machines

MDPI

Article

A Physics-Guided Dual-Stream Vibration Feature Fusion Network for Chatter-Induced Surface Mark Diagnosis in Wafer Thinning

Heng Li [1,2], Hua Liu [2,*], Liang Zhu [2], Xiangyu Zhao [1,2], Lemiao Qiu [1] and Shuyou Zhang [1]

[1] School of Mechanical Engineering, Zhejiang University, 866 Yuhangtang Rd, Hangzhou 310058, China; liheng@jsjd.cc (H.L.); zhaoxiangyu@jsjd.cc (X.Z.); qiulm@zju.edu.cn (L.Q.); zsy@zju.edu.cn (S.Z.)
[2] Zhejiang Qiushi Semiconductor Equipment Co., Ltd., 500 Shunda Rd, Hangzhou 311100, China; zhuliang@jsjd.cc
* Correspondence: liuhua@jsjd.cc

Abstract

Ultra-precision thinning of hard and brittle materials like monocrystalline silicon demands high dynamic stability in thinning spindle. To address the challenge of accurately detecting subtle spindle chatter anomalies in industrial environments characterized by high noise and limited data, this paper proposes a physics-guided dual-stream attention fusion transfer network (PG-AFNet). First, a physics-guided signal preprocessing method was developed. Using variational mode decomposition (VMD) and continuous wavelet transform (CWT) masking, one-dimensional dynamic features and high-frequency regions of interest (ROIs) rich in transient impact features were extracted. Second, the PG-AFNet architecture was designed. By introducing an attention mechanism, it achieves deep integration of one-dimensional purely dynamic sequences with two-dimensional spatiotemporal visual textures to capture surface damage features caused by subtle vibrations. Finally, systematic validations were conducted using a real silicon wafer thinning dataset with 197 real samples. By overcoming small-sample limitations via physical augmentation, PG-AFNet achieved an 82.45% (86.64% after data augmentation) diagnostic accuracy, significantly outperforming traditional baselines. Furthermore, a large-scale cross-load validation on the diverse CWRU dataset yielded an exceptional 99.68% accuracy under mixed-load conditions, conclusively verifying the model's robust domain generalization. Lastly, a rigorous ablation study explicitly quantified the indispensable contributions of the physics-guided dual-stream architecture and attention fusion. This research provides a feasible theoretical foundation for intelligent surface quality monitoring in semiconductor hard-brittle material processing.

Keywords: wafer thinning; spindle vibration analysis; variational modal decomposition; continuous wavelet transform; physics-guided dual-stream network; feature fusion

Check for updates

Academic Editors: Danko Brezak, Tomislav Staroveški and Miho Klaić

Received: 17 March 2026
Revised: 2 April 2026
Accepted: 4 April 2026
Published: 7 April 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) license.

1. Introduction

As semiconductor technology shifts from "More Moore" to "More than Moore" [1,2], ultra-precision wafer thinning has emerged as the critical link connecting front-end fabrication with advanced packaging. To achieve extreme die integration and optimal heat dissipation, 12-inch monocrystalline silicon wafers must be reduced from a thickness of 775 μm down to 50 μm or even thinner [3]. However, this abrupt reduction in thickness causes the mechanical rigidity of the wafer to plummet. Compounded by the inherently

hard and brittle nature of silicon, the grinding process becomes extremely sensitive to dynamic disturbances from the machine tool spindle. Consequently, self-excited regenerative chatter frequently occurs. Even minor spindle vibrations can generate periodic spiral-shaped tool marks on the wafer surface, which serve as early physical indicators of internal structural damage and directly compromise the final chip yield [4]. These chatter-induced topographic defects are not only detrimental to the bare wafer but also devastating to the subsequent chemical mechanical polishing (CMP) process. Specifically, as systematically demonstrated by Tao et al. [5], grinding-induced periodic grooves drastically reduce the effective mechanical contact area between the polishing pad asperities and the wafer surface. This topographic deviation heavily concentrates the contact stress on the groove peaks, while leaving the valleys in a near-contact state where the pad fails to transmit cutting forces. Consequently, the pad cannot effectively remove the chemical passivation film generated within these geometric defects, which fundamentally disrupts the local chemical–mechanical synergy and severely degrades the uniformity of the material removal rate. Driven by these stringent manufacturing demands, overcoming the quality control bottlenecks in thinning large-size ultra-thin wafers is of paramount importance. Shifting from traditional post-process inspections towards zero-defect online monitoring has become a critical challenge requiring urgent resolution in the semiconductor equipment field.

To achieve reliable online monitoring, a deep comprehension of the mapping relationship between structural dynamics and surface topography is fundamental. Notably, the dynamic instability encountered in wafer thinning is physically analogous to the well-documented challenges in machining thin features, such as thin walls and thin floors. In this broader domain, fundamental and pioneering works have been established by Urbikain, Campa, and their co-researchers. They have made significant breakthroughs in modeling dynamic cutting forces and predicting regenerative chatter stability lobes using advanced mathematical frameworks, including Chebyshev polynomials [6,7] and homotopy perturbation methods [8]. Their theoretical insights have been seamlessly encapsulated into comprehensive predictive tools, such as the Turn+ MATLAB-based software (R2023b or higher)for dynamic turning, chatter analysis, and surface roughness prediction [9]. Furthermore, to effectively validate these complex dynamic models in confined industrial setups, their recent studies successfully deployed mini-size accelerometers [10] to precisely capture high-frequency transient responses without altering the structural mass. Inspired by these dynamic modeling and sensor deployment strategies in thin-feature machining, our study translates these mechanical condition monitoring concepts into the domain of ultra-precision semiconductor wafer thinning.

While these rigorous mechanistic models provide a profound theoretical foundation, they are often based on idealized rigid-body or steady-state cutting assumptions. Consequently, it remains challenging to accurately capture the transient nonlinear behavior of spindle systems under complex, time-varying operating conditions. Additionally, the intensive computational cost of simulation calculations makes it difficult to meet the strict real-time requirements. To overcome these limitations, online sensor-based monitoring techniques, particularly utilizing spindle vibration signals, have been actively developed [11]. As pointed out by Teti et al. [12], multi-sensor fusion and advanced signal processing hold great potential in precision machining monitoring. However, in ultra-precision wafer thinning processes, the signal characteristics corresponding to incipient surface damage are exceptionally weak. These marginal chatter signatures are easily masked by coolant splashes and profound background noise from the air-bearing spindle. Traditional threshold-based diagnostic methods rely heavily on time-domain statistical features, such as root mean square (RMS) and Kurtosis, but often suffer from severe delays. Usually, the

alarm is triggered only after serious periodic defects have already formed on the wafer surface, resulting in the scrapping of high-value wafers [13].

Driven by foundational breakthroughs in representation learning [14], data-driven diagnostic frameworks, particularly convolutional neural networks (CNNs), have demonstrated exceptional end-to-end feature extraction capabilities in monitoring general rotating machinery [15]. Despite these architectural advancements, directly deploying conventional deep networks to ultra-precision wafer thinning remains severely bottlenecked by three practical challenges. Firstly, the scarcity of anomalous samples restricts model convergence. Unlike standard bearing diagnostics supported by massive public datasets, acquiring labeled chatter samples during actual wafer production is highly destructive and economically prohibitive, making deep architectures inherently prone to severe overfitting [16]. Secondly, severe domain shifts hinder feature generalization. Bridging the significant feature distribution gap between general rotating equipment and specialized air-bearing spindle systems remains theoretically challenging [17]. Finally, the lack of model interpretability impedes industrial trust. Purely data-driven black-box models lack kinematic constraints, failing to provide explicit evidence that the extracted representations correspond to genuine physical defect frequencies rather than stochastic background noise [18].

To overcome these barriers, bridging the physical principles of spindle dynamics with the hierarchical learning mechanisms of CNNs is essential. Operational condition monitoring of air-bearing spindles fundamentally involves pattern recognition of non-stationary time series. From a dynamic perspective, abnormal vibrations in wafer-thinning spindle systems are essentially forced or self-excited vibrations, including those caused by regenerative chatter or grinding wheel imbalance. As highlighted by Lei et al. [19], the mathematical representations of these signals, much like gear meshing or rotor friction, are amplitude- or frequency-modulated waveforms. Surface damage or contact instability inevitably introduces periodic transient impulses into the continuous vibration profile. Crucially, this dynamic response characteristic is physically homologous across all types of rotating machinery. Within the framework of deep learning, neural networks exhibit a unique hierarchical distribution of features [20]. Fundamental visual and signal processing research indicates that shallow convolutional layers primarily function as generalized Gabor filters. They are capable of accurately capturing low-level, domain-invariant geometric features, such as transient pulses, rising edges, and periodic textures [21]. As the network depth increases, these representations become increasingly abstract and highly tailored to the specific failure modes of the target equipment. Therefore, exploiting this physical homology by leveraging shallow networks as universal vibration feature extractors provides a robust algorithmic foundation to overcome data scarcity and domain shifts in high-precision spindle monitoring.

Based on the above analysis, to address those challenges, this paper aims to propose a novel online monitoring approach integrating grinding dynamics mechanisms with deep transfer learning. Unlike previous single-dimensional studies, this paper introduces the grinding kinematic mechanisms as prior knowledge into the feature extraction process, proposing a physics-guided dual-stream attention fusion transfer network (PG-AFNet). This approach reveals the frequency-domain mapping relationship between vibration and tool marks by constructing a dynamic model, guiding adaptive signal preprocessing. In this study we address the small-sample challenge through cross-domain transfer learning strategies, incorporating attention mechanisms and visualization techniques to verify the physical consistency of the model.

2. Thinning Mechanism and Dynamics Modeling

2.1. Process Workflow and Thinning Mechanism of the Wafer Thinning Machine

Backside thinning of wafers is a complex nonlinear process involving random cutting with multiple abrasive particles. In this ultra-precision back thinning configuration, the monocrystalline silicon wafer is secured onto a rotating porous ceramic chuck via vacuum suction. The thinning tool, typically a diamond cup wheel attached to the ultra-precision spindle, rotates at a high speed (ω_s), while the wafer chuck rotates independently at a lower speed (ω_w). Notably, the rotational axes of the spindle and the chuck are purposely offset. The cup wheel is positioned such that its cutting abrasive rim covers the center to the outer edge of the wafer. As the spindle feeds downward along the z-axis at a constant feed rate (v_f), the entire surface of the wafer is uniformly thinned through the continuous relative cross-rotational motion, as shown in Figure 1. A typical thinning process consists of two steps. The rough grinding stage (Z1) uses a coarse-grit grinding wheel (e.g., #320-grit) combined with a high feed rate to rapidly remove most of the silicon material. This stage primarily causes brittle fracture, leaving a deep subsurface damage layer (SSD) on the wafer surface. The fine grinding stage (Z2) switches to a fine-grit grinding wheel (e.g., #2000-grit) and reduces the feed rate. The primary objective of the Z2 stage is to achieve the target thickness. At the same time, this stage controls the final surface roughness R_a and total thickness variation (TTV), and removes the SSD layer left over from rough grinding. During silicon wafer thinning, because silicon wafers are typically hard and brittle materials, they generate intense and highly fluctuating normal and tangential grinding forces at the wheel–wafer contact interface. Due to the closed-loop stiffness of the machine tool, these dynamic cutting forces propagate directly upward into the spindle bearing system. When the excitation energy couples with the weak modal stiffness of the spindle, self-excited regenerative chatter is easily triggered. This chatter not only inflicts severe periodic tool marks on the wafer surface but also leaves distinct high-frequency transient impact signatures in the spindle's vibration signals, which forms the physical basis for our condition monitoring approach.

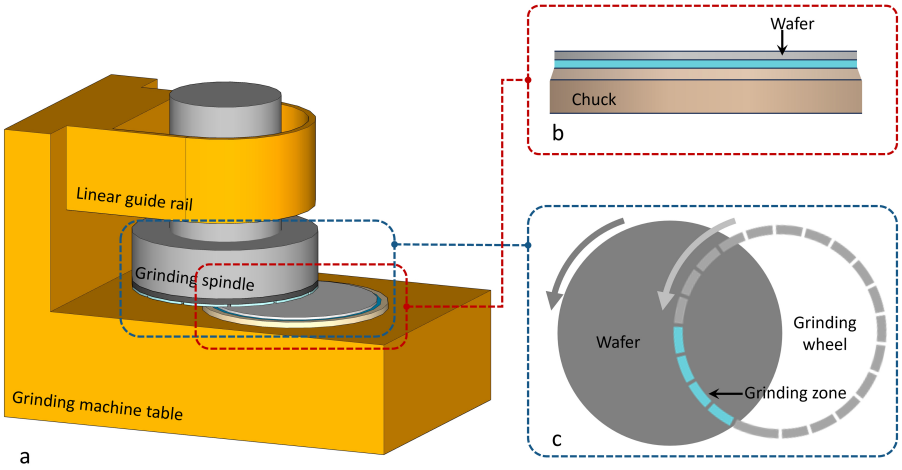


Figure 1. Schematic diagram of self-rotating grinding process principle: (**a**) thinning machine structure; (**b**) wafer adsorption and fixing; (**c**) grinding area.

2.2. Principle of Material Removal in Silicon Wafer Thinning

Single crystal silicon, as a typical brittle material, exhibits a grinding removal mechanism that depends on the cutting depth of the abrasive particles. According to the brittle-plastic transition theory proposed by Bifano [22], there exists a critical depth of cut d_c. During the brittle regime, when the instantaneous cutting depth of a single abrasive

grain reaches a certain threshold $h(t) > d_c$, material removal primarily occurs through crack propagation and spalling. At this stage, the surface primarily consists of pits and microcracks, and the surface quality appears unsatisfactory. For ductile regime removal $h(t) < d_c$, silicon material exhibits metal-like plastic flow behavior, where material is removed through shear slip. This process yields a mirror-like finish without damage. In an ideal precision grinding process, process parameters are set to ensure the vast majority of abrasive grains operate within the plastic range. When abnormal vibrations occur in the spindle system, dynamic cutting forces cause the instantaneous cutting depth $h(t)$ to fluctuate violently. Once the peak of these fluctuations exceeds a certain threshold d_c, the local machining mode shifts into the brittle range, resulting in periodic deep-seated damage striations on the surface.

2.3. Kinematics and Vibration Coupling Model

To analyze the specific effects of spindle vibration on surface topography, it is necessary to establish equations describing the relative motion trajectories of the polishing particles on the wafer surface. A Cartesian coordinate system is established with the center of the wafer as the origin. The motion trajectory (x_i, y_i) of any polishing particle G_i relative to the wafer can be described as a cycloidal motion:

$$\begin{cases} x_i(t) = R_s \cos(\omega_s t + \theta_i) + e\cos(\omega_w t) \\ y_i(t) = R_s \sin(\omega_s t + \theta_i) + e\sin(\omega_w t) \end{cases} \tag{1}$$

where R_s is the radius of the grinding wheel, e is the eccentricity between centers of the grinding wheel and wafer, and θ_i is the initial phase angle of the abrasive particle. After introducing z-axis spindle vibration, the motion of the abrasive particles transitions from a 2D plane to oscillation in 3D space. Given that the spindle is subjected to harmonic vibration with a frequency of f_v and an amplitude of A_v, the actual instantaneous cutting depth $H(t)$ is the superposition of the nominal feed rate and the vibration displacement:

$$H(t) = v_f \cdot t + A_v \sin(2\pi f_v t + \phi) - \text{ElasticRecovery} \tag{2}$$

where v_f represents the axial feed rate; ElasticRecovery denotes the elastic recovery of the material. For hard and brittle materials such as silicon, this factor can be neglected and is omitted in subsequent analyses. When the vibration frequency f_v satisfies a specific relationship $f_s = \omega_s/2\pi$ with the spindle frequency, a distinctive texture pattern forms on the surface. During synchronous vibration $(f_v = k \cdot f_s)$, the abrasive particle trajectory coincides in each rotational cycle, causing waveform superposition and forming distinct radial striations. During asynchronous vibration $(f_v \approx f_n)$, regenerative vibration occurs when the vibration frequency approaches the system's natural frequency. In this state, the cutting trajectory of the abrasive particles undergoes a phase shift relative to its previous trajectory, resulting in a time-varying modulation of the cutting depth. This modulation effect manifests spatially as spiral-shaped tool marks. According to the kinematic equations, the vibration frequency f_v is directly proportional to the spatial wavelength λ on the wafer surface:

$$\lambda \approx \frac{v_{linear}}{f_v} \tag{3}$$

where v_{linear} is the linear velocity at the abrasive grain. Specific frequency bands in the spindle z-axis vibration signal are the root cause of the deterioration in surface quality (i.e., the flutter frequency and its harmonics). This kinematic relationship establishes a direct mapping from the vibration frequency f_v to the spatial wavelength λ. It provides the physical basis for selecting the target frequency bandwidth in Section 3.2.

3. Physics-Guided Signal Pre-Processing Framework

Given the non-stationary properties of the wafer thinning process and the strong noise disturbances in industrial environments, feeding the raw signal directly into a deep neural network makes it difficult to extract key damage features. Based on the analysis proposed in Section 2, this section presents a signal preprocessing framework that incorporates physical prior knowledge. This framework consists of three stages: process-semantic-based adaptive segmentation, variational modal decomposition (VMD) for denoising, and time–frequency reconstruction of regions of interest (ROI).

3.1. Process-Aligned Signal Capture Strategy

In actual production, the thinning machine follows a predefined process recipe. The thinning process is divided into different stages: blank cutting, rough grinding (P1), fine grinding (P2), and air cutting (P3), followed by polishing. Because the data acquisition system operates independently of the PLC control system, the raw vibration signals may exhibit nonlinear drift on the time axis due to trigger delays or network transmission. Localized unstable cutting conditions may occur around the P2 stage, which can introduce additional noise. Therefore, vibration signals must be segmented and extracted according to process stages to ensure subsequent model accuracy. As shown in Figure 2, the strategy for aligning spindle load feedback with process timing is as follows:

(1) Sequence localization: directly read the process status flags from the machine's PLC to precisely determine the theoretical start and end times (T_{start} and T_{end}) of the precision grinding stage (P2). (2) Transient rejection: considering the process transitions at the beginning and end of stage P2, potential non-steady-state processes may introduce atypical vibration characteristics. A stability window is introduced within the coarse alignment interval to buffer δ_in and δ_out between T_{start} and T_{end}. The valid signal segment x_{ROI} is defined as $x_{ROI} = x(t)|_{t\in[t_{start}+\delta_{in},\, t_{end}-\delta_{out}]}$. All samples are uniformly resampled to a fixed length n to ensure consistent data dimensionality for input into the deep neural network.

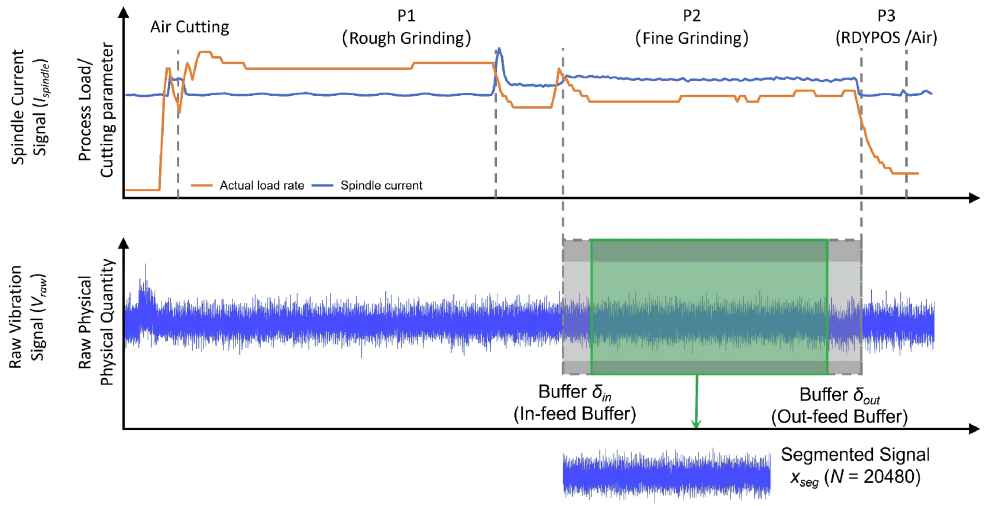


Figure 2. Process-aligned signal capture strategy (The dashed lines in the figure indicate the division of thinning process stages).

3.2. Bandwidth-Constrained Adaptive VMD Denoising

To separate the weak impact components induced by grinding texture from background noise, VMD is introduced. However, the traditional maximum Kurtosis criterion easily fails when confronted with strong periodic background signals (such as the main shaft rotational frequency and its harmonics). The background signal from normal machining (IMF1, 937 Hz) exhibits extremely high energy and negative Kurtosis, masking

the high-frequency components (IMF4, 2500 Hz) that genuinely contain damage features. Therefore, this paper proposes a band-constrained max-Kurtosis criteria. Based on grinding dynamics analysis, the frequency range primarily responsible for regenerative chatter and microscopic surface damage occurs at $\omega_{mask} = [500, 4500]$ Hz (spindle speed of approximately 2300 rpm). Calculate the peak height values of the modal components at the center frequency $f_c \in \omega_{mask}$, and select the intrinsic mode function (IMF) with the highest peak height to reconstruct the dynamic flow input x_{dyn}:

$$x_{dyn} = u_{k^*},\ k^* = \arg\max_k \{\text{Kurtosis}(u_k) \mid f_{c,k} \in \omega_{mask}\} \tag{4}$$

This strategy successfully circumvents low-frequency rotational interference and ultra-high-frequency electronic noise, enabling targeted extraction of subtle fault characteristics. VMD adaptively decomposes the signal x_{seg} into k IMF components $u_k(t)$ with specific center frequencies ω_k. In the VMD algorithm, the quadratic penalty factor is set to its empirical baseline of 2000 according to the classic variational framework [23].

To prevent mode mixing and ensure the stable extraction of high-frequency transient features, a comprehensive statistical analysis was conducted across the entire dataset. For the decomposition parameter K [24], the mean of the extracted maximum Kurtosis values after VMD was compared for K ranging from 2 to 6, as shown in Figure 3. When $K < 4$, under-decomposition occurs, and the chatter signatures remain coupled with background noise, reducing the Kurtosis. When $K > 4$, over-decomposition occurs, and the physical chatter mode is artificially split into multiple spurious IMFs, fragmenting the transient energy. The optimal value $K = 4$ adaptively extracts the purest high-frequency transient impact component (x_{dyn}) and achieves the highest Kurtosis without mode fragmentation, as shown in Figure 4a. Its constrained variational problem is formulated as follows:

$$\min_{\{u_k\},\{\omega_k\}} \left\{ \sum_{k=1}^{K} \| \partial_t [(\delta(t) + \frac{j}{\pi t}) * u_k(t)] e^{-j\omega_k t} \|_2^2 \right\} \tag{5}$$

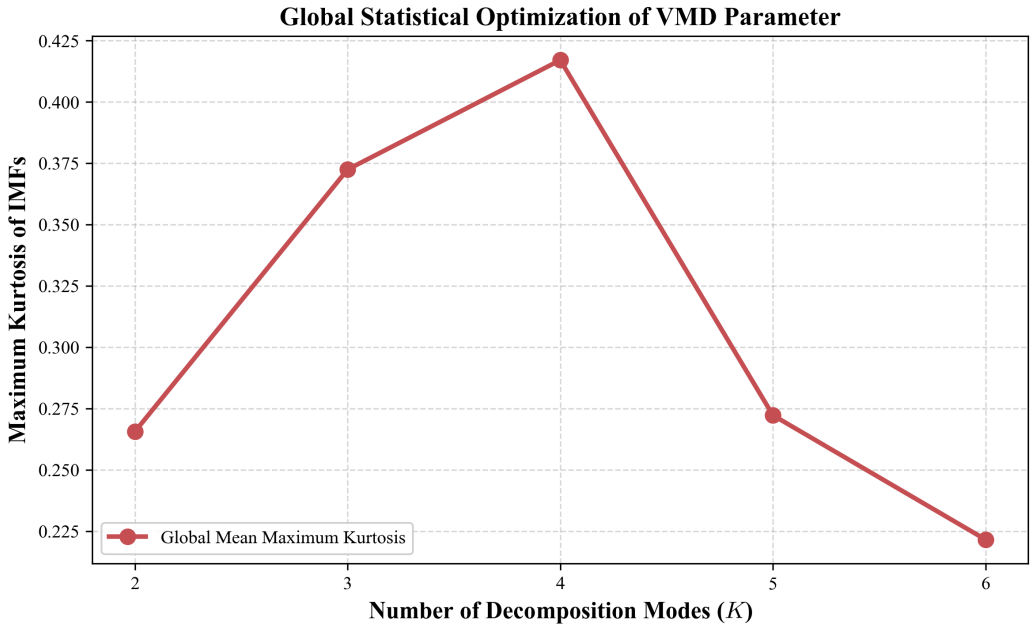


Figure 3. Global Statistical Optimization of VMD Parameter.

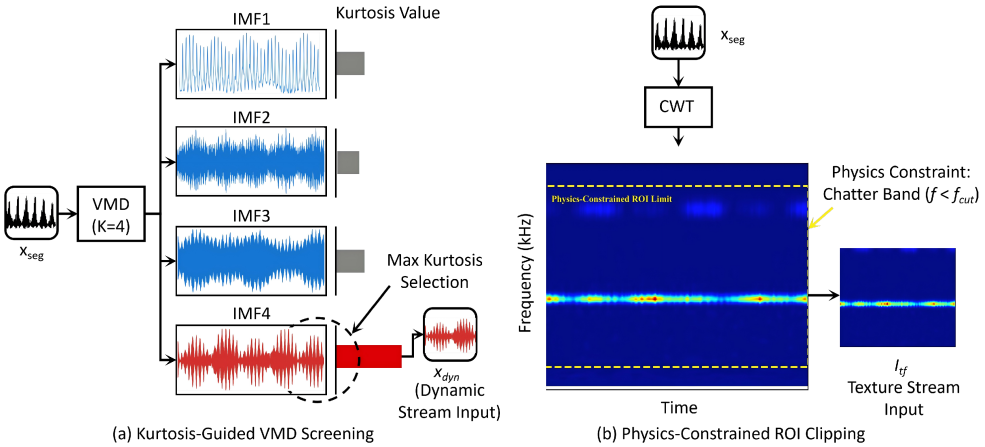


Figure 4. Decomposition and noise reduction of spindle vibration signals based on VMD and CWT. (**a**) Bandwidth-constrained adaptive VMD denoising; (**b**) physics-constrained ROI clipping.

3.3. Physics-Constrained Time–Frequency ROI Reconstruction

To capture the time-varying texture features of the grinding process, the signal undergoes a continuous wavelet transform (CWT). Specifically, the complex Morlet wavelet (cmor3-3) is selected for two critical reasons. First, from a physical perspective, its exponentially decaying harmonic morphology perfectly matches the transient impulse responses generated by mechanical chatter. Second, from a mathematical perspective, characterized by its balanced bandwidth parameter ($F_b = 3$) and center frequency parameter ($F_c = 3$), cmor3-3 provides an optimal joint time–frequency resolution. This prevents severe temporal smearing and frequency leakage, thereby generating highly discriminative spatial textures for the subsequent network:

$$W(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} x_{seg}(t) \psi^* \left(\frac{t-b}{a} \right) dt \tag{6}$$

where a is the scaling factor and b is the translation factor. Based on the dynamic derivation in Section 2, the spindle chatter frequency $f_{chatter}$ causing tool marks and its higher-order harmonics are primarily distributed within the frequency band $[0,\ f_{cut}]$ ($f_{cut} = 4.5$ kHz at a spindle speed of 2300 rpm).

Since the sampling frequency of the actual data acquisition system is 10 kHz, according to the Nyquist sampling theorem, the effective upper limit of the analysis frequency is 5 kHz. Therefore, the frequency axis of the CWT time–frequency plot is clipped to $[0,\ f_{cut}]$. This physical constraint not only eliminates the high-frequency background region lacking information but also significantly increases the pixel proportion of texture features within the image.

To eliminate high-frequency noise and reduce computational redundancy, this paper implements physically constrained ROI clipping on the scalogram. After frequency-domain clipping, only the scalogram coefficients within the range $[0, f_{cut}]$ are retained. The cropped coefficient matrix S_{roi} undergoes energy normalization using Min–Max normalization. It is then mapped to a 2D matrix I_{tf} of size 256×256 and serves as input to the texture stream in the dual-stream network, as shown in Figure 4b.

4. PG-AFNet

To address the scarcity of abnormal samples in thinning machines and the poor explainability of existing models, this chapter proposes the PG-AFNet. This network consists of three core modules: a dual-stream feature extractor based on parameter transfer, a multimodal mechanism fusion module, and a fault classifier, as shown in Figure 5.

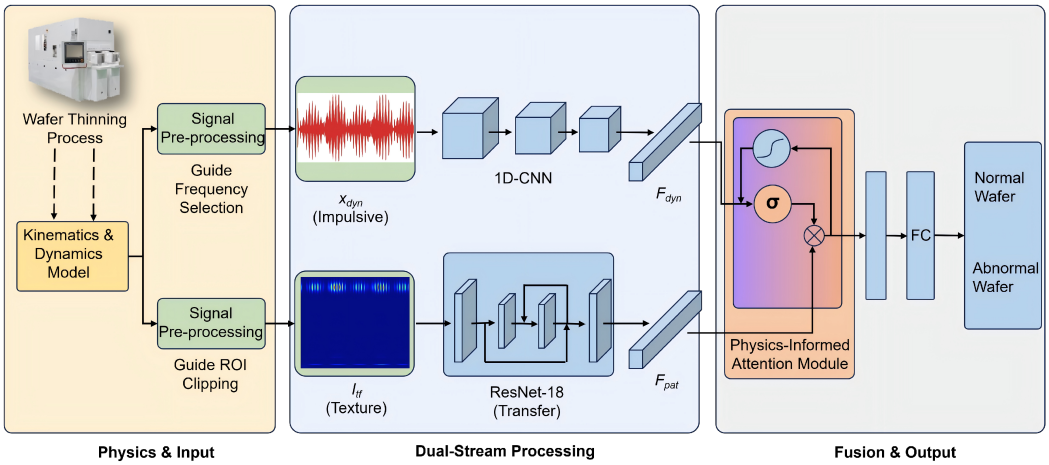


Figure 5. The overall architecture of the proposed PG-AFNet.

4.1. Source-Domain Transfer Learning Strategy

Based on the previous analysis, the key to constructing an efficient feature extractor lies in the quality and quantity of source domain data, rather than complete structural consistency of the equipment. In order to develop a robust shallow universal feature extractor, it is essential to select a standard dataset characterized by a large sample size, comprehensive operational coverage, and high signal-to-noise ratio. Given the limited availability of labeled damage samples obtained from wafer thinning experiments (target domain, D_t), directly training deep networks is highly prone to overfitting. The Case Western Reserve University (CWRU) dataset contains an exceptionally rich array of rotational modulation and transient shock characteristics, widely regarded as a universal benchmark in the field of rotating machinery fault diagnosis [25]. Numerous studies have demonstrated that deep transfer learning architectures, pre-trained on standardized vibration databases, exhibit excellent cross-domain diagnostic performance when transferred to complex industrial equipment, including planetary gearboxes and wind turbine rotors [26,27]. Although bearings and wafer-thinning spindles differ in their mechanical structures, they share a high degree of similarity in their fundamental vibration mechanisms. The CWRU dataset is selected as the source domain (D_s), serving as a rich repository of generic rotating machinery vibration features. Parameter-based transfer learning can use the layered feature extraction capabilities of CNNs, with the shallow layers primarily extracting general edge, texture, and transient impact features. By transferring the parameters of a pre-trained shallow neural network, it is initialized as a general-purpose feature extractor for the spindle thinning monitoring network. This method enables efficient detection of subtle spindle failures at both the physical mechanism and data feature levels. The transfer strategy is as follows (Figure 6):

(1) Pre-training: train the ResNet-18 network on D_s to learn generic spatio-temporal texture features;
(2) Parameter freezing: freeze the weights of the pre-trained model's first three convolutional layers (Conv1, Pool1, ResBlock1, ResBlock 2) to serve as a fixed feature extractor;
(3) Fine-tuning: only fine-tune the high-level semantic layers (ResBlock r3, ResBlock 4) and fully connected layers to adapt to the unique blade pattern feature distribution of thinning machines.

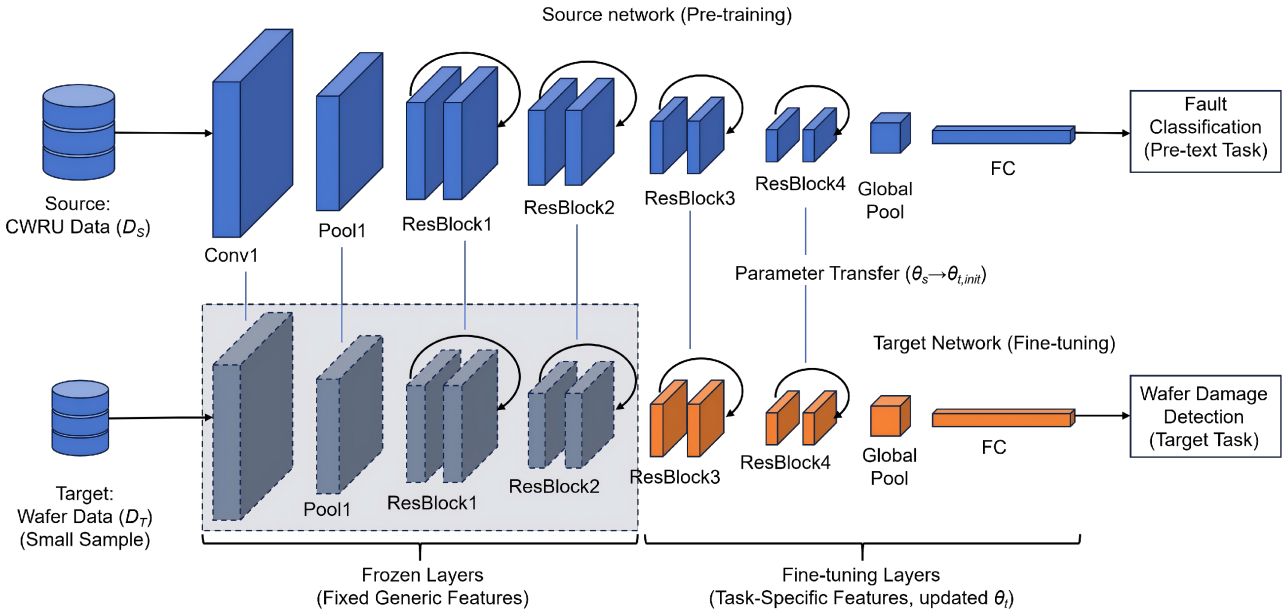


Figure 6. Schematic of the proposed transfer learning strategy.

4.2. Dual-Stream Feature Extraction Architecture

While VMD preprocessing effectively removes industrial background noise and highlights primary transient impacts, any band-decomposition method inherently carries the risk of omitting fault frequencies or high-order harmonics. To mitigate this risk, the proposed PG-AFNet adopts a two parallel processing streams, as shown in Figure 7. The 2D stream (Stream B) operates entirely independently of the VMD-based Stream A, and by applying CWT directly to the raw signal, it serves as an architectural safeguard that preserves the global, unfiltered time–frequency energy distribution. Even if the VMD inadvertently filters out certain subtle fault-related components, the critical fault-relevant information is still retained to the greatest extent possible within the parallel 2D stream.

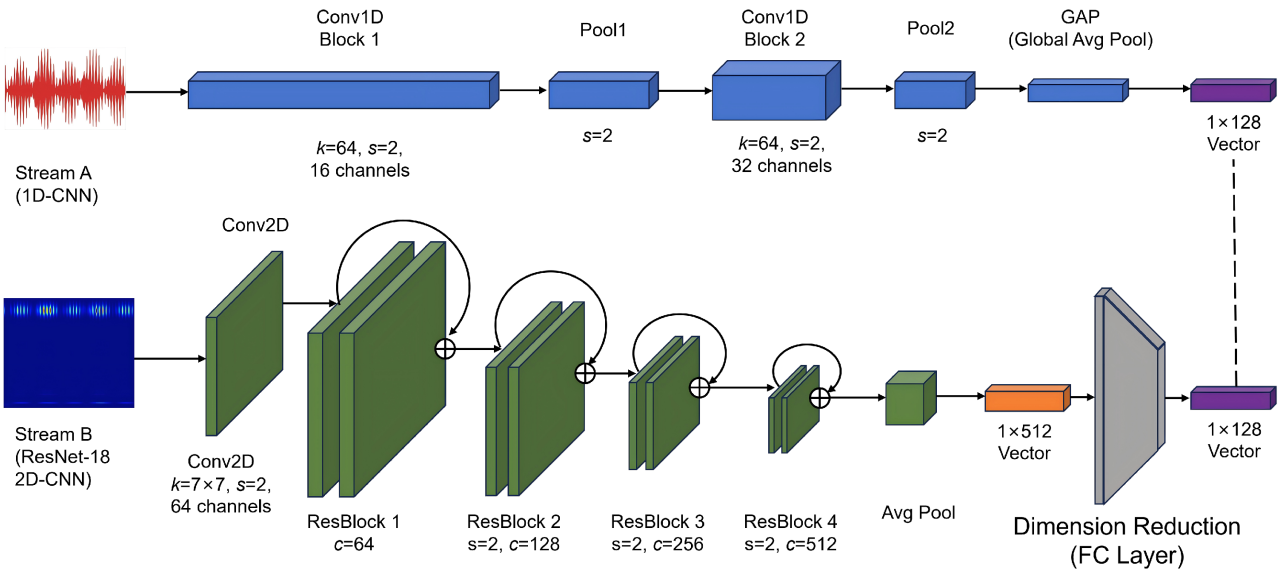


Figure 7. Detailed structure of the dual-stream feature extractor.

(1) Stream A: temporal dynamic stream. Designed to capture transient impact features caused by vibrations. Uses a custom lightweight one-dimensional CNN (1D-CNN).

Input: signal $x_{dyn} \in \mathbb{R}^{1\times 2048}$ reconstructed via VMD.

Structure: Consists of three convolutional blocks. The first layer uses large convolutional kernels (Kernel = 64) to capture long-range patterns, while the subsequent layers use small convolutional kernels (Kernel = 3) to extract local details. Output: feature vector $F_{dyn} \in \mathbb{R}^{1\times128}$ obtained after global average pooling.

(2) Stream B: time–frequency pattern stream. Designed to capture periodic texture features induced by chatter. Uses a fine-tuned ResNet-18 backbone network.

Input: ROI-cropped time–frequency input $I_{tf} \in \mathbb{R}^{3\times224\times224}$.

Structure: frozen layers extract low-level textures; fine-tuned layers extract specific fault patterns.

Output: feature vector $F_{pat} \in \mathbb{R}^{1\times512}$ after pooling layer.

4.3. Physics-Guided Attention Fusion

Traditional feature fusion methods typically rely on blind vector concatenation, which lacks physical interpretability. The proposed fusion strategy in this paper is based on grinding dynamics and has physical guidance and constraints (Figure 8). The VMD and ROI cropping at the input stage constitute hard constraints. The proposed network model explicitly excludes high-frequency noise and low-frequency rigid-body motion, extracting features only within the chatter frequency bandwidth determined by the dynamic equations. This is equivalent to mapping prior physical insight into implicit attention masks for a neural network. During the thinning process, grinding damage presents in two physical forms: transient impacts caused by abrasive particle fracture and steady-state chatter caused by insufficient system stiffness. Based on extracted features, stream A exhibits sensitivity to transient impact, while stream B demonstrates sensitivity to steady-state chatter.

In this paper, a gated attention module is designed to serve as an arbitrator. By reducing the dimensionality of texture features F_{pat} to $F'_{pat} \in \mathbb{R}^{1\times128}$ through a fully connected layer, a joint feature $F_{joint} = [F_{dyn}, F'_{pat}]$ is constructed. The attention weight α is calculated as follows:

$$\alpha = \sigma\big(\mathbf{W}_2 \cdot \mathrm{ReLU}(\mathbf{W}_1 \cdot F_{joint} + b_1) + b_2\big) \tag{7}$$

where σ represents the Sigmoid activation function, $\mathbf{W}_1, \mathbf{W}_2$ denote learnable weight matrices, and b_1 and b_2 are bias. The final fused feature F_{final} is computed as follows:

$$F_{final} = \alpha \cdot F_{dyn} + (1 - \alpha) \cdot F'_{pat} \tag{8}$$

This mechanism enables the network to dynamically switch between dynamic impact-dominated and texture-dominated failure modes, greatly enhancing robustness against subtle early-stage failures. When detecting impact damage (e.g., grinding wheel hard spot impacts, characterized by high steepness), the network automatically increases the weight ($\alpha \to 1$) of the 1D dynamic flow (F_{dyn}); When diffuse damage (such as regenerative chatter, manifested as specific textures) is detected, the network focuses on the features ($\alpha \to 0$) of the 2D texture flow (F_{pat}). This mechanism ensures that the model's decision logic remains highly consistent with the actual grinding damage generation mechanism, addressing the lack of interpretability in traditional stitching fusion methods.

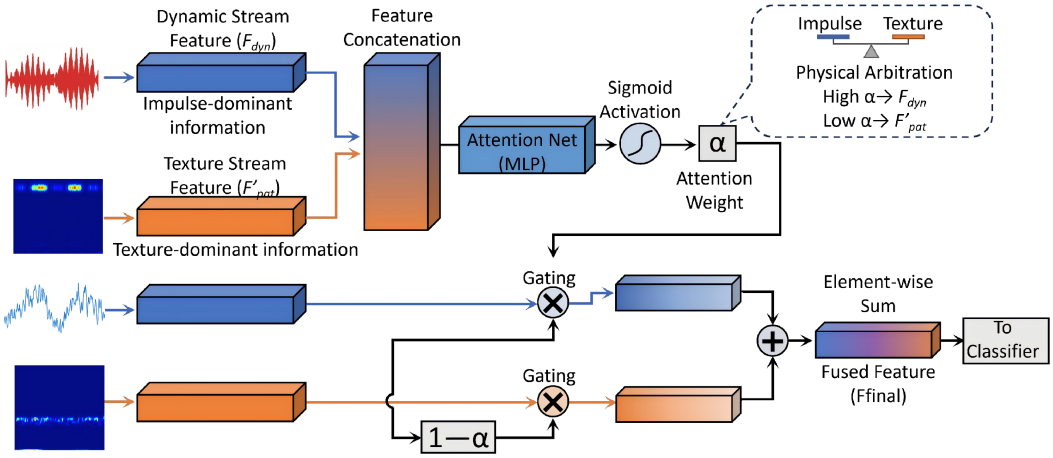


Figure 8. Diagram of the physics-guided attention fusion model.

4.4. Loss Function and Optimization

Given the scarcity of abnormal samples in wafer manufacturing (class imbalance) and the over-fitting risk in small-sample training, standard cross-entropy loss is no longer suitable. This paper proposes a hybrid objective function:

$$L_{total} = L_{Focal} + \lambda ||\mathbf{w}||^2 \tag{9}$$

(1) Focal Loss: by reducing the weight of simple samples (normal wafers), it forces the model to focus on difficult-to-classify abnormal samples:

$$L_{Focal}(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \tag{10}$$

where p_t represents the model's predicted probability, and the focal parameter γ is set to 2.

(2) Label smoothing: introducing label smoothing regularization converts hard labels into soft labels, preventing the model from becoming overly confident on limited data and thereby enhancing its generalization capability.

5. Experiments and Data Analysis

5.1. Experimental Platform and Data Acquisition

This study conducted sample collection on a 12-inch silicon wafer thinning machine in a laboratory environment, using 12-inch silicon wafers as the workpieces. During the thinning process, the spindle of the thinning machine is set to rotate at approximately 2300 r/min (corresponding to a rotational frequency of about 38.3 Hz). To accurately capture the weak abnormal vibrations of the spindle during the thinning process, a high-frequency piezoelectric accelerometer was selected for the experiment. The sensor model is CXT-YD-3193 (Zeda instruments, Hangzhou, China), a triaxial sensor with an axial sensitivity of 100 mV/g ±5 % and maximum support for 40 kHz synchronous sampling. The sensor was fixed to the spindle housing of the thinning machine (analyzing the z-axis), as shown in Figure 9c. It is worth noting that the grinding of monocrystalline silicon using a diamond cup wheel exhibits a relatively slow tool wear rate. To strictly control the experimental variables and prevent severe domain shift caused by long-term tool degradation, all samples in this dataset were collected during the stable steady-state wear phase of the grinding disc. Under this condition, the tool state is physically considered quasi-static.

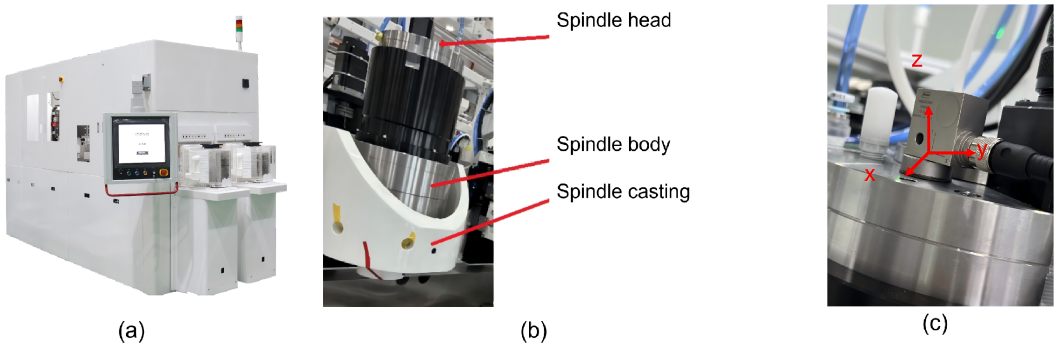


Figure 9. (**a**) Wafer thinning machine; (**b**) wafer thinning machine spindle and experimental test bench; (**c**) accelerometer mounting position.

During the data acquisition process, the vibration signals were captured at a sampling frequency of 10 kHz, yielding an effective analysis bandwidth of 5 kHz according to the Nyquist–Shannon sampling theorem. Physically, the abnormal transient impacts induced by spindle vibration typically emerge within the 2 to 5 kHz band, as the dominant regenerative chatter frequencies inherently lock onto the structural natural frequencies of the air-bearing spindle (usually between 1.5 kHz and 4.0 kHz). To empirically justify the sufficiency of this sampling configuration and verify that no informative high-frequency components were excluded, a Fast Fourier Transform (FFT) spectral analysis of a typical severe chatter signal is presented in Figure 10. As shown in the spectrum, the primary dynamic impact energy is intensely concentrated within the 2.0 to 4.5 kHz band. Notably, the spectral amplitude rapidly decays to the baseline noise floor well before reaching the 5 kHz. This spectral distribution confirms that any components above 5 kHz are overwhelmingly dominated by stochastic background noise rather than critical defect signatures. Consequently, setting the sampling rate to 10 kHz guarantees the complete and alias-free preservation of high-frequency transient diagnostic features, while effectively preventing the computational hardware overload and data redundancy associated with excessively high sampling rates, thereby optimizing the framework for real-time online monitoring.

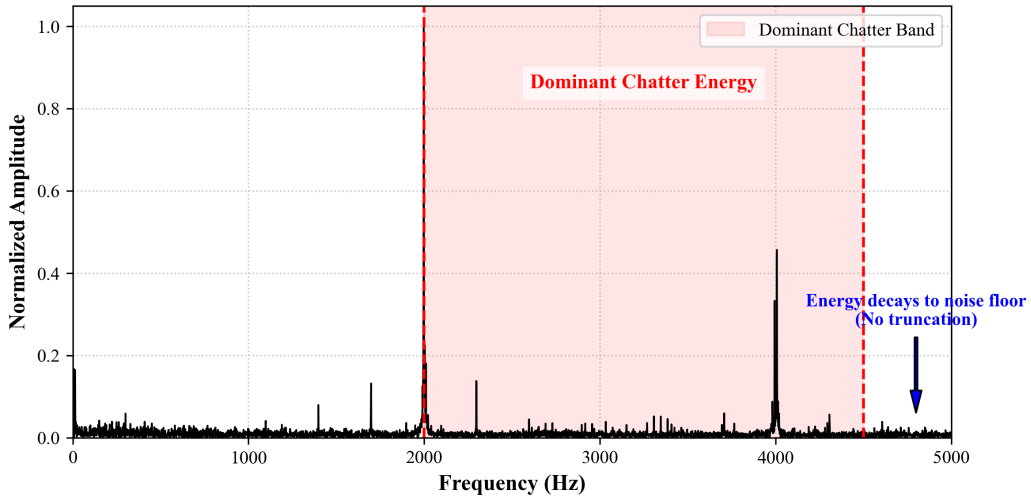


Figure 10. Frequency spectrum analysis of a sample signal.

During the pre-shipment debugging and test-run process of the thinning machine, the surface quality of experimental samples is calibrated through high-intensity visual inspection under strong light and R_a measurements. Figure 11a illustrates sample specimens demonstrating the ideal degree of tool marks, ranging from coarse to smooth. Samples

with smooth surfaces roughness $R_a < 10$ nm (a semiconductor customer manufacturer standard) and no visible tool marks (or very light) were classified as Class 0 (Figure 11a); those exhibiting pronounced chatter patterns or machining defects were classified as Class 1 (Figure 11b,e). During the testing phase, 197 sets of actual measurement data were collected, comprising 108 valid samples and 89 abnormal samples (during April of 2025. The conditions for obtaining samples in actual semiconductor production environments are extremely stringent and challenging).

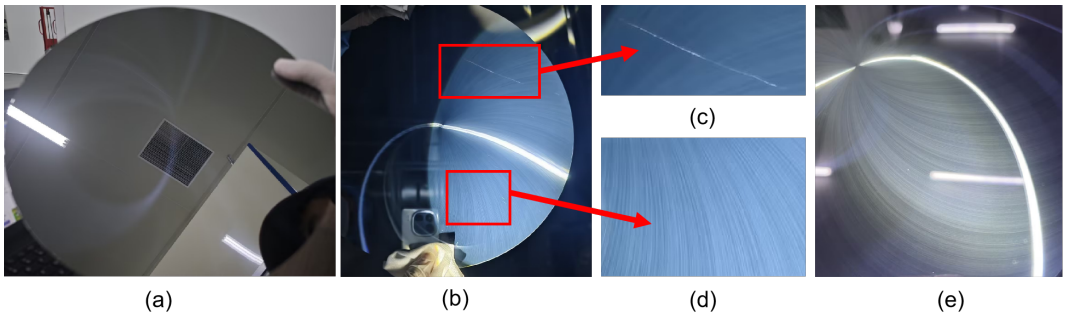


Figure 11. (**a**) Wafer surface is generally good with minor tool marks; (**b**) Tool marks are relatively prominent; (**c**) Localized zoomed-in scratch; (**d**) Localized zoomed-in tool mark; (**e**) Surface tool marks are deep.

5.2. Sample Pre-Processing and Time–Frequency Feature Extraction

Feature extraction and transformation are performed on the collected spindle vibration data. As shown in Figure 12, the time-domain waveform of the raw vibration signal exhibits strong randomness and modal aliasing, making it difficult to observe periodic impact components related to tool wear degradation from the time-domain amplitude. Adaptive processing of the raw signal using VMD (Figure 12). Through VMD, the overlapping signals are decoupled into multiple IMF components with independent center frequencies. Using a peak-ratio maximization criterion, the dominant low-frequency spindle fundamental frequency components, which exhibit superior amplitude but are insensitive to faults, are eliminated. This process precisely locates and extracts the preferred IMF component, which is rich in high-frequency transient impact information. This component preliminarily reveals a pulse envelope in the time domain that correlates with the spindle rotation cycle.

To prevent potential information loss caused by VMD filtering and to preserve the holistic time–frequency energy distribution, the raw segmented signal directly undergoes CWT processing, as shown in Figure 13a. While VMD has isolated the low-frequency strong noise, some high-frequency white noise remains scattered within the wide frequency band. These redundant features are highly prone to triggering dimensionality disaster and overfitting when input into deep convolutional networks. A physical mask is applied at the two-dimensional time–frequency scale to eliminate frequency bands inconsistent with the physical bandwidth of machining chatter, precisely excising the specific ROI, as shown in Figure 13b. After CWT time–frequency mapping, and physical ROI excision, the faint vibration time–frequency texture features are ultimately extracted (i.e., the high-energy bright bands in Figure 13b). The extracted ROI tensor serves as one of the model inputs for PG-AFNet.

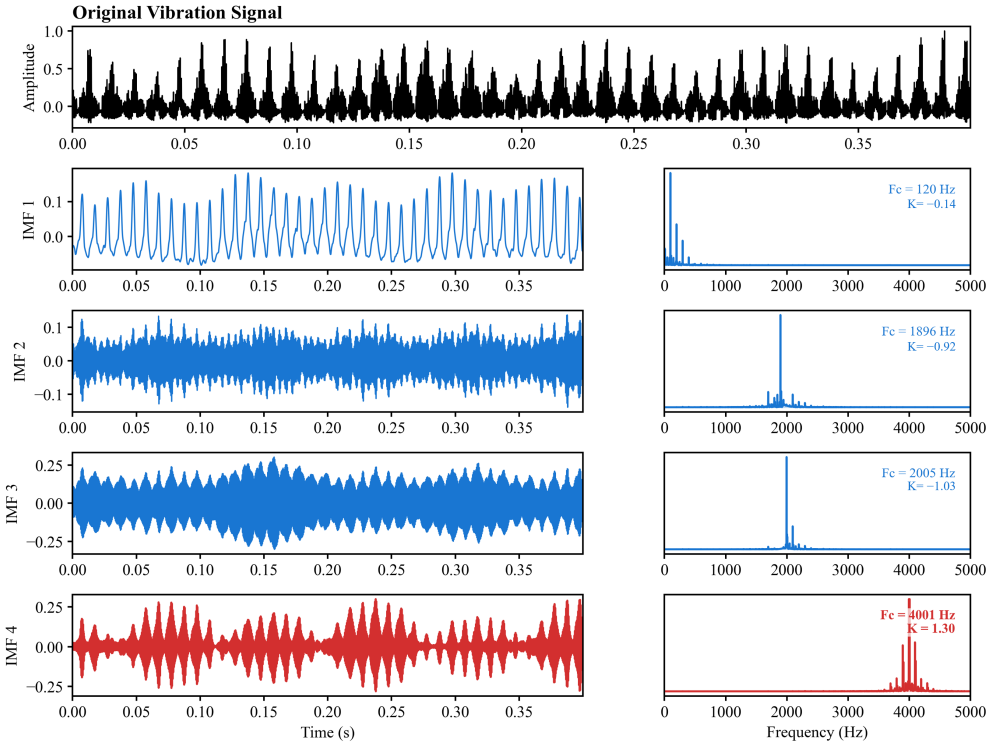


Figure 12. Comparison of original vibration signal and VMD adaptive decomposition results.

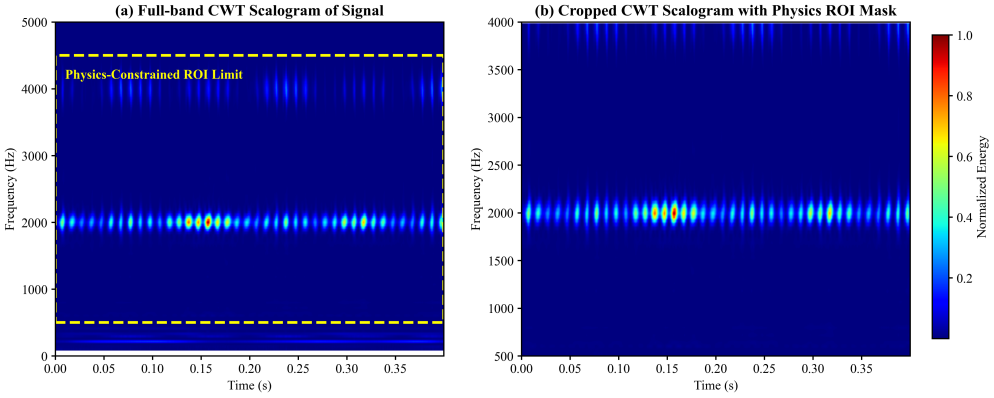


Figure 13. Comparison of CWT and ROI extraction: (**a**) full-band CWT scalogram of signal; (**b**) cropped CWT scalogram with Physics ROI mask.

5.3. Verification and Comparison

To validate the algorithmic performance of the proposed PG-AFNet method in the silicon wafer thinning process, we conducted experiments to perform comparative analyses across four dimensions: methodology, sample augmentation, benchmark validation and ablation experiments.

5.3.1. Detailed Structure and Parameter Settings of Comparison Algorithms

To objectively evaluate the performance boundaries of different feature extraction methods under conditions of high noise and small sample sizes, three representative benchmark models were selected for comparison with the proposed PG-AFNet. All deep learning models were trained independently in the PyTorch (v.2.11.0) framework using uniform hyperparameters (e.g., batch size = 16, initial learning rate $lr = 5 \times 10^{-4}$, identical cosine-smoothed annealing decay, etc.):

(1) SVM: Normalize the raw vibration signals and manually extract six-dimensional time-domain features: RMS, Kurtosis, mean curvature, peak value, variance, and peak-to-peak value; construct a feature vector. The classifier employs an RBF kernel function and applies heuristic category weights to mitigate the issue of sparse samples.
(2) 1D-CNN: The model directly takes normalized 1D time-series signals as input. The first layer employs a large-scale convolution kernel (Kernel size = 32, Stride = 2), followed by alternating stacks of convolution layers and max-pooling layers. The network utilizes the GELU activation function and applies Dropout ($p = 0.4$) before the fully connected layer to suppress overfitting.
(3) Simplified ResNet-18: To circumvent the curse of dimensionality caused by small sample sizes, this paper implements a simplified ResNet-18 using truncated transfer learning. The network retains only an extremely shallow structure (from conv1 to layer2), with input being the ROI tensor extracted after CWT processing.
(4) PG-AFNet (the model proposed in this paper): This model integrates 1D temporal data with 2D ROI feature inputs. Following dual-stream attention fusion, the optimization process employs the AdamW optimizer (with strong L2 regularization at 0.05) and introduces Focal Loss ($\gamma = 2.0$) with adaptive class weights. This method effectively reduces the tendency of the model to favor dominant categories, thereby improving the recall rate for samples with subtle defects.

5.3.2. Physical Level Sample Augmentation Strategy

During algorithm validation, the samples in the training set undergo augmentation (generating approximately 800 samples). Based on empirical best practices in intelligent fault diagnosis, we carried out a 4- to 5-fold expansion of the original dataset, yielding approximately 800 training samples to achieve an optimal balance, i.e., sufficient data diversity for improved generalization while explicitly avoiding overfitting or semantic drift in deep neural networks caused by excessively large artificial samples. To avoid data leakage, all augmentation operations were only applied to the training set. The test set remained untouched to ensure unbiased evaluation.

(1) Random amplitude scaling is used to simulate energy variations resulting from different cutting depths and minor fluctuations in spindle load; a scaling factor of $\lambda_{sig} - U(0.85, 1.15)$ is employed.
(2) Resampling is performed using time shifts, and the original long vibration signal is trimmed using random overlapping windows (shifting the starting point by $\pm$130 sample points).
(3) Dynamic background noise is introduced to simulate environmental interference by superimposing Gaussian white noise with a standard deviation of 1–5% of the original amplitude onto the signal, without causing data distortion.

5.3.3. Large-Scale and Diverse Dataset Validation Setup

To thoroughly evaluate the generalization capacity and prevent overfitting of the proposed PG-AFNet on larger and more diverse industrial datasets, the standard CWRU dataset was introduced for cross-load validation. Specifically, the evaluation was formulated as a comprehensive 10-class fault diagnosis task, encompassing one normal baseline state and nine distinct defect states (comprising inner race, outer race, and rolling element faults at varying severity levels). To strictly simulate diverse and variable working conditions, the dataset was evaluated under four independent load conditions (0 HP, 1 HP, 2 HP, and 3 HP), as well as a highly challenging mixed-load condition (ALL_HP). For absolute statistical fairness, a precise resampling strategy was employed to guarantee that the training sample capacity for each condition was strictly fixed at 3000 samples

(i.e., exactly 300 samples per class). Furthermore, in the mixed-load ALL_HP condition, identical fault types across different loads were categorized under the exact same class label. This rigorous setup forces the network to ignore load-induced amplitude fluctuations and learn load-invariant physical fault kinematics. Additionally, traditional shallow machine learning models (e.g., SVM) were excluded from this section, as they lack the parameter capacity and computational efficiency required to process such large-scale, high-variance mixed distributions, allowing the comparative analysis to focus purely on high-capacity deep learning architectures (1D-CNN, ResNet-18, and PG-AFNet).

5.3.4. Ablation Experiments

To quantify the individual contributions of the core modules in the proposed PG-AFNet, a comprehensive ablation study was conducted. The core modules mainly include physics-guided preprocessing, the dual-stream architecture, and the attention fusion mechanism. Based on the setup in Section 5.3.1, 5 model variants were developed. The evaluation was performed on the wafer thinning dataset through 100 independent cross-validation runs. The 5 variants and the input data configurations are as follows:

Model A (1D-CNN only): Uses only the VMD-purified 1D vibration signal;

Model B (ResNet only): Uses only the 2D ROI-cropped CWT scalogram;

Model C (Dual-stream w/o Physics): Concatenates the raw 1D signal and CWT scalogram without any physical prior constraints;

Model D (Dual-stream w/o Attention): Concatenates the VMD-purified 1D signal and the ROI-cropped CWT scalogram using rigid feature concatenation;

PG-AFNet (Proposed): The complete architecture employing physical preprocessing and the physics-informed attention mechanism for dynamic fusion.

5.3.5. Implementation Details and Fair Comparison Setup

To ensure a fair, rigorous, and completely unbiased comparative analysis, all benchmark models and the proposed PG-AFNet were executed under strictly identical experimental conditions. Firstly, to eliminate the influence of data distribution variance, all models shared the exact same dataset splits, governed by identical random seeds across the 100 independent cross-validation runs. Secondly, the identical data augmentation strategy (as described in Section 5.3.2) was applied to the training sets of all deep learning baselines (1D-CNN and ResNet), guaranteeing that the performance disparities originated purely from architectural superiority rather than the quantity of training data. Finally, equal effort was devoted to hyperparameter tuning for all baseline models to ensure they reached their respective optimal convergence states. Specifically, the parameters of the SVM (penalty factor C and RBF kernel parameter γ) were rigorously optimized via Grid Search. For all deep learning models, identical training protocols were adopted, including the optimizer, a consistent batch size, and an early-stopping mechanism based on validation loss to prevent overfitting while guaranteeing full convergence.

5.4. Experimental Results

The original data was randomly split into training and testing sets at a ratio of 75% to 25%. To mitigate statistical randomness caused by small sample splits, all experiments were repeated 100 times using independent random cross-validation splits. The performance metrics ultimately selected were accuracy (Acc) and F1-score. The F1-score comprehensively evaluates both precision and recall, providing a more objective measure of a model's overall classification capability, particularly in cases of mild class imbalance.

Comparative Analysis of Experimental Results

(1) In the method comparison experiments, the statistical results of each baseline model are shown in Table 1 and Figure 14. In terms of model accuracy, 1D-CNN performed the worst (63.33%), even slightly lower than the SVM based on manual features (68.33%). This result may seem unreasonable, but it aligns with successful experiences in the field of small-sample industrial fault diagnosis [28,29]. This is attributable to the support vector mechanism of SVMs, which enables them to achieve better performance under sample scarcity. Furthermore, 1D-CNNs are easily trapped in local optimization when handling strong noise coupling. Single-stream ResNet achieved an average accuracy of 73.33%. The proposed PG-AFNet attained an average accuracy of 82.45% ± 2.70%, significantly outperforming all comparison methods by efficiently capturing subtle time–frequency features. The inefficiency of 1D-CNN and the performance limitations of SVM indicate that when high-frequency abnormal chatter features strongly couple with low-frequency fundamental vibrations of the main shaft, end-to-end learning directly relying on time-domain amplitude data can become trapped in local optima. The attention fusion approach, which uses both 1D time-series data and 2D feature inputs, effectively preserves the vibration feature information selected by VMD.

Table 1. Comparison of algorithmic performance in thinning experiments.

Methods	Accuracy (%)	Acc_std	F1-Score	F1_std
SVM	68.33	±2.1%	0.671	±0.023
1D-CNN	63.33	±4.5%	0.628	±0.048
ResNet	73.33	±3.8%	0.721	±0.040
PG-AFNet (ours)	82.45	±2.7%	0.807	±0.041

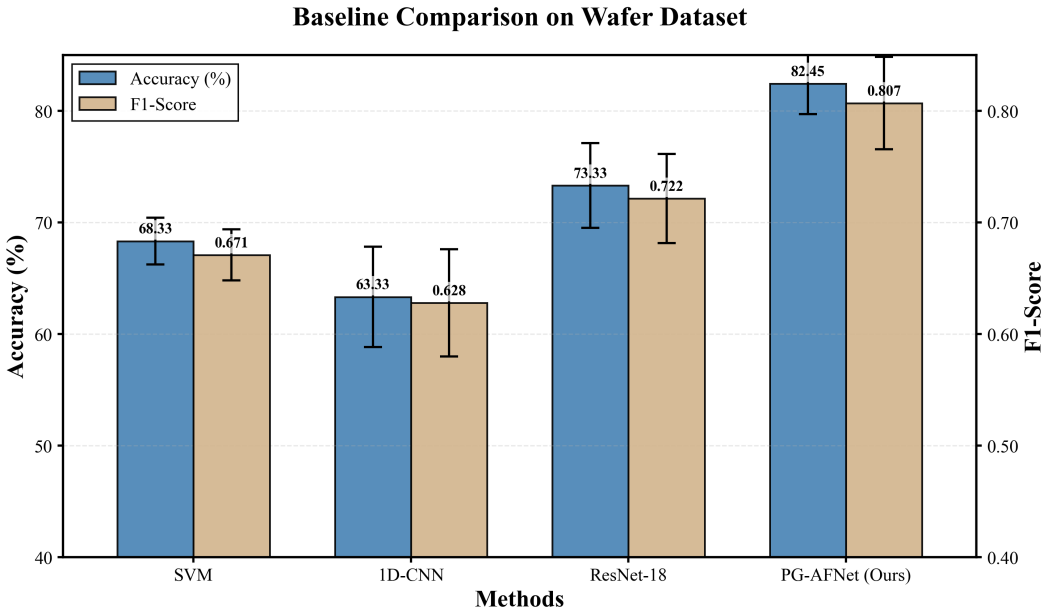


Figure 14. Performance statistics of different algorithms on the raw thinning dataset.

(2) After introducing sample augmentation strategies during the sample augmentation comparison process, the experimental results are shown in Table 2 and Figure 15. From an algorithmic perspective, all models benefit from performance improvements driven by data augmentation, yet significant performance disparities persist. Traditional 1D-CNN and SVM, constrained by shallow structures or 1D representations, easily encounter feature extraction capacity bottlenecks when handling high-noise, strongly coupled signals even

with increased samples. The single-stream ResNet achieves an accuracy improvement through deep 2D convolutions (from 73.3% to 78.4%). The proposed PG-AFNet method achieves an average accuracy of 86.6% ± 1.6%. This shows that the fusion mechanism adopted by PG-AFNet maximizes the extraction of subtle high-frequency vibration features from the data. It demonstrates superior performance and robustness in environments with strong noise and small sample sizes. From the perspective of data augmentation, all models achieved varying levels of performance improvement. The accuracy of the SVM based on manually designed features increased from 68.3% to approximately 71.2%, while the 1D-CNN improved from 63.3% to approximately 67.5%. Furthermore, regarding the improvement of the physical data augmentation, the statistical results across the 100 independent cross-validation runs demonstrate that the model did not exhibit significant performance degradation, i.e., no instances of negative transfer. This also indicates that the strictly constrained augmentation boundaries established in Section 5.3.2 functioned as a safe regularizer. It significantly compressed the standard deviation (from ±2.7% to ±1.6%).

Table 2. Comparison of algorithms before and after augmentation.

Methods	Original Data		Augmentation	
	Accuracy (%)	Acc_std	Accuracy (%)	Acc_std
SVM	68.33	±2.1%	71.20	±1.8%
1D-CNN	63.33	±4.5%	67.50	±4.1%
ResNet	73.33	±3.8%	78.40	±3.2%
PG-AFNet (ours)	82.45	±2.7%	86.64	± 1.6%

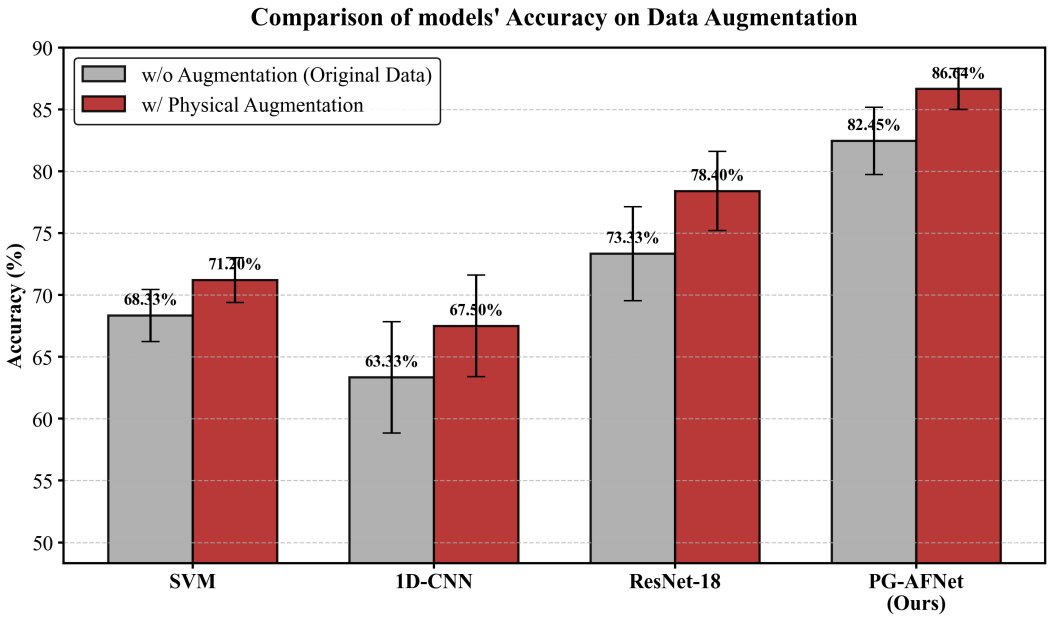


Figure 15. Comparison of models' accuracy before and after augmentation.

(3) The statistical results of the 100 independent cross-validation runs on the large-scale CWRU dataset are comprehensively presented in Figure 16. The results explicitly demonstrate the superiority and robustness of the PG-AFNet under large and diverse data distributions. Under isolated load conditions (0 HP to 3 HP), the baseline 1D-CNN exhibited noticeable fluctuations, dropping to 92.44% under 1 HP. In contrast, both the 2D ResNet-18 and the proposed PG-AFNet maintained high accuracy. However, PG-AFNet consistently outperformed ResNet-18, achieving near-perfect diagnostic accuracies

ranging from 98.77% to 99.70% across all individual loads. The robustness of the proposed architecture is most evident in the standard deviation matrix. Single-stream networks like the 1D-CNN exhibited severe instability when encountering specific loads (e.g., yielding a massive ±4.69% standard deviation under 2 HP). PG-AFNet, leveraging the dual-stream physical attention fusion, significantly compressed the standard deviation to a remarkably narrow range of ±0.31% to ±1.66%. This proves that the physics-guided preprocessing effectively isolates essential impact features, preventing the network from catastrophically failing on difficult randomized splits. The most rigorous test of model generalization is the ALL_ HP mixed-load condition, which intrinsically introduces severe rotational speed fluctuations and background noise drift. Impressively, PG-AFNet achieved an exceptional accuracy of 99.68% ± 0.68% in this diverse scenario. This conclusively demonstrates that rather than merely memorizing load-specific background profiles (overfitting), the physics-informed dual-stream architecture successfully extracts the fundamental, load-invariant physical signatures of the faults. This significantly addresses the overfitting concerns typically associated with small-sample training and confirms the model's reliability for complex, variable-condition mass production environments.

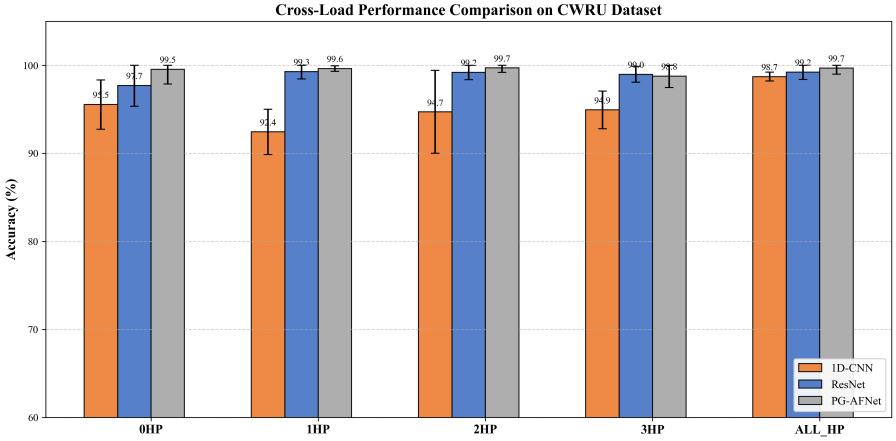


Figure 16. Performance statistics of different algorithms on the large-scale CWRU dataset.

(4) The statistical comparison results of the ablation study (mean accuracy ± standard deviation) are presented in Table 3. Under the same model architecture, Model C directly uses raw signals and full-band CWT textures, achieving an accuracy of 72.82% ± 5.6%. Compared to Model C, Model D introduces VMD and ROI cropping as physical mechanisms, improving accuracy to 76.08% (+3.2%) and reducing the standard deviation to ±4.6%. This indicates that physics-guided preprocessing effectively refines features and reduces noise. Model D uses a simple dual-stream architecture with rigid concatenation. Although it outperforms the single-stream Models A and B, its improvement over Model B is less than 3%. This suggests that while the dual-stream mechanism provides additional diagnostic information for thinning data, simple concatenation has inherent limitations. Assigning equal importance to both data types may cause feature interference, meaning the model may fail to learn the true features, thereby limiting further performance gains. The proposed PG-AFNet introduces a physics-informed attention module. It increases the average accuracy to 82.45%, which is 6.37% higher than Model D, and greatly reduces the standard deviation to ±2.7%. This shows that the attention mechanism resolves feature interference. Instead of treating both streams equally, it dynamically adjusts and assigns adaptive weights to the most reliable feature representations for each sample. This fully unlocks the potential of the dual-stream architecture and improves diagnostic stability. In summary, the ablation study shows that the performance gains of PG-AFNet do not come

from simply stacking deep learning layers. Instead, they rely on the synergy among physics-guided preprocessing, dual-stream architecture, and attention mechanism. Removing any of these modules may lead to performance loss or statistical instability. This confirms that the highly integrated architecture is scientifically effective for complex industrial wafer thinning environments.

Table 3. Comparison of ablation study on different model variants.

Model Variant	1D Stream Input	2D Stream Input	Fusion Strategy	Accuracy (%)
Model A	VMD Purified Signal	–	–	63.33 ± 4.5
Model B	–	ROI-Cropped CWT	–	73.33 ± 3.8
Model C	Raw 1D Signal	Full-band CWT	Concatenation	72.82 ± 5.6
Model D	VMD Purified Signal	ROI-Cropped CWT	Concatenation	76.08 ± 4.6
PG-AFNet	VMD Purified Signal	ROI-Cropped CWT	Attention	82.45 ± 2.7

6. Conclusions and Future Work

6.1. Research Conclusion

This paper addresses the challenge of detecting weak chatter in wafer thinning processes under conditions of high noise and small sample sizes. Research was conducted from two core dimensions: signal physical mechanism purification and deep feature fusion, resulting in the proposal and validation of the PG-AFNet model. The main research conclusions are as follows:

1. An effective pre-processing method based on mechanism-based noise reduction and physical masking has been established. To address the extremely strong fundamental frequency background noise in industrial spindles, the proposed VMD optimal mode selection and CWT ROI extraction method successfully achieves physical isolation between high-frequency white noise and low-frequency environmental interference. This approach effectively overcomes the feature extraction blind spot of end-to-end deep learning under extremely low signal-to-noise ratios, providing the model with high-purity intrinsic input.
2. A dual-stream attention fusion architecture for capturing weak features has been constructed. The proposed PG-AFNet successfully integrates 1D temporal features with two-dimensional spatiotemporal texture features. Comparative experiments demonstrate that this dual-stream architecture is more adept at capturing the weak, high-frequency impact features associated with machining degradation in hard brittle materials than either a single temporal model (1D-CNN) or a visual model (ResNet), significantly improving the classification lower bound.
3. In the validation of the measured wafer thinning dataset, engineering strategies at the code deployment level, including truncated network transfer and physical data augmentation, successfully mitigated high-dimensional overfitting. The physical augmentation fundamentally eliminated catastrophic statistical bad cases, boosting the diagnostic accuracy to 86.64% with significantly compressed variance. Furthermore, a systematic ablation study explicitly quantified the network's internal mechanics, proving that the remarkable performance gains stem from the indispensable synergy of physics-guided preprocessing, the complementary dual-stream architecture, and attention-based dynamic fusion, rather than rigid deep learning structures alone. Finally, the large-scale cross-load validation on the CWRU dataset conclusively confirmed the model's capacity for load-invariant feature learning. By achieving an exceptional accuracy of 99.68% under highly challenging mixed-load conditions (ALL_HP), PG-AFNet exhibits outstanding statistical stability and domain general-

ization, showcasing immense potential for reliable on-site deployment in complex, variable-condition mass production environments.

6.2. Research Outlook

Although the proposed PG-AFNet achieved certain results on wafer thinning datasets under laboratory conditions, future research can be further deepened in the following directions to address the complex and variable semiconductor mass production environment:

First, this study currently relies primarily on spindle vibration acceleration signals. However, during wafer thinning, brittle materials like silicon wafers are influenced by multiple coupled factors. In actual mass production, conditions such as coolant flow rate and instantaneous cutting depth are dynamically time-varying. Future research will integrate the synchronous monitoring of complex process parameters, including vibration, dynamic cutting force, and cutting temperature, to push the upper limits of monitoring accuracy in challenging environments.

Second, despite the promising results, this study currently focuses on chatter diagnosis under a relatively stable tool state. In actual long-term industrial production, continuous thinning wheel wear (tool degradation) will eventually alter the dynamic cutting coefficients and induce a gradual drift in the background vibration distribution. Future research must address this temporal dynamic by developing algorithms capable of maintaining diagnostic robustness across the entire lifecycle of the grinding tool, ensuring reliable chatter detection from the initial sharp state to the severe wear state.

Finally, exploring theoretical augmentation boundaries and unsupervised domain adaptation. To address the data distribution drift caused by the aforementioned tool wear and variable process parameters, we will explore unsupervised domain adaptation methods based on Generative Adversarial Networks (GANs). Furthermore, while the empirical physical augmentation in this study successfully avoided catastrophic 'bad cases', the theoretical upper boundaries of data augmentation remain an interesting open question. Using GANs will not only enable the systematic investigation of augmentation limits to guarantee data trustworthiness but also facilitate the seamless transfer of models trained under laboratory conditions to complex production lines without the need for additional labeling.

Author Contributions: Conceptualization, H.L. (Heng Li) and H.L. (Hua Liu); methodology, H.L. (Heng Li) and X.Z.; software, H.L. (Heng Li); validation, H.L. (Heng Li); formal analysis, H.L. (Heng Li), H.L. (Hua Liu) and L.Z.; investigation, H.L. (Heng Li) and X.Z.; resources, H.L. (Hua Liu) and L.Z.; data curation, H.L. (Heng Li) and X.Z.; writing—original draft preparation, H.L. (Heng Li); writing—review and editing, H.L. (Hua Liu) and L.Q.; visualization, X.Z.; supervision, L.Z., L.Q. and S.Z.; project administration, H.L. (Hua Liu) and L.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

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

Conflicts of Interest: Heng Li, Hua Liu, Liang Zhu and Xiangyu Zhao were employed by Zhejiang Qiushi Semiconductor Equipment Co., Ltd. Other authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

PG-AFNet	physics-guided dual-stream attention fusion transfer network
VMD	variational modal decomposition
CWT	continuous wavelet transform
ROI	regions of interest
AI	artificial intelligence
CNN	convolutional neural networks
1D	one-dimensional
SSD	subsurface damage layer
TTV	total thickness variation
RMS	root mean square
IMF	intrinsic mode function
CWRU	Case Western Reserve University
1D-CNN	one-dimensional CNN

References

1. Lau, J.H. Recent Advances and Trends in Advanced Packaging. *IEEE Trans. Components Packag. Manuf. Technol.* **2022**, *12*, 228–252. [CrossRef]
2. Yuan, J.; Lyu, B.; Hang, W.; Deng, Q. Review on the progress of ultra-precision machining technologies. *Front. Mech. Eng.* **2017**, *22*, 158–180. [CrossRef]
3. Yan, H.; Zhang, H.; Cao, S.; Wang, C. Study on Nano-Grinding Characteristics and Formation Mechanism of Subsurface Damage in Monocrystalline Silicon. *Micromachines* **2025**, *16*, 976. [CrossRef]
4. Chen, S.; Cheung, C.; Zhang, F.; Zhao, C. Three-dimensional modelling and simulation of vibration marks on surface generation in ultra-precision grinding. *Precis. Eng.* **2018**, *53*, 221–235. [CrossRef]
5. Tao, H.; Zeng, Q.; Liu, Y.; Zhao, D.; Lu, X. Effects of grinding-induced surface topography on the material removal mechanism of silicon chemical mechanical polishing. *Appl. Surf. Sci.* **2023**, *631*, 157509. [CrossRef]
6. Urbikain, G.; Olvera, D.; de Lacalle, L.N.L.; Elías-Zúñga, A. Stability and vibrational behaviour in turning processes with low rotational speeds. *Int. J. Adv. Manuf. Technol.* **2015**, *80*, 871–885. [CrossRef]
7. Urbikain, G.; Campa, F.J.; Zulaika, J.J.; de Lacalle, L.N.L.; Alonso, M.A.; Collado, V. Preventing chatter vibrations in heavy-duty turning operations in large horizontal lathes. *J. Sound Vib.* **2015**, *340*, 317–330. [CrossRef]
8. Urbikain, G.; Olvera, D.; de Lacalle, L.N.L.; Elías-Zúñga, A. Spindle speed variation technique in turning operations: Modeling and real implementation. *J. Sound Vib.* **2016**, *383*, 384-396. [CrossRef]
9. Urbikain-Pelayo, G.; Olvera-Trejo. D; de Lacalle, L.N.L.; Elías-Zúñga, A. Turn+: A MATLAB-based software for dynamic turning, chatter analysis, and surface roughness prediction. *SoftwareX* **2025**, *32*, 102401. [CrossRef]
10. Olvera, D.; Elías-Zúñga, A.; Martínez-Romero, O.; de Lacalle, L.N.L.; Martínez-Romero, H.; Siller, H.R.; Pineda, M.W. Improved predictions of the stability lobes for milling cutting operations of thin-wall components by considering ultra-miniature accelerometer mass effects. *Int. J. Adv. Manuf. Technol.* **2016**, *86*, 2139–2146. [CrossRef]
11. Kumar, A.; Vasu, V. Tool wear prediction using multi-sensor data fusion and attention-based deep learning. *Prod. Eng.* **2025**, *19*, 455–471. [CrossRef]
12. Teti, R.; Jemielniak, K.; O'Donnell, G.; Dornfeld, D. Advanced monitoring of machining operations. *CIRP Ann.* **2020**, *69*, 573–596. [CrossRef]
13. Wang, J.; Ma, Y.; Zhang, L.; Gao, R.; Wu, D. Deep learning for smart manufacturing: Methods and applications. *J. Manuf. Syst.* **2018**, *48*, 144–156. [CrossRef]
14. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
15. Jardine, A.K.S.; Lin, D.; Banjevic, D. A review on machinery diagnostics and prognostics implementing condition-based maintenance. *Mech. Syst. Signal Process.* **2006**, *20*, 1483–1510. [CrossRef]
16. Snell, J.; Swersky, K.; Zemel, R. Prototypical networks for few-shot learning. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4080–4090. [CrossRef]
17. Ben-David, S.; Blitzer, J.; Crammer, K.; Alex, K.; Fernando, P.; Vaughan, J.W. A theory of learning from different domains. *Mach. Learn.* **2010**, *79*, 151–175. [CrossRef]
18. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef] [PubMed]

19. Lei, Y.; Yang, B.; Jiang, X.; Jia, F.; Li, N.; Nandi, A. Applications of machine learning to machine fault diagnosis: A review and roadmap. *Mech. Syst. Signal Process.* **2020**, *138*, 106587. [CrossRef]
20. Yosinski, J.; Clune, J.; Bengio, Y.; Lipson, H. How transferable are features in deep neural networks? *Adv. Neural Inf. Process. Syst. (NeurIPS)* **2014**, *27*, 3320–3328. [CrossRef]
21. Zeiler, M.D.; Fergus, R. Visualizing and Understanding Convolutional Networks. In Proceedings of the 13th European Conference on Computer Vision (ECCV), Zurich, Switzerland, 6–12 September 2014; Part I, pp. 818–833. [CrossRef]
22. Bifano, T.G.; DePiero, D.K.; Golini, D. Chemomechanical effects in ductile-regime machining of glass. *Precis. Eng.-J. Int. Soc. Precis. Eng. Nanotechnol.* **1993**, *15*, 238–247. [CrossRef]
23. Dragomiretskiy, K.; Zosso, D. Variational Mode Decomposition. *IEEE Trans. Signal Process.* **2014**, *62*, 531–544. [CrossRef]
24. Habbouche, H.; Amirat, Y.; Benkedjouh, T.; Benbouzid, M. Digital twin-based gearbox fault diagnosis using variational mode decomposition and dynamic vibration modeling, *Measurement* **2025**, *246*, 116669. [CrossRef]
25. Smith, W.; Randall, R. Rolling element bearing diagnostics using the Case Western Reserve University data: A benchmark study. *Mech. Syst. Signal Process.* **2015**, *64*, 100–131. [CrossRef]
26. Chatterjee, J.; Dethlefs, N. Deep learning with knowledge transfer for explainable anomaly prediction in wind turbines. *Wind Energy* **2020**, *23*, 1693–1710. [CrossRef]
27. Michau, G.; Fink, O. Unsupervised transfer learning for anomaly detection: Application to complementary operating condition transfer. *Knowl.-Based Syst.* **2021**, *216*, 106816. [CrossRef]
28. Fernández-Francos, D.; Martínez-Rego, D.; Fontenla-Romero, O.; Alonso-Betanzos, A. Automatic bearing fault diagnosis based on one-class v-SVM. *Comput. Ind. Eng.* **2013**, *11*, 136–154. [CrossRef]
29. Wu, Y.; Dai, J.; Yang, X.; Shao, F.; Gong, J.; Zhang, P.; Liu, S. The Fault Diagnosis of Rolling Bearings Based on FFT-SE-TCN-SVM. *Actuators* **2025**, *14*, 152. [CrossRef]

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