance.

On the other hand, Table 5 demonstrates that for both the encoder and decoder modules (models 2 and 4), increasing the number of layers leads to a steady improvement in the model's R^2 for prediction. This indicates that deeper feature representations significantly enhance the model's predictive capability. At the same time, error metrics such as RMSE and MAE decrease significantly as the number of layers increases. This suggests that the model can effectively capture the complex patterns of tool wear in deeper structures, thereby reducing prediction errors.

However, as the number of layers in the encoder and decoder modules increases, the model's computational complexity also rises, leading to increased time costs per step. For the attention module (model 3), compared with the traditional multi-head attention, the cross-attention module demonstrates significant advantages in predictive performance and computational efficiency. These results indicate that cross-attention is more effective in capturing the interaction relationships between tool wear features.

Fig. 8 presents the prediction results after removing four different modules. As shown in model 2, adding the dual-channel Transformer module improves the model's ability to fit the tool wear curve. The original model exhibits some deviations when fitting the curve, but the subsequent dual-channel Transformer model more accurately captures the trends and features of the tool wear curve, thereby reducing prediction errors.

Model 3 shows that traditional cross-attention achieves lower prediction accuracy than custom cross-attention. The prediction curve fits poorly, and the prediction error is relatively high, highlighting the role of our weighted fusion module.

From model 4, it can be observed that when the hierarchical LSTM module is properly configured, the model's predicted curve more closely matches the actual data. This improvement is particularly evident during the initial and normal wear stages, where the predicted curve almost completely matches the actual values. However, as shown in Table 5, increasing the number of layers, from a single layer to two layers, does not significantly enhance the model's predictive performance metrics. Only when the number of layers increases to three, can the prediction metrics reach their highest values, at which point the predicted curve nearly perfectly matches the actual curve. This is because the third layer of the hierarchical LSTM module integrates the outputs from the first and second layers, enabling more efficient feature fusion and information sharing, resulting in a superior fit to the experimental tool wear curve.

5 Conclusions

This work presents CdualTAL, an innovative tool-wear prediction algorithm that fundamentally addresses the dual challenges of multi-domain feature noise and long-range dependency modeling. Unlike conventional methods that either concatenate features or apply standard attention mechanisms, CdualTAL introduces a heterogeneous paradigm that synergistically combines feature isolation, adaptive fusion, and hierarchical decoding. The algorithm establishes a new benchmark in industrial prognosis by achieving state-of-the-art performance with an average R^2 of 0.983 and an RMSE of 4.373 across three PHM Society datasets.

The core contributions of CdualTAL are fourfold. First, the correlation-adaptive feature selection module uses the sliding window maximal information coefficient to adaptively identify and isolate high-predictivity features while suppressing redundant noise. By continuously re-evaluating feature correlations within temporal windows, it dynamically optimizes the feature subset, ensuring robustness against sensor drift and wear-stage transitions, which is critical for handling noise phases such as the blunting failure stage. Second, the dual-channel Transformer encoder processes strong features and global multi-domain features in parallel, preventing cross-interference while extracting discriminative patterns. Third, the custom cross-attention mechanism dynamically balances strong and weak features through feature-type-aware weighting, enhancing focus on critical signals while judiciously integrating complementary information from weak features. Finally, the hierarchical LSTM decoder with residual connections captures deep temporal dependencies, enabling precise modeling of wear evolution across initial, normal, and rapid stages.

These components collectively form an integrated architecture that addresses the ubiquitous feature interference problem in multi-sensor systems. Comprehensive validation against 11 state-of-the-art methods confirms the superiority of CdualTAL. Ablation studies further demonstrate the indispensability of each module, highlighting the role of the dynamic feature selection module in reducing noise during extreme wear phases compared with static alternatives. The algorithm's high R^2 and low error metrics validate its ability to model complex wear trajectories.

Despite these advancements, some limitations remain. Due to the sparse availability of features, the model exhibits a slight performance decline during the initial wear phase, and the computational overhead increases with the depth of the encoder and decoder. Additionally, although CdualTAL effectively mitigates sensor noise through dynamic feature selection, challenges persist during extreme wear phases. Future work will integrate topological signal processing techniques to derive noise-invariant representations and explore topological mathematics to extract noise-invariant input features, enhance real-time application capabilities through lightweight architectures, and extend the framework to other industrial monitoring tasks such as bearing fault diagnosis.

Acknowledgments

This work was supported by the Shandong Provincial Key Research and Development Program (No. 2024CXPT011) and the National Key Research and Development Program of China (No. 2024YFB3312302).

Author contributions

Na LI and Zhendong LIU designed the research. Xiao WANG processed the data. Xiao WANG and Jiamin JIANG drafted the paper. Yanjie WEI helped organize the paper. Zhendong LIU revised and finalized the paper.

Conflict of interest

All the authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declaration on the use of generative AI tools

The authors declare that they have not used any generative AI tool during the preparation of this paper.

References

Cai WL, Zhang WJ, Hu XF, et al., 2020. A hybrid information model based on long short-term memory network for tool condition monitoring. *J Intell Manuf*, 31(6): 1497-1510. https://doi.org/10.1007/s10845-019-01526-4

Chan YW, Kang TC, Yang CT, et al., 2022. Tool wear prediction using convolutional bidirectional LSTM networks. *J Supercomput*, 78(1):810-832. https://doi.org/10.1007/s11227-021-03903-4

Chehrehzad M, Kecibas G, Besirova C, et al., 2024. Tool wear prediction through AI-assisted digital shadow using industrial edge device. *J Manuf Process*, 113:117-130. https://doi.org/10.1016/j.jmapro.2024.01.052

Chen XP, Wang R, Khalilian-Gourtani A, et al., 2024. A neural speech decoding framework leveraging deep learning and speech synthesis. *Nat Mach Intell*, 6(4):467-480. https://doi.org/10.1038/s42256-024-00824-8

Duan J, Hu C, Zhan XB, et al., 2022. MS-SSPCANet: a powerful deep learning framework for tool wear prediction. *Rob Comput Integr Manuf*, 78:102391. https://doi.org/10.1016/j.rcim.2022.102391

Duan J, Zhang X, Shi TL, 2023. A hybrid attention-based paralleled deep learning model for tool wear prediction. *Expert Syst Appl*, 211:118548. https://doi.org/10.1016/j.eswa.2022.118548

Fawzi A, Balog M, Huang A, et al., 2022. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610(7930):47-53. https://doi.org/10.1038/s41586-022-05172-4

Gao KP, Xu XX, Jiao SJ, 2022. Measurement and prediction of wear volume of the tool in nonlinear degradation process based on multi-sensor information fusion. *Eng Fail Anal*, 136:106164. https://doi.org/10.1016/j.engfailanal.2022.106164

Guo H, Lin X, Zhu KP, 2022. Pyramid LSTM network for tool condition monitoring. *IEEE Trans Instrum Meas*, 71:2509511. https://doi.org/10.1109/TIM.2022.3173278

He ZP, Shi TL, Xuan JP, 2022. Milling tool wear prediction using multi-sensor feature fusion based on stacked sparse autoencoders. *Measurement*, 190:110719. https://doi.org/10.1016/j.measurement.2022.110719

Hou KL, Li RY, Liu XL, et al., 2025. Swin-fusion: an adaptive multi-source information fusion framework for enhanced tool wear monitoring. *J Manuf Syst*, 79:435-454. https://doi.org/10.1016/j.jmsy.2025.02.003

Huang ZW, Zhu JM, Lei JT, et al., 2020. Tool wear predicting based on multi-domain feature fusion by deep convolutional neural network in milling operations. *J Intell Manuf*, 31(4):953-966. https://doi.org/10.1007/s10845-019-01488-7

Huang ZW, Shao JJ, Zhu JM, et al., 2024. Tool wear condition monitoring across machining processes based on feature transfer by deep adversarial domain confusion network. *J Intell Manuf*, 35(3):1079-1105. https://doi.org/10.1007/s10845-023-02088-2

Kejriwal M, Kildebeck E, Steininger R, et al., 2024. Challenges, evaluation and opportunities for open-world learning. *Nat Mach Intell*, 6(6):580-588. https://doi.org/10.1038/s42256-024-00852-4

Khoshouei M, Bagherpour R, Yari M, 2024. A smart look at monitoring while drilling (MWD) and optimizing using acoustic emission technique (AET). *Sci Rep*, 14(1): 19766. https://doi.org/10.1038/s41598-024-70717-8

Kumar AS, Agarwal A, Jansari VG, et al., 2025. Realizing on-machine tool wear monitoring through integration of vision-based system with CNC milling machine. *J Manuf Syst*, 78:283-293. https://doi.org/10.1016/j.jmsy.2024.12.004

Li WY, Fu HY, Han ZY, et al., 2022. Intelligent tool wear prediction based on Informer encoder and stacked bidirectional gated recurrent unit. *Rob Comput Integr Manuf*, 77:102368. https://doi.org/10.1016/j.rcim.2022.102368

Lu SF, Zhu YJ, Liu S, et al., 2022. A tool wear prediction model based on attention mechanism and IndRNN. Int Joint Conf on Neural Networks, p.1-7. https://doi.org/10.1109/IJCNN55064.2022.9889794

Marani M, Zeinali M, Songmene V, et al., 2021. Tool wear prediction in high-speed turning of a steel alloy using long short-term memory modelling. *Measurement*, 177:109329. https://doi.org/10.1016/j.measurement.2021.109329

Max K, Kriener L, Pineda García G, et al., 2024. Learning efficient backprojections across cortical hierarchies in real time. *Nat Mach Intell*, 6(6):619-630. https://doi.org/10.1038/s42256-024-00845-3

Mo CQ, Huang K, Ji HX, 2024. A lightweight and precision dual track 1D and 2D feature fusion convolutional network for machinery equipment fault diagnosis. *Sci Rep*, 14(1):31666. https://doi.org/10.1038/s41598-024-81118-2

Oppliger J, Denner MM, Küspert J, et al., 2024. Weak signal extraction enabled by deep neural network denoising of diffraction data. *Nat Mach Intell*, 6(2):180-186. https://doi.org/10.1038/s42256-024-00790-1

Ou JY, Liu HK, Huang GJ, et al., 2021. Tool wear recognition based on deep kernel autoencoder with multichannel signals fusion. *IEEE Trans Instrum Meas*, 70:3521909. https://doi.org/10.1109/TIM.2021.3096283

Qiao HH, Wang TY, Wang P, et al., 2018. A time-distributed spatiotemporal feature learning method for machine health monitoring with multi-sensor time series. *Sensors*, 18(9):2932. https://doi.org/10.3390/s18092932

Shah M, Vakharia V, Chaudhari R, et al., 2022. Tool wear prediction in face milling of stainless steel using singular generative adversarial network and LSTM deep learning models. *Int J Adv Manuf Technol*, 121(1-2):723-736. https://doi.org/10.1007/s00170-022-09356-0

Shah M, Borade H, Dave V, et al., 2024. Utilizing TGAN and ConSinGAN for improved tool wear prediction: a comparative study with ED-LSTM, GRU, and CNN models. *Electronics*, 13(17):3484. https://doi.org/10.3390/electronics13173484

Shi CM, Luo B, He SP, et al., 2020. Tool wear prediction via multidimensional stacked sparse autoencoders with feature fusion. *IEEE Trans Ind Inform*, 16(8): 5150-5159. https://doi.org/10.1109/TII.2019.2949355

Si ZK, Si SM, Mu DQ, 2024. Efficient tool wear prediction in manufacturing: BiLPReS hybrid model with performer encoder. *Arab J Sci Eng*, 49(11):15193-15204. https://doi.org/10.1007/s13369-024-08943-5

Turbé H, Bjelogrlic M, Lovis C, et al., 2023. Evaluation of post-hoc interpretability methods in time-series classification. *Nat Mach Intell*, 5(3):250-260. https://doi.org/10.1038/s42256-023-00620-w

Wang CG, Wang GP, Wang T, et al., 2024. Exploring the processing paradigm of input data for end-to-end deep learning in tool condition monitoring. *Sensors*, 24(16): 5300. https://doi.org/10.3390/s24165300

Wang JJ, Yan JX, Li C, et al., 2019. Deep heterogeneous GRU model for predictive analytics in smart manufacturing: application to tool wear prediction. *Comput Ind*, 111:1-14. https://doi.org/10.1016/j.compind.2019.06.001

Wang S, Yu ZL, Guo YQ, et al., 2022. A CNN-LSTM-PSO tool wear prediction method based on multi-channel feature fusion. *Mech Eng Sci*, 4(2):39-48. https://doi.org/10.33142/mes.v4i2.9086

Wu CH, Wu FZ, Qi T, et al., 2021. Fastformer: additive attention can be all you need. https://doi.org/10.48550/arXiv.2108.09084

Yan BL, Zhu LD, Dun YC, 2021. Tool wear monitoring of TC4 titanium alloy milling process based on multi-channel signal and time-dependent properties by using deep learning. *J Manuf Syst*, 61:495-508. https://doi.org/10.1016/j.jmsy.2021.09.017

Yu XN, Wang HF, Wang JR, et al., 2024. A common feature-driven prediction model for multivariate time series data. *Inform Sci*, 677:120967. https://doi.org/10.1016/j.ins.2024.120967

Zhao R, Wang JJ, Yan RQ, et al., 2016. Machine health monitoring with LSTM networks. 10th Int Conf on Sensing Technology, p.1-6. https://doi.org/10.1109/ICSensT.2016.7796266