

Article

Unsupervised Tool State Classification Based on Multi-Modal Fusion

Xuanbo Liu, Jiaqi Zhou *, Caixu Yue, Ting Sun, Haojie Huang and Xingyu Chen

Key Laboratory of Advanced Manufacturing and Intelligent Technology, Ministry of Education, Harbin University of Science and Technology, Harbin 150080, China; 15504529967@163.com (X.L.); yuecaixu@hrbust.edu.cn (C.Y.); thenewepoch0116@163.com (T.S.); huanghaojie200604@163.com (H.H.); 18302494895@163.com (X.C.)

* Correspondence: zhoujiaqi@hrbust.edu.cn

Abstract

As a critical process in the manufacturing of core components for equipment, milling directly influences workpiece quality and production efficiency through the process of tool wear. However, current monitoring methods face limitations in extracting comprehensive wear features from multi-source signals, while data-driven approaches are constrained by scarce wear samples and imbalanced stage distribution in practical applications. To address these challenges, this paper presents an unsupervised tool wear state classification method based on multi-modal feature fusion. The proposed approach establishes a dual-channel feature extraction model that processes vibration signals and workpiece surface texture images separately, enabling complementary-information integration through a fusion strategy. An autoencoder architecture performs unsupervised dimensionality reduction and reconstruction of the fused high-dimensional features, extracting low-dimensional representations strongly correlated with wear states. By optimizing feature discriminability through reconstruction error minimization, the method effectively mitigates the impact of limited labeled data on model performance. Experimental validation using cemented carbide milling cutters demonstrates that the proposed method achieves a classification accuracy of 93.33% under unsupervised conditions, significantly outperforming conventional unsupervised approaches. This research offers a viable approach for intelligent perception and precise analysis of cutting processes, presenting practical value for advancing equipment condition monitoring, process optimization, and predictive maintenance within intelligent manufacturing systems.

Keywords: multi-modal fusion; unsupervised learning; tool condition monitoring

Academic Editor: Kai Cheng

Received: 13 May 2026

Revised: 8 July 2026

Accepted: 9 July 2026

Published: 14 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.

1. Introduction

In milling operations, tool wear is an inevitable and critical factor that has a direct impact on both machining quality and efficiency [1]. Severe tool wear not only leads to compromised dimensional accuracy and degraded surface quality of workpieces, but may also result in catastrophic tool failure or other accidents, undermining productivity and causing economic losses [2]. Therefore, establishing a high-precision Tool Condition Monitoring (TCM) system is essential to accurately identify the tool wear state and provide early warnings, thereby ensuring the stability of precision machining processes.

In recent years, TCM methods have been primarily categorized into indirect and direct approaches [3]. The indirect method determines the tool wear state by collecting various physical signals during machining, such as vibration, cutting force, spindle power, acoustic emission, and current signals, and it is widely employed in industrial applications [4]. Researchers commonly combine time-domain and frequency-domain analysis techniques to extract relevant features from these signals and establish correlation models with tool wear. In contrast to indirect approaches, the direct method employs machine vision to capture images of the tool wear area, enabling an intuitive assessment of wear condition [5]. This approach minimizes interference during signal transmission and exhibits strong noise immunity. However, single-modal monitoring methods possess inherent limitations: signal data are susceptible to environmental noise and variations in working conditions, resulting in insufficient feature stability, and while image data offer visual clarity, they often fail to reflect continuous wear progression. Given these constraints, multi-modal fusion has emerged as a promising direction for enhancing the performance of TCM systems. This approach compensates for the limited feature coverage of single-modal data by integrating complementary information from multiple data modalities. Furthermore, the existing literature has demonstrated the advantages of multi-modal fusion methods in enhancing the performance of TCM systems. Zhang et al. performed decision-level fusion after acoustic emission and cutting force signals, demonstrating that the complementary use of dual sensors outperforms single-sensor monitoring [6]. Kou et al. classified tool wear states by fusing vibration, current, and infrared image data. They transformed vibration and current signals into images using the GADF method before conducting data-level fusion, achieving a recognition accuracy of 91% and validating the effectiveness of multi-sensor integration [7]. Hou et al. converted signal data into images and constructed a dual-channel model combining a residual network with a gated recurrent unit, which was then fused with the original signal to enhance prediction performance [8]. In another study, Yuan et al. extracted features from current signals using variational mode decomposition, and they obtained texture features from machined surface images using rotation-invariant uniform local binary patterns (RULBPs) and grayscale gradient co-occurrence matrices. These features were fused to establish correlations with tool wear [9]. Collectively, these studies confirm that multi-modal fusion can substantially improve the ability of monitoring systems.

In the development of TCM algorithms, deep learning methods have attracted considerable attention due to their powerful feature extraction capabilities. However, the heavy reliance of traditional supervised learning on labeled data significantly constrains their practical applicability. The labeling of tool wear state requires offline measurement and empirical judgment, which is time-consuming and labor-intensive, and it is difficult to achieve application in large-scale production scenarios [10]. Unsupervised learning offers a promising alternative by exploring inherent patterns and distributions within the data itself [11], thereby enabling state classification without the need for labeled samples.

Unsupervised learning approaches for TCM primarily encompass clustering algorithms [12], dimensionality reduction techniques [13], and autoencoders [14]. Yan et al. developed a Robust Convolutional Autoencoder (RCAE) tailored for tool anomaly detection in industrial environments subject to noise interference [15]. Addressing classification challenges, Jiang et al. proposed a framework integrating unsupervised Transfer Component Analysis (TCA) with Support Vector Machines (SVMs), achieving notable improvement in classification accuracy [16]. For unsupervised detection and segmentation of surface defects, Yang et al. introduced a Residual Pooling Autoencoder (RPAE) that offers high precision [17]. While these methods have shown promising results, their feature extraction processes exhibit certain limitations. The multiple transformations involved in deriving deep features may introduce deviations from the original data characteristics.

Furthermore, existing autoencoder models are predominantly designed for single-modal data, making it challenging to effectively capture cross-modal correlation features.

Despite the advances reviewed above, several critical limitations remain in existing tool wear state classification methods. First, most existing fusion strategies rely on simple concatenation or decision-level integration, which often fail to fully exploit the complementary information across heterogeneous modalities. Second, supervised deep learning methods require large amounts of labeled data, yet obtaining reliable wear labels is time-consuming and expensive in real-world production environments. Third, while unsupervised learning has shown promise in TCM, most existing unsupervised models are designed for single-modal data and are not well-suited for capturing cross-modal correlations. These limitations motivate us to develop an unsupervised classification framework that integrates multi-modal features at the feature level and learns a compact, wear-sensitive latent representation via autoencoder-based reconstruction.

To address these challenges, this paper proposes an unsupervised tool wear state classification method based on multi-modal feature fusion. The innovation of the proposed method lies in the following:

1. Handcrafted statistical features from vibration signals and deep visual features from surface textures are concatenated prior to dimensionality reduction. This preserves modality-specific discriminative information and enables effective integration of heterogeneous sensor data.
2. The autoencoder is trained solely to minimize reconstruction error. It produces wear-sensitive embeddings that capture the underlying structure of multi-modal features.
3. The learned latent space is directly compatible with off-the-shelf clustering algorithms. This allows automatic tool wear state discrimination without supervised fine-tuning or additional classifier training.

This is the first systematic investigation of feature-level multi-modal fusion combined with unsupervised autoencoder-based clustering for TCM. By eliminating manual annotation and leveraging complementary sensor data, the proposed method offers a practical solution for intelligent wear state identification in machining environments.

2. Unsupervised Learning Model Based on Multi-modal Fusion

In TCM, both vibration signals and workpiece surface texture data exhibit clear evolutionary patterns as tool wear progresses. Vibration signals provide continuous, high-frequency dynamic information in the time domain, while surface texture data offer discrete yet reflective insights into the spatial distribution characteristics of the machined surface. These two modalities capture wear-related information from temporal dynamic and spatial morphological dimensions respectively, establishing a complementary relationship. Their integration enables a more comprehensive feature set for precise identification of tool wear state.

2.1. Multi-Modal Data Feature Fusion

In the current era of big data, multi-modal feature fusion has gained increasingly widespread application. Wan et al. proposed a tool wear recognition method based on empirical wavelet transform (EWT) and a Kolmogorov–Arnold Network (KAN), which integrates multi-sensor signals (cutting force, acceleration, and acoustic emission) through multi-scale convolutional networks and attention mechanisms, achieving recognition accuracy exceeding 99.9% [18]. However, challenges persist in current research: most fusion approaches rely on ensemble models or simple concatenation strategies, which do not adequately account for the heterogeneity of multi-modal data in milling. There remains considerable room for improvement in both fusion accuracy and efficiency [19]. The core objective of this approach is to integrate feature information from different modalities,

thereby uncovering richer and more comprehensive feature representations that ultimately enhance task performance in terms of both accuracy and robustness. Multi-modal fusion strategies are generally categorized into three primary approaches: data-level fusion, feature-level fusion, and decision-level fusion.

Data-level fusion operates at the raw data layer by directly integrating initial data from different modalities [20]. While this approach preserves complete original feature information, it also introduces challenges such as excessive data volume and sensitivity to noise [21]. Therefore, such methods are more suitable for specific scenarios where datasets are limited in scale and strong inter-modal correlations exist. Typical algorithms in this category primarily include classical machine learning methods based on estimation theory, statistical inference, information theory, and related principles.

Feature-level fusion operates by first extracting feature vectors from each modality independently and subsequently integrating them [22]. This approach reduces dimensionality while preserving salient characteristics, thereby enhancing the performance of downstream learning tasks [23]. However, it faces several key challenges, including feature alignment across modalities, processing of heterogeneous data structures, and the design of effective fusion strategies. Therefore, this method is particularly suitable for applications that require leveraging the prominent features of each modality while managing computational complexity.

2.2. Feasibility Analysis of Feature Fusion

During the milling process, vibration signals and surface texture features exhibit complementary characteristics that represent tool wear states, thereby providing comprehensive information for monitoring and analysis [24]. To validate this complementarity, we employ t-distributed stochastic neighbor embedding (t-SNE) for high-dimensional feature visualization [25]. The axes in Figure 1 represent the relative positions of data points in the low-dimensional embedding space produced by t-SNE. This technique computes probability-based similarities between high-dimensional samples and projects them into a low-dimensional space while preserving these similarity relationships. Figure 1 presents the t-SNE visualizations of feature distributions from three different sources, namely the original data, vibration-based features, and surface texture-based features, along with their corresponding dimensionality reduction results.

As shown in Figure 1a, the sample points of the original data exhibit substantial overlap in the feature space, indicating that raw data without feature extraction provides poor separability. In Figure 1b, the vibration-based model correctly classifies most samples, but a certain proportion remains misclassified, suggesting that while vibration features capture significant wear-related information, they are insufficient for complete discrimination. Figure 1c shows that the surface texture model also exhibits recognition capability; however, its performance is affected by dynamic variations in the machining process and machine tool stiffness. The scatter plot in Figure 1c reveals that texture features are sensitive to external disturbances, which limits their standalone effectiveness. Notably, when the misclassified samples from the surface texture model in Figure 1c are reclassified using the vibration signal model (Figure 1d), most of them are correctly identified, demonstrating that vibration features can compensate for the weaknesses of texture features. Similarly, the misclassified samples from the vibration signal model in Figure 1b are largely correctly identified when reclassified through the texture model (Figure 1e). This reciprocal correction confirms that the two modalities provide complementary information.

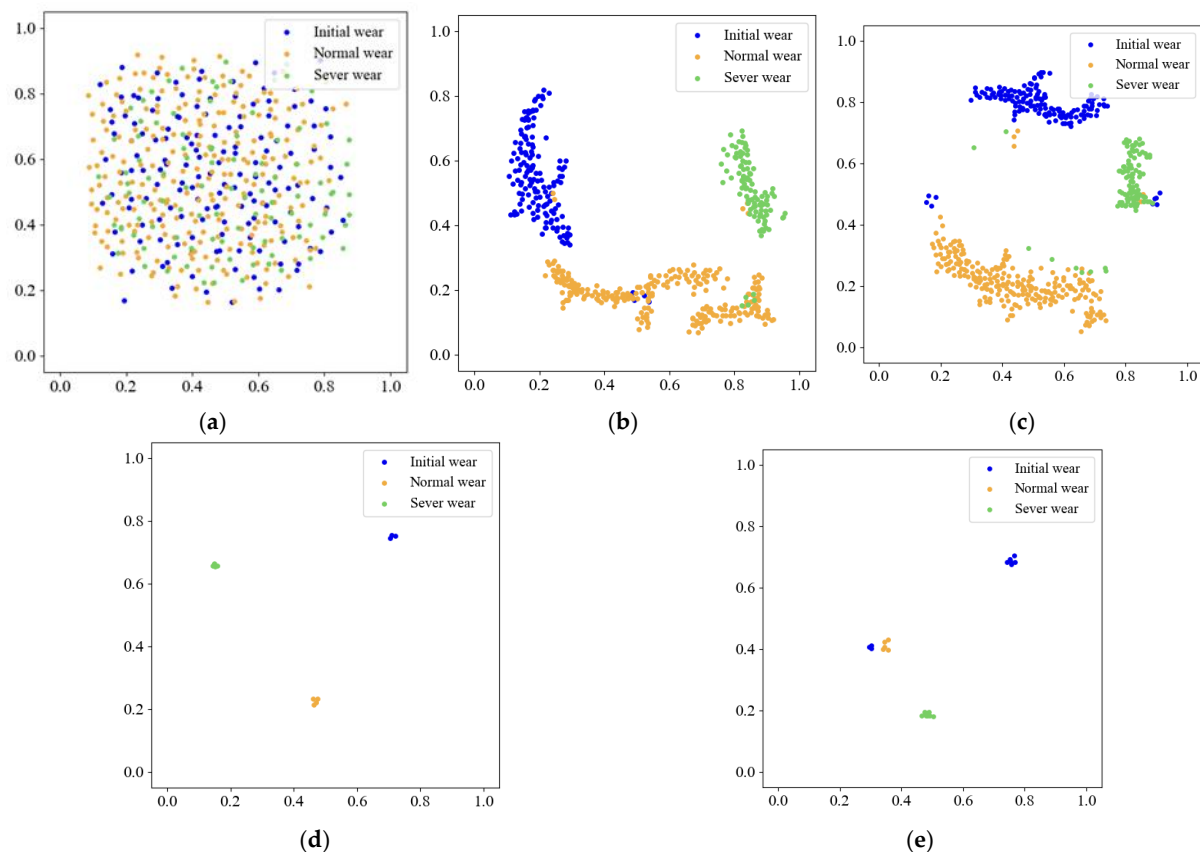


Figure 1. Feasibility analysis of feature fusion ((a) Input scatter plot, (b) signal output scatter plot, (c) image output scatter plot, (d) image supplementary scatter, (e) signal supplementary scatter).

These experimental results confirm that the fusion of vibration signals and surface texture features effectively leverages the complementary information between the two modalities, corrects erroneous classifications made by individual models, and substantially enhances the overall accuracy of tool wear state identification.

3. Tool Condition Monitoring Methodology

3.1. Feature Extraction

In TCM, when using vibration signals as input, it is essential to extract statistical features from the original signal to achieve dimensionality reduction, eliminate redundancy, and preserve critical wear-related information. This study extracts vibration signal features by computing statistical measures—including the mean, standard deviation, maximum, minimum, root mean square, and waveform factor—to convert the time-domain waveform into a low-dimensional feature vector. This process effectively captures the trend, fluctuation, and extreme value characteristics associated with tool state changes. The resulting feature vector has a size of $b \times 18$, and the corresponding calculation formulas are provided in Equations (1)–(7) [7]. These basic quantities are used to compute the statistical features listed above, such as the mean, standard deviation, root mean square, and waveform factor, where b denotes the batch size, the number of samples processed simultaneously.

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i \quad (1)$$

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2} \quad (2)$$

$$x_{\max} = \max \{x_1, x_2, \dots, x_N\} \quad (3)$$

$$x_{\min} = \min \{x_1, x_2, \dots, x_N\} \quad (4)$$

$$RMS = \sqrt{\frac{1}{N} \sum_{i=1}^N x_i^2} \quad (5)$$

$$SF = \frac{RMS}{\frac{1}{N} \sum_{i=1}^N |x_i|} \quad (6)$$

$$F_v = [\mu_1, \sigma_1, x_{\max,1}, x_{\min,1}, \dots, \mu_k, \sigma_k, x_{\max,k}, x_{\min,k}] \in R^{b \times 18} \quad (7)$$

where N is the total number of data points within the signal segment; x_i is the amplitude of the i point in the signal; μ is the mean value of the signal segment; σ is the mean standard deviation, measuring the spread of the signal; RMS is the mean root mean square, representing the signal's effective energy; SF is the mean shape factor, defined as the ratio of RMS to the mean absolute value; and F_v is the mean the vibration feature vector.

For the extraction of workpiece surface-texture features, the proposed method employs ResNet18. This architecture follows the standard ResNet-18 implementation in PyTorch's torchvision library (PyTorch 2.3.0), consisting of four residual blocks cascaded together—each containing two convolutional layers with skip connections. The number of channels increases with network depth (64, 128, 256, 512), enabling hierarchical feature abstraction. The network retains this design to leverage its hierarchical feature abstraction capability and to ensure reproducibility. Finally, the spatial dimension is compressed by the global average pooling layer, and the $b \times 512$ -dimensional feature tensor representing the texture information is the output. The dimension transformation is shown in Equation (8).

$$F_i = \text{ResNet-18}(I) \in R^{b \times 512} \quad (8)$$

The exact implementation details of the ResNet-18 feature extractor used in our code are as follows. The network is initialized with weights from torchvision models. Features are extracted from the layer immediately preceding the global average pooling layer, yielding a 512-dimensional vector per image. The entire backbone was fine-tuned end-to-end during training. This decision was made because the pre-trained weights provide a strong starting point, and with a low learning rate the model adapts to the specific texture characteristics without overfitting. All input images are resized to 224×224 pixels, scaled to $[0,1]$, and normalized channel-wise using the ImageNet mean and standard deviation.

To achieve effective integration of multi-modal data features, this study employs a feature-level fusion strategy for combining vibration signals and surface texture characteristics. Specifically, the vibration feature vector ($b \times 18$) and texture feature tensor ($b \times 512$) are concatenated along the feature dimension to form a fused feature matrix ($b \times 530$),

with the mathematical formulation provided in Equation (9). Notably, the feature extractor parameters remain frozen during this fusion process, and the resulting fused features are subsequently fed into the autoencoder module as pretrained inputs. The complete framework is illustrated in Figure 2.

$$F_f = \text{Concat}(F_v, F_i) \in R^{b \times 530} \quad (9)$$

where F_f denotes the fused feature matrix with dimension $b \times 530$. This matrix serves as the input to the subsequent autoencoder module.

Figure 2 illustrates the two-stage workflow of the proposed method: the training phase and the inference phase.

Training phase: The fused multi-modal feature matrix F_f is fed into an encoder to obtain a low-dimensional latent representation Z ; a decoder reconstructs Z to F_r . F_r represents the reconstructed feature matrix. The model is optimized by minimizing the reconstruction loss, which in this work is the mean squared error (MSE).

Inference phase: The trained encoder extracts latent features from new samples. These features are clustered using K-means (with the number of clusters set to 3, corresponding to the three wear states). The Hungarian algorithm is then applied to optimally map the cluster labels to the actual tool wear states, enabling unsupervised classification.

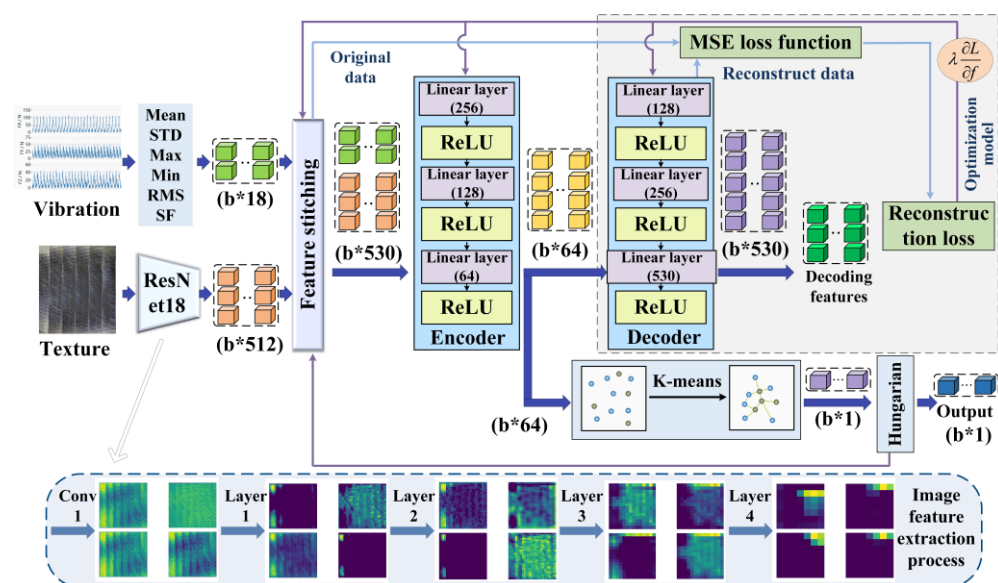


Figure 2. Unsupervised learning module.

The architecture of the encoder and decoder components is detailed in Figure 3. The encoder processing flow operates as follows: First, the concatenated $b \times 530$ -dimensional feature vector is fed into the initial linear layer, which projects the input dimension from 530 to 256 through linear transformation using a weight matrix and bias vector, as formulated in Equation (10) [15]. This is followed by nonlinear processing through a ReLU activation function, which enhances the model's nonlinear representation capability by setting negative values to zero while preserving positive inputs, as specified in Equation (11) [15].

The subsequent linear layer then reduces the feature dimension from 256 to 128, again followed by ReLU activation. Finally, a third linear layer further compresses the dimension to 64, a latent dimension chosen as a trade-off between compression rate and information retention. Preliminary experiments showed that reducing the latent dimen-

sion to 32 led to a noticeable drop in classification accuracy, while increasing it to 128 yielded no significant improvement but increased training cost. This trade-off is consistent with common practice in autoencoder-based monitoring tasks. Through this sequential processing, the encoder ultimately generates $b \times 64$ -dimensional low-dimensional features, effectively achieving compression and essential information extraction from the original high-dimensional input.

$$y = \omega x + b \quad (10)$$

$$f(x) = \max(0, x) \quad (11)$$

In the formula, ω is the weight matrix; x is the input; and b is the bias vector.

The decoder's primary role is to reconstruct the original $b \times 530$ -dimensional input from the $b \times 64$ low-dimensional features generated by the encoder. This reverse mapping is achieved through a series of linear layers and ReLU activation functions that progressively expand the feature dimension from 64 back to 530. By leveraging the compact representations learned in the low-dimensional latent space, this reconstruction process effectively restores the original feature dimensionality while preserving essential information.

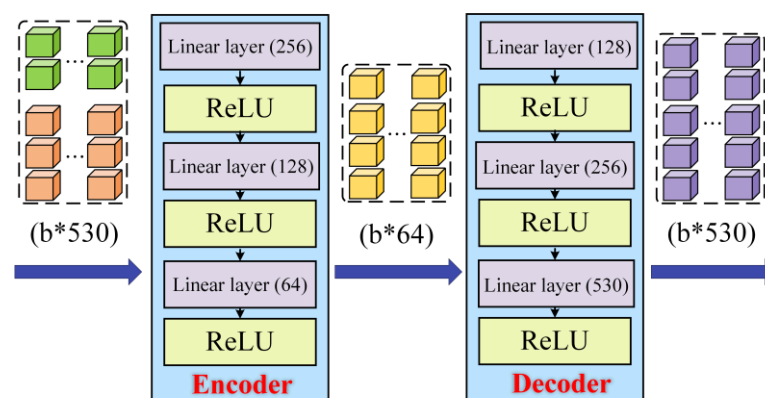


Figure 3. Autoencoder module.

To evaluate the model's capability to reconstruct input data, the mean square error (MSE) is adopted as the reconstruction loss function. This function measures the discrepancy between the original $b \times 530$ -dimensional feature vector derived during feature extraction and the reconstructed feature vector generated by the decoder. A lower MSE value indicates a smaller overall discrepancy between the two vectors, meaning the reconstructed output more closely approximates the original input and reflects a stronger model fitting capacity. The formula for computing the reconstruction loss is given in Equation (12) [17].

$$L_r = \frac{1}{n} \sum_{i=1}^n \left(x_i - \hat{x}_i \right)^2 \quad (12)$$

In the formula, x_i represents the result of feature extraction; $\hat{x}_i$ represents the output of the decoder; and n represents the feature dimension.

In addition, L_r denotes the reconstruction loss value; X is the original fused feature vector (dimension $b \times 530$) extracted from the multi-modal data; $\hat{X}$ is the reconstructed feature vector produced by the decoder; and d is the total feature dimension

($d = 530$). A lower L_r indicates that the reconstructed output more closely approximates the original input, reflecting a stronger model fitting capacity.

3.2. Training and Inference Stages

The proposed unsupervised learning framework comprises two distinct phases, namely an autoencoder training stage and a clustering-based inference stage, as illustrated in Figure 2. During the training phase, the multi-modal feature matrix $F_f \in R^{b \times 530}$ is fed into an encoder to obtain a low-dimensional latent representation $Z \in R^{b \times 64}$, which is then reconstructed by a decoder into $\hat{F}_f$. The training objective is to minimize the reconstruction loss L_r via the Adam optimizer, as expressed in Equation (13) [17]. Through this process, network parameters are updated iteratively to learn discriminative feature representations.

$$L_r = \frac{1}{n} \sum_{i=1}^n (F_{f,i} - \hat{F}_{f,i})^2 \quad (13)$$

In the inference phase, the trained encoder is frozen and used to extract latent features from new samples. These features are subsequently grouped using the K-means clustering algorithm with the cluster number set to three, corresponding to the predefined wear states. Finally, the Hungarian algorithm is applied to map the resulting cluster labels to the actual tool wear states, enabling classification without labeled data. This two-stage design ensures effective feature learning while providing a clear workflow for practical deployment.

In unsupervised learning, the inherent advantage of the autoencoder lies in its independence from labeled data. Through its encoder–decoder architecture, the model learns the essential structure of input data and generates efficient nonlinear feature representations. The training objective is to minimize the loss function between the input data and reconstructed output, while the network parameters are updated using iterative optimization algorithms. This approach enables the autoencoder to capture core data characteristics and achieve high-fidelity reconstruction, thereby significantly enhancing its capability in feature learning and representation.

4. Results and Discussion

4.1. Experimental Setup

The milling experimental platform and data acquisition setup are illustrated in Figure 4. A 508-III steel workpiece measuring $200 \times 200 \times 100$ mm was processed by dry end milling using a five-flute cemented carbide face-milling cutter with a diameter of 40 mm. The tool wear experiment was conducted on a three-axis CNC machining center (VDL-1000E).

The milling parameters employed in the machining experiments are summarized in Table 1. To acquire the necessary monitoring data, vibration signal acquisition and surface texture imaging were synchronized with the machining process. The workpiece was securely mounted on the machine worktable, with its datum plane accurately aligned parallel to the machine coordinate system through precise adjustment. A dial indicator, used in conjunction with the worktable rotation function, enabled precise tool alignment.

Vibration signals along the X, Y, and Z axes were measured using a Donghua DH5981 dynamic signal acquisition system. The system provides 8 synchronous channels with 24-bit A/D resolution, a dynamic range better than 110 dB, and an input voltage range of ± 10

V. All three channels were sampled continuously at a fixed rate throughout the entire cutting process. All acquired signals were conditioned through a charge amplifier and digitized by the acquisition hardware. The sampling frequency was set to 10 kHz for all channels, and raw data were stored in CSV format for subsequent analysis.

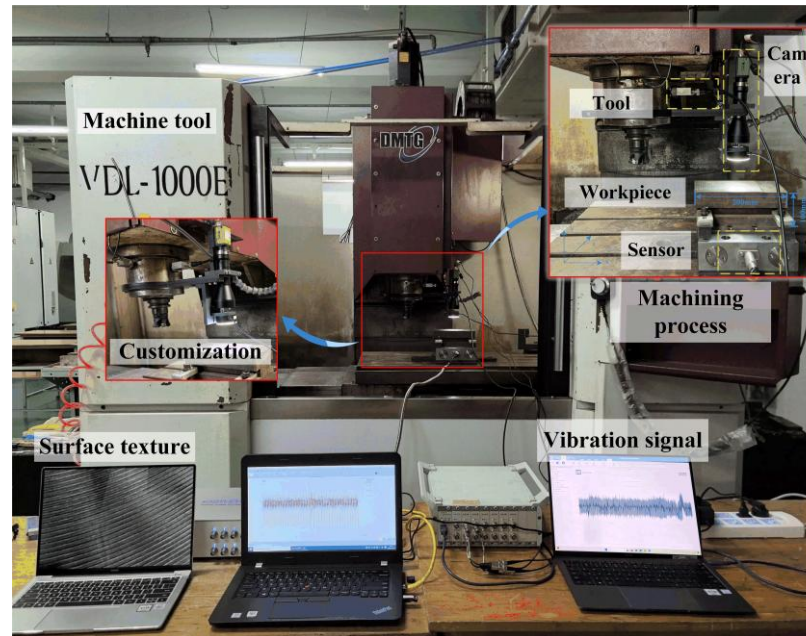


Figure 4. Experimental equipment.

Surface texture images were captured with a MARS-1230-23U3M/C industrial camera (resolution 4096×3000 pixels, pixel size $3.45 \mu\text{m}$) equipped with a CA-LMHE0510 lens. It is important to note that image acquisition was not performed during cutting. Instead, after each cutting pass, when the tool entered the retraction phase, the worktable was brought to a complete stop and image triggering occurred. This dwell-time strategy eliminates motion blur and ensures image quality comparable to offline measurements. The acquisition parameters were as follows: The exposure time was 20 ms. The image acquisition was triggered automatically by the CNC system's "cutting pass completed" signal, ensuring consistent timing and eliminating manual intervention. Since the workpiece is completely stationary during image capture, the machining feed rate does not affect image quality (the feed rate used in this experiment was 0.1 mm/r , as listed in Table 1). Under the current cutting parameters (workpiece length 200 mm, feed rate 0.1 mm/r , spindle speed 800 r/min), each cutting pass takes approximately 25 s. The image acquisition cycle—including worktable deceleration, stopping, image capture, and re-acceleration—requires approximately 2 s. This confirms that the dwell-time strategy is practical and does not significantly affect machining efficiency.

Two independent imaging systems were used in this study: the surface imaging system described above (mounted near the spindle) to capture the machined surface, and a separate ultra-depth-of-field microscope (VHX-1000) to measure the flank wear width VB offline, which served as the ground-truth for tool wear state labeling. The two systems are functionally distinct and do not overlap. To enhance data collection efficiency and ensure reliability, a dedicated fixture system was designed. The industrial camera, optical lens, and light source were integrated into the machine tool spindle, establishing a stable spatial relationship with both the tool and workpiece, as illustrated in Figure 4. During the machining process, image acquisition of the machined surface texture was triggered automatically when the tool completed each cutting pass and entered the retraction phase, following preset sampling interval parameters.

Table 1. Experimental cutting parameters.

Category	Machining Parameter
Rotational Speed	800 r/min
Cutting Speed	108.17 m/min
Feed Rate	0.1 mm/r
Depth of Cut	0.3 mm
Width of Cut	15 mm

To establish accurate correspondence between vibration signals and surface images, the continuously acquired vibration data were segmented into windows of 3000 sampling points each (equivalent to 300 ms). The start time of each signal window was synchronized with the timestamp of the most recent surface image acquisition. This synchronization strategy ensures that every surface image frame is paired with a precisely aligned 300 ms signal segment. Since tool wear evolves gradually over time, the impact of minor temporal misalignments—within 50 ms—on wear trend prediction remains negligible.

The proposed method was implemented using the PyTorch framework with Python as the programming language and an RTX 3090 GPU as the hardware platform. The model was optimized with the Adam optimizer, using an initial learning rate of 0.008. All surface texture images were resized to 224×224 pixels, with the batch size set to 64 and training conducted for 300 epochs. To ensure fair comparison, identical training parameters were maintained across all models during the experimental process.

4.2. Labeling and Evaluation Metrics

Following each cutting cycle, tool wear images were captured using an industrial camera (MARS-1230-23U3M/C) equipped with a CA-LMHE0510 lens. The flank wear width (VB) of the identical cutting edge was subsequently measured using an ultra-depth-of-field microscope.

The three wear stages were defined based on the measured flank wear width (VB) progression. After each cutting pass, VB was measured using an ultra-depth-of-field microscope (VHX-1000). The measured values were plotted against the number of cutting passes, and the slope changes in the wear curve were examined. Two characteristic inflection points were identified from the experimental data. The first inflection point, where the wear rate decreases from rapid to gradual, occurs at approximately $VB = 80 \mu\text{m}$; the second inflection point, where the wear rate begins to accelerate sharply, occurs at approximately $VB = 120 \mu\text{m}$. Based on these observed inflection points, the tool wear process was divided into three states: initial wear ($VB \leq 80 \mu\text{m}$), normal wear ($80 \mu\text{m} < VB \leq 120 \mu\text{m}$), and severe wear ($VB > 120 \mu\text{m}$). This classification reflects the actual wear progression observed in our experiments. This classification scheme, illustrated in Figure 5, ensures objective and reproducible state labeling. For validation, ground-truth labels from an independent dataset were mapped to cluster assignments using the Hungarian algorithm to enable quantitative evaluation without supervised training.

In tool wear state classification, model performance must be comprehensively assessed using multiple evaluation metrics. In this study, ground-truth tool wear labels from a dataset independent of the training set were used to externally validate the clustering results obtained through unsupervised learning. The Hungarian algorithm was applied to establish an optimal mapping between the cluster assignments and true labels. Standard classification metrics were then employed to objectively evaluate the effectiveness of the unsupervised method in identifying tool wear states without using labeled data during training. The confusion matrix provides a visual representation of classification per-

formance by comparing the mapped predicted labels against the actual labels across all wear state categories.

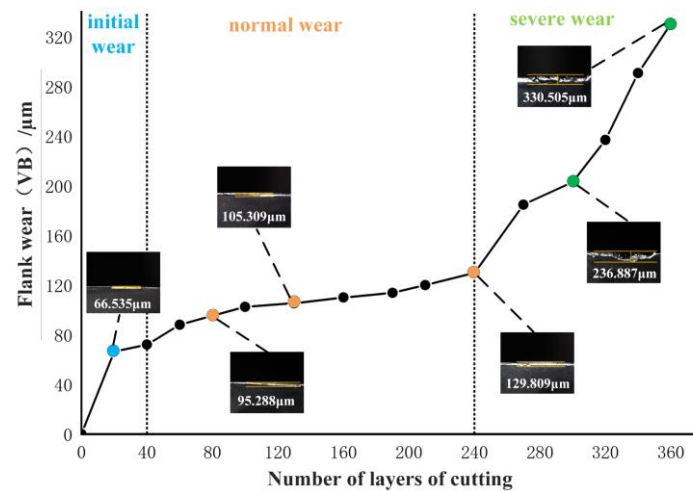


Figure 5. Tool wear state division.

The performance of the proposed method is evaluated using the following four metrics [13]. Accuracy (Acc) reflects the proportion of correctly classified samples out of the total number of samples. Precision (Pre) denotes the ratio of correctly predicted positive samples to all samples predicted as positive for a specific wear state, reflecting the reliability of the model's positive predictions. Recall (ReC) measures the proportion of correctly identified positive samples out of all actual positive samples for a particular wear state, indicating the model's ability to detect true positive instances. The F₁-score (F₁) represents the harmonic mean of precision and recall, serving as a balanced metric. An ideal classifier aims to achieve an F₁-score of 1.

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \quad (14)$$

$$Pre = \frac{TP}{TP + FP} \quad (15)$$

$$ReC = \frac{TP}{TP + FN} \quad (16)$$

$$F_1 = 2 \cdot \frac{Pre \times ReC}{Pre + ReC} \quad (17)$$

In the formula, TP refers to the number of positive cases correctly identified; FP refers to the number of positive examples misidentified; FN refers to the number of positive samples that were incorrectly classified as negative; and TN refers to the number of negative samples that were correctly classified as negative.

4.3. Experimental Result Analysis

Upon completion of model training, the test set was utilized to evaluate the performance of tool wear state classification under identical experimental conditions. Focusing on the task of tool wear state classification in machining applications, this paper systematically compares the classification performance of four unsupervised learning methods across different wear states. The evaluated methods include the conventional Gaussian

Mixture Model (GMM), hierarchical clustering, K-means clustering, and the autoencoder-based approach proposed in this work.

Table 2 compares the classification performance of different unsupervised methods. Among conventional algorithms, GMM achieves the lowest accuracy (84.72%). Under the autoencoder framework, vibration-only attains 91.00% accuracy, outperforming texture-only (78.50%). The proposed multi-modal fusion method achieves the highest accuracy (93.33%), which is computed as the ratio of correctly classified samples to the total number of test samples. Based on the confusion matrix in Figure 6d, the total number of correctly classified samples across the three wear states is 336 out of 360 test samples, giving an accuracy of 93.33%. This result confirms that feature fusion significantly improves wear state classification.

Table 2. Unsupervised learning performance of different models.

Model	Acc	Pre	ReC	F ₁
GMM	84.72% ± 2.1	82.80% ± 2.3	85.04 ± 1.9	83.88% ± 2.0
HC	88.33% ± 1.8	85.46% ± 2.0	88.61 ± 1.7	86.98% ± 1.8
K-Means	85.00% ± 2.4	81.69% ± 2.5	84.89 ± 2.2	83.23% ± 2.3
Autoencoder (vibration)	91.00% ± 1.5	89.50% ± 1.6	91.00% ± 1.4	90.24% ± 1.5
Autoencoder (texture)	78.50% ± 2.2	76.50% ± 2.4	79.00% ± 2.1	77.73% ± 2.2
Autoencoder (fusion)	93.33% ± 1.3	92.16% ± 1.4	93.63 ± 1.2	92.84% ± 1.3

As reflected in the confusion matrix in Figure 6a, significant misclassification occurs between the initial wear and normal wear states. This limitation arises from GMM's fundamental assumption of Gaussian-distributed data, which constrains its capability to model complex nonlinear relationships in high-dimensional feature space. Given the multifaceted nature of tool wear characteristics during cutting processes, which are poorly represented by simple Gaussian distributions, the model fails to accurately capture the actual data distribution, consequently leading to increased classification errors.

The hierarchical clustering algorithm achieved an accuracy of 88.33%, demonstrating moderate overall performance. As shown in the confusion matrix of Figure 6b, significant misclassification is observed across wear states. This limitation stems from the algorithm's high sensitivity to distance metrics between samples. Given that boundaries between different tool wear states are often blurred by noise interference and feature overlap, the resulting distortions in similarity measurements lead to reduced clustering accuracy.

The K-Means clustering algorithm achieved an accuracy of 85.00%. Analysis of Figure 6c reveals substantial overlap between the initial wear and normal wear states, with a misclassification rate of approximately 15%. Specifically, 18 initial wear samples were erroneously classified as normal wear, while 12 normal wear samples were misidentified as initial wear. This limitation primarily stems from the algorithm's dependence on random initialization of cluster centroids, coupled with the inherent complexity of tool wear data distribution that contradicts the algorithm's underlying assumptions, consequently compromising classification accuracy.

The method proposed in this chapter achieves a notably higher accuracy of 93.33% under the same experimental conditions, significantly outperforming the other methods. As observed in Figure 6d, misclassification is substantially reduced across all wear states. Specifically, 145 samples of initial wear and 136 samples of normal wear are correctly classified, demonstrating the efficacy of the multi-modal fusion autoencoder approach.

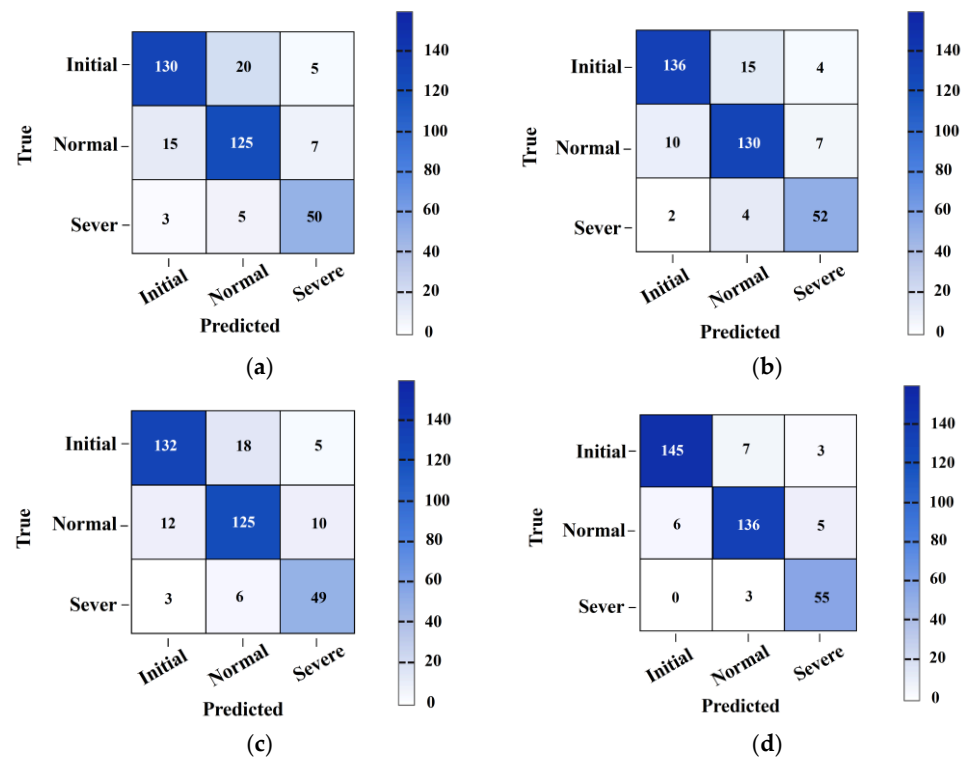


Figure 6. Confusion matrix of different algorithms ((a) GMM, (b) HC, (c) K-Means, (d) autoencoder).

The histogram in Figure 7 presents the algorithm performance comparison. The error bars, derived from the standard deviation across multiple independent experiments, indicate the stability of the algorithms. The autoencoder architecture developed in this study effectively achieves cross-modal feature fusion between vibration signals and surface texture images. The encoder compresses the multi-modal input data into a low-dimensional feature representation, while the decoder reconstructs the original data while preserving critical information. This approach allows the model to capture more discriminative tool wear characteristics, thereby significantly improving the accuracy of wear state classification.

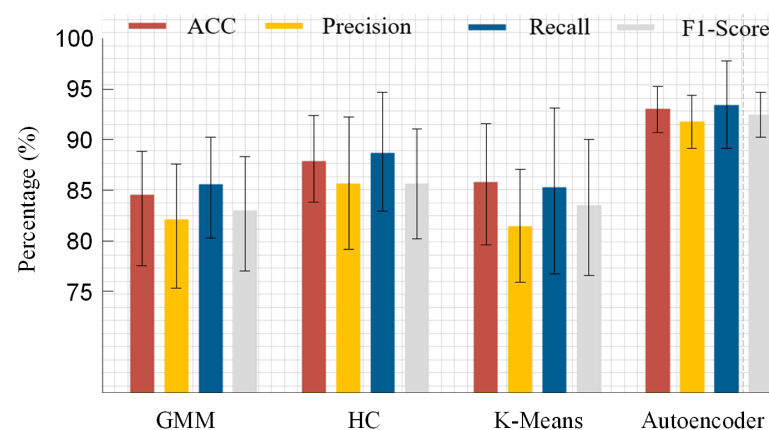


Figure 7. Histograms of different algorithms (with error bars).

The error bars in Figure 7 represent the standard deviation of the classification accuracy across five independent experimental runs. Each run was conducted with a distinct random seed (42, 123, 2026, 1997, 101), governing the weight initialization, the stratified 8:2 data split, and the initial centroid selection in K-means clustering. The observed variability primarily stems from the stochastic nature of model initialization, the optimization

process, and the clustering procedure, collectively reflecting the stability of the model under different initialization conditions.

Figure 8 shows the loss curves during training. The loss values plotted are dimensionless, as they represent the reconstruction error (mean squared error between the input and reconstructed features) computed entirely within the feature space. The blue solid line represents the loss on the train set, and the orange dashed line represents the loss on the test set. Both curves drop rapidly in the first 50 epochs and then converge, staying close to each other. This indicates that the model does not overfit and has good generalization ability.

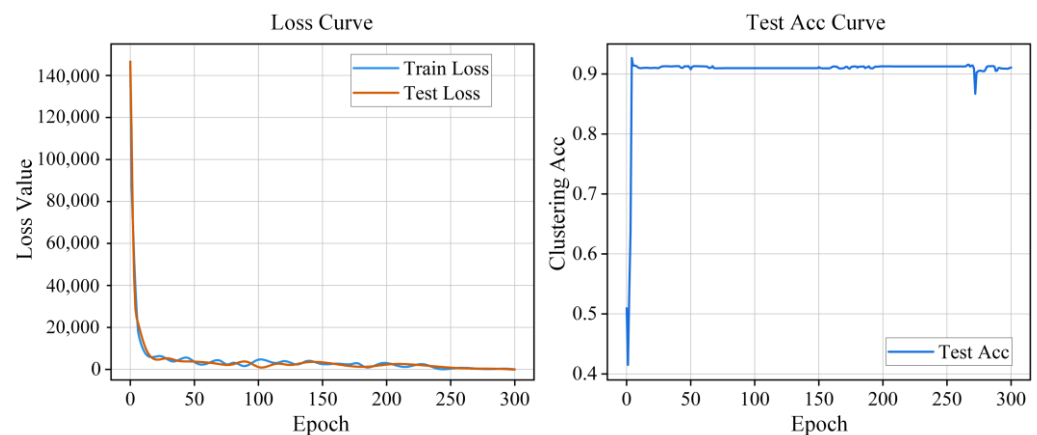


Figure 8. Training and test loss curves. The vertical axis represents the loss value (dimensionless scalar in feature space), and the horizontal axis represents the training epoch.

5. Discussion

Although the proposed method achieves promising classification performance under controlled laboratory conditions, several practical challenges must be considered before it can be reliably deployed in real industrial manufacturing environments. This section provides a candid assessment of these limitations and discusses potential countermeasures, thereby offering practical guidance for future engineering implementations.

1. Coolant and chip interference:

The experiments in this study were conducted under dry cutting conditions to avoid the confounding effects of coolant on vibration signals and image quality. However, in actual production environments, coolant is widely used for heat dissipation, lubrication, and chip removal. Coolant may adhere to the camera lens or workpiece surface, causing image blur, refraction, or reflection artifacts that degrade the quality of texture feature extraction. Additionally, metal chips generated during cutting may obstruct the field of view or scratch the machined surface, compromising image consistency and introducing outliers in the dataset.

To mitigate these issues in practical deployment, we recommend the following countermeasures: (1) installing an air knife or compressed air blowing device directed at the camera lens to prevent coolant droplet accumulation; (2) adding a short compressed air cleaning step before each image acquisition to remove chips and coolant residues from the workpiece surface; and (3) using oil-resistant protective windows on the camera lens with regular cleaning cycles. These measures are already widely adopted in industrial machine vision systems for machining monitoring (e.g., in-process part inspection and tool wear detection) and have been proven effective in maintaining image quality under wet cutting conditions.

2. Sampling frequency:

The proposed method adopts a “once-per-pass” dwell-time acquisition strategy, this overhead is acceptable for medium- to low-speed finishing operations, which are the primary application scenarios for high-precision tool wear monitoring.

However, for high-speed machining applications where a single cutting pass may be completed in a few seconds (e.g., spindle speed $> 10,000$ r/min, feed rate > 1 mm/r), the 2 s acquisition overhead would become a significant bottleneck, reducing production efficiency by 30% or more. In such cases, the current dwell-time strategy would be impractical. To address this limitation, future work will investigate continuous-motion image acquisition with short-exposure-time strobe lighting (e.g., < 1 ms exposure) and global-shutter cameras to eliminate motion blur without stopping the worktable. Alternatively, synchronizing image capture with the machine’s position encoder could enable acquisition during rapid traversing movements, further reducing the time penalty. For the present study, however, the dwell-time strategy was deliberately chosen to prioritize image quality and measurement repeatability, which are essential for method validation.

3. Illumination stability:

Industrial lighting conditions are often subject to variations caused by ambient light changes, lamp aging, and shadows from machine tool components or operators. Inconsistent illumination can lead to significant variations in surface texture image intensity and contrast, thereby affecting the consistency of features extracted by the ResNet-18 backbone. Although we applied ImageNet-based normalization to input images, this does not fully compensate for uneven local illumination or shadowing effects.

To enhance robustness in real-world deployment, we used a fixed-intensity ring light; however, industrial deployment would benefit from a feedback-controlled lighting system that adjusts intensity based on image histogram analysis. Additionally, periodic reference—white calibration and the use of optical diffusers to homogenize light distribution—can further reduce illumination-related variability. These considerations are particularly important for long-term deployment in production workshops.

4. Installation and calibration:

The proposed method relies on precise geometric calibration between the camera, the tool, and the workpiece. The camera must maintain a fixed pose relative to the machined surface to ensure consistent texture feature extraction. In practical deployment, vibrations from adjacent machining operations, thermal expansion of machine tool components, or accidental tool collisions may disturb the camera’s mounting position, invalidating the calibration parameters and introducing systematic errors in feature extraction.

To ensure reliable online monitoring in industrial settings, we suggest: (1) designing a dedicated rigid fixture with vibration-damping mounts for the camera system; (2) implementing automatic periodic verification using a known reference pattern (e.g., a checkerboard) to detect pose deviations; and (3) incorporating a recalibration routine that can be executed automatically during tool change cycles, minimizing manual intervention. In our laboratory setup, the camera was manually calibrated before the experiment and its pose was verified after each cutting pass. For industrial deployment, an automated calibration procedure is essential to maintain measurement consistency over extended production runs.

5. Universality:

Although this study focuses specifically on face milling cutters with cemented carbide inserts, the proposed framework is conceptually transferable to other tool types (e.g., end mills, turning tools, drilling tools) with minor adjustments to the sensor arrangement and feature extraction components. The key requirement is that both vibration signals and surface texture images can be acquired synchronously, which is feasible for most machin-

ing processes where the tool or workpiece can be accessed for imaging during retraction phases.

6. Conclusions

To address the challenges of inadequate data integration and incomplete feature extraction in TCM, this paper introduces an unsupervised tool wear state classification approach based on multi-modal feature fusion. The proposed method extracts temporal features from vibration signals and spatial features from surface texture images, subsequently integrating the deep representations of both modalities. Furthermore, an autoencoder architecture is employed to train the model on unlabeled data, where minimization of reconstruction loss enables the learning of highly correlated features, ultimately achieving accurate identification of wear states. Experimental results from cemented carbide milling cutter wear tests demonstrate the method's effectiveness, with the main conclusions summarized as follows:

1. The proposed framework effectively integrates statistical features from vibration signals with visual features extracted by ResNet18, implementing feature compression and reconstruction through an autoencoder architecture. By optimizing the model through reconstruction loss minimization, the system learns discriminative feature representations that significantly enhance tool wear characterization.
2. t-SNE visualization demonstrates complementary relationships between misclassified samples from different modalities in the feature space. This complementarity confirms that multi-modal fusion substantially enhances the accuracy of wear state identification.
3. Evaluation based on confusion matrices and multiple performance metrics shows that the proposed method achieves a classification accuracy of 93.33% under unsupervised learning conditions, compared to 84.72% for GMM, 88.33% for HC, and 85.00% for K-Means. The F₁-score reaches 92.84%, further confirming the robustness of the proposed approach.
4. The experimental results further reveal that the multi-modal fusion approach substantially reduces misclassification rates across all wear stages compared to single-modal or conventional clustering methods. This observation indicates that the fusion of vibration and texture features provides complementary information that enhances the separability of different wear stages, supporting the effectiveness of the proposed unsupervised framework for tool condition monitoring.

Although the paper specifically addresses wear monitoring in carbide milling tools, the proposed framework can be readily adapted to other tool types. With minor modifications to the sensor arrangement and feature extraction components, the system can be efficiently deployed for cost-effective online wear monitoring across diverse machining centers. This adaptability underscores its significant potential for practical industrial deployment, process optimization, and predictive maintenance within intelligent manufacturing systems.

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

Funding: This research was funded by the Joint Fund of the National Natural Science Foundation of China for Enterprise Innovation and Development (U24B2060).

Institutional Review Board Statement: The content studied in this study belongs to the field of metal processing, and does not involve humans or animals. This study strictly follows the accepted principles of ethical and professional conduct.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: The authors thank DeepSeek for assistance in language polishing. AI tools were not used for data analysis, modeling, result interpretation, or conclusion formation. The authors assume full responsibility for the work.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported.

References

1. Cao, F.; Zhang, X.; Dai, J.; Wen, L.; Chang, T.; Mei, C.; Zhang, Z. A lightweight model for tool wear amount segmentation fusing the attention mechanism and U2Net network. *Measurement* **2026**, *280*, 121725. <https://doi.org/10.1016/j.measurement.2026.121725>.
2. Guo, Y.; Sun, Y.; Wu, K. Research and development of monitoring system and data acquisition of CNC machine tool in intelligent manufacturing. *Int. J. Adv. Robot. Syst.* **2020**, *17*, 12. <https://doi.org/10.1177/1729881419898017>.
3. Sun, W.; Zhang, D.; Luo, M. Machining process monitoring and application: A review. *J. Adv. Manuf. Sci. Technol.* **2021**, *1*, 2021001. <https://doi.org/10.51393/j.jamst.2021001>.
4. You, Z.; Meng, Y.; Li, D.; Zhang, Z.; Ren, M.; Zhang, X.; Zhu, L. Adaptive detection of tool-workpiece contact for nanoscale tool setting based on multi-scale decomposition of force signal. *Mech. Syst. Signal Process.* **2024**, *208*, 111000. <https://doi.org/10.1016/j.ymssp.2023.111000>.
5. Guo, R.; Zhang, H.; Zhi, X.; Wei, Y. Research on on-machine measurement of tool wear based on machine vision. *J. Lanzhou Univ. Technol.* **2024**, *50*, 33–41.
6. Zhang, K.; Yuan, H.; Nie, P. A method for tool condition monitoring based on sensor fusion. *J. Intell. Manuf.* **2015**, *26*, 1011–1026. <https://doi.org/10.1007/s10845-015-1112-y>.
7. Kou, R.; Lian, S.; Xie, N.; Lu, B.-E.; Liu, X.-M. Image-based tool condition monitoring based on convolution neural network in turning process. *Int. J. Adv. Manuf. Technol.* **2022**, *119*, 3279–3291. <https://doi.org/10.1007/s00170-021-08282-x>.
8. Hou, W.; Wang, J.; Wang, L.; Zhang, S. Novel tool wear prediction method based on multimodal information fusion and deep subdomain adaptation. *Mech. Syst. Signal Process.* **2025**, *224*, 112128. <https://doi.org/10.1016/j.ymssp.2024.112128>.
9. Yuan, J.; Liu, L.; Yang, Z.; Bo, J.; Zhang, Y. Tool wear condition monitoring by combining spindle motor current signal analysis and machined surface image processing. *Int. J. Adv. Manuf. Technol.* **2021**, *116*, 2697–2709. <https://doi.org/10.1007/s00170-021-07366-y>.
10. Liu, B.; Wu, W.; Zhao, D.; Wang, P.; Pei, C. Research on tool wear state prediction based on encoder-decoder structure. *Mini-Micro Syst.* **2025**, *46*, 1530–1536. <https://doi.org/10.20009/j.cnki.21-1106/TP.2025-0092>.
11. Yang, Y.; Guo, Y.; Huang, Z.; Chen, N.; Li, L.; Jiang, Y.; He, N. Research on the milling tool wear and life prediction by establishing an integrated predictive model. *Measurement* **2019**, *145*, 178–189. <https://doi.org/10.1016/j.measurement.2019.05.009>.
12. Wang, G.; Zhang, Y.; Liu, C.; Xie, Q.; Xu, Y. A new tool wear monitoring method based on multi-scale PCA. *J. Intell. Manuf.* **2019**, *30*, 113–122. <https://doi.org/10.1007/s10845-016-1235-9>.
13. Lei, W.; Yan, H.; Li, Q.; Li, Y.; Zheng, P. Unsupervised tool anomaly detection based on AGRU autoencoder. *Mach. Tool. Hydraul.* **2024**, *52*, 30–37.
14. Shao, X.; Zhao, Z.; Nie, X.; Zhang, Y. Unsupervised tool anomaly detection method based on spatiotemporal feature extraction. *Comput. Integr. Manuf. Syst.* **2025**, *31*, 877–887. <https://doi.org/10.13196/j.cims.2023.0517>.
15. Yan, S.; Shao, H.; Xiao, Y.; Liu, B.; Wan, J. Hybrid robust convolutional autoencoder for unsupervised anomaly detection of machine tools under noises. *Robot. Comput.-Integr. Manuf.* **2023**, *79*, 102441. <https://doi.org/10.1016/j.rcim.2022.102441>.
16. Jiang, Z.; Ma, Y. Fault classification method based on unsupervised transfer component analysis and support vector machines. *Comput. Integr. Manuf. Syst.* **2023**, *29*, 3066. <https://doi.org/10.13196/j.cims.2023.09.018>.
17. Yan, S.; Chen, Y. Unsupervised anomaly segmentation based on residual max-pooling autoencoder. *Comput. Integr. Manuf. Syst.* **2024**, *30*, 2069. <https://doi.org/10.13196/j.cims.2021.0843>.

18. Wan, M.; Cao, M.W.; Wang, S.; Zheng, X. Recognition of tool wear in milling using Kolmogorov-Arnold networks and multi-sensor feature fusion. *Chin. J. Mech. Eng.* **2026**, *39*, 100146. <https://doi.org/10.1016/j.cjme.2025.100146>.
19. Huang, Z.W.; Shao, J.J.; Guo, W.C.; Li, W.; Zhu, J.; He, Q.; Fang, D. Tool Wear Prediction Based on Multi-Information Fusion and Genetic Algorithm-Optimized Gaussian Process Regression in Milling. *IEEE Trans. Instrum. Meas.* **2023**, *72*, 2516716. <https://doi.org/10.1109/TIM.2023.3280531>.
20. Kong, W.L.; Li, H. Combining Adaptive Time-Series Feature Window and Stacked Bidirectional Lstm for Predicting Tool Remaining Useful Life without Failure Data. *Int. J. Adv. Manuf. Technol.* **2022**, *121*, 7509–7526. <https://doi.org/10.1007/s00170-022-09771-3>.
21. Hua, J.Q.; Li, Y.G.; Liu, C.Q.; Wan, P.; Liu, X. Physics-Informed Neural Networks with Weighted Losses by Uncertainty Evaluation for Accurate and Stable Prediction of Manufacturing Systems. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *35*, 11064–11076. <https://doi.org/10.1109/TNNLS.2023.3247163>.
22. An, H.; Wang, G.; Dong, Y.; Yang, K.; Sang, L. Tool Life Prediction Based on Gauss Importance Resampling Particle Filter. *Int. J. Adv. Manuf. Technol.* **2019**, *103*, 4627–4634. <https://doi.org/10.1007/s00170-019-03934-5>.
23. Hanachi, H.; Yu, W.N.; Kim, I.Y.; Liu, J.; Mechefske, C.K. Hybrid Data-Driven Physics-Based Model Fusion Framework for Tool Wear Prediction. *Int. J. Adv. Manuf. Technol.* **2019**, *101*, 2861–2872. <https://doi.org/10.1007/s00170-018-3157-5>.
24. Herrera-Granados, G.; Misaka, T.; Herwan, J.; Komoto, H.; Furukawa, Y. An Experimental Study of Multi-Sensor Tool Wear Monitoring and Its Application to Predictive Maintenance. *Int. J. Adv. Manuf. Technol.* **2024**, *133*, 3415–3433. <https://doi.org/10.1007/s00170-024-13959-0>.
25. Wu, X.; Zhang, C.; Li, Y.; Huang, W.; Zeng, K.; Shen, J.; Zhu, L. Researches on Tool Wear Progress in Mill-Grinding Based on the Cutting Force and Acceleration Signal. *Measurement* **2023**, *218*, 113234. <https://doi.org/10.1016/j.measurement.2023.113234>.

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.